RNA Structures and their Molecular Evolution in HIV; Evolution of Robustness in RNA Structures and Theoretical Systems

by
Kristen Kamerath Dang

A Dissertation submitted to the faculty of the The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering (Bioinformatics and Computational Biology).

Chapel Hill
2008

Approved by:

Carol N. Lucas, Chair

Christina L. Burch, Advisor

Ronald I. Swanstrom, Advisor

Shawn Gomez, Reader

Alexander Tropsha, Reader

Abstract

Kristen Kamerath Dang

RNA Structures and their Molecular Evolution in HIV; Evolution of Robustness in RNA Structures and Theoretical Systems.
(Under the direction of Christina L. Burch and Ronald I. Swanstrom.)

The known functions of RNA structures have expanded of late, such that RNA is considered a more active player in molecular biology. The presence of RNA secondary structure in a sequence should constrain evolution of its constituent nucleotides because of the requirement to maintain the base-pairing regions in the structure. In a previous work, we found support for this hypothesis in nine molecules from various organisms, the exception being a structure found in a protein-coding region of the HIV-1 genome. In this work, I examine the interaction of constraints imposed by RNA structures and host-induced hypermutation on molecular evolution in HIV-1. I conclude that RNA structures in HIV do evolve via compensatory evolution, but that hypermutation can obscure the expected signal. Since RNA's known roles have increased, so have the methods for identification and prediction of RNA structures in genetic sequence. I use a method adapted for searching in multiple coding regions to identify conserved RNA structures throughout the HIV-1 and HIV-2 genomes. I find evidence for several new, small structures in HIV-1, but evidence is less strong for HIV-2. Finally, I consider the evolution of robustness, the property of phenotypic constancy, using RNA structures and two other theoretical model systems. I find that pervasive environmental variation can select for environmental and genetic robustness in all three systems, and conclude that it may be a generic mechanism for the evolution of robustness.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

APOBEC	Apolipoprotein B mRNA editing catalytic polypeptide
cPPT/PPT	central polypurine tract / polypurine tract
HIV	Human Immunodeficiency Virus
κ	kappa, the transition-transversion rate ratio: μ_{Ti}/μ_{Tv}
RNA	Ribonucleic Acid
RRE	Rev Response Element
SIV	Simian Immunodeficiency Virus
TAR	Tat Response Region

Chapter 1

Introduction

1.1 Functions of RNA

RNA is classically known for its roles in transporting and translating protein-coding gene sequences. It fulfills these functions as messenger, transfer, and ribosomal RNA molecules. Beginning with the discovery of the catalytic properties of ribozymes (e.g.[118]), it was apparent that RNA had a more varied role in molecular biology. For example, recent work has illuminated RNA's roles in gene regulation as microRNAs[37] and riboswitches[10]. In viruses, RNA structures have many roles, including as packaging signals that ensure viral genomes are packaged into viral capsids. They also recruit host cell ribosomes to viral transcripts to allow translation of viral proteins. Thus, rather than a passive role as a medium for transporting and translating protein coding genetic sequence, RNA is recognized as an active player in molecular biology.

1.2 Method development in identifying conserved RNA secondary structure

Due no doubt to the expanding role of RNA in molecular biology, activity has been high in the development of bioinformatics methods to predict RNA secondary and tertiary structure (see [15] for a review of relevant reviews). Zuker's mfold algorithm[120] is the classic one for predicting the lowest free-energy structure for a given sequence. This is helpful if one already knows that a given sequence harbors some functional structure, but needs to get a precise

prediction for the structure. It also assumes that the minimum free-energy structure is in fact the functional structure. In many cases, rather than asking what is the structure for a given sequence, one wishes to know whether there is evidence of any conserved structure in a given sequence(s) or even in a whole genome. Though some attempt has been made to answer the latter question by looking for structures with significantly lower-than-average free energies, it has been shown that the folding energies of functional RNA structures are generally not significantly different from background levels, and cannot be used to distinguish regions of functional structure from non-conserved structure[90]. This kind of structure identification and prediction requires a different approach, as described by Meyer[72].

Another type of RNA structure prediction takes into account the information in a multiple sequence alignment about mutation patterns consistent with conservation of RNA structure. These methods, known as covariation methods, identify likely stem regions by finding columns in an alignment containing positions that can form base pairs and that co-vary, such that single-position mutants are never (or rarely) observed. RNAalifold is a program that combines thermodynamic predication with covariation information[111]. In a review by Gardner[32] of several types of prediction programs, including single-sequence prediction methods such as mfold, other covariation methods, and methods that both align sequences and predict structures, RNAalifold was one of the most successful. However, these covariation methods do not incorporate evolutionary models that account for substitution rate and distance between sequences in an alignment, therefore they can misestimate the significance of co-varying mutations. In addition, these methods require that enough sequence diversity be present in the alignment that several co-variations can be observed per stem.

1.3 Selection for RNA structures creates heterogeneous rates of evolution across sites

A different method, one that surpassed RNAalifold's performance for some datasets in Gardner's study, is Pfold[53], a program that, like RNAalifold, uses information from a sequence alignment to predict structure. Instead of looking for co-variation in an alignment, Pfold uses differences in rates between positions to make its predictions. Thus in theory it can make predictions

in alignments where few or no co-variations are observed as long as there is enough diversity for rate differences between positions to be reliably assessed[76]. Pfold is one of several RNA prediction methods (e.g. [81]) to take advantage of the observation of a difference in rate (or, in the same vein, variability[103]) between nucleotides in base-paired regions of RNA structures and those that are single-stranded, whether in a loop or a non-structured region. The presence of RNA secondary structure in a sequence should constrain evolution of its constituent nucleotides because of the requirement to maintain the base-pairing regions in the structure. If structures are functionally important, single mutations that destabilize the structure by eliminating a base-pair bond in a stem region will not often become fixed in a population except with the simultaneous fixation of a compensating mutation on the opposite side of the stem that restores the bond. Thus, stem positions must simultaneously fix two mutations, a slower process than fixation of a single mutation. Innan and Stephan use data from an earlier study of secondary structure in a non-coding region of the Drosophila genome[81] to estimate of the rate of evolution in RNA base-pairing regions, which they find to be roughly two times lower than the background rate[44].

In a prior work, I collaborated with Jen Knies, Todd Vision, Noah Hoffman, Ron Swanstrom and Christina Burch to test the idea that evolution in RNA structures occurs by fixation of double mutants, a process known as compensatory evolution[48, 14]. We predicted that due to compensatory evolution we should observe an elevated ratio of transition to transversion mutation rates in stem-forming regions of structures. In a stem region base-pair, a transition mutation can be compensated only by another transition, whereas transversions require compensation by transversions. Since transitions occur more frequently than transversions[110], compensation of transversion mutations is expected to be a rare event compared to transitions. Thus, the ratio of the two types of mutations will be exaggerated in stem regions. Specifically, the ratio in stems will be the square of the ratio in loops. We tested this prediction in ten structures, including mitochondrial tRNAs, rRNAs and viral structures such as the HIV Rev response element, the hepatitis C virus cis-acting replication element, and the pestivirus internal ribosome entry site. Our prediction for an elevated transition-transversion rate ratio in stems was supported for nine of ten structures[51]. We concluded that compensatory evolution does operate in RNA structures and that it produces a specific, quantitative effect on molecular

evolution, consistent with the idea that maintenance of RNA structures constrains evolution at participating sites. In chapter 2, I follow up on this project by examining the interaction of compensatory evolution and host-induced hypermutation in the HIV-1 genome, the one example from the prior study that did not agree with our predicted result for structures experiencing compensatory evolution.

1.4 Overlapping coding regions and multiple constraints

Many viruses, including HIV, have very small genomes, possibly limited by their error-prone polymerase[41]. Some theorize that this limitation on length has led to the development of a more compact genome, by overlapping coding for proteins, RNAs, and other elements[42]. Overlapping may confer another advantage: if the messages overlap such that particular sites are of high importance in both, this reduces the number of sensitive targets that can be affected by mutation, particularly when mutation rate is high[87]. In a computational simulation study, Hogeweg and Hesper showed that multiple coding preferentially evolved in situations with medium-to-high mutation rates, and particularly those with recombination[40], situations applicable to RNA viruses. A recent survey of RNA viruses found that amount of gene overlap is significantly inversely correlated with genome size[5]. Regardless of the origin, the fact remains that coding overlap does occur, in viruses, bacteria[46] and eukaryotes[64], creating multiple constraints on the evolution of a single position. Huynen simulated evolution of RNA structures and found that dual protein-RNA coding still allows for evolution of differing structures, but reduces the ability to fine-tune structural evolution[42].

1.5 Prediction of RNA secondary structure in coding regions

Modeling techniques for molecular evolution have been developed to address among-site rate heterogeneity[116]. Typically, when examining nucleotide sequences, such variability is modeled using a gamma distribution. Newer methods implicitly acknowledge the presence of multiple coding by modeling variation in synonymous rates of coding sequences[54], suggesting that other forces may be constraining these sites that were previously thought to evolve neutrally.

And at least one RNA prediction method has been developed for finding structures in multiple coding regions[83]. This method uses several evolutionary models to account for the different combinations of constraints placed on protein and RNA dual-coding sequences[82]. In chapter 3, I use this method to search for new RNA structures in the (mostly protein-coding) genomes of HIV-1 and HIV-2.

1.6 Evidence suggests RNA structures are robust

In Huynen's simulation study of RNA structure evolution, structures evolved with selection for a specific secondary structure were more similar to their neighbor structures, (structures that differ from the evolved structure by point mutations), than were random structures. This means that the structures evolved to a region in sequence space where mutations have less of an effect on structure, allowing it to maintain a more-or-less constant phenotype. This property, phenotypic constancy in the face of mutational perturbation, is commonly referred to as robustness. In biological terms, we can think of robustness as an organism's ability to maintain a constant level of performance despite genetic or environmental perturbations. For example, many genetic sequences produce proteins that assume similar tertiary structures with equal functionality despite their mutational differences. Robustness can be achieved through several mechanisms (reviewed in [49] and [107]), including redundancy and as an emergent system property. The advantages of robustness are self-evident: an organism that can withstand perturbations while maintaining a high fitness is more likely to survive. This advantage can also be a disadvantage. If an organism is unable to change its phenotype in response to environmental or genetic pressure, it may be unable to adapt to new conditions. Recently, many studies have claimed to show evidence of robustness in various biological and theoretical systems (for a review, see [22, 49]), some suggesting that robustness is an evolved property of these systems. In particular, several studies suggest that conserved RNA structures are robust compared to non-conserved[108, 9] or artificially selected[74] structures.

1.7 How did robustness evolve?

Given that many biological phenotypes, such as RNA secondary structures, may exhibit evolved robustness, how did natural selection produce this property? Since organisms are thought to experience more environmental than genetic perturbations during their lifetimes, it seems likely that they will evolve to be robust to environmental changes[33]. Indeed, theory[109] supports the evolution of environmental robustness by selection. The origins of genetic robustness are less clear. One possibility for the evolution of genetic robustness, considered extensively by G. Wagner[109], is that it will increase as a direct result of selection for it. This is the "adaptive" hypothesis discussed by de Visser et al[22]. As argued by G. Wagner, in order for genetic robustness to evolve directly, there must be sufficient genetic variation present in the population that will favor selection for robustness to the potential phenotypic changes caused by the genetic variation. Otherwise, the benefits conferred by genetic robustness will be of little selective value. G. Wagner showed that it requires more genetic variation than is maintained in a population at mutation-selection balance to evolve genetic robustness. Flatt[29] argues that other sources – environmental heterogeneity, pleiotropy, epistasis, and heterosis – can maintain the necessary genetic diversity that will increase selection for genetic robustness, but he concludes that stabilizing selection is too weak for evolution of genetic robustness.

Typically, studies of robustness evolution are conducted using theoretical models. Laboratory experiments of robustness can be difficult due to the need to obtain a good fitness measure and to implement multiple different perturbations and measure their effects. Many of the theoretical studies use RNA structures and sequences, since this is one system where the relationship between genotype, phenotype, and fitness is known (with some caveats) and can be calculated, assuming that the lowest free energy structure is the most fit one, and that it is the one assumed by the sequence.

I participated in a prior computational study that demonstrated that sexual recombination can select for robustness, even if the mutation rate is very low[3]. I wrote computer code to implement a test of the idea that mutations that eventually become fixed in sexually-recombining populations are less deleterious to the population that exists at the time of their occurrence than are those that fix in asexual populations. Thus, populations that sexually recombine will

retain those mutations that are least likely to have negative interactions with other genetic elements, leading to a robust population. Another mechanism has been proposed for the evolution of genetic robustness which suggests genetic robustness is a correlated response to selection for environmental robustness[109, 33]. This is de Visser et al's "congruent" hypothesis[22], which requires that environmental and genetic robustness are achieved through the same biological mechanisms. That is, if an organism is robust to an environmental perturbation such as temperature, the same methods of protecting it from thermal changes will also protect it from mutations[70, 11]. If the same mechanisms that respond to environmental perturbations also confer genetic robustness, there is no need to postulate a separate theory of evolution of genetic robustness. I explore this hypothesis in chapter 4, using data from three different theoretical model systems, including an RNA sequence-structure simulation.

Chapter 2

Compensatory evolution, hypermutation, and RNA secondary structure in HIV

2.1 Abstract

The mammalian protein family APOBEC3 is known to cause elevated rates of Guanine-to-Adenine mutations in retroviral genomes. This activity is thought to be a defense mechanism evolved to protect host genomes against pervasive endogenous retroviruses. The extent of its action against HIV in vivo is a matter of some debate, as is the pattern of hypermutation it creates in the HIV genome. Here I show indirect evidence in support of the action of APOBEC3 proteins against HIV-1 *in vivo* by examining patterns of evolution in regions of conserved RNA structure. In a previous work, we demonstrated that the presence of conserved RNA structure predictably constrains molecular evolution, except in the RRE structure in the HIV-1 genome. Given known characteristics of APOBEC3 protein activity – that it induces transition mutations only in single-stranded regions of nucleic acid sequence – I can explain deviation from expected patterns of molecular evolution for regions of RNA structure such as the RRE. The amount of deviation for a particular RNA structure indicates the amount of hypermutation present in that region of the genome. I examined several known and predicted RNA structures in HIV-1 for adherence to the pattern predicted for RNA structures. I also extended my analysis to look at elevation of transition-transversion rate ratios across the whole genome. The results support the action of APOBEC3 proteins against HIV-1 sequences, and suggest the presence of two patterns of hypermutation in the genome.

2.2 Introduction

There are many sources of variation in the rate of evolution across the HIV genome. Some sources are genome-wide, such as the differing level of conservation among nucleotides in a codon. Others are restricted to certain regions, such as in the variable regions in the *env* gene, where positive selection leads to high rates of evolution due to their roles interacting with host defenses. Selection on protein sequence is the most obvious source of variation in molecular evolution, but selection can also act specifically on the nucleotide sequence to conserve regulatory regions, RNA secondary structures, or other important functional nucleotides. For example, the polypurine tracts are conserved for their role in genome replication and the *gag-pro-pol* frameshift slippery sequence and RNA stem-loop structure are conserved for their role in gene regulation. Finally, some positions have an elevated rate of evolution because their underlying mutation rate in increased by APOBEC3G and APOBEC3F, two mammalian proteins that have been shown to mutate retroviral genomes[34, 20]. This chapter examines the interaction of two of these constraints: conservation of RNA secondary structure and APOBEC3-induced hypermutation.

I participated in a previous study[51] where we showed that the presence of conserved RNA structures predictably constrains molecular evolution in a variety of molecules and organisms in a specific way, leading to an increase in the transition-transversion mutation rate ratio. This prediction is derived from an assumption that evolution in RNA structures occurs through a compensatory process, whereby substitutions in stem regions occur in combinations of two, one on each side of the stem, in order to maintain the base-pair interaction[14]. The double mutant is thus selection-neutral, because it does not disrupt formation of the secondary structure. However, due to the chemical requirements of forming a base pair, both mutations must be either transitions (purine→purine or pyrimidine→pyrimidine) or transversions (purine→pyrimidine or pyrimidine→purine). Thus, any inherent difference in the rate of these types of mutations is exaggerated in RNA structure stem regions, where the mutations must occur in pairs. Specifically, we predicted that the ratio of the transition and transversion rates in stem regions should be the square of its value in loop regions, which should be unconstrained by structure. We found widespread adherence to the prediction, except in a well-known conserved RNA struc-

ture in HIV-1, the Rev response element. There we found the opposite result: the rate ratio was higher in loops compared to stems. This means that either transition mutations were elevated or transversion mutations were suppressed in loops compared to stems.

We speculated that this deviation from the prediction for compensatory evolution could be caused by APOBEC3 proteins, a mammalian protein family known to defend against retroviruses by deaminating cytidine residues in single-stranded minus-strand DNA during viral replication[34]. These residues are then complemented by adenine when the plus strand is synthesized, causing a G→A transition mutation in the genome. Since these proteins increase the number of transition mutations, and only operate on single-stranded (i.e. non-stem) DNA, their action could explain the increased transition-transversion rate ratio in loops compared to stems. This explanation assumes that genome regions conserved for RNA structure also fold in the single-stranded DNA, protecting the stem regions from deamination. Potentially then, both compensatory evolution of RNA structures and APOBEC3-induced hypermutation elevate the transition-transversion rate ratio, but the former works on stem positions, while the latter works on loops. This difference of effect should be most apparent in structures that are heavily hypermutated. If I know the distribution of hypermutation across the genome, I should be able to predict to what extent a structure in that region will deviate from the compensatory evolution model.

