A Phylogenetic Method for Detecting Positive Epistasis in Gene Sequences and Its Application to RNA Virus Evolution

Beth Shapiro, Andrew Rambaut,* Oliver G. Pybus,* and Edward C. Holmes*†

*Department of Zoology, Oxford University, Oxford, United Kingdom; and †Center for Infectious Disease Dynamics, Department of Biology, Mueller Laboratory, The Pennsylvania State University

RNA virus genomes are compact, often containing multiple overlapping reading frames and functional secondary structure. Consequently, it is thought that evolutionary interactions between nucleotide sites are commonplace in the genomes of these infectious agents. However, the role of epistasis in natural populations of RNA viruses remains unclear. To investigate the pervasiveness of epistasis in RNA viruses, we used a parsimony-based computational method to identify pairs of co-occurring mutations along phylogenies of 177 RNA virus genes. This analysis revealed widespread evidence for positive epistatic interactions at both synonymous and nonsynonymous nucleotide sites and in both clonal and recombining viruses, with the majority of these interactions spanning very short sequence regions. These findings have important implications for understanding the key aspects of RNA virus evolution, including the dynamics of adaptation. Additionally, many comparative analyses that utilize the phylogenetic relationships among gene sequences assume that mutations represent independent, uncorrelated events. Our results show that this assumption may often be invalid.

Introduction

Epistatic interactions among mutations play an important role in molecular evolution (Michalakis and Roze 2004). Epistasis arises when 2 mutations have a combined effect on fitness other than what would be expected if they acted independently. Epistasis can be either positive or negative, depending on the direction of deviation from multiplicative fitness effects. Negative epistasis between deleterious mutations (also known as synergistic epistasis) is a key aspect of some theories of the evolution of recombination (Kondrashov 1988). Alternatively, positive epistasis between deleterious mutations may increase genetic robustness by reducing their combined effect on fitness (Sanjuan et al. 2004).

Despite the importance of both forms of epistasis in evolution, little is known about the prevalence of epistatic interactions in natural populations, largely due to difficulties in quantifying epistatic effects and limitations imposed by the experimental context. For example, if positive and negative epistatic interactions were equally common, as observed among deleterious mutations in the bacterium *Escherichia coli* (Elena and Lenski 1997), the sum of the observed deviations from expected fitness would be negligible. Here, we consider epistatic interactions in a phylogenetic context, which enables us to undertake a broad comparative survey of the extent and nature of these interactions in natural populations.

RNA viruses are ideal for exploring the nature of epistatic interactions among mutations because they comprise compact genomes, usually less than 15 kb in length, often with overlapping reading frames and functional secondary structure, both of which are predicted to lead to strong epistasis. Using in vitro fitness assays, 2 recent studies have identified positive epistasis in experimental populations of vesicular stomatitis virus (VSV) (Sanjuan et al. 2004), a negative-sense RNA virus, and human immunodeficiency virus type 1 (HIV-1) (Bonhoeffer et al. 2004), a retrovirus. Both studies compared the fitness of viruses with single and multiple artificially derived mutations, and both found evidence that most interactions among deleterious mutations exhibited positive epistasis. This has important implications for the understanding of the evolution of sex, suggesting that RNA recombination is not an adaptation to purge deleterious mutations from viral genomes (Bonhoeffer et al. 2004). However, the generality of these results to viruses other than VSV and HIV-1 remains unexplored.

Because RNA viruses have very high rates of mutation and substitution (Drake and Holland 1999; Jenkins et al. 2002), a fundamental and potentially detectable consequence of positive epistasis will be the evolution of compensatory changes in gene sequences. Experimental populations of bacteria and viruses have been shown to recover fitness through such compensatory changes (Burch and Chao 1999), and simulations of RNA folding have predicted widespread compensatory mutations (Wilke et al. 2003).