While it is known that certain APOBEC3 proteins (APOBEC3G and APOBEC3F) can edit the HIV genome, the pattern of hypermutation they create in the HIV genome is a matter of some debate. In a study of editing in wild-type and Vif-deleted infections of HOS.CD4.X4 cells, Yu and colleagues[117] found the highest amounts of editing in the *env* and *nef* regions of the Vif-deleted viruses (see Figure 1 for gene locations). The viral protein Vif is known to counteract the effects of APOBEC3, by prohibiting it from being packaged with the viral genome. In viruses containing a functional Vif, the editing was highest in the *pro* gene and the upstream two-thirds of the *env* gene. However, the intervening region was not analyzed. The authors proposed a single gradient of editing across the genome, with the amount of editing depending on the length of time the minus-strand DNA spends in a single-stranded state. Based largely on the results of the Vif-deleted viruses, they proposed that the *nef* region immediately upstream of the polypurine tract (PPT) should experience the most hypermutation. Though their study

did not examine this region, they noted that they also expected a drop in hypermutation immediately downstream of the central polypurine tract (cPPT).

These studies are complemented by several studies of sequences extracted from infected patients. Kieffer and colleagues[47] examined patterns of mutation from the *protease* and RT genes taken from resting CD4+ T cells of HAART patients, finding that each patient contained a minority of hypermutated genomes. The variability in editing consistent with APOBEC3G across the sequenced regions was large, with some sites showing G→A mutations in more than 80% of sequences, while the median fraction of sequences experiencing G→A mutation was 19%. Suspene et al[101] examined full-length genomes from several studies in two ways: by comparing the number of edited nucleotides per position in a hypermutated and reference sequence, and by calculating a product-substrate ratio for APOBEC3G/F target sites across the genome. For one HIV-1 group O hypermutated sequence, both types of analysis showed a distinct pattern of two mutation gradients across the genome, with peaks immediately upstream of both polypurine tracts. For a separate collection of 29 sequences, all obtained via a database search for hypermutated annotations, mostly consistent but less distinct results were obtained. In the most heavily mutated sequences, the *pro* and *pol* genes generally showed the greatest amount of editing, while editing was also high but less pronounced in the *env* and *nef* genes. The authors conclude theses results are consistent with dual origins of plus-strand DNA synthesis, one at each polypurine tract. Regions immediately upstream (relative to the plus strand) of the polypurine tracts are expected to be single-stranded and susceptible to APOBEC3G/F editing the longest. Generally consistent with these results were findings by Pace et al[78], who examined editing of sequences of varying length from 127 patients. They calculated several metrics of editing by comparing mutated positions to the population consensus. Considering only G→A mutations, regardless of di-nucleotide motif context, they found results consistent with a two-peak model, though the peak in the *env-nef* region was lower than for the *pro-pol* region. No peaks were observed when the metric was changed to look for AOBEC3G-specific target motif editing.

All of the above studies support elevated rates of APOBEC3G/F mutations in the *pro-pol* region, and three of them support elevated rates in the *env-nef* region (Keiffer's study did not include this portion of the genome). Both studies that examined the full genome found

a drop in APOBEC3-editing following the central polypurine tract, but the relative height of the peaks in the *pro-pol* vs. *env-nef* region are not clear. There seems to be general support for a correlation between amount of hypermutation and length of time a position spends as single-stranded DNA.

Here we measure adherence to the compensatory evolution model for RNA secondary structures across the HIV-1 genome in order to determine whether instances of high deviation correlate with regions of high hypermutation. We compare the results in unedited and hypermutated genomes for regions of known and predicted structure. We also look at the pattern of the transition-tranversion rate ratio across the whole genome. We discuss implications for the presence of conserved structures and for the pattern of APOBEC3 action across the genome.

2.3 Methods

2.3.1 Alignments and phylogenies

An alignment of non-recombinant group M subtype reference sequences[60] was obtained from the Los Alamos HIV database (`http://www.hiv.lanl.gov/content/index`). Minor manual editing was done to improve the alignment and to adjust it to match the numbering of the HXB2 reference sequence. To examine hypermutated sequences, I downloaded from GenBank all the HIV-1 group M sequences from a previous study[101], excluding circulating recombinant forms. I aligned them to the HXB2 reference sequence using the FFT-NS-2 method of MAFFT[45] and manual editing. This resulted in a total of 23 sequences in the hypermutated alignment, plus HXB2. Phylogenetic trees were inferred for each alignment using Tree-Puzzle[92], with the GTR+ γ (4) model (Figure 2). Accurate parameter estimation was used, with quartet sampling used for substitution process and the neighbor-joining tree used for rate variation.

2.3.2 Transition-transversion rate ratio in specific structures

The transition-transversion rate ratio, κ, was estimated for specific structures as described previously[51], with one modification for equilibrium frequencies. Briefly, the HKY85 model of nucleotide substitution was separately fit to stem and loop sites for each structure using HyPhy[55]. Given the expected differences in APOBEC3-induced mutation rates at stems and

loops, equilibrium nucleotide frequencies were calculated separately for them. Rate heterogeneity was modeled using a discretized gamma distribution and four rate categories. Confidence intervals are calculated from the Fisher information matrix by HyPhy, assuming asymptotic normality of the estimator. Stem and loop designations for each position were obtained from the highest-probability fold of the given region using the phylo-grammar (see chapter 3 for explanation of phylo-grammar predictions).

2.3.3 Sliding-window analysis of κ

The local transition-transversion rate ratio was estimated across the genome using a sliding window of 150 nucleotides and an increment of 30 nucleotides. κ was estimated separately for each window using a common phylogenetic tree and γ-distributed rate variation as described above. To estimate the rate ratio for non-stem positions only, gaps were created at all positions with a high probability of being a stem (i.e. pairing probability greater than 0.8 – see chapter 3). These columns are subsequently treated as missing information in the rest of the analysis. Due to the heterogeneous distribution of stem sites, the actual number of non-stem sites in each 150-nucleotide window used to estimate κ is variable.

2.4 Results

2.4.1 κ for individual structures in hypermutated and normal alignments

I chose seven regions of predicted structure in which to examine the transition-transversion rate ratio. They include two known structures: the Rev response element and the 5' non-coding region. Other structures were predicted as part of a whole-genome analysis of HIV-1 using a phylo-grammar (see chapter 3). I measured κ, the rate ratio, for each structure in two alignments: the reference sequences for the HIV-1 group M subtypes and a separate collection of group M hypermutated sequences that have been analyzed previously[101]. The rate ratios for paired and unpaired sites for all structures are shown in Figure 3. If the structures are unaffected by APOBEC3-induced hypermutation, I expect they will adhere to the prediction for compensatory evolution, $\kappa_{stems} = \kappa_{loops}^2$, which is indicated by the solid line in Figure 3. On the other hand, hypermutation should cause elevated values of κ in the loops, but not in

the stems, since double-stranded regions such as stems should be protected from APOBEC3 mutation. Thus, structures experiencing more hypermutation should be further to the right in Figure 3 compared to non-hypermutated structures, but no difference in the vertical direction is expected due to hypermutation.

No structure from either alignment has confidence intervals that overlap the expected value for sites undergoing compensatory evolution, shown by the solid line, although one hypermutated structure comes close. All of the structures fall to the right of this line, and several are well to the right of the dashed line, indicating an elevated κ_{loops} compared to κ_{stems}. The dashed line indicates $\kappa_{stems} = \kappa_{loops}$, a result expected when constraints on molecular evolution do not differ between sites identified as stems and loops. This could happen due to misidentification of some stem sites as loops, or vice versa. It could also result from incorrectly predicting a structure where none in fact exists. Thus, one or two of the structures that have both points falling very near this line are possibly not real structures, or some of their positions have been misidentified. In Figure 3b, the same data is shown, but with lines connecting the two alignments for the same structure. For visual clarity, confidence intervals are not shown. From this figure, it can be seen that four hypermutated structures are further to the right in the graph compared to the same structures measured in the reference sequences, indicating that they have higher κ_{loops}. These hypermutated structures also have elevated κ_{stems} compared to the reference sequences, but the difference is not as great as for κ_{loops}. The other three structures have higher κ_{stems} in hypermutated as opposed to reference sequences, but little difference is seen in κ_{loops}. Overall, some consistent trends are apparent, particularly that no structure meets the prediction for compensatory evolution, and several of the hypermutated structures are more deviant from the compensatory evolution prediction than the reference sequences.

Some prior studies of APOBEC3 action suggest that different regions of the genome experience different amounts of hypermutation as a result of the amount of time those regions spend as single-stranded DNA during replication[101, 117]. If this is so, I expect the structures in regions that are single-stranded the longest during replication will have the highest transition-transversion rate ratios for their loops compared to the other structures because they should have the most opportunity for hypermutation. In contrast, the (double-stranded) stem sites should be relatively unaffected by hypermutation. To examine whether there is a location-

specific effect on κ in loops, I plotted the κ values for loops and stems versus the position of their structure in the HIV-1 genome (Figure 4a). Given the results of prior studies, I expect the highest κ_{loops} values to be in the *pro-pol* region, followed by the *env-nef* region. Due to the effect of the central polypurine tract, I expect a dip in κ_{loops} immediately downstream of it. Observed values of κ_{loops} in both the reference sequences and the hypermutated sequences do appear elevated in the *pro-pol* region compared to the rest of the genome, and to a lesser extent in the *env-nef* region. Also, a dip in κ_{loops} consistent with the position of the central polypurine tract is observed. Interestingly, κ_{loops} in the last half of the genome is not particularly elevated compared to the rest of the genome, in contrast to what would be expected if hypermutation was most dense in this region, as suggested in[117]. Compared to κ_{loops}, κ_{stems} is less variable across the genome, particularly in the reference sequences, which is consistent with the hypothesis that stem regions are protected from APOBEC3 mutations.

A more precise characterization of the relationship between genome location and deviation of the transition-transversion rate ratios from the prediction for compensatory evolution is plotted in Figure 4b. The deviation, $\sqrt{\kappa_{stems}} - \kappa_{loops}$, quantifies how much κ_{loops} differs from its predicted value from the compensatory model described in[51], assuming that κ_{stems} is unaffected by hypermutation. A high deviation could be indicative of a large amount of APOBEC3-induced hypermutation, because it signifies an elevated rate of transitions in single-stranded regions, characteristics consistent with APOBEC3 mutation. The hypermutated sequences generally show equal or greater deviation than the reference sequences, which is consistent with their status as hypermutated. Deviation for both is greatest in the *pro-pol* region. There is also a peak in deviation at the *env-nef* region, though less so for the reference sequences. This suggests support for a double gradient of hypermutation peaking at each of the polypurine tracts.

2.4.2 κ across the whole genome

To observe the effect of hypermutation on all sites, whether in a structure or not, I performed a sliding-window analysis of κ across the HIV-1 genome for both alignments. κ was estimated for overlapping 150-nucleotide windows across the genome. The results are shown in Figure 5a. I compared the results for this analysis with a similar one where I first removed all likely stem

sites from the alignment by replacing those columns with gaps (Figure 5b). In both analyses, it is apparent that there are two distinct patterns of κ values, one represented by positions 1200-5000, and the other by positions 450-1200 and 5000-9600. In the former, κ is generally higher and more variable, with several high peaks. In the latter, κ is lower and nearly flat, with a few small peaks. In both graphs, κ is elevated for the hypermutated sequences compared to the reference sequences throughout the genome. This is likely an effect of genome-wide hypermutation in those sequences. The difference between the κ for the two alignments is generally greater in the upstream portion of the genome.

The peaks in κ may represent regions of RNA structure in the genome. The RRE is visible in both graphs as a modest peak between 7700-8000. Compensatory evolution would predict that this peak is a result of a higher κ_{stems} compared to κ_{loops}, but previous results and this study show that in fact $\kappa_{loops} \geq \kappa_{stems}$ in the RRE. Several other peaks in κ correspond to areas of known or predicted structure throughout the genome. The peak at ˜ 2000 corresponds to the location of the known *gag-pro-pol* frameshift signal. Peaks at (roughly) 1250, 7000, and 8500 match elevated regions of pairing probability (see chapter 3). As with the RRE, compensatory evolution would suggest that these peaks are visible due to their high κ_{stems}. However, the increase in κ for most of the peaks in 5(b) compared to 5(a) suggests that $\kappa_{loops} > \kappa_{stems}$ for these structures also. Since the only difference in the data between 5(a) and 5(b) is the removal of likely stem sites, the increase of κ in 5(b) suggests that the removed sites had lower κ. This would be the case if hypermutation was prevalent in the loops of these structures. Assuming, then, that a higher κ indicates a larger amount of hypermutation, it appears that loops experience more hypermutation than do non-structured regions. This can be seen, for example, by comparing the elevated κ for the RRE with baseline κ observed throughout most of the downstream half of the genome (Figure 5b). This suggests that loops are more accessible to APOBEC3 proteins than are non-structured regions.

2.5 Discussion

The results of our prior study[51] suggested that the Rev response element RNA structure did not evolve via compensatory evolution, since we did not observe the expected relationship be-

tween κ_{loops} and κ_{stems}. It now appears that this deviation was caused by APOBEC3-induced hypermutation, which greatly increased κ_{loops}. As the regions of highest deviation also correspond to places where APOBEC3 hypermutation is most prevalent (i.e. the *pro-pol* region), it is reasonable to conclude that the deviation is caused by elevated rates of mutations in the loops of structures, not by lack of adherence to compensatory evolution by the stems. Compensatory evolution likely does occur in these structures, and in places where hypermutation is low, their deviation from the prediction is small (Figure 4b). Both compensatory evolution of RNA structures and APOBEC3 hypermutation cause elevated transition-transversion rate ratios, but APOBEC3 has the stronger effect, as seen by the higher κ_{loops} compared to κ_{stems} for several of the individual structures (Figure 4a). The genome-wide κ analysis also supports this conclusion, since the most of the highest κ values occur after stem sites are removed (Figure 5b).

Several methods have been used to study the pattern of APOBEC3 hypermutation across the genome. Most involve counting the number of differences from a consensus sequence[78, 47] and/or calculating a ratio of observed target nucleotides or dinucleotides to product (di)nucleotides[101, 78]. In contrast, the method used here uses sequence relationships described by a phylogenetic tree to infer substitution rate parameters describing the evolution of the sequences. I use the HKY85 model of evolution[35], which allows two categories of rate – one for transitions and one for transversions – and estimates their relative rate ratio as a single parameter: κ. This model allows me to observe the effect of APOBEC3 on the parameter, because APOBEC3 increases transition rates, but not transversions. I could also use a model that estimates the rate parameters for all types of mutations separately, so that I could look at the rate of the specific mutation caused by APOBEC3, but estimation of more parameters also leads to more error in the estimation of each parameter. The HKY85 model is the simplest model that allows me to estimate a separate parameter for transition mutations and to estimate different equilibrium frequencies for each nucleotide. Compared to the methods that count differences from a consensus sequence, this method has the distinct advantage that it is easy to use. All that is required is a good alignment and phylogeny. It also takes advantage of the information in the phylogeny to estimate rates, rather than counts. However, the method is not good for precisely identifying hotspots of APOBEC3 mutation on a very small scale, since the estimation of κ becomes noisy as the window size is reduced.

My results for deviation from compensatory evolution (Figure 4b) and for κ across the genome (Figure 5) suggest a pattern for APOBEC3 activity with a broad peak in the *pro-pol* region, tapering off near the central polypurine tract. This is consistent with the above-mentioned studies that show elevation of in this region compared to the *gag* gene. The genome-wide analysis does not support a second broad peak in the *env-nef* region (suggested by [117, 101]). The study by Pace[78] also finds no support for a second peak in the *env-nef* region. My data do support isolated regions of elevated κ in this region, most corresponding to regions of structure (Figure 4a and Figure 5). The predominant hypothesis explaining the distribution of hypermutation across the genome is that the amount of APOBEC3 activity in a given genome region is proportional to the time the region spends as single-stranded DNA during replication[101, 117] (but see [23]). For my data to be consistent with this hypothesis, it must be assumed that the downstream region spends less time than the upstream region as single-stranded DNA, in contrast to the model proposed by Suspene[101]. One possible cause is if the upstream region contains more conserved secondary structure that takes longer to unfold for replication, though this is not likely given that structural predictions do not support substantially more structure here (see chapter 3). Another possibility is that the strand transfer of the plus-strand (strong stop) DNA that initiates at the polypurine tract is slow compared to initiation of replication at the central polypurine tract. This would create a proportional difference in timing of initiation or strand transfer that could leave the upstream segment single-stranded longer than the downstream segment.

Several peaks in κ in Figure 5 correspond to regions of known and predicted structure. It is interesting, then to note that loops appear to be more hypermutated than non-structured regions. This suggests that loops in conserved structures are more accessible to APOBEC3 than are non-structured regions, a reasonable possibility if the loops are projected away from the rest of the genome by stems and are therefore less sterically hindered substrates. Another possibility is that much of the minus-strand DNA forms transient, non-conserved structures that protect otherwise non-structured regions from hypermutation, while leaving the loops in conserved structures open. In our prior work, we suggested that κ could be used as a simple diagnostic for validating predicted structures, with regions of high κ indicating conserved structures due to the effect of compensatory evolution in stems. However, due to the complicating signal of

hypermutation, the elevated κ is likely a result of elevated κ_{loops}, rather than elevated κ_{stems}. Since κ_{loops} appears to be elevated in loops compared to non-structural regions, it may still be a useful diagnostic for structure in retroviral sequences that are affected by APOBEC3.

2.6 Acknowledgments

I thank Jen Knies and Jerry Jeffrey for discussions and analysis of preliminary data.

Chapter 3

Conserved RNA structures in HIV-1 and HIV-2

3.1 Abstract

Lentiviral genomes are known to contain several functional RNA structures, including the Rev Response Element (RRE), the highly-structured 5' non-coding region which includes the Tat Responsive Region (TAR), and the stem-loop ribosomal frameshift site within the overlapping region of the *gag* and *pro-pol* reading frames. These structures have been characterized in HIV-1 and to some extent in HIV-2 using thermodynamic or covariation predictions, chemical assays, and imaging methods. No strong evidence has been published supporting other comparable structures in the remainder of the genome. Here I use a phylogenetic stochastic grammar method to predict conserved structures throughout the HIV-1 and HIV-2 genomes. My results replicate the known structures, and suggest a handful of previously unknown ones. These results are robust to methodological parameter adjustments and some appear to be conserved between HIV-1 and HIV-2 and their nearest simian relatives (SIVcpz and SIVsm, respectively). The most significant difference between the HIV-1 and HIV-2 results is at the 5' end, where HIV-2 is predicted to contain a structure approximately twice the size of the corresponding structure in HIV-1. This 5' HIV-2 structure also overlaps the *gag* gene by approximately 400 nucleotides, compared to approximately 50 nucleotides of overlap into *gag* in HIV-1. Other conserved structures are found throughout the genome, but their functions are unknown. The results suggest that RNA structures are largely conserved across the primate lentiviruses, but

that some differences exist between the two human viruses. Also, genetic analysis of these new structures to determine function can now be undertaken.

3.2 Introduction

In recent years, as the number of known functional RNA structures has expanded, we have begun to ascribe to RNA a more prominent role in molecular biology. In viruses, RNA structures are known to function in genome packaging, regulation of transcript splicing, translation, and RNA transport, among others. Structures can also provide protection from RNase and, in the case of DNA structures formed by retroviral genomes, from the cytidine deaminase host defense (APOBEC) proteins. Sequences that produce these structures are not genes in the traditional sense, but the structures constitute functional phenotypes on which selection can act. In this sense, the structures and their corresponding sequences describe another genetic code, one that encodes RNA structures rather than polypeptides.

Most of the well-known RNA structures in viruses, such as the internal ribosome entry sites and packaging signals, are found in non-protein-coding regions, but some structures that exist in coding regions are known. Sequences that encode both an RNA structure and a protein represent an interesting group of sequences that have multiple constraints on their molecular evolution. Silent nucleotide substitutions that might otherwise be considered selection-neutral to the relevant coding region can be deleterious if they affect RNA structure folding. Indirect evidence for the presence of structures in coding regions has been inferred by studies demonstrating reduced rates or numbers of synonymous substitutions, for example in mRNAs in mammals[12] and in viral genomes[96]. Bioinformatics studies using co-variation and/or thermodynamic prediction have predicted structures in the coding regions of a variety of viruses, including hepatitis C and G viruses[103, 102], other Flaviviridae[102], and the Picornaviridae[114, 96]. Some of the predictions for HCV have been supported by enzymatic mapping[104] and by site-directed mutagenesis followed by studies of viral replication [69]. A separate, but similar study used site-directed mutagenesis and characterization of viral replication to show support for the role of a proposed stem-loop structure in regulating translation initiation in dengue virus[16].

Similarly, in retroviruses, most work has examined structure in the 5' non-coding region

(e.g. [4, 80, 7]), but two structures in coding regions are well known. The Rev response element (RRE), located in the *env* gene of lentiviral genomes, consists of approximately 350 nucleotides comprising a long stem-loop and several smaller stem-loops that branch off the main loop. It forms in unspliced and singly-spliced transcripts, is recognized and bound by the viral protein Rev, and the protein-RNA complex is then exported from the nucleus. Elements of the RRE structure in HIV-1 have been studied using NMR and crystallography (see [63] for review of the RRE in lentiviruses). In addition, the entire structure has been predicted using many computational/sequence analysis methods[38, 65, 59, 13]. Another known structure in a coding region is the *gag-pro-pol* frameshift stem-loop that causes the ribosome to slip from the *gag* reading frame to the *pro-pol* reading frame. Other studies have suggested the presence of additional structures within coding regions[39, 86], but none of these structures have have been consistently supported.

HIV-1 and HIV-2 are related primate lentiviruses resulting from transmissions of different simian immunodeficiency viruses (SIVcpz from chimpanzees and SIVsm from sooty mangabeys, respectively) into the human population (Figure 6). They share a similar genome organization with differences in the placements of the accessory genes (*vpr/vpu/vpx*) and approximately 50% nucleotide identity. Previous work has described differences in the RNA structures of the two viruses, again focusing mostly on the 5' non-coding region and the RRE. The TAR of HIV-2 is proposed to be a double-branched stem-loop, based on the evidence from several thermodynamic predictions and covariation analyses,[6, 8, 57], whereas the HIV-1 structure is unbranched. A recent study suggests that the TAR region of the HIV-2 transcript can assume more than one form, and proposed an alternative extended stem structure using a combination of enzymatic cleavage, gel shift mobility, and structural prediction techniques[79]. The HIV-2 dimerization initiation site (stem-loop 1), also located upstream of the *gag* gene, is proposed to be longer and to have more stem sections than its HIV-1 counterpart[56]. Differences in stem and loop lengths are also proposed for the stem-loop structure containing the primer binding site and the poly-A signal[8], but the overall structure of both is similar between HIV-1[80] and that predicted for HIV-2[8]. Predictions for the structure of the HIV-2 RRE also show overall similarity to the HIV-1 structure, though the organization and relative length of the branches is somewhat different[13, 59].

Here I compare predicted RNA structures derived from RNA-Decoder, a phylogenetic stochastic grammar, for the complete genomes of HIV-1 and HIV-2. This method builds on developments in structure prediction via machine learning methods (grammars) for single sequences by combining the method with evolutionary information in a multiple sequence alignment and phylogeny[82]. Its predictions are based both on sequence patterns and differences in evolutionary rates in base-paired versus non-base-paired regions. Essentially, regions of conserved RNA base-pairs are expected to evolve more slowly than non-pairing regions due to the necessity of maintaining the pairing interaction[44, 100, 82]. If positions with slower rates also show sequence patterns characteristic of those observed for base-paired regions in the training data set, such as the presence of matching base pairs, the method identifies these positions as stem positions with a high probability. To investigate the accuracy of the structural predictions, I also compare the results of RNA-Decoder to a chemical analysis of the entire HIV-1 genome[112] and note regions of congruence.

3.3 Materials and Methods

3.3.1 Grammar predictions of structure

I used the RNA-Decoder program, a phylo-grammar method, to identify regions of conserved structure in the HIV-1 and HIV-2 genomes. Grammars and evolutionary models have been previously combined into a single program called Pfold[52, 53]. An assumption of Pfold is that the primary constraint on the rates of evolution in the input sequence alignment is the presence or absence of RNA secondary structure; it does not take protein coding constraints into account. Nearly all (approx. 95%) of the HIV genome is protein-coding, and it is well-known that the nucleotides within a codon evolve at different rates, the fastest being the third position. In order to accurately observe rate differences between nucleotide positions that are due to RNA structure, RNA-Decoder takes into account the rate differences expected based on position within the codon. This method has been validated on HCV structures, where it was shown to have a significantly better signal to noise relationship than other methods, including thermodynamic methods, co-variation methods, and even Pfold. It has also been used to search for structures in the human genome[73, 84].

RNA-Decoder takes a multiple-sequence alignment, a phylogenetic tree, and grammar parameters as input. For this study, I used grammar parameters derived from a previous training on hepatitis C virus[82], tRNAs, and rRNAs. There are two possible outputs from RNA-Decoder depending on the mode in which it is run. Scanning experiments output the pairing probability for each nucleotide position, which is effectively the probability of that position being in a stem in any possible structure containing it, given the phylogenetic tree and the grammar structural model. More precisely, the pairing probability for position i in alignment D with the phylogenetic tree T is:

$$\sum_{k} P(\pi_i = k|M)P(D|\pi, T, M) \tag{3.1}$$

where k is over all stem structural labels (i.e. right-stem, left-stem), π is the structure, M is the grammar model parameters, and $P(\pi_i = k|M)$ is the posterior probability that position i has the specific structural label k, given the grammar, and is calculated via the inside-outside algorithm[24]. In Bayesian terms, $P(\pi|M)$ is the prior probability of structure π. $P(D|\pi, T, M)$ is the alignment probability, which is calculated using the Felsenstein algorithm[27]. Folding experiments output the single highest-probability fold for an input sequence, which is calculated via the CYK algorithm[24].

Using the scan mode, I performed a sliding window analysis of pairing probability across the entire genomes. In order to accommodate as many pairing interactions as possible, I used the largest scan window size the program would permit (1300 nucleotides), and spaced the scans at 300-nucleotide intervals. Pairing probabilities for each scan were combined using R into a prediction for the entire genome, for each site taking the maximum pairing probability of the multiple overlapping predictions for that site. I also used the fold mode to fold entire genomes in overlapping 800-nucleotide pieces. Folding and scanning jobs were executed on the University of North Carolina's Linux cluster. Accession numbers for all sequences used are listed in Table 1. I wrote MATLAB code to convert the multiple sequence alignment files into the specific input format required by RNA-Decoder, which is essentially a matrix transpose of the standard alignment layout.

All HIV-1 predictions were made using the alignment of non-recombinant group M subtype

reference sequences (excluding subtype G, which has lately been shown to be a recombinant[1]) obtained from the Los Alamos HIV database, with minor manual editing. The input file also requires that each nucleotide be assigned a number to indicate its position within a codon. Codon positions in overlapping genome regions were designated according to the first member of the following pairs: *gag-pro, pol-vif, vpr-vif, vpr-tat, rev-tat, env-vpu, env-tat2, env-rev2.* Given the differences in nucleotide content and evolution patterns in the two halves of the HIV genome, I scanned the genome in two sections, upstream and downstream, that overlapped in the *vif* gene. This allowed use of separate phylogenetic trees for each scan with branch lengths calculated according to the rates of evolution of those genome regions.

All HIV-2 predictions were made using a reduced version of the HIV-2 web alignment available from the Los Alamos site (downloaded Dec. 6, 2007). The alignment was adjusted to the coordinates of its reference sequence from SIVmac, MAC239 (GenBank accession: M33262), but the MAC239 sequence was not included in the analysis. Codon positions in overlapping regions were designated according to the reading frame of the first member of the following pairs: *gag-pro, pol-vif, vif-vpx, vpr-tat1, tat1-rev1, env-tat2, env-rev2, env-nef.* As for the HIV-1 genome, the alignment was scanned using separate trees for the upstream and downstream sections.

3.3.2 Phylogenies and evolutionary rate estimation

All phylogenetic trees were generated by Tree-Puzzle[92], using the GTR+γ (4) model, 10,000 puzzling steps, accurate parameter estimation, and other default settings. The phylogenetic tree for the first half of the HIV-1 genome was built on the third codon positions of the *gag*, *pro*, *pol*, and *vif* genes and the 5' non-coding region, while the tree for the second half was built on the third positions of *vif, vpr, rev, vpu, env,* and *nef* genes and the 3' non-coding region. For HIV-2, the upstream section tree was inferred from the third codon positions of the *gag*, *pro*, *pol*, and *vif* genes and the 5' non-coding region, while and the downstream tree was inferred from the third positions of the *vpx, vpr, tat1, env,* and *nef* genes and the 3' non-coding region. The topology of the trees for each genome did not significantly change between the upstream and downstream trees. All trees are shown in Figure 7. I characterized the amount and distribution of divergence in the phylogenetic trees using three measures: total tree length, median pairwise

distance, and maximum pairwise distance (Table 2). Total tree length is the sum of all branch lengths in the tree and was calculated using the BioPerl Tree module[98]. Median and maximum pairwise distances were calculated by HyPhy[55] using the branch lengths from the Tree-Puzzle trees.

I used HyPhy to estimate relative rates of evolution across the HIV-1 genome. Rates were estimated using the GTR model with beta-gamma rate distribution and eight categories.

3.3.3 Comparing grammar and chemical-thermodynamic predictions

I compared my results to a biochemical analysis of RNA structure in the HIV-1 genome[112] that annotates each sequence position as either stem or non-stem. Since the pairing probability output from RNA-Decoder is not binary, I chose a cutoff value to enable comparison of its results to those of the chemical-thermodynamic method for each position. I aligned the chemical-thermodynamic-derived structural annotation of the NL4-3 genome to the pairing probability results for the HXB2-based genome alignment used for the RNA-Decoder input. I used the R module ROCR[97] to plot the Matthews correlation coefficient across various cutoffs for the pairing probability:

$$\phi = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{3.2}$$

Instances of agreement between the two methods are considered true positives (TP) if both methods predict a stem and true negatives (TN) if loop. False positives (FP) are cases where RNA-Decoder predicts stem but the chemical-thermodynamic does not, while false negatives (FN) are the converse. I also examined agreement between the two methods across various cutoffs, where agreement is defined as the fraction of positions with matching structural designations between the two methods, but concluded that this measure was not appropriate for the dataset because it does not take into account the fractions of false positives and false negatives and thus gives a similar accuracy result across a wide range of cutoff values. I compared the results of the correlation analysis in the RRE and the 5' non-coding regions (Figure 8) for two sets of grammar parameters – a more conservative and a more liberal set. The more liberal set showed a significantly better correlation for the RRE than did the conservative set, while in

the 5' non-coding region there was little difference. I selected the liberal parameter set since most of the genome is protein-coding, as is the RRE, and the parameters that do best in the RRE region were expected to perform best in the rest of the coding parts of the genome. For the same reason, the cutoff value for calling a position a stem in RNA-Decoder (0.8) was also selected to be the value that gave the highest correlation in the RRE region.

3.4 Results

3.4.1 Conserved structures in the HIV-1 M Group predicted by RNA-Decoder

I examined conserved structures in the HIV-1 group M genomes by using RNA-Decoder to calculate the pairing probability for each nucleotide position (Figure 9, large panels). Rather than choose a single subtype to study, I included all of the non-recombinant subtypes of group M in order to derive structure predictions that were relevant for all subtypes. All positions in this alignment are numbered according to the HXB2 genome (accession number K03455). The pairing probability represents the probability of the given position being in a base-paired stem region of any possible structure containing the position. The effect of evolution rate on pairing probability can be seen by comparing the top and bottom panels of Figure 9. Since structure constraints are expected to slow evolution, regions of strong structure signal, such as the RRE, should have a low rate. The reverse pattern can also be observed, where fast-evolving regions have very low pairing probability.

RNA-Decoder's grammar transition parameters were previously trained on hepatitis C virus structures[82]. To determine whether those parameters were appropriate for HIV, I examined the pairing probability results for the most well-described structures in HIV-1: the 5' non-coding region, the RRE, and the *gag-pro-pol* frameshift site structure (see Table 3 for structure locations). These structures are effective positive controls because they represent a range of structure complexity: the simple stem-loop structure of the *gag-pro-pol* frameshift site, the series of stem-loops in the 5' non-coding region of the genome (including TAR), and the large, complex, 350-nucleotide RRE in *env*. Our results clearly show strong predictions for conserved structures in each of these regions, with average pairing probability greater than 0.6 across

all stem and loop sites for each structure. The highest-probability fold that is calculated by RNA-Decoder for the RRE is similar to other published predicted structures, and most base-pairs have a reliability (i.e. posterior probability) greater than 80% (Figure 10). These results suggest that the method parameters are appropriate for structure identification in HIV-1 and related viruses with similar mutation rates and patterns.

The RRE is the largest cohesive structure observed in the results for all the coding sequence. A search for other regions with a similar magnitude of pairing probability identified several regions. There are small predicted structural elements, particularly in the *pro* and *pol* genes, and a series of peaks at the end of *env* and the middle of the *nef* gene. There are also several regions where a distinct lack of structure is predicted, as evidenced by the several sharp peaks of low pairing probability. These typically correspond to regions of high evolution, including the *env* hyper-variable regions 1, 4, and 5 (located near positions 6600, 7400, and 7600, respectively). These regions could be selected to lack structure to increase their rate of evolution[58] or for other functional reasons, such as translation speed. The window size used for the genome scans was large enough to pick up all pairing interactions within 1000 nucleotides of each other, with the exception of pseudoknots, which are not predicted by RNA-Decoder. These results suggest that structures in the coding region are either smaller than the RRE or have interactions that span more than 1000 nucleotides and thus are not detected.

To get an idea of the type of structures represented by the regions of elevated pairing probability, I examined the highest-probability structure reported by RNA-Decoder for several regions (Figure 10). Most predictions are typically for an extended stem-loop, with a base stem of at least two nucleotides and including a few bulges or smaller, branching stem-loops. Some predictions show stems with as few as two base-pairs, but these are unlikely to be thermodynamically stable and probably do not form *in vivo*. Although RNA-Decoder is clearly capable of predicting structures as large as the RRE, the highest-confidence new folds are much smaller structures.