Compensatory changes may also be caused by interactions between natural selection and protein structure. For example, compensatory changes could occur in viral epitopes, which are short peptides presented on the surface of infected cells that are recognized by the immune system. If a mutation that alters epitope recognition also compromises viral protein structure or function, then a compensatory mutation may be required to restore overall fitness. Subsequently, that pair of mutations may increase in frequency in the population (by selection or drift) and be detected in a sufficiently large sample of sequences. In contrast, negative epistasis among deleterious mutations generates unfit individuals and will therefore be practically undetectable through the analysis of gene sequence data.

To test whether positive epistasis is an important feature of RNA virus evolution, we undertook a comparative analysis of 177 RNA virus genes sampled from natural populations, including retroviruses and positive-sense, negative-sense, and double-stranded viruses. We implemented a phylogenetic test for positive epistasis, quantified as the occurrence of pairs of mutations that are only observed to change in concert—never independently—on gene trees (here called compensatory or co-occurring changes; illustrated in fig. 1). Clearly, this is a highly conservative test as it will not count advantageous pairs of mutations that have already been fixed in the population.

Key words: coevolution, phylogenetics, RNA viruses, viral evolution, epistasis.

E-mail: beth.shapiro@zoo.ox.ac.uk.

Mol. Biol. Evol. 23(9):1724–1730. 2006
doi:10.1093/molbev/msl037
Advance Access publication June 14, 2006

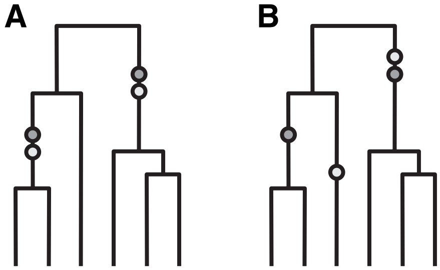

FIG. 1.—Detection of compensatory mutations between 3 positions (sites) in an alignment. Each circle represents a mutation occurring at 1 of the 3 sites. In tree A, site 1 and site 2 are considered to be compensatory because at least 2 mutations occur at both sites, and every mutation at site 1 is accompanied by a mutation at site 2. In tree B, the 2 sites are not compensatory.

The analysis was conducted in 3 stages. First, the overall extent of epistasis in RNA viruses was estimated by calculating the number of sites involved in compensatory interactions for each data set. Second, we inferred the phylogenetic distribution of the observed compensatory mutations. Last, we mapped their genomic position and investigated the distance along the genome between each pair of co-occurring changes. At each stage, nucleotide and protein alignments were evaluated separately, and the significance of the results was evaluated by comparing the observed data with parametric simulations. Strikingly, we found evidence for positive epistasis in all major groups of RNA viruses, with the majority of compensatory interactions spanning short sequence regions.

Materials and Methods

We compiled, from GenBank, alignments of 177 structural gene sequences representing 152 different RNA viruses and 30 virus families. Each sequence was taken from a different individual, and excessively passaged sequences were avoided. Alignments ranged from 468 to 2,741 nt (median 940) in length and included 10–75 (median 26) sequences. A detailed description of each alignment is available in the Supplementary Material online. A maximum likelihood (ML) phylogeny was then estimated for each alignment under the General Time Reversible (GTR) + I + Γ model using PAUP* (Swofford 2003).

Given the phylogenetic tree of each data set, the branch on which each mutation occurred was reconstructed using ACCTRAN parsimony reconstruction (Swofford and Maddison 1987) along the ML tree (positions with gaps were excluded from the analysis). In our analysis, pairs of mutations were identified as co-occurring if 1) each branch with a mutation at one site had a corresponding mutation at the other site and 2) this co-occurrence of mutations happened more than once in the phylogeny (i.e., on at least 2 independent branches; see fig. 1). For each co-occurring mutation, we recorded 1) the codon position (i.e., 1st, 2nd, and 3rd), 2) whether it was a synonymous or nonsynonymous nucleotide mutation, 3) whether the mutation occurred on an internal or external phylogeny branch, 4) the nucleotide or amino acid position of the mutation along the sequence, and 5) the distance (in nucleotides or amino acids) between the co-occurring mutations. This parsimony reconstruction was performed using both nucleotide and protein alignments.

Because the GTR + I + Γ model does not take into account the possibility of covariation among nucleotide sites, lineages that contain the same covarying sites may cluster together in the phylogeny as each site will be evaluated independently. This will reduce the number of branches in the estimated phylogeny on which these 2 changes will occur. Because our model only counts sites as covarying if they change simultaneously along at least 2 branches in the phylogeny, the overall effect of using this model to infer trees will be to reduce the number of observed covarying sites, making our test highly conservative.

For each data set, the total number of co-occurring pairs and the genomic distance between them were compared with those expected by chance through simulation. To assess statistical significance, parametric bootstrapping was employed as follows: For each virus phylogeny, ML branch lengths were first estimated using CODEML (Yang 1997) under 3 models: (i) a codon substitution model (Yang and Nielsen 1998) in which the relative rate of transitions/transversions, κ, and the relative rate of nonsynonymous/synonymous changes, ω, were coestimated (for comparison to nucleotide data sets), (ii) the WAG model of amino acid evolution (Whelan and Goldman 2001) with Γ-distributed rate heterogeneity among sites (for comparison to amino acid data sets), and (iii) the WAG model without Γ-distributed rate heterogeneity (also for comparison to amino acid data sets). Rate variation in model (i) was modeled as a consequence of differences in nonsynonymous and synonymous rates of evolution.

Three thousand alignments were simulated for each real data set (1,000 for each of the models (i), (ii), and (iii) described above) using the estimated branch lengths and substitution model parameters. Simulations were performed using the program Evolver in the PAML package (Yang 1997). Mutations in the resulting 531,000 (3,000 × 177) simulated data sets were then reconstructed using the parsimony method described above. For each simulation, the number of co-occurring mutations and the genomic distance between them were estimated in the same manner as for the real data. The number of co-occurring mutations observed in a real virus alignment was considered significant if it was greater than that observed in >95% of the 1,000 corresponding simulated alignments (each of the 3 models was evaluated separately). False discovery rate correction (Benjamini and Hochberg 1995) was used to account for the possibility of false positives under multiple testing.

Results

Measuring the Extent of Epistasis in RNA Viruses

When compared with the parametric simulations, 55 of the 177 RNA virus nucleotide alignments (31%) showed significantly more co-occurring changes than expected by chance, after correcting for false-positive discoveries (table 1). Furthermore, when the proportion of such sites was pooled across all 177 data sets, there was a highly significant excess of epistatic sites in the real data sets in comparison to the simulated ones ($P < 0.001$; table 1; depicted in fig. 2). This result is particularly notable given the

Table 1
Total Proportion of Sites across All 177 Data Sets That Are Involved in Compensatory Interactions

Alignment Type	Real Data Sets	Simulated Data Sets (mean of 1,000 replicates)	Significantly Higher Proportion of Sites in the Real Data Sets? (*P* value)[a]	Number of Real Data Sets with Significantly Higher Proportion of Sites[b]
Nucleotide	0.027	0.023	Yes ($P < 0.001$)	55 (31%)
Amino acid (with rate heterogeneity)	0.007	0.007	No	13 (7%)
Amino acid (without rate heterogeneity)	0.007	0.008	No	44 (25%)

[a] Significance is derived from the probability that a simulated data set has a higher proportion of compensatory sites than the real data set.

[b] False discovery rate correction (Benjamini and Hochberg 1995) was used to correct for multiple testing.

conservative nature of the test used. Fewer RNA virus data sets were found to contain significant epistasis when sequences were compared at the amino acid level, and the total proportion of sites involved in compensatory interactions was also not significant when compared against alignments simulated using either model (ii) or (iii) (table 1). Adding among-site rate heterogeneity to the amino acid simulations reduced the number of significant amino acid alignments from 44 to 13. This is expected as increasing rate heterogeneity will increase the probability that variable sites contain multiple mutations.