3.4.2 Structure conservation across phylogenetic distance

To determine how well the structure signals described above were conserved across larger phylogenetic distances, I performed a similar genome scan of pairing probability using an alignment

containing HIV-1 group N and O genomes and SIVcpz genomes. The pairing probability for this scan and the difference between the HIV-1 group M scan are reported in Figure 11. The signals for the well-known structures, the 5' non-coding region, the frameshift, and the RRE, are markedly diminished compared to the results for the HIV-1 group M sequences. Previous work on RNA-Decoder[82] and Pfold[53] demonstrated that the prediction accuracy decreases with phylogenetic distance, so it was not unexpected that the signals were reduced. In fact, the distance between some sequences exceeds 1.0 substitution per site, so the results for this alignment may be compromised by evolutionary saturation. In spite of this, some common elements between the two sets of results can be observed. These are places where the difference plot shows a value of zero and include some of the unstructured regions observed previously, such as positions near 1100-1200 and 7250-7300 as well as some in variable regions at 6600-6700 and 7400-7500. Interestingly, the region near positions 8500-8600, which gave a weak conserved structure signal for group M, shows a distinct lack of structure when the more distant sequences are added. This suggests a possible region of group M-specific structure, or, conversely, that this region is selected to lack structure in the more distant sequences. Some of the other regions of predicted new structures described above for group M do show weak signals here also, but with a little variation in the exact positions that could be due to structural evolution.

3.4.3 Comparison with chemical mapping techniques

I compared the structural predictions generated by RNA-Decoder for the HIV-1 genome with those produced by a chemical technique that assesses reactivity of single nucleotides with 1-methyl-7-nitroisotoic anhydride. Previous studies have shown that this method effectively distinguishes between single- and double-stranded (or base-paired) nucleotides due to the difference in reactivity of the 2'-hydroxyl group of the ribose: single-stranded (unpaired) nucleotides are substantially reactive compared to paired nucleotides[71]. These reactivity values are incorporated into a thermodynamic-based structure prediction program[67] to create a prediction for a single genome. I compare my results to the chemical-thermodynamic structure prediction for the NL4-3 HIV-1 genome (GenBank accession AF070521)[112]. The output of the chemical-thermodynamic prediction method is a structural annotation for each position that indicates its pairing state (paired or unpaired) and its pairing partner, if applicable. The output can be

thought of as a binary indicator of whether each position is in a stem. I assessed the correlation between the two methods by performing a sliding window analysis across the entire genome. In each window, the correlation between structural annotations of the two methods is reported (Figure 12).

In known structures, the methods agree well. In the 5' non-coding region, which includes the TAR stem-loop and several other heavily studied stem-loop structures, RNA-Decoder agrees with the chemical-thermodynamic annotations for 80% of the total positions and the correlation is 0.55 (Table 3). A previous study using Pfold, a predecessor of RNA-Decoder, also showed good predictions in this non-coding region[21]. The methods also agree well in the two well-known structures in the coding area of the genome: the RRE and the *gag-pro-pol* frameshift signal. RNA-Decoder agrees with chemical-thermodynamic annotations for 79% and 100%, respectively, of the positions in those structures, with correlations of 0.53 and 1.0 (Table 3).

There are other regions of good correlation throughout the remainder of the genome, though none approach the size of the RRE. The size of most regions of high correlation suggests the presence of isolated stem-loop structures, rather than branched, multi-stem structures such as the RRE. In order to identify which regions have the strongest joint prediction between the two methods, I chose a threshold correlation value and looked for clusters of three or more adjacent points (consisting of 30 or more nucleotides) falling at or above the threshold. The threshold (correlation ≥ 0.5, the blue dashed line in Figure 12) was chosen to be slightly below the correlation observed in the known structural regions (i.e. the RRE, frameshift, and 5' non-coding region). This analysis identifies several regions that both methods predict to contain some structure (Table 3). These regions comprise about 10% of the genome. Those falling below the threshold, but having a positive correlation, comprise another 45%. Another type of agreement is in regions that both methods predict to lack structure. These are identified by the lack of signal from both methods as well as lack of a correlation value. By definition, the correlation is undefined when there are no positive (i.e. stem) predictions from either method. The largest jointly-predicted unstructured regions are located near positions: 850-900, 4700, 4850, 6600-6700, 7250-7300. Regions of undefined correlation comprise 14% of the genome, but they include both regions of agreement and disagreement between the methods.

In regions where both methods predict structure, I give strong confidence to the joint pre-

diction. The two prediction methods use completely independent sources of information to make their predictions, so the overlapping of their results is a strong indicator. Regions of disagreement fall into two categories based on which method predicts presence and which method predicts absence of structure. The regions of strongest disagreement include positions near: 5050, 5350, 6100, 7050, 8250, 8500, 9450. Some of them are identified in Figure 12 as points where the correlation is less than zero, which represents a worse-than-random correlation. They represent 32% of the genome. Another portion of them are in the 14% of the genome mentioned above where the correlation could not be calculated due to lack of positive predictions from one of the methods. In cases where the chemical-thermodynamic method predicts structure but RNA-Decoder does not, it is likely that the structure predicted by the chemical-thermodynamic method exists, but is not a conserved, functional structure. In the opposite case, when RNA-Decoder predicts a structure not observed by chemical-thermodynamic, there are two likely explanations: (1) the structure is conserved, but forms only in a specific step in the viral life cycle, or (2) the prediction by RNA-Decoder is incorrect. Predictions by RNA-Decoder are dependent on rates of evolution at individual sites, since base-paired sites are expected to evolve more slowly than non-paired sites. Although the method takes constraints of protein-coding into account, places of very strong coding conservation can give a false signal to RNA-Decoder, as can constraints on primary sequence for reasons other than protein coding or RNA structure. An outstanding example of a likely false positive by RNA-Decoder are the strong peaks in pairing probability between 4600-4700 and at 4800. These are likely a result of the sequence conservation of the cPPT region at approximately 4790, which is conserved for its role in replication. Since this type of sequence conservation is not accounted for by protein coding, RNA-Decoder assumes it is caused by conservation of RNA structure and makes a false prediction.

3.4.4 Structure predictions for HIV-2 using RNA-Decoder

I conducted a similar genome scan across the HIV-2 genome using an input alignment containing sequences from four subtypes. All positions are numbered according to the MAC239 reference sequence. While HIV-1 has spread globally and diversified into at least nine recognized subtypes, HIV-2 has fewer subtypes and a more restricted geographic spread[85]. However, its

subtypes are analogous to HIV-1 groups, since both likely originated from separate simian-to-human transmission events. Thus, the HIV-2 input alignment, while having fewer sequences and therefore less total tree length than either HIV-1 alignment (Table 2), contains enough sequence diversity to give a pairing probability signal equal in magnitude to the HIV-1 result in the RRE and the 5' non-coding region. Adding sequences from SIVsm and SIVmac, the most closely related SIV species, did little to change the pairing probability signal (data not shown).

A comparison of the smoothed pairing probabilities for both HIV-1 and HIV-2 is shown in Figure 13. Except for the RRE and 5' non-coding regions, the HIV-2 data has a weaker overall signal compared to HIV-1 in that fewer peaks have pairing probability greater than 50%. Fewer full-length sequences of different subtypes are available for HIV-2, so the input alignment may not sample the sequence diversity as broadly across the genome as does the HIV-1 alignment. One of the longest congruences between the two signals is in the 5' non-coding-region, where the signals are well-matched for most of the 400 nucleotides. This is immediately followed by one of the largest regions of incongruence: in HIV-1, the pairing probability drops for about 300 nucleotides after the beginning of the *gag* gene while the HIV-2 signal remains high for about 400 nucleotides after the beginning of *gag* and then falls to a sharp valley of low pairing probability. Both signals have a peak at the frameshift signal, though the HIV-1 peak is much sharper. Both also show a strong, extended peak for the RRE of approximately the same magnitude and length. There is also a good match in signal though most of the *env* gene upstream of the RRE, with the exception of positions 7000-7200 (MAC239 numbering), where HIV-2 has a strong, somewhat extended peak that is absent in HIV-1. Lack of structure is suggested by both signals for much of the middle of the *pol* gene (3800-4900 MAC239).

The pairing probability signals of HIV-1 and HIV-2 differ throughout most of gag. In addition to those differences mentioned above, there is also a distinct flat region of no signal in HIV-2 at 1800-1900 (MAC239) that is not echoed in the HIV-1 data. In the *pro-pol* genes there is some correspondence of peaks and valleys, but in particular there does not seem to be support in HIV-2 for the novel structure region predicted in HIV-1 at 3200-3300 (HXB2). The very sharp, tall peak in in the cPPT region of HIV-1 is much diminished in HIV-2, though the downstream adjacent peaks are of similar magnitude and extent. Signal agreement is particularly hard to characterize in the accessory gene region between *pol* and *env* and also in the *nef* gene because

the genomes have different arrangements in these areas. In HIV-1, the *env* and *nef* genes do not overlap, while there is an 167-nt overlap in HIV-2. Thus, the sharp peak in pairing probability between positions 9000-9100 (MAC239) is quite different from the HIV-1 signal in *env* at the corresponding location (approx. 8650 HXB2), but is just slightly offset from a similar peak in the HIV-1 signal for *nef* at the corresponding location (approx. 8750 HXB2). The signal on either side of the peak is quite low compared to the corresponding regions in the HIV-1 *env* gene, where there are several consecutive peaks. The difference in the signals at the end of the *env* gene could be due to the difference in evolutionary constraint in the two sequences: the HIV-1 sequence encodes only one protein here, compared to HIV-2 which has overlapping coding regions. The HIV-1 sequence thus has a reduced constraint compared to HIV-2, and is more able to also encode a functional secondary structure. This should also be the case in the first part of the *nef* gene, and could explain the larger region of elevated pairing probability in HIV-1 compared to HIV-2 in the 5' part of *nef*.

3.5 Discussion

In the HIV-1 genome, no new structures that approach the size of the RRE are predicted in the coding regions by RNA-Decoder. However, several smaller structures are predicted, including one each in *gag*, *pro*, *pol*, *vif*, and *nef*, while *env* possibly contains two. A few of these same structures are also supported by a chemical-thermodynamic assay of nucleotide reactivity. Most of these new structures are not strongly predicted at the corresponding locations in HIV-2, except for the structure in *vif* and possibly one in *env*. A small structure region is predicted upstream of RRE in HIV-2 that is not predicted for HIV-1. The largest difference between the two genomes is the large region of structure that continues from the 5' non-coding region into the first 350 nucleotides of the *gag* gene. Figure 14 summarizes the structure predictions for each genome, including those that agree with the chemical-thermodynamic method.

I have compared the predictions of a phylo-grammar to those of a chemical-thermodynamic method. Each uses different information to make structural predictions, so their predictions are entirely independent of each other (i.e. $P(A|B) = P(A)$). The chemical-thermodynamic method uses the chemical reactivity of each nucleotide in a single sequence to constrain the

thermodynamic folding process. It predicts both large and small known HIV-1 structures, such as the RRE and the frameshift signal, but characterizes much of the genome as large independently-folding semi-structured regions. Its predictions are like a snapshot of one moment in the viral life cycle, and may not find structures that form at another step, such as in particular spliced transcripts. In contrast, the phylo-grammar method predicts conserved structures that in theory can be functional at any time of the virus life cycle. These conserved structures are identified by differences among nucleotides in rate of evolution and the presence of sequence patterns consistent with base-pairing. This method uses pre-estimated evolutionary and structural model parameters, which are derived from known hepatitis C virus structures, tRNAs and rRNAs. Like all machine learning methods, the quality of its predictions is dependent on how well the training dataset (i.e. the hepatitis C virus structures, rRNAs and tRNAs) represents the testing dataset, or in this case, the HIV-1 and HIV-2 structures. Since the method finds the RRE, the frameshift and the structures in the 5' non-coding region with high confidence, I conclude that the method's parameters are appropriate for finding other, similar structures that might exist elsewhere in the genomes of HIV-1 and HIV-2. However, it may be weak at finding structures that are substantially different from the known structures, such as the large folded regions predicted by the chemical-thermodynamic method. Possibly some of the large folded regions predicted by the chemical-thermodynamic method are conserved to some extent, but exhibit very different folding patterns such as absence of a main stem and very long-range pairing that are not characteristic of the grammar's training set. Most of the high-confidence structures RNA-Decoder identifies are smaller, isolated structures with a main stem, features that also describe the RRE, frameshift, and 5' non-coding structures. Thus the strongest correspondence between the two methods might be expected in structures that share these features.

I examined conservation of RNA structure across a range of phylogenetic distances (Table 2 and Figure 7). In the alignment containing HIV-1 groups M, N, O and SIVcpz sequences, which was the most divergent alignment, the pairing probability signal was weak in areas of known structure (RRE, frameshift, 5' non-coding), compared to the less-divergent HIV-1 group M alignment. The decline in signal strength could be due to group-specific structural evolution. For example, if selection acts differently on the structures in group M compared to

group O, they will not all fold into a single high-confidence structure, nor will the constraints on rate be the same for each position. This will result in a weaker pairing probability for positions with conflicting roles in the different groups. Given the conservation of the RRE across species[63], it seems unlikely that major differences in the RRE would occur between HIV-1 groups, causing such a reduction in signal. Another possibility is rate saturation in the phylogenetic tree. Trees are inferred from third-codon positions in order to calibrate the relative rates of other positions to the fastest rate[82]. If the amount of divergence present in the alignment exceeds one substitution per site, then divergence estimates become inaccurate because multiple substitutions cannot be accounted for. This leads to inaccurate rate estimates for positions, which can also give weaker signal. Since the maximum pairwise distance in the most divergent HIV-1 alignment exceeds one, this is the more likely explanation for the weaker signal observed in that alignment. This explanation is consistent with previous results shown for RNA-Decoder where the prediction error increases with increasing total tree length[82], though total tree length does not correspond directly with divergence. Another argument against group-specific structural evolution is the strength of signal observed in the HIV-2 alignment. Even when SIVsm and SIVmac sequences were included (data not shown), the signal in the RRE region for this alignment was comparable to that observed in the HIV-1 group M alignment, which suggests conservation of structure across the human-simian transmission. Future studies should then focus on sequence alignments with maximum divergence less than one substitution per site to avoid saturation and false predictions. This could be accomplished for SIVcpz sequences by excluding very distant (or all) HIV-1 sequences from the alignment.

In contrast to the most divergent HIV-1 alignment, the HIV-2 alignment produced equal signal magnitude in the known structure regions of the RRE and the 5' non-coding region. The signal at the frameshift site was somewhat diminished. However, few other regions in HIV-2 have high pairing probability, suggesting that either HIV-2 has fewer conserved structures in the rest of the genome or the input alignment was insufficiently diverse to sample rate changes across the whole genome, and therefore missed some structural predictions. Structural evolution is a third possibility; as mentioned above, there could be subtype-specific selection giving conflicting results per position. Since the signal strength improves somewhat in the *env* region, which is the most variable in the genome, it suggests that lack of sequence diversity might be the cause of

lower signal in other regions of the genome. As more full-length sequences from other subtypes of HIV-2 become available, the analysis could be repeated with an expanded alignment. Despite the diminished signal, it is still easy to discern regions of congruence and incongruence with the HIV-1 signal. Further investigations will be required to demonstrate the particular roles of any diverging structures in the biology of their respective virus.

3.6 Acknowledgments

I thank Joe Watts and Kevin Weeks for sharing the results of their chemical analysis of the HIV-1 genome with me.

Table 1: Sequences used.

Alignment	Accession numbers	Organism/Group/Subtype
HIV-1 group M	AF004885	subtype A
	AF069670	subtype A
	U51190	subtype A
	AF484509	subtype A
	AF286238	subtype A
	AF286237	subtype A
	K03455 (HXB2)	subtype B
	AY423387	subtype B
	AY173951	subtype B
	AY331295	subtype B
	U52953	subtype C
	U46016	subtype C
	AF067155	subtype C
	AY772699	subtype C
	K03454	subtype D
	AY371157	subtype D
	AY253311	subtype D
	U88824	subtype D
	AF077336	subtype F
	AF005494	subtype F
	AF075703	subtype F
	AJ249238	subtype F
	AY371158	subtype F
	AJ249236	subtype F
	AJ249237	subtype F
	AF377956	subtype F
	AF190127	subtype H
	AF190128	subtype H
	AF005496	subtype H
	AF082394	subtype J
	AF082395	subtype J
	AJ249235	subtype K
	AJ249239	subtype K
HIV-1 & SIVcpz	sequences from grp M alignment	group M
	AY532635	group N
	AJ006022	group N
	AJ271370	group N
	L20587	group O
	L20571	group O
	AY169810	group O
	AJ302647	group O
	U42720	SIVcpz
	AF103818	SIVcpz
	AF447763	SIVcpz

HIV-2	AY509259	subtype A
	AY509260	subtype A
	U38293	subtype A
	M30502 (BEN)	subtype A
	U22047	subtype A
	J04498	subtype A
	J04542	subtype A
	AF082339	subtype A
	Z48731	subtype A
	J03654	subtype A
	D00835	subtype A
	M15390	subtype A
	U27200	subtype B
	L07625	subtype B
	X61240	subtype B
	AB100245	subtype B
	AF208027	subtype G
	AY530889	subtype U

Table 2: Evolutionary information in sequence alignments.

	Number of sequences	Total tree length	Median pairwise distance	Maximum pairwise distance
HIV-1 group M reference subtypes (no G)	33	fh = 2.5 lh = 3.7	fh = 0.29 lh = 0.44	fh = 0.37 lh = 0.55
HIV-1 M, N, O and SIVcpz	43	fh = 5.8 lh = 9.2	fh = 0.31 lh = 0.46	fh = 1.44 lh = 2.39
HIV-2 subtypes A, B, G, U	18	fh = 2.6 lh = 3.1	fh = 0.44 lh = 0.55	fh = 0.83 lh = 0.96

Pairwise distance values are in terms of substitutions per site. “fh” = first half of genome; “lh” = last half of genome.

Table 3: Regions of agreeing structure predictions in HIV-1, using both RNA-Decoder and chemical-thermodynamic predictions.

Region (HXB2 numbering)	Name or coding context	Base paired nucleotides*/ region size	chemical-predicted base-paired nucleotides	Average pairing probability	Correlation** with chemical
466:797	5’ non-coding	174/332	194/332	0.6693123	0.5546707
1420:1450	*gag*	10/31	15/31	0.3513013	0.5746117
2099:2126	*gag-pro* frameshift	24/28	24/28	0.823658	1.0
2460:2510	*pro*	23/51	28/51	0.6450999	0.6630435
2550:2620	*pol*	20/71	31/71	0.5164635	0.711343
3220:3280	*pol*	32/61	31/61	0.5558255	0.8363837
6300:6330	*vpu-env*	13/31	18/31	0.4351924	0.5897436
7360-7410	*env*	16/51	25/51	0.4224268	0.6049816
7709:8061	RRE	190/353	229/353	0.6956017	0.5325899
8460:8490	*env/rev*	15/31	15/31	0.7314326	0.6125
9200:9240	*nef*	16/41	27/41	0.4692163	0.5760658

* = base-paired nucleotides calculated as number positions with pairing probability greater than the cutoff (0.8).

** Matthews correlation coefficient is calculated as described in Methods.

Chapter 4

Antagonistic pleiotropy plays a role in the congruent evolution of genetic robustness

4.1 Abstract[1]

Genetic robustness, the ability to produce high fitness phenotypes in the face of mutations, has been observed in many biological systems but the mechanism by which it evolves is still unclear. In general, genetic robustness is not expected to evolve by the direct action of natural selection, because there is too little standing genetic variation observed in natural populations to favor the evolution of a mechanism to mask that variation. As a result, evolutionary biologists have proposed the congruent hypothesis' that genetic robustness generally evolves as a correlated response to selection for environmental robustness (the ability to produce high fitness phenotypes in the face of a variable environment). However, the empirical support for this hypothesis is limited to a few specific molecules, like chaperone proteins, where the mechanism of both genetic and environmental robustness is known. We test the generality of the congruent hypothesis by monitoring the evolution of genetic robustness in computational models of regulatory networks, RNA secondary structure, and digital organisms while imposing either a constant or variable environment. Consistent with the congruent hypothesis, all models showed significant evolutionary responses in genetic robustness only in variable environments. We show that this common outcome resulted because the evolutionary response to environmental variability

[1]**This chapter has been submitted for publication with the following authors: Kristen K. Dang, Matt C. Cowperthwaite, Christina L. Burch.**

was to eliminate antagonistically pleiotropic phenotypes, phenotypes that were advantageous in some environments but costly in others. Genetic robustness increased, as a result, because fewer loci were used to encode the reduced number of phenotypes. Mutations in the remaining loci no longer affected fitness. Surprisingly, we did not observe the accumulation of greater unexpressed, or masked, genetic variation in populations that evolved higher genetic robustness. Combined, our results suggest that the congruent hypothesis for the evolution of genetic robustness is likely whenever the environment is variable, and illustrate that the accumulation of unexpressed genetic variation depends on the mechanism by which genetic robustness evolves.

4.2 Introduction

In any biological or engineered system, a property of interest is the system's ability to function properly when one or more components are damaged or altered. In biological terms, we can think of this property, termed robustness, as the organism's ability to produce a high fitness phenotype despite mutations (genetic robustness) or changes in the environment (environmental robustness). In biological systems, genetic and environmental robustness are manifest at every level of biological organization, from the organization of the genetic code to protein folding to developmental underpinnings of the phenotype [107]. Understanding the evolutionary origins of the robustness that characterizes biological systems has been a major goal of recent theoretical and empirical investigations [33].

Although the evolution of environmental robustness is well understood [109], the evolutionary origins of genetic robustness are less clear. Three hypotheses have been proposed to explain the evolution of genetic robustness [22]. The intrinsic hypothesis' posits that genetic robustness is an intrinsic property of high fitness phenotypes; the adaptive hypothesis' posits that genetic robustness evolves as a direct response to selection; and the congruent hypothesis' posits that genetic robustness evolves as a correlated response to selection acting on environmental robustness. Although at least some of the genetic robustness that characterizes biological systems is intrinsic[22], the focus of the current paper is to distinguish between adaptive and congruent origins of genetic robustness.

Disagreements over the origin of genetic robustness stem from the major difference be-

tween genetic and environmental robustness. Whereas environmental variation is abundant, the amount of genetic variation maintained in most natural populations is thought to be insufficient for natural selection to favor masking that variation. Thus, on largely theoretical grounds, the adaptive hypothesis is thought to be less likely than the congruent hypothesis [33, 70]. However, it has been difficult to confirm this view with empirical evidence from biological systems [70].

Experimental attempts to determine whether genetic robustness is an evolved property of biological systems generally fail because a critical experimental design requirement can not be met [22]. In particular, the demonstration that genetic robustness has evolved would require the observation of a long-term evolutionary process and the ability to compare ancestral and derived genotypes. In addition, demonstrating the mechanism by which genetic robustness evolved would require a comparison of evolution in experimental (selects for robustness) and control (does not select for robustness) conditions. Although comparative investigations of existing species are able to examine robustness differences that resulted from long-term evolutionary processes, they suffer from the inability to compare ancestral and derived genotypes, or to compare evolution in experimental and control conditions. In contrast, laboratory evolution experiments are able to compare ancestral and derived genotypes, and to compare the evolution of robustness in experimental and control environments, but usually suffer from an inability to monitor evolution over a sufficiently long time period, but see [75]. As a result, the only evidence of evolved genetic robustness comes from investigations of known robustness mechanisms, in which it is possible to infer the ancestral state and the proximate mechanism by which robustness evolved.

In particular, the two cases in which genetic robustness has been shown to be an evolved property of a biological system began with investigations of mechanisms known to confer robustness to high temperature environments. Naturally-occurring RNA secondary structures are more thermally stable and genetically robust than most alternative sequences that fold into the same structure [108, 74], and computational investigations confirm the existence of a genetic correlation between these two traits [2]. In addition, molecular chaperones such as hsp90 and GroEL ensure proper protein folding in the face of both high temperature and mutations [26]. Thus, the few examples in which genetic robustness has been shown to be an evolved property

of the system share the characteristic that the same mechanism that confers environmental robustness also confers genetic robustness, suggesting that the congruent mechanism operates at least in some cases.

However, it is not known whether the mechanistic link between environmental and genetic robustness in the known examples is characteristic only of these specific molecules, or whether it is a general characteristic of complex biological systems. Here we examine the general applicability of the congruent hypothesis for the evolution of genetic robustness using three computational models that differ widely in how their component parts combine to determine fitness models of gene regulatory networks, digital organisms, and RNA secondary structure. For each model, we conducted evolution experiments in silico to determine whether genetic robustness evolves as a correlated response to selection for environmental robustness. Although the proximate mechanisms governing robustness in these models may not perfectly capture biological reality, examination of these models should address whether the congruent hypothesis for the evolution of genetic robustness holds regardless of the proximate mechanism that underlies genetic robustness.

4.3 Experimental Design

As discussed above, testing experimentally whether and why complex biological systems evolve to become genetically robust is difficult because it requires the ability to conduct long-term evolution experiments, to compare ancestral and evolved genotypes, and to compare evolution in experimental and control conditions. Since these requirements are difficult to meet even in the fastest evolving biological model organisms (but see [75]), we followed the approach of several previous studies [61, 2, 94] and monitored the evolution of robustness in computational models of biological systems. We chose to examine three particular models gene networks, digital organisms, and RNA secondary structure because all three have been used in previous investigations of evolved robustness, and because the proximate mechanisms that govern robustness in the three models differ.

In each model, we determined whether genetic robustness evolves as a correlated response to selection for environmental robustness by monitoring the evolution of environmental and genetic

robustness during adaptation to a variable environment. Because we defined the nature and frequency of environmental and genetic perturbations during the adaptation, we could measure environmental and genetic robustness using a representative sample of the perturbations that evolved populations actually experienced. We compared the robustness of ancestral and evolved genotypes to confirm that the observed robustness was an evolved property of the models, and we compared the magnitude of evolved robustness in variable and control environments to distinguish between the congruent and direct hypotheses for the evolution of genetic robustness.

4.3.1 Regulatory network model

This model, first described by A. Wagner[106], consists of regulatory elements that jointly determine the expression states of R interacting genes. The model is implemented as an R x R matrix of weights, $\mathbf{w}$, whose elements $w(i,j)$ represent the effect of the gene j on the expression of gene i, and a vector of binary expression states, $\vec{s}$, whose elements $s(i)$ indicate whether the proteins are present, $s(i) = 1$, or absent, $s(i) = -1$. Each row $w(i)$ in the weight matrix is analogous to a promoter region for gene i. The initial expression state values and regulatory weights for a network are randomly generated to simulate a random initial environment ($\vec{s}$) and genotype ($\mathbf{w}$). Networks go through a process termed development in which the expression vector $\vec{s}$ changes in an iterative process according to the following equation:

$$\vec{s}_{t+1} = \frac{2}{1 + e^{-a(\mathbf{w} \cdot \vec{s}_t)}} - 1 \tag{4.1}$$

where $\vec{s_t}$ is the expression vector at iteration t, $\vec{s_t} = 0$ is the initial vector, and a is a scale constant. If the network obtains a stable expression – an invariant expression vector – within $t = 100$ iterations, it is considered viable and has the opportunity to reproduce in the next generation. During evolutionary simulations, networks experience mutations – random changes to values in the genotype matrix $\mathbf{w}$ – and environmental perturbations – random changes to values in the initial expression state vector $\vec{s}$ – at the beginning of the development period in each generation. Since all stable networks are equally viable, their fitness is determined by the fraction of stable offspring they produce[3]. Networks are thus selected for their ability to produce offspring that achieve stable expression states in the face of the perturbations –

environmental or genetic – that they experience during the evolutionary simulations.

4.3.2 Digital organisms

We used AVIDA, an artificial life software that has been used for several experiments regarding genome complexity and robustness (e.g. [61, 62, 113]). The software creates worlds populated by self-replicating computer programs that replicate many behaviors of living organisms without explicitly modeling chemical reactions. Genomes are modeled explicitly as a list of instructions, which work together when executed to perform logic functions. The digital organisms compete for central processing unit (CPU) time rather than food, and their environment can be set up to reward them with extra CPU time (or punish them by removing it) for executing particular logic functions. Fitness is a function of the organism's replication efficiency and the amount of CPU time it acquires by performing logic functions [62]. In AVIDA, the environment is defined by the CPU rewards associated with the various logic functions. We assigned positive reward values to seven logic functions, such that organisms performing these functions received increased CPU time. Mutations were implemented by replacing one of the genome's instructions with one of the other 25 possible instructions chosen at random, and environmental perturbations were implemented by decreasing one or more reward values, sometimes causing particular logic functions to be penalized rather than rewarded. AVIDA organisms are thus selected for their ability to replicate efficiently and maximize acquisition of CPU time in the face of the environmental and genetic perturbations.

4.3.3 RNA secondary structure

RNA secondary structure has been well developed as a model for investigating the evolution of environmental and genetic robustness. Empirical investigations indicate that natural RNA secondary structures are characterized by both environmental robustness (measured as thermodynamic stability) and genetic robustness (measured as the fraction of mutations that have no effect on the minimum free-energy structure) [108, 74], and computational investigations indicate that genetic robustness can evolve as a byproduct of selection for thermodynamic stability[2]. We built on the previous computational investigations, making modifications only to improve our ability to distinguish the congruent and direct hypotheses for the evolution of

genetic robustness, and to ensure that we could directly compare the results of the gene network, digital organisms, and RNA secondary structure models. Most importantly, we expanded the investigation to include a wider variety of secondary structure targets than had been examined previously.

Following the approach of Ancel and Fontana [2], we investigated the evolution of robustness in RNA secondary structures by monitoring evolution of populations of nucleotide sequences where fitness is based on an explicit biological model of RNA folding that compares the minimum free-energy structure of an RNA sequence to a target secondary structure [30]. In contrast to the gene network and digital organisms models, environmental perturbations were not modeled explicitly. Instead, selection for environmental robustness was imposed (or prevented) by selecting for thermostability of the minimum free-energy structure in addition to its structural similarity to the target (or selecting only for structural similarity to the target)[2]. Thus, environmental robustness was measured as thermostability, by estimating the fraction of time spent in the minimum free-energy structure, not as the average fitness following an environmental perturbation. To improve comparisons with the gene network and digital organisms models, we changed the measure of genetic robustness from the genetic neutrality metric used by Ancel and Fontana to the mean relative fitness (i.e. mean structural similarity) after a single mutation.

4.4 Results

4.4.1 Correlation between environmental and genetic robustness in selection-naive systems

For each model, we examined the correlation between environmental and genetic robustness among large samples of high fitness individuals, chosen randomly with respect to robustness. We generated samples of individuals that had a high fitness in the constant (or control) environment because selection for high fitness in the environment most often experienced is expected to be stronger than selection for robustness to either environmental or genetic perturbations. Basically, we assessed the correlation between environmental and genetic robustness among only those individuals that had a real possibility of contributing to adaptation. High fitness individuals were chosen in such a way that selection for environmental or genetic robustness should

not have affected their probability of being included in the sample. The sampled individuals were selection-naive in this sense.

We produced the random samples for each model as follows. For the networks model, we generated 10,000 random networks, requiring only that each network achieved a stable gene expression pattern. Randomly generated digital organisms rarely exhibit high fitness, thus, obtaining a sample of high fitness organisms from randomly generated genomes proved computationally prohibitive. Instead, we generated a sample of high fitness genomes by allowing 76 populations founded by the same starting genotype to adapt independently to a constant environment (no selection for environmental robustness) with a low mutation rate (no selection for genetic robustness) for 2.65 million updates. For the RNA secondary structure model, we used inverse folding (see Methods) to generate collections of molecules whose minimum free-energy structure matched one of 17 different target structures, but were otherwise random with respect to RNA sequence. These sequences were not selected based on thermostability, rather, they were selected only because they folded into a desired structure. We generated 2,000-9,000 sequences for each of the targets, depending on sequence length of the target.

In each model, we measured genetic robustness as the mean fitness of individuals carrying a single mutation relative to unmutated individuals. In the networks and digital organisms models, we measured environmental robustness in an analogous manner: as the mean fitness in perturbed environments relative to unperturbed environments. In the RNA model, environmental robustness was measured as the thermostability of the minimum free-energy structure. All three models showed a positive correlation between environmental and genetic robustness (Figure 15), and the correlation was statistically significant for both the gene network and RNA secondary structure models. The lack of statistical significance in the digital organisms model probably resulted both because the relationship between environmental and genetic robustness is weak in this model, and because we were computationally prevented from obtaining a larger sample of high fitness, but otherwise random, individuals. The strength of the relationship between environmental and genetic robustness varied substantially among the three models (digital organisms: Kendall's $\tau = 0.06$; gene networks: Kendall's $\tau = 0.14$; various RNA secondary structures, $0.60 <$ Pearson's r < 0.75). In addition, the average robustness exhibited by random high fitness individuals varied substantially. Gene networks exhibited high genetic and

environmental robustness, whereas digital organisms exhibited low genetic and environmental robustness. RNA secondary structures fell in the middle, at least in terms of genetic robustness. Because we used different measures, environmental robustness could not be directly compared.

4.4.2 Evolution experiments

To confirm that genetic robustness evolves in these models as a correlated response to selection for environmental robustness, we evolved pairs of experimental and control populations. Experimental populations evolved in an environment that imposed selection for environmental robustness (e.g. a variable environment) and control populations in an environment that did not impose such selection (e.g. a constant environment). For each model, we evolved 31-50 pairs of populations (see Methods), using as large a population size (N) as was computationally feasible for each model ($500 < N < 1000$). We imposed low per-genome mutation rates ($U = 1/N$) and asexual reproduction to limit direct selection for genetic robustness [109, 107, 3, 113]. Each population was evolved long enough to allow a response to selection. The experiment length varied between the models (range = 400-250,000 generations) and was chosen based on preliminary experiments that assessed the time needed for a response in each model (data not shown).

All models showed a significant increase in genetic and environmental robustness in the variable experimental environments (Figure 16). We can be sure that the evolved robustness resulted from selection for environmental, and not genetic, robustness because control populations differed from experimental populations only in the intensity of selection for environmental robustness (the intensity of selection for genetic robustness was identical), and none of the models showed a significant change in genetic or environmental robustness in the constant control environments (Figure 16). Thus, genetic robustness evolved in all three models only as a correlated response to selection for environmental robustness. Consequently, and as expected from the weak genetic correlations between environmental and genetic robustness in these models, the gene network and digital organisms models both showed a weaker response in genetic robustness than in environmental robustness (the analogous comparison in the RNA secondary structure model is not meaningful because environmental and genetic robustness are not measured on the same scale).

4.4.3 Proximate mechanisms

The AVIDA software provides the means for monitoring evolution of both the phenotype, i.e. the performance of individual logic functions, and the genotype, i.e. the contribution of each instruction (locus) in the genome to the production of phenotypes. We monitored both of these characteristics, and found that the populations in the variable environment evolved to perform fewer logic functions than those in the constant environment and to use fewer instructions to perform those functions (Figure 17). Recall that all populations were first subject to a preadaptation period in the constant environment. Populations that then continued to evolve in the constant environment retained all of the logic functions gained during the preadaptation. In contrast, populations that were then subject to evolution in the variable environment lost the ability to perform 0.64 logic functions, on average.