Because multiple mutations at the same site are a common occurrence in RNA viruses, we also calculated the total proportion of reconstructed changes that were involved in epistatic interactions across all data sets (table 2). This statistic distinguishes more evolutionary events than the proportion of sites statistic (table 1) and is therefore more powerful. This "proportion of reconstructed changes" statistic indicated a significant excess of compensatory changes for both the nucleotide alignments ($P < 0.001$) and the amino acid alignments, when compared with simulations without rate heterogeneity ($P < 0.001$). As before, adding rate heterogeneity to the amino acid simulations led to a reduction in statistical significance.

The Phylogenetic Distribution of Co-occurring Changes

Next, we investigated the phylogenetic position of the reconstructed mutations that had been identified as co-occurring. Table 2 shows the total proportion of such mutations across all 177 data sets that occurred on internal or external phylogeny branches. In comparison to the parametric simulations, there was a significant excess of co-occurring changes on both internal and external branches for the nucleotide alignment and for the rate heterogeneity–free amino acid analysis (table 2). For the amino acid analysis incorporating rate heterogeneity, a significant excess of co-occurring mutations was found only on internal branches (table 2). ACCTRAN parsimony reconstruction will force mutations to occur more deeply in the tree, which could result in a bias toward mutations along internal rather than external branches. However, as the same parsimony reconstruction was performed on the real and simulated data, any potential methodological bias would not have affected our results.

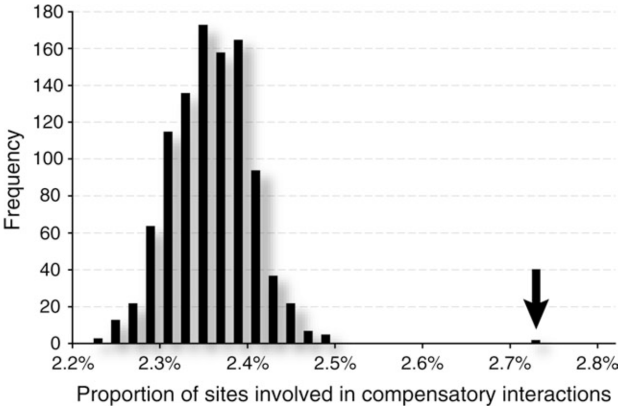

FIG. 2.—Frequency distribution of the total proportion of nucleotide sites involved in compensatory interactions for the 1,000 simulation replicates (in which each replicate contained 177 simulated alignments). The equivalent value for the 177 real RNA virus alignments is indicated by the arrow. This figure represents the results summarized in the top row of table 1.

We then compared the probability that co-occurring mutations were identified along internal rather than external branches. Importantly, in the real data sets, compensatory mutations were significantly more likely to be found on internal branches than on external branches, in comparison to the parametric simulations (table 3). This was tested by calculating the total ratio of such mutations that fell on either external or internal branches. The excess of co-occurring changes on internal branches was significant for both the nucleotide alignments and for the amino acid alignments without rate heterogeneity (table 3).

The Genomic Distribution of Co-occurring Changes

Finally, we mapped the genomic position of each reconstructed mutation that had been identified as co-occurring. The spatial distribution of compensatory interactions along the genome was also significantly different between the real and simulated data sets (fig. 3). In the real data, epistatic interactions were more likely to occur between pairs of mutations that were located close together in the sequence alignment. This pattern was observed at both the nucleotide and amino acid levels (both with and without rate heterogeneity). At the nucleotide level, the majority of interacting pairs of mutations occurred within just 3 bases of each other. However, significantly more interacting pairs than were expected by chance were found at most distances (fig. 3*A*). At the amino acid level, the majority of interacting pairs were separated by fewer than 15 amino acids (fig. 3*B*).