We assessed evolved changes in the number of instructions (loci) involved in the performance of logic functions by deleting each instruction in the genome, one at a time. In a genome of size 70, we found that the loss of logic functions in the variable environment was accompanied by the use of 3.2 fewer instructions (15.6 compared to 18.8) in the evolved variable-environment populations compared to their preadapted ancestor. We also examined genetic diversity in the populations by measuring the Hamming distance between all genotypes in each population. We found that the populations evolved in the variable environment had less genetic diversity than those in the constant environment (Figure 18).

4.5 Discussion

In this paper we present the results of evolutionary simulations designed to test the general relevance of the congruent hypothesis for the evolution of genetic robustness. We followed the approach of a number of recent studies and monitored the evolution of robustness in computational models. Most of these studies focused on unusual scenarios such as high mutation rates that are expected to impose direct selection for genetic robustness[113]. In contrast, we built on the work of Ancel and Fontana[2], and investigated whether environmental variability a common characteristic of natural environments was alone sufficient to produce genetic robustness. We imposed controlled conditions in which environmental variability was the only selective force

capable of producing evolved robustness and, thereby, demonstrated conclusively that genetic robustness evolved as a correlated response to this selection. The consistency of this result across three computational models that differ both in the extent to which they mimic biological reality and in the manner in which their component parts combine to determine fitness suggests that the congruent hypothesis contributes to the evolution of genetic robustness regardless of the mechanistic basis of environmental robustness.

Our first piece of evidence for the congruent evolution of genetic robustness highlights the advantage that investigations of computational models bring to the study of evolved robustness. We demonstrated a positive correlation between environmental and genetic robustness in a single trait (fitness) among a collection of high fitness genotypes in two of the three models and a weak non-significant one in the digital organisms model. The power of this demonstration comes from the abilities, when investigating computational models, to use the appropriate measure of genetic robustness (robustness to spontaneous mutation), and to measure robustness in collections of high fitness genotypes. The analogous demonstration has not been possible in biological systems because it is difficult to measure genetic robustness in a single genotype, much less in numerous genotypes. As a result, most studies of biological systems have investigated the correlation between environmental and mutational robustness among different traits within the same genotype [99, 89]. This approach minimizes the difficulty of obtaining measures of mutational robustness, but doesn't capture the parameter of interest: the correlation between environmental and mutational robustness in a single trait among different genotypes. The few previous studies of biological systems that examined the latter correlation compared wild type genotypes either to genotypes in which a known robustness mechanism had been knocked out (e.g. GroEL:[26] egfr: [25]), or to genotypes carrying unknown spontaneous deleterious mutations (e.g. [11]). Because the mutated genotypes in these studies were not the product of selection, the relevance of these comparisons for the evolution of genetic robustness is unclear[36].

By investigating genetic robustness in computational models, we were also better able to measure other aspects of the phenotype and genotype as populations evolved. In addition to measuring the average effect of mutations on fitness, we could accurately estimate the shape of the distribution of mutation effects on fitness (as in [18]), and we could measure the production

of key fitness-determining phenotypes. We discovered that monitoring evolution of the mutation effect distribution and the production of key phenotypes provided insight into the mechanistic basis of environmental and genetic robustness in each model.

The proximate mechanism by which environmental and genetic robustness jointly evolve in RNA secondary structure has been explored elsewhere[2]. Environmental robustness, or thermostability, is achieved by increasing the stability of stems in the target structure and/or by decreasing the stability of alternative folds. Essentially, thermostability is achieved by reducing the likelihood that stem nucleotides can find stable binding partners elsewhere in the RNA molecule. As a result, the minimum free energy structure becomes more stable than any alternate structure, and nucleotides that are not involved in binding interactions become less likely to affect the relative stability of the target structure, even when they are mutated. Thus, environmental variability selects for use of less of the genome to produce the desired phenotype, and genetic robustness emerges from the resulting increase in neutrality, the proportion of mutations that do not perturb the lowest free energy structure.

In AVIDA we found that environmental robustness was achieved by performing fewer logic functions, i.e. encoding fewer phenotypes. The reason that populations in variable environments evolved to perform fewer logic functions is intuitive. Although logic functions were rewarded in the constant environment, the amount of the reward varied among individuals in the variable environment, and was sometimes negative. That is, specific logic functions were sometimes penalized in the variable environment. The evolution of environmental robustness via the performance of fewer logic functions depended on the way in which we implemented genotype by environment interactions, and required that logic functions were advantageous in some environments but not others (data not shown). Because these environmentally robust populations used fewer genome commands to produce a smaller number of logic functions, they also exhibited genetic robustness. This situation is exactly analogous to what is observed in RNA secondary structure. Environmental variability selected for use of less of the genome to produce fitness-related phenotypes, and genetic robustness emerged from the resulting increase in neutrality.

In the artificial gene networks, the only way genetic robustness could evolve was through increasing neutrality. Therefore, the fact that environmental variability resulted in increasing

neutrality does not distinguish between alternative mechanisms for the congruent evolution of genetic robustness, but it does mean that environmental variability selected for use of less of the genome in the production of phenotypes and of fitness just as in the other models.

Although the three models are mechanistically different, a similar pattern emerged in each. Genetic robustness evolved by minimizing the number of loci used to encode phenotypes (or fitness). Thus, a major contributor to increasing genetic robustness was increasing neutrality, the proportion of loci in which mutations have no effect. Why does environmental variability select for neutrality? Our analysis of the phenotype and genotype evolution in the AVIDA model gave the clearest answer, suggesting that the only requirement for the congruent evolution of genetic robustness was antagonistic pleiotropy the kind of genotype by environment interaction in which particular alleles (and phenotypes) are advantageous in some environments, but deleterious in others. In a variable environment, antagonistic pleiotropy selected against the production of particular phenotypes, and parts of the genome were made nonfunctional as a result. The consistent evolution of increasing neutrality as a result only of environmental variability, indicates that antagonistic pleiotropy had a causal role in the evolution of genetic robustness in all three models.

Perhaps, the observation that antagonistic pleiotropy played a role in the evolution of genetic robustness should not have been surprising because such genotype by environment interactions are central to the evolution of environmental robustness. Essentially, if environmentally induced phenotypic variation reduces fitness, then there will be selection to minimize the environmental sensitivity of the phenotype [109]. If there are genotype by environment interactions, such selection will favor genotypes that produce similar phenotypes across environments [36]. Further, if environmentally induced phenotypic variation reduces fitness sufficiently, then natural selection may be expected to favor a genotype that produces a stable phenotype over a genotype that produces a variable phenotype, even if the latter achieves a higher fitness in the most common environment [88]. Our observations match these expectations. In the artificial life model, reduced envrionmental variability in fitness was achieved by ceasing to produce certain phenotypes, which in turn, reduced fitness in the most common (i.e. constant) environment. However, we were surprised that the production of a similar fitness across environments was achieved, in each model, by eliminating phenotypes with variable fitness effects entirely. We could not

have predicted that the strength of antagonistic pleiotropy (i.e. the cost of producing certain phenotypes in certain environments) would be sufficiently large to favor the elimination of such phenotypes in all three models.

Thus, we extend the findings from biological systems in an important way. Investigations in biological systems provided a critical first step by demonstrating the existence and mechanistic basis of evolved robustness in both RNA[2] and protein folding[91]. Our results confirm that the congruent evolution of genetic robustness is not characteristic only of these particular aspects of biological systems. Instead, the congruent hypothesis appears to be relevant for any aspect of the system that experiences antagonistic pleiotropy. Since natural populations of biological organisms are characterized by environmental variability and by an abundance of antagonistic pleiotropy[17, 28, 77], our results suggest that genetic robustness should evolve as readily in biological organisms as they evolved here in computational models. Combined, these observations provide strong support for the congruent hypothesis that genetic robustness generally evolves as a correlated response to selection for environmental robustness.

Finally, an underappreciated consequence of the congruent evolution of genetic robustness is that it may obscure our ability to predict differences in the amount of standing genetic variation between robust and non-robust populations. Thinking that we could demonstrate the accumulation of higher levels of genetic variation in robust compared to non-robust populations (as suggested by [107]), we compared the pairwise genetic diversity in the digital organisms populations evolved in constant and variable environments. We were initially surprised to find less genetic diversity in the robust populations, but quickly realized that the reduced genetic variation was a direct result of the fact that robustness evolved as a consequence of environmental variability[36]. In addition to selecting for robustness, the environmental variability ensured that genotypes with strong genotype by environment interactions were eliminated by natural selection. Although known robustness mechanisms clearly do allow the accumulation of unexpressed genetic variation, it is worth considering whether that characteristic applies specifically to mechanisms like hsp90 that buffer nearly every phenotype of the organism including those that do not experience strong genotype by environment interactions. In a situation like the one described in this paper, where strong genotype by environment interactions coupled with environmental variability are required for the evolution of genetic robustness, we posit that

these same factors would ensure minimal accumulation of unexpressed genetic variation.

4.6 Methods

4.6.1 Regulatory networks model

All networks were generated using software described previously[3] with default parameter values. We generated 50 networks and used them to found two sets of clonal populations (of size $N = 500$). One set experienced environmental perturbations, where each generation we changed the value of one or more randomly-selected expression state values of each offspring prior to development. Those that attained a stable expression state following this perturbation were allowed the chance to reproduce in the next generation. We conducted this experiment three times, using a different number of perturbations to the expression state values (1, 3, or 5) of one of the population sets for each repetition. Since there was no qualitative difference in the results for different numbers of perturbations, we only report results for 5 perturbations.

We measured environmental and genetic robustness in both sets of populations every 10 generations. At each timepoint, we made 75 environmental or genetic perturbations to each network one at a time and then measured the percentage of perturbations that resulted in unstable gene expression (i.e. fitness = 0). These perturbations were only for testing purposes and were not maintained in the population. The perturbations were equivalent to those experienced by the populations evolving in the variable environment. That is, if the populations were evolved with 3 perturbations to their expression state values at each generation, we tested their robustness by making 3 perturbations 75 times. The genetic perturbations, or mutations, were tested by mutating a randomly-chosen single element of the genotype matrix. The median fitness effect of 75 such mutations is reported.

Since the data were not normally distributed, particularly for the environmental robustness, we analyzed the data using non-parametric methods. Normal-approximation significance values of the Wilcoxon signed rank test were calculated using the medians of each population (constant and variable pairs) as data.

4.6.2 Digital organisms

All experiments were performed using AVIDA version 2.3 (obtained by request from C. Ofria) compiled with Cmake version 2.0.6 for Linux (configuration files available upon request). We generated founder genotypes by seeding approximately 90 populations of size $N = 900$ with the software-provided, minimally functional default genotype, which we modified to reduce its genome size to 70 instructions. We evolved these populations at a per-instruction mutation rate (AVIDA genesis file parameter "COPY_MUT_PROB") of $\mu = 0.00001587$ ($\mu * 70 = 1/N = U$) for 2.65 million updates in a constant environment with a constant reward value for performing any of seven logic functions. Insertion and deletion mutations were not permitted in this experiment. Following this pre-adaptation period, 31 of the populations had evolved the ability to perform 3 or more logic functions. We did not pursue further the remaining 59 populations. We considered them insufficiently adapted to the AVIDA environment.

We took the highest-fitness genotype from each of these 31 populations and used it to found two sets of clonal ($N = 900$) populations that were evolved for approximately 16,000 generations. For this experiment, we changed the environment reward structure such that organisms were rewarded or penalized only for functions they performed at the end of pre-adaptation period. One set of populations was evolved in a constant environment where they were consistently rewarded for performing logic functions and the other was subjected to environmental perturbations several times each generation. Environmental perturbations were implemented as random transient decreases to the phenotype reward structure, analogous to daily fluctuations in rainfall around a seasonal average. These perturbations temporarily decrease the CPU time, and hence the fitness, for any organisms that perform the logic function for which the reward is decreased.

In order to avoid developing a recurrent pattern of environmental perturbations to which the second set of populations could adapt, we chose the perturbations such that each reward value had a $1/k$ chance of being perturbed at each update, where $k =$ the number of functions performed by that population's founder. The perturbations were implemented as transient decreases in the function reward values and the magnitude of the perturbations was chosen from an exponential distribution with the scale parameter (the mean) equal to the function

reward value. The perturbed values reverted to the original values after 2 updates, or less than 1 generation.

For both sets of populations, we calculated the effect of a perturbation as the median of the ratio of perturbed to unperturbed fitness for all perturbations. Using the landscape analysis function in the software, we generated all single point mutations for each genotype and measured their effect on fitness. We could not exhaustively sample all environments, so we generated a random sample of 40 environments that were perturbed from the standard constant environment in the same way as described above. Fitness of each genotype was measured in all 40 environments. Non-parametric tests were used as described above for networks. To measure the number of functions performed and the fraction of the genome used to perform tasks, we used the average modularity analysis function in the software, which performs sequential knockouts of each genome instruction and tests the effect on performance of logic functions. To measure genetic diversity, we used the hamming function, which calculates the hamming distance between all genotypes in the population.

Computations for the networks and digital organisms models were executed on a Linux cluster maintained by the University of North Carolina at Chapel Hill's research computing group.

4.6.3 RNA

RNA molecules carry tiny electrostatic charges that allow them to fold into extensive secondary and tertiary structures. For genes in which the RNA molecule itself is the functional product (e.g. tRNA, rRNA), structure is important for function and has been highly conserved during the evolutionary history of these genes. We cannot yet predict RNA tertiary structure, but we can rapidly predict the secondary structure of RNA molecules using thermodynamic minimization [120, 119, 66]. The thermodynamic minimization approach is reasonably accurate for short RNA molecules; the approach does not, however, consider pseudoknots or other non-canonical interactions.

Predicted RNA folding has been extensively used to build computational simulations of evolving populations [2, 30, 43, 31, 105]. Here, we used a discrete-generation population model that was developed to make straightforward comparisons to existing theory [19]. In this model,

evolution proceeds exclusively by point mutations, the mutation rate is equal for all nucleotides, and the population size is held constant. At each discrete generation, individuals replicate in proportion to their fitness to produce the next generation.

To measure fitness, we used a hyperbolic decaying function, $f(\sigma)$, to calculate a selective value based on how well a structure, σ, matched the target shape:

$$f(\sigma) = \frac{1}{\alpha + (d(\sigma, t)/L)^{\beta}} \tag{4.2}$$

where α and β are scaling constants, $d(\sigma, t)$ is the Hamming distance between σ and the target shape (t), and L is the length of the sequence. The values $\alpha = 0.01$ and $\beta = 1$ were chosen to produce the hyperbolic decaying shape of the selective-value function and to maintain consistency with prior work [2, 19, 31, 18]. By scaling the distance with a hyperbolic decaying function, we modeled strong selection for target structure (i.e. few shapes function well).

In the control environment, we used the "simple" fitness function. In this scenario, fitness is solely a function of the similarity of the minimum free energy (mfe) structure to the target structure. The robustness of the mfe structure is not considered. The fitness of an individual is therefore the value of $f(\sigma)$ with σ as the mfe shape. For the experimental populations, we used a "plastic" fitness model [2, 19, 18]. The plastic model considers phenotypic plasticity whereby RNA molecules may assume alternative shapes, which are nearly as stable as the mfe structure. We refer to this ensemble of alternative shapes as the suboptimal repertoire, and predict it using an extension to the standard thermodynamic folding algorithms [115]. We predict all shapes within 3 kcal/mol of the ground state shape, which is roughly equivalent to breaking two G-C bonds. The approach of [115] permits computation of the partition function Z for a molecule's suboptimal repertoire:

$$Z = \sum e^{-\Delta G_s/KT}, \tag{4.3}$$

where ΔG_s is the free energy of shape s, K is the Boltzmann constant, and T is the temperature [68]. The Boltzmann probability p of a shape s is then

$$p_s = \frac{-\Delta G_s/KT}{Z}. \tag{4.4}$$

The Boltzmann probability is precisely the probability of finding an RNA molecule in shape s (a shape in its suboptimal repertoire) in a large sample of identical RNA molecules; it also approximates the amount of time an RNA molecule spends in s. This model assumes thermodynamic equilibration and no kinetic barriers to transitioning between shapes [68].

In the plastic model, fitness is measured as the sum of the selective values of all shapes in a molecule's suboptimal repertoire, each weighted by its Boltzmann probability: $\sum f(\sigma)p_s$ [2, 19, 18]. We are therefore selecting for RNA molecules that stably fold into shapes that match the target shape.

We begin each simulation with one of three isogenic populations of N individuals, which were poorly adapted to their environment. Each population was then allowed to separately adapt to one of fifteen targets under both the simple and plastic fitness functions for 250,000 generations, for a total of 45 populations. Our simulation maintained a constant population size of $N = 1000$, and used a genomic mutation rate of $U = 0.001$ (the reciprocal of the population size). Genome length slightly varied across simulations to match the length of the target shape, but was approximately 105 (+/- 5) nucleotides.

We used inverse folding to produce a set of pseudorandom, non-adapted RNA molecules that fold into a particular minimum free energy structure, but do not necessarily have a high degree of thermodynamic stability. Inverse folding is commonly used to produce RNA molecules that fold into a desired shape [93]. The program RNAinverse in the ViennaRNA software package initially divides the target shape into several smaller regions and the starting sequence into segments, which each correspond to a small region of the target structure. Each segment of the starting sequence is individually optimized through single base changes or compatible base-pair changes. Once all of the separate regions of the starting sequence have been individually optimized the full sequence is created and further optimized.

4.7 Acknowledgments

I thank Matt Cowperthwaite for performing the RNA simulations and drafting methods language for them. I thank Charles Ofria and David Bryson for providing code, compilation assistance and troubleshooting support for AVIDA software.

Chapter 5

Discussion

Many programs exist for the prediction of non-(protein)coding RNA genes, yet RNA-Decoder is apparently the only one specifically adapted for predicting structures that exist in protein-coding regions. While overlapping coding for RNA and proteins may not occur often in organisms other than viruses with small genomes, it is still an important feature of viral genomes and for this reason warrants further methodological development. RNA-Decoder performs well enough to recognize the RRE and the *gag-pro-pol* frameshift signal in HIV-1 and HIV-2, but additional training may improve its predictions for other structures and other viruses. RNA-Decoder comes pre-trained, but without any method for re-training the grammar transition parameters. Presumably this feature is included in the source code, but the code is not well documented and is difficult to compile, due (among other things) to some inconsistencies with current compilers. These difficulties could be overcome using a separate tool specifically for training phylo-grammars[50]. The availability of such software for parameter estimation opens the possibility of re-training the existing grammar on different structures or developing a new grammar without having to modify existing RNA-Decoder software. It would be interesting to use known HIV structures in the training dataset and then re-predict the HIV genome and that of other lentiviruses. For example, most structures predicted for the HIV-1 and HIV-2 genomes are relatively simple stem-loop structures that are anchored by a main stem. No large, complex structures are predicted with high reliability, particularly ones that do not have a single prominent stem, as does the RRE. This could mean that no such structures in fact exist, or that the training dataset was not diverse enough to allow their prediction. These possibilities could

be tested by modifying the training dataset to exclude simple stem-loops or to include complex structures, even structures that are not entirely validated. Results may suggest whether lentiviral or other viral genomes are capable of harboring other complex RNA structures, or whether they are predominantly small stem-loops.

The evolutionary parameters of RNA-Decoder are more straightforward to adjust, and another potential modification is to fit the underlying evolutionary model[82] to an alignment of the sequences of interest to obtain evolutionary parameters specific to those sequences. This could reasonably be done for groups of related viruses that are likely to share evolutionary parameters, such as retroviruses.

In the prior work on compensatory evolution[51], we discussed another possibility for refining RNA structure prediction software: using the difference in transition-transversion rate ratio instead of the difference of rates among sites as evidence for RNA secondary structure. Another possibility is to model variation in synonymous rates per codon as a signal for conservation of RNA structure. Typically, the rate of evolution at synonymous sites in protein coding regions of genes is used as a proxy for the neutral, or background, rate because these types of mutations do not change the encoded amino acid and therefore are not expected to be under selection. In multiple coding regions, the synonymous rates should be depressed because mutations that are neutral to the protein may have an effect on the RNA. Previous work along these lines shows reduced variability of synonymous sites as evidence for the presence of RNA secondary structure in those regions[95, 103]. However, these studies did not use an explicit model of codon or nucleotide evolution, as in[54]. Model-based tests for variability of synonymous rates by Simon Frost's group found evidence for among-site variation in over 400 alignments of virus genes (`http://www.viralevolution.org/`). Variation in synonymous rates cannot automatically be assumed to imply the presence of RNA secondary structure, but does at least indicate the presence of multiple constraints. Likelihood methods could be used to test for variability of synonymous rates, and to choose the number of such rates to estimate, as described in[54]. These rates could be used instead of the current, parameter-rich model used in RNA-Decoder. Although such an approach might lose some specificity because it does not model the specific effect of RNA structure on rates at individual positions, it may require estimation of fewer parameters than RNA-Decoder currently does. Also, the parameters could

be estimated without a priori knowledge of structure in an alignment. Potentially, either of these two signals, transition-transversion rate ratio or synonymous rates, could be combined with a structural grammar using a Bayesian scoring method as in RNA-Decoder.

Using a new prediction method that incorporates some of the above adjustments, a large-scale analysis of conserved RNA structures in viruses (or certain sub-groups) could be undertaken and stored in a web-accessible database. Studies near this scale have already been attempted using thermodynamic folding software[39, 102, 114]. A comparable version using the latest sequence data, alignment and phylogeny methods, and predictions based on evolutionary rate information and grammar parameters would be a large but feasible undertaking and a good complement to data from thermodynamic and/or biochemical prediction methods and would potentially spur additional such studies. Such a survey would also provide data to assess the frequency of multiple coding and to quantify its constraints on molecular evolution. Thinking broadly, such an approach could be expanded to bacteria, archaea, and/or eukaryotes in order to characterize differences among the groups in use of multiple coding.

With respect to the issue of conserved structures in HIV-1 and HIV-2, it may also be of interest to predict structures for the most closely related simian viral genomes separately from the human viruses. There are enough sequences of SIVcpz to make a reliable dataset for a separate analysis of structure in this virus, and SIVsm in combination with SIVmac may also produce a large enough dataset. Such an investigation could use the same grammar transition parameters and evolutionary parameters as used for HIV-1 and HIV-2, so the only additional work would be to align the sequences and infer phylogenetic trees. Prior work on the RRE across lentiviral species suggests that this structure is present in other members of this viral group, but with some structural differences[63]. If prediction methods can suggest other such differences between the human viruses and their most closely-related simian viruses, they may suggest opportunities for biochemical structure probing to validate structures and additional work to elucidate any functional differences.

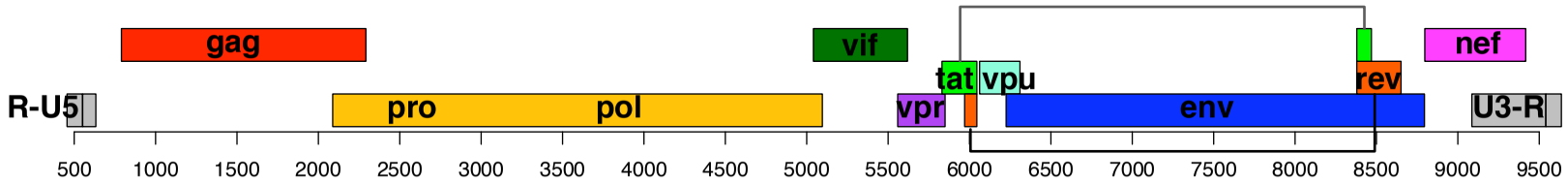


Figure 1: The HIV-1 genome organization is shown using the HXB2 reference genome numbering. Genes are shown in colored boxes, arranged vertically according to their reading frame. The LTR non-coding regions (U5, R, U3) are shown as gray boxes at either end of the genome.

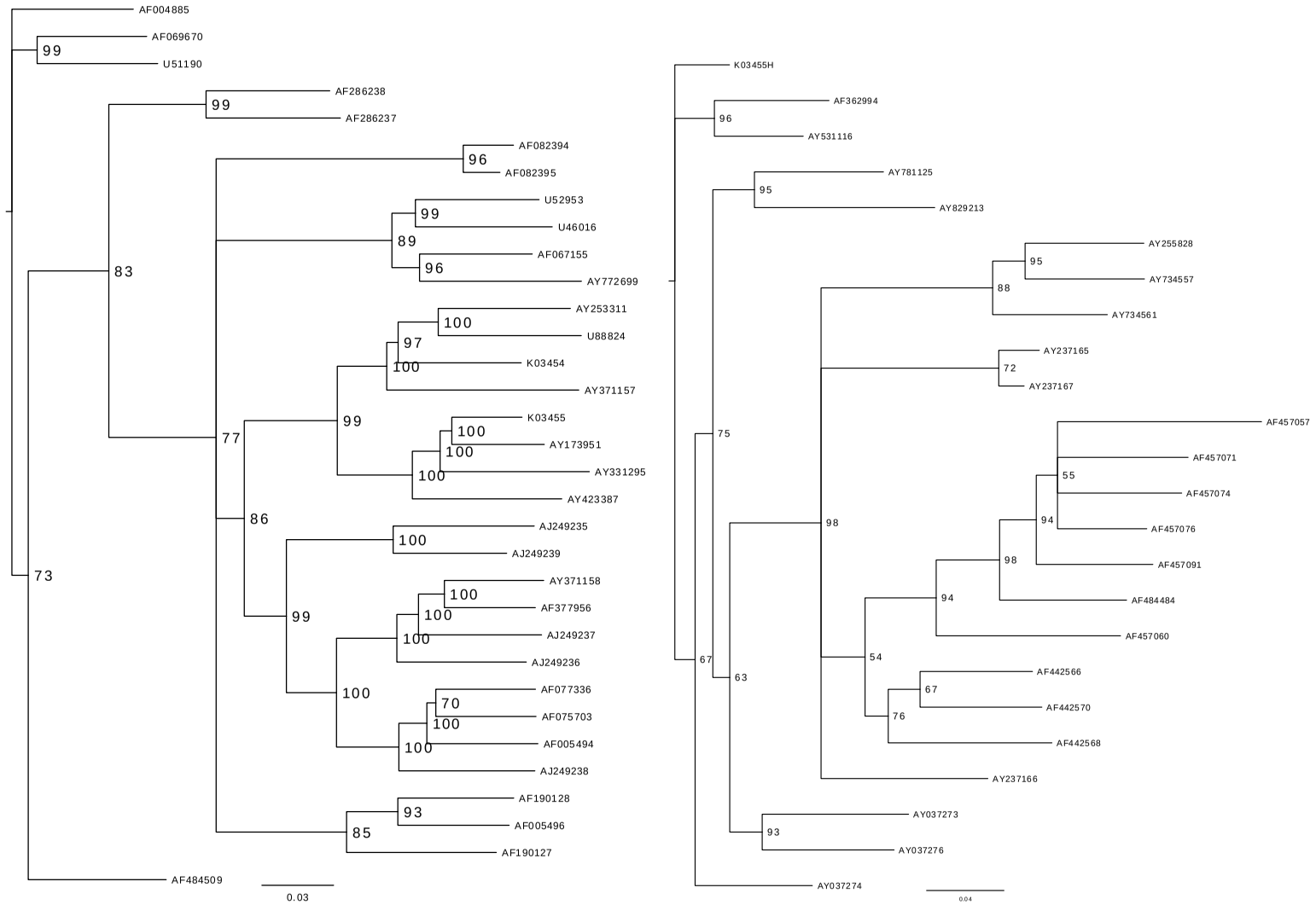


Figure 2: Phylogenetic trees inferred from alignments. Trees were generated using the GTR+γ model of evolution. Branch support is indicated. (left) Reference subtype sequences; (right) hypermutated sequences.

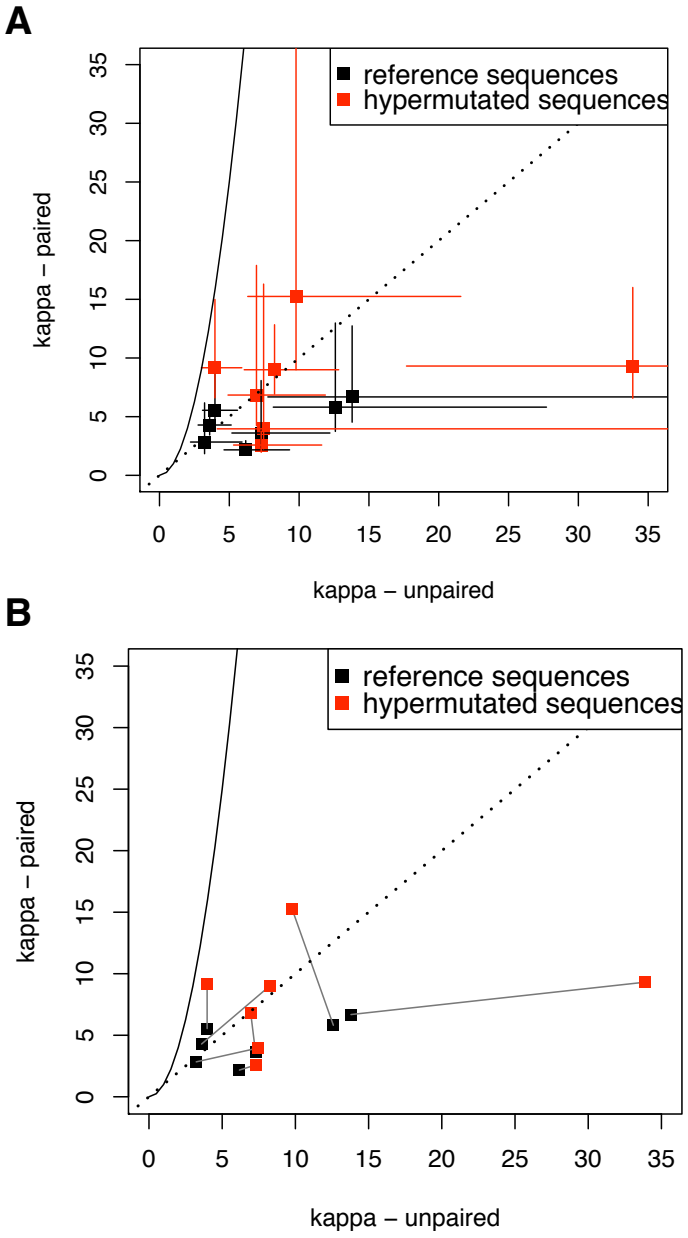


Figure 3: Transition-transversion rate ratios in stem and loop sites for selected structures.The solid like represents $\kappa_{stems} = \kappa_{loops}^2$, the prediction for structures evolving via compensatory evolution. The dashed line represents $\kappa_{stems} = \kappa_{loops}$, the null expectation. (a) Estimates of κ are shown with their 95% confidence intervals for several known and predicted structures in HIV-1. Stem and loop sites were designated by predictions from a phylo-grammar. (b) The same data with lines connecting values measured for the same structure.

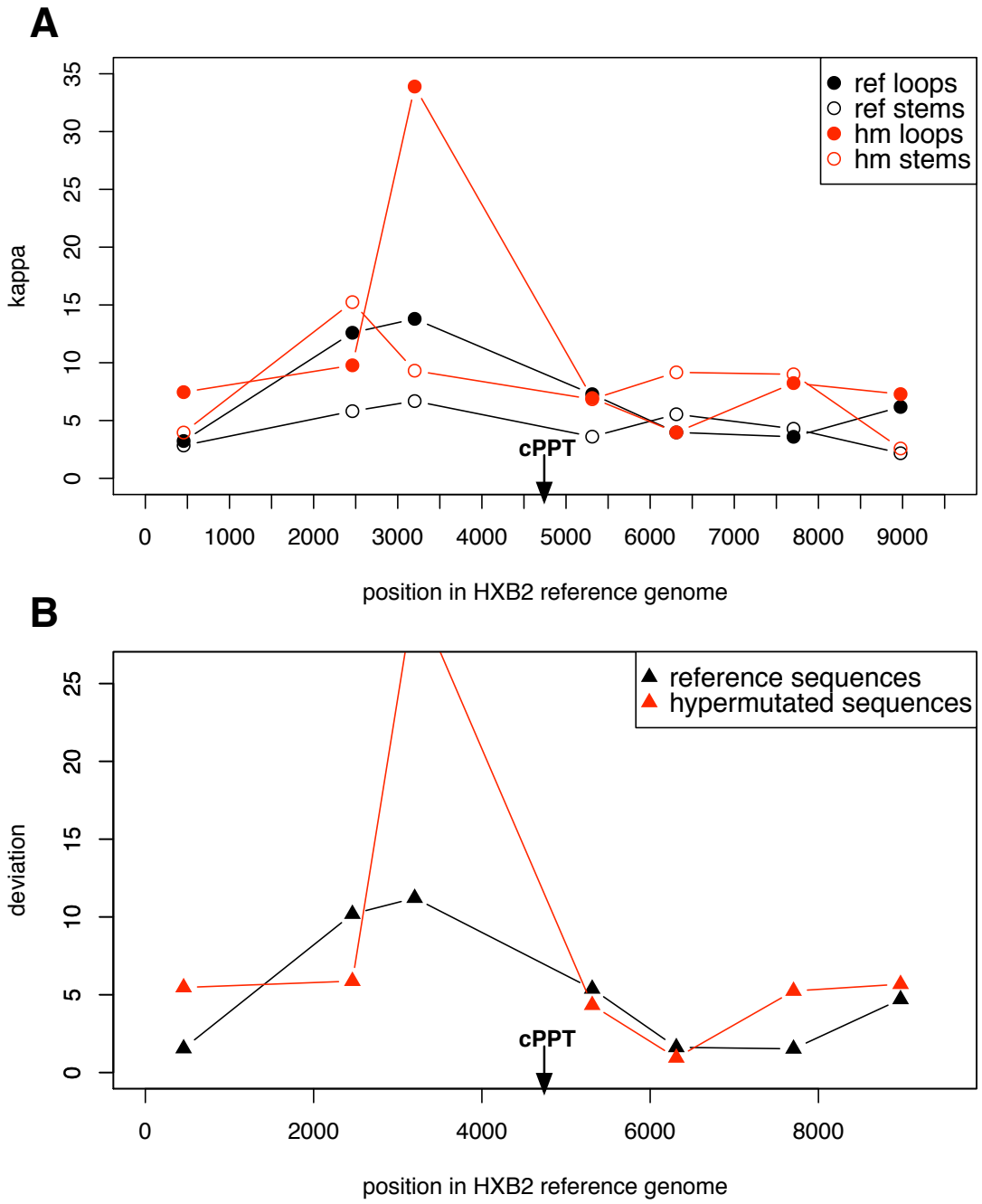
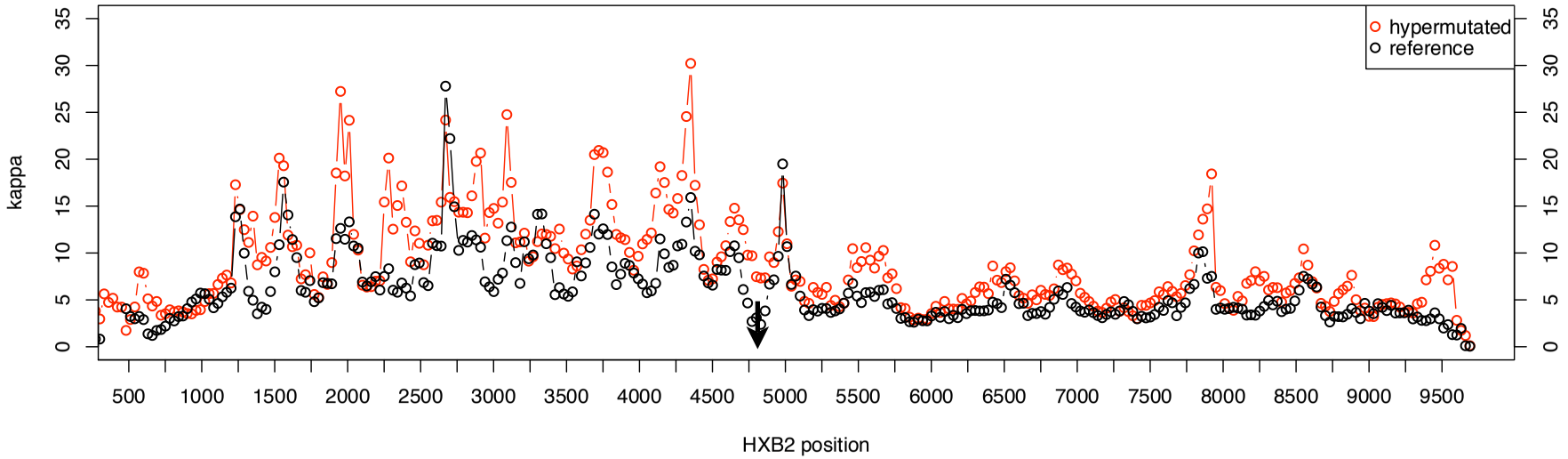