Discussion

Despite the highly conservative nature of our parsimony-based method, the analysis conducted here revealed that positive epistasis is commonplace at the nucleotide level in natural populations of RNA viruses. Importantly, the

Table 2
Total Proportion of Reconstructed Mutations across All 177 Data Sets That Are Involved in Compensatory Interactions

Alignment Type	Real Data Sets	Simulated Data Sets (mean of 1,000 replicates)	Significantly Higher Proportion of Sites in the Real Data Sets? (*P* value)[a]
Nucleotide	0.11	0.084	Yes ($P < 0.001$)
Internal branches only	0.11	0.082	Yes ($P < 0.001$)
External branches only	0.1	0.083	Yes ($P < 0.001$)
Amino acid (with rate heterogeneity)	0.039	0.035	No
Internal branches only	0.042	0.034	Yes ($P = 0.02$)
External branches only	0.034	0.034	No
Amino acid (without rate heterogeneity)	0.039	0.025	Yes ($P < 0.001$)
Internal branches only	0.042	0.023	Yes ($P < 0.001$)
External branches only	0.034	0.024	Yes ($P < 0.001$)

[a] Significance is derived from the probability that a simulated data set has a higher proportion of compensatory changes than the real data set.

majority of paired mutations that were found occurred along internal phylogeny branches, indicating that they have reached a relatively high frequency in the population. These mutations are, therefore, likely to be either selectively neutral or beneficial, as would be expected if the mutations were compensatory. The frequency of co-occurring mutations was not strongly correlated with most aspects of viral biology, including genomic segmentation, host type, or viral family. However, when viruses with different genome architectures were compared, approximately 2.5 times more compensatory evolution was found in negative-sense than in positive-sense viruses (Kolmogorov–Smirnov test, $P = 0.001$). Because negative-sense viruses are largely clonal (Chare and Holmes 2004), recombination is unlikely to be responsible for the independent occurrence of multiple mutations and, therefore, for the pattern observed in this analysis.

Overall, our results suggest that, for RNA viruses, positive epistasis may be more important for nucleotide sequence evolution than for protein sequence evolution. Surprisingly, the majority of co-occurring nucleotide mutations were synonymous in that they did not result in a corresponding amino acid change. In fact, the probability that both mutations in a co-occurring pair were synonymous was 0.75 (2,766 out of 3,674 total pairs), compared with only 0.04 (130 of 3,119 average total pairs) in the simulations. Conversely, 96% of pairs in the simulated data sets contained at least one nonsynonymous mutation. Although previous phylogenetic studies have shown nonneutral evolution of synonymous sites in particular RNA virus species (e.g., GB virus C; Simmonds and Smith 1999), our results demonstrate that this is a general trend across RNA viruses.

Crucially, the epistatic changes that we document are spatially restricted. In the nucleotide analysis, a large proportion of paired mutations occur within 3 nucleotides of each other (although significantly more interactions among nucleotides occur at most distances, in comparison to the parametric simulations; fig. 3*A*). In contrast, the spatial distribution of interacting pairs in the amino acid analysis is strongly skewed toward short distances, with most pairs of co-occurring changes separated by fewer than 15 amino acids (fig. 3*B*). This result strongly points toward the importance of structural constraints in RNA virus evolution, even within coding regions (see Simmonds et al. 2004; Thurner et al. 2004). The genomic proximity of compensatory changes may be explained by localized RNA or protein structure, such as hairpin loops or alpha helices, which impose selective constraints on otherwise beneficial mutations. Tightly linked compensatory changes have been implicated previously in the escape cytotoxic T-lymphocyte responses in both HIV and influenza A virus (Kelleher et al. 2001; Rimmelzwann et al. 2005).

Localized structure may also be important in maintaining single-stranded RNA–binding domains in RNA viruses, as observed in untranslated regions (e.g., Gougas et al. 2004), structural genes (e.g., Isogai and Yoshikawa 2005), and nonstructural genes (e.g., Filmore et al. 2002). RNA-binding domains have been shown to play an important role in viral

Table 3
Internal/External Branch Ratio of Reconstructed Mutations That Are Involved in Compensatory Interactions

Alignment Type	Real Data Sets	Simulated Data Sets (mean of 1,000 replicates)	Significantly Higher Ratio in the Real Data Sets? (*P* value)[a]
Nucleotide	1.24	1.01	Yes ($P < 0.001$)
Amino acid (with rate heterogeneity)	1.11	0.97	No
Amino acid (without rate heterogeneity)	1.11	0.75	Yes ($P < 0.001$)

[a] Significance is derived from the probability that a simulated data set has a higher proportion of compensatory changes than the real data set.

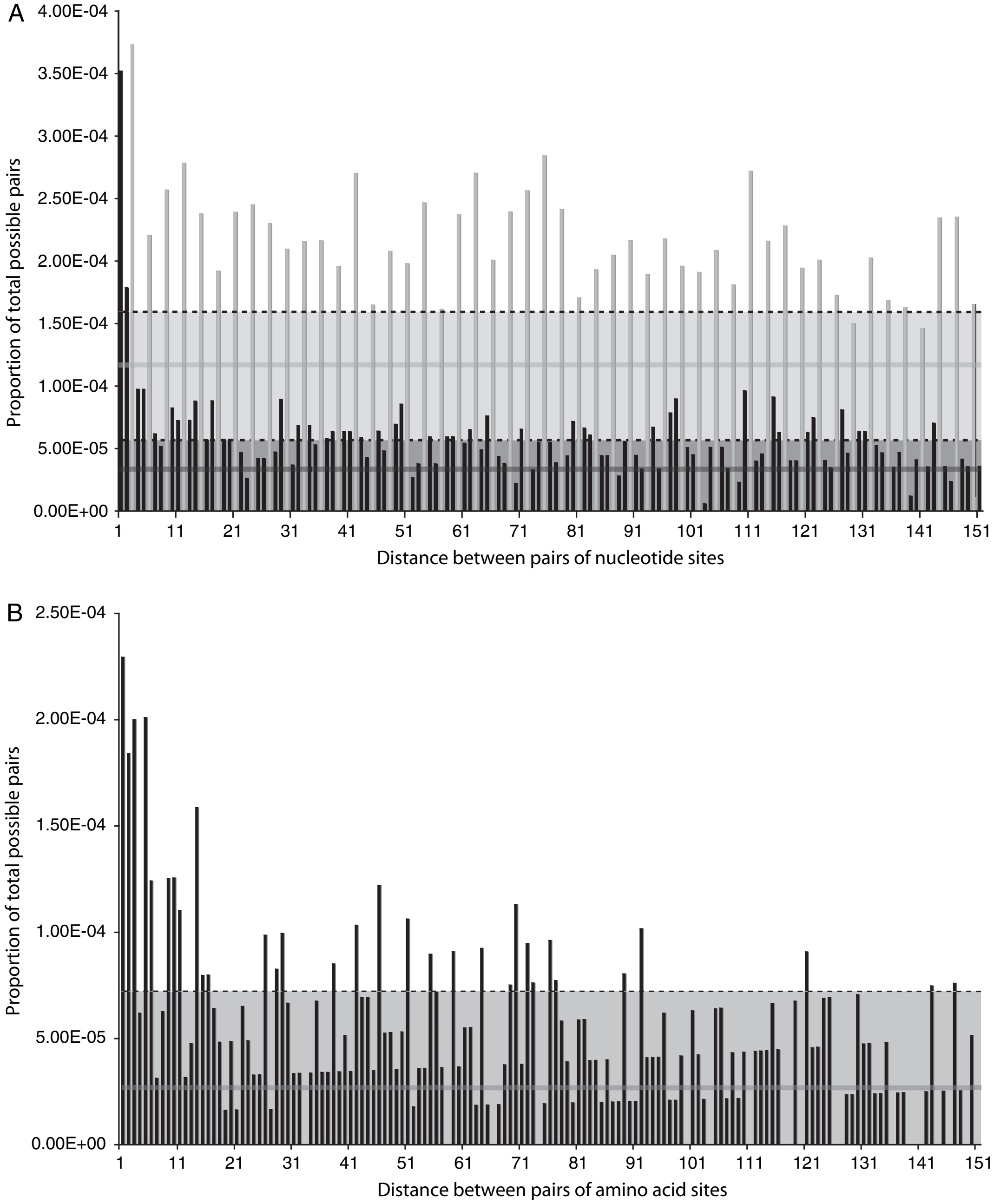

FIG. 3.—The spatial distribution of pairs of compensatory mutations for (*A*) the nucleotide alignments and (*B*) the amino acid alignments. The *x* axis depicts the number of sites along the genome between each pair of coevolving sites. The *y* axis shows the proportion of the total number of possible pairs of sites at each distance that have been identified as coevolving. (Note that the number of possible pairs decreases as the distance between them increases.) Values from the real data set are shown as vertical bars. The 95% highest value (shaded area) and average value (solid horizontal lines) obtained from the 1,000 parametric simulations are also given. In the nucleotide analysis, the majority of coevolving mutations occurred at third codon positions, resulting in a higher proportion of coevolving pairs of sites at distances that are multiples of 3. In (*A*), values for these distances are distinguished as light gray bars. Additionally, in (*A*), the 95% highest and average values obtained from the simulations were calculated separately for distances that are multiples of 3 (gray) and all other distances (black).

replication and assembly (Filmore et al. 2002; Gougas et al. 2004), viral persistence (Simmonds et al. 2004), and the protection of viral RNA from RNAse attack (Taliansky and Robinson 2003). The ubiquity of RNA-binding nucleoproteins in negative-sense RNA viruses provides a possible explanation for the higher frequency of co-occurring changes in these viruses (as noted above). However, the biological role of localized structural constraints is likely to depend on specific aspects of viral life-history strategies; so further work focusing on specific virus groups is evidently necessary to differentiate the factors involved.

Although more viruses were identified as showing significant positive epistasis in amino acid sequences when rate heterogeneity was excluded from the simulations, the addition of this parameter provided a better approximation of the real data in >95% of virus genes (data not shown). Rate heterogeneity alone, however, was not sufficient to account for the total extent of observed co-occurring mutations, suggesting that more sophisticated substitution models are required to better explain protein sequence evolution in RNA viruses.

Many comparative analyses that utilize phylogenetic relationships among gene sequences assume that mutations represent independent and unlinked events. When this assumption is invalid, as our results show may be often, phylogeny estimation will be adversely affected. For example, if a pair of sites is evolving together but assumed to be evolving independently, then any phylogenetic information resulting from changes at these sites will be counted twice, resulting in overestimates of confidence in well-supported phylogenetic clades and underestimates of confidence in alternate arrangements (Jow et al. 2002). Although progress has been made in developing more realistic models of RNA evolution, current methods that allow for nonindependence between sites require either compensatory sites to be identified a priori (Jow et al. 2002; Hudolet et al. 2003) or nonindependence to occur only among neighboring sites (Lunter and Hein 2004; Siepel and Haussler 2004), which is insufficient to explain the covariation in RNA virus genes observed here. However, phylogenetic methods that allow evolutionary rates to be autocorrelated along gene sequences may reasonably approximate RNA virus evolution when the locations of interacting sites are unknown. Such a method was previously implemented by Felsenstein and Churchill (1996) but has not been widely used. Ideally, a method is required that both explicitly models the evolution of covarying sites and estimates their frequency as well as the phylogeny itself. It should be noted, however, that the total number of sites that were conclusively identified as being compensatory is small. Although this could be due, at least in part, to the extremely conservative nature of our test, the true amount of compensatory evolution is likely to vary greatly among species and genes, in turn having more or less significant effects on phylogeny estimation.

The precise evolutionary effects of positive epistasis are complex and still to be determined and are influenced by factors such as the number of sites involved, the heterogeneity of the environment, and the balance between mutation, selection, and drift. Depending on the scenario, positive epistasis can either reduce the burden of deleterious mutations or act as a constraint to adaptive evolution. In RNA viruses, epistasis may therefore be central to epidemiologically important phenomena as host switching, immune escape, and the development of antiviral resistance.

Supplementary Material

Supplementary Material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank A. Drummond and C. Holmes for advice on the statistical analysis. We also thank P. H. Harvey for discussion and comments on the manuscript. This work was supported by the Wellcome Trust and the Royal Society.

Funding to pay the Open Access publication charges for this article was provided by the Department of Zoology, Oxford University.

Literature Cited

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300.

Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ. 2004. Evidence for positive epistasis in HIV-1. Science 306:1547–50.

Burch CL, Chao L. 1999. Evolution by small steps and rugged landscapes in the RNA virus phi 6. Genetics 151:921–7.

Chare ER, Holmes EC. 2004. Selection pressures in the capsid genes of plant RNA viruses reflect mode of transmission. J Gen Virol 85:3149–57.

Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. Proc Natl Acad Sci USA 96:13910–13.

Elena SF, Lenski RE. 1997. Test of synergistic interactions among deleterious mutations in bacteria. Nature 390:395–8.

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

Filmore GC, Lin H, Li JK-K. 2002. Localization of the single-stranded RNA-binding domains of Bluetongue virus nonstructural protein NS2. J Virol 76:499–506.

Gougas LM, Filman DJ, Hogle JM, Gehrke L. 2004. Cofolding organizes alfalfa mosaic virus RNA and coat protein for replication. Science 306:2108–11.

Hudolet C, Gowri-Shankar V, Jow H, Rattray M, Higgs PG. 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol Phylogenet Evol 28:241–52.

Isogai M, Yoshikawa N. 2005. Mapping the RNA-binding domain on the apple chlorotic leaf spot virus movement protein. J Gen Virol 86:225–9.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol 54:152–61.

Jow H, Hudolet C, Rattray M, Higgs PG. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol Biol Evol 19:1591–601.

Kelleher AD, Long C, Holmes EC, et al. (18 co-authors). 2001. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T-lymphocyte responses. J Exp Med 193:375–85.

Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336:435–40.

Lunter G, Hein J. 2004. A nucleotide substitution model with nearest-neighbor interactions. Bioinformatics 20:i216–23.

Michalakis Y, Roze D. 2004. Epistasis in RNA viruses. Science 306:1492–3.

Rimmelzwann GF, Berkhoff EG, Nieuwkoop NJ, Smith DJ, Fouchier RA, Osterhaus AD. 2005. Full restoration of viral fitness by multiple compensatory mutations in the nucleoprotein of influenza A virus cytotoxic T-lymphocyte escape mutants. J Gen Virol 86:1801–5.

Sanjuan R, Moya A, Elena SF. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci USA 101:8396–401.

Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468–88.

Simmonds P, Smith DB. 1999. Structural constraints on RNA virus evolution. J Virol 73:5787–94.

Simmonds P, Tuplin A, Evans DJ. 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. RNA 10:1337–51.

Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sunderland, MA: Sinauer Associates.

Swofford DL, Maddison DR. 1987. Reconstructing ancestral character states under Wagner parsimony. Math Biosci 87:199–229.

Taliansky ME, Robinson DJ. 2003. Molecular biology of umbraviruses: phantom warriors. J Gen Virol 84:1951–60.

Thurner C, Witwer C, Hofacker IL, Stadler PF. 2004. Conserved RNA secondary structures in Flaviridae genomes. J Gen Virol 85:1113–24.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–9.

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

Yang ZH. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6.

Yang ZH, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–18.

Peter Lockhart, Associate Editor

Accepted June 1, 2006