Figure 4: Relationship of genome position to transition-transversion rate ratios. (a) The same data shown in Figure 3 is plotted versus the locations of the structures in the HIV-1 genome. Positions are according to the HXB2 reference sequence. (b) The deviation of the data shown in (a) from the compensatory prediction is shown. The negative of the deviation is plotted, such that values greater than zero indicate higher κ_{loops} than expected. See text for explanation of deviation. The position of the central polypurine tract (cPPT) is indicated.


kappa
HXB2 position
hypermutated
reference
0
5
10
15
20
25
30
35
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500

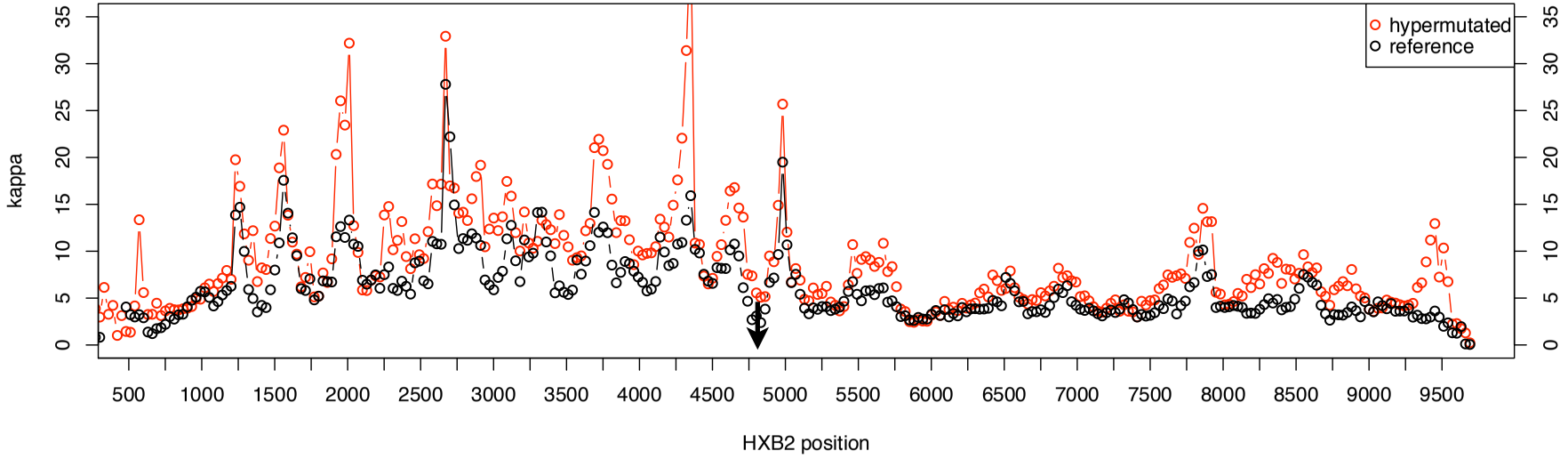

kappa
HXB2 position
hypermutated
reference
0
5
10
15
20
25
30
35
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500

Figure 5 *(preceding page)*: Sliding-window analysis of transition-transversion rate ratio. κ is plotted for overlapping 150-nucleotide windows across the genome for both the reference sequences alignment and the hypermutated alignment. All positions in the alignment were used for the top panel; positions with a high likelihood of being in a stem were removed for the bottom panel.

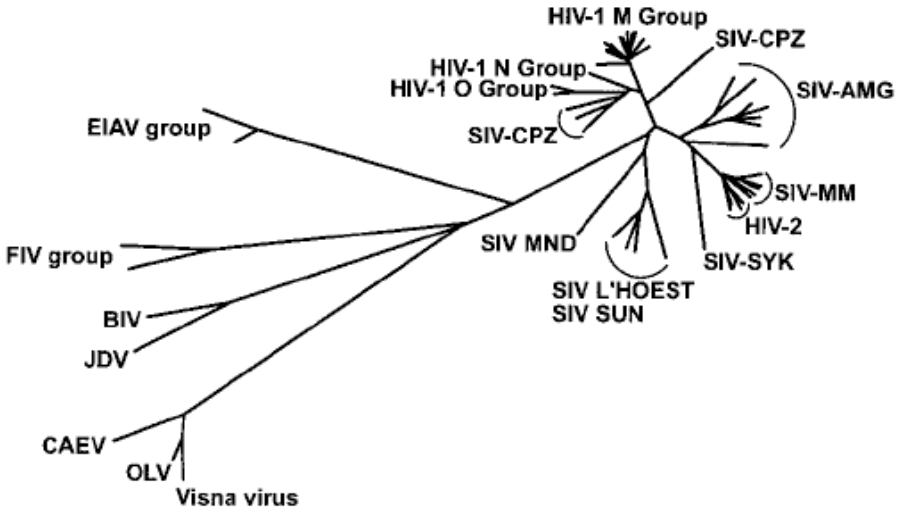
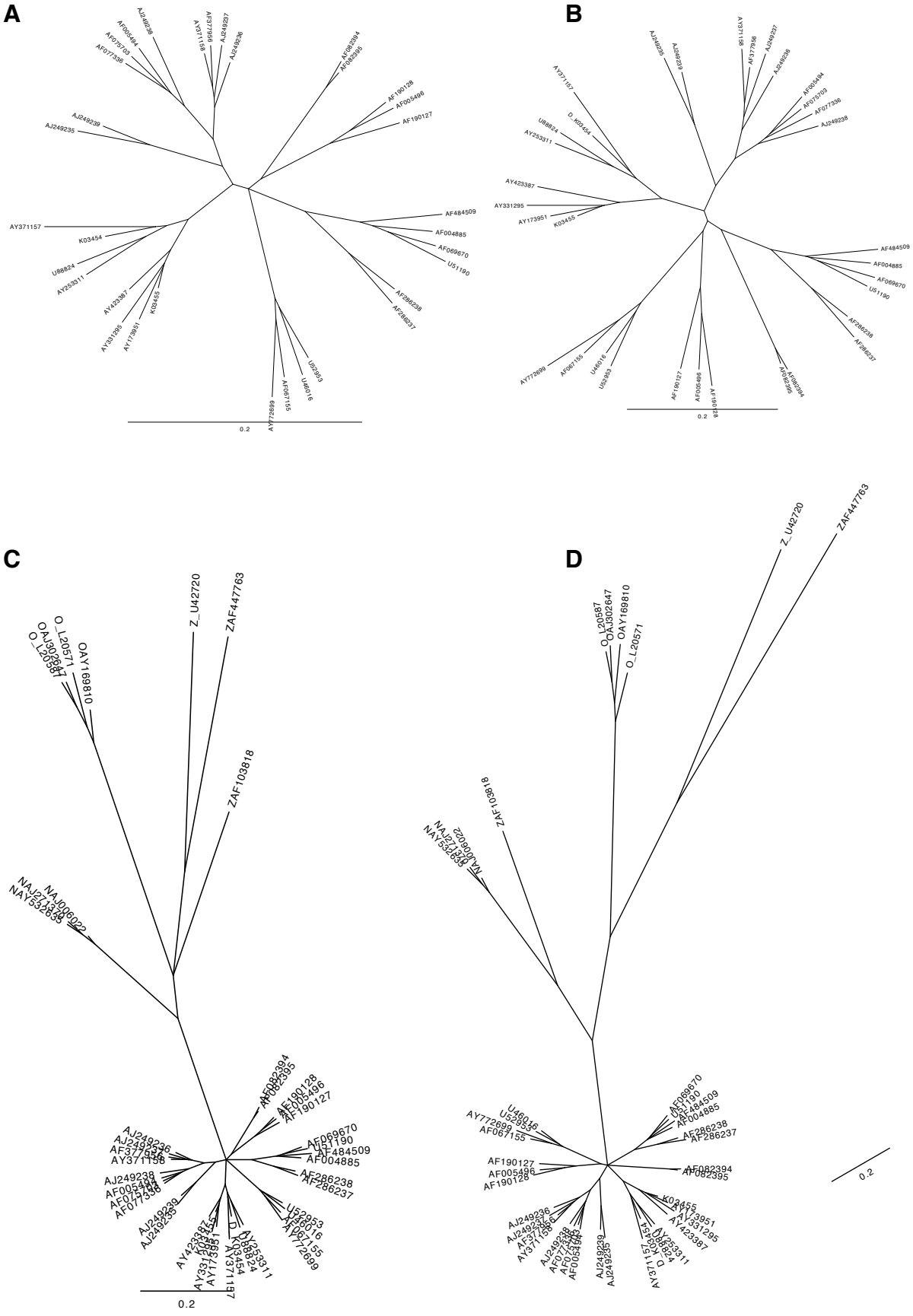


Figure 6: Phylogenetic tree for lentiviruses, inferred from the *gag* gene, as shown in [63].


A
AJ249238
AF005494
AF075703
AF077336
AF227967
AY371158
AJ249237
AJ249236
AF082394
AF082395
AF190128
AF005496
AF190127
AJ249239
AJ249235
AF484509
AF004885
AF069670
U51190
AF286238
AF286237
U52953
U46016
AF067155
AY772699
AY371157
K03454
U88824
AY253311
AY423387
AY331295
AY173951
K03455
0.2
B
AJ249238
AJ249239
AY371158
AF377936
AJ249237
AJ249236
AF005494
AF075703
AF077336
AJ249238
AY371157
D_K03454
U88824
AY253311
AY423387
AY331295
AY173951
K03455
AF484509
AF004885
AF069670
U51190
AF286238
AF286237
AY772699
AF067155
U46016
U52953
AF190127
AF005496
AF190128
AF082394
AF082395
0.2
C
Z_U42720
ZAF447763
O_L20587
OAJ302647
O_L20571
OAY169810
ZAF103818
NAJ271336
NAY532639
NAJ006022
AF082394
AF082395
AF190128
AF005496
AF190127
AJ249236
AJ249237
AF377936
AY371158
AF069670
U51190
AF484509
AF004885
AJ249238
AF005494
AF075703
AF077336
AF286238
AF286237
AJ249239
AJ249235
U52953
U46016
AF067155
AY772699
AY423387
AY331295
AY173951
K03455
AY253311
U88824
D_K03454
AY371157
0.2
D
Z_U42720
ZAF447763
O_L20587
OAJ302647
OAY169810
O_L20571
ZAF103818
NAJ271336
NAY532639
NAJ006022
AF069670
U51190
AF484509
AF004885
AF286238
AF286237
U46016
U52953
AY772699
AF067155
AF190127
AF005496
AF190128
AF082394
AF082395
K03455
AY173951
AY331295
AY423387
AY253311
U88824
D_K03454
AY371157
AJ249236
AJ249237
AF377936
AY371158
AJ249238
AF005494
AF075703
AF077336
AJ249239
AJ249235
0.2

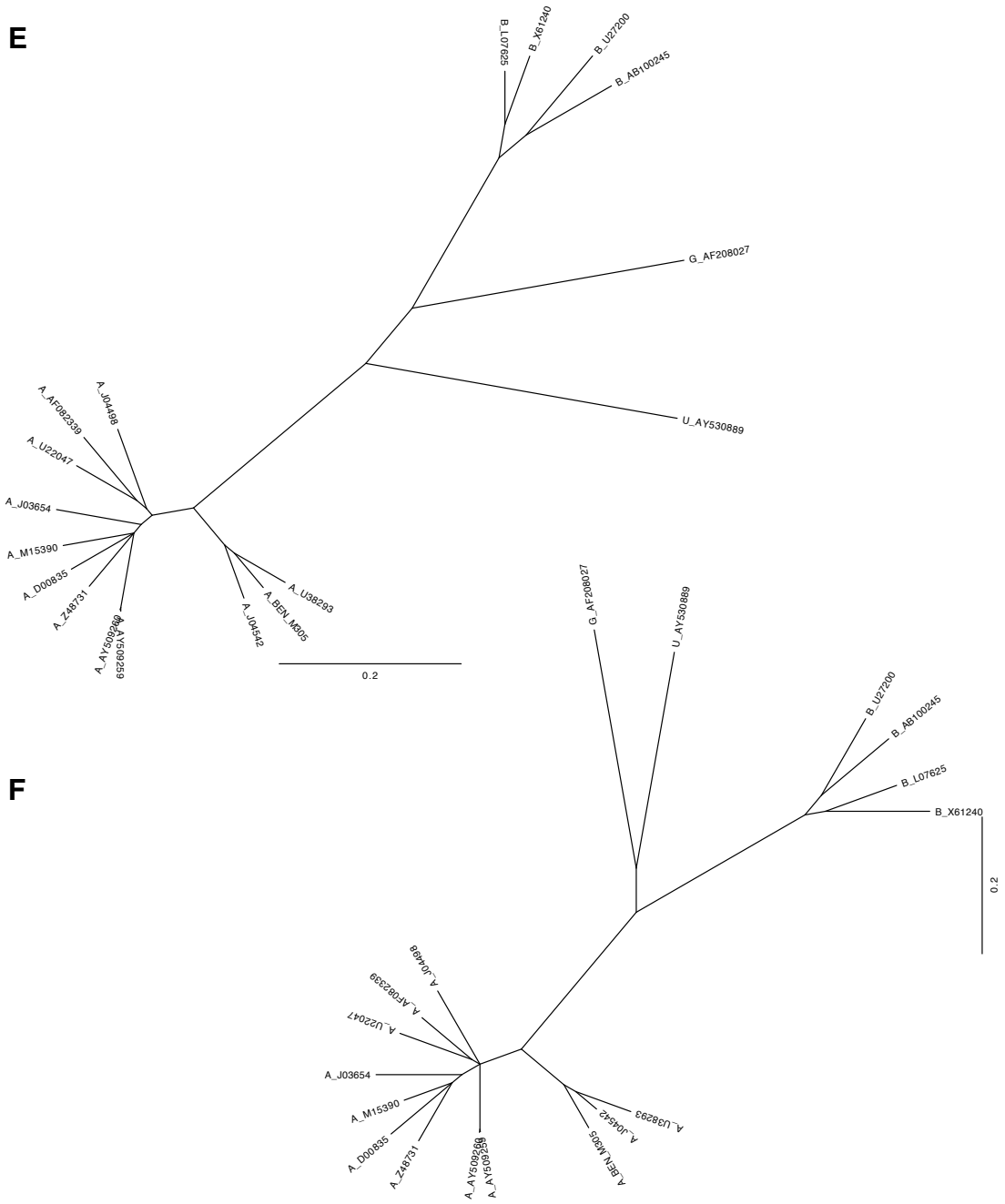


Figure 7: Phylogenetic trees inferred from HIV-1 and HIV-2 genome alignments. To account for the differences in nucleotide content and evolutionary rate in different parts of the HIV genome, two trees were inferred for each genome, one for the upstream and downstream regions. The trees were inferred from third positions of the relevant coding regions. See methods for details. Shown are the phylogenies for the following alignments: (a-b) HIV-1 group M upstream and downstream, respectively, (c-d) HIV-1 groups M, N and O and SIVcpz upstream and downstream, respectively, and (e-f) HIV-2 upstream and downstream, respectively.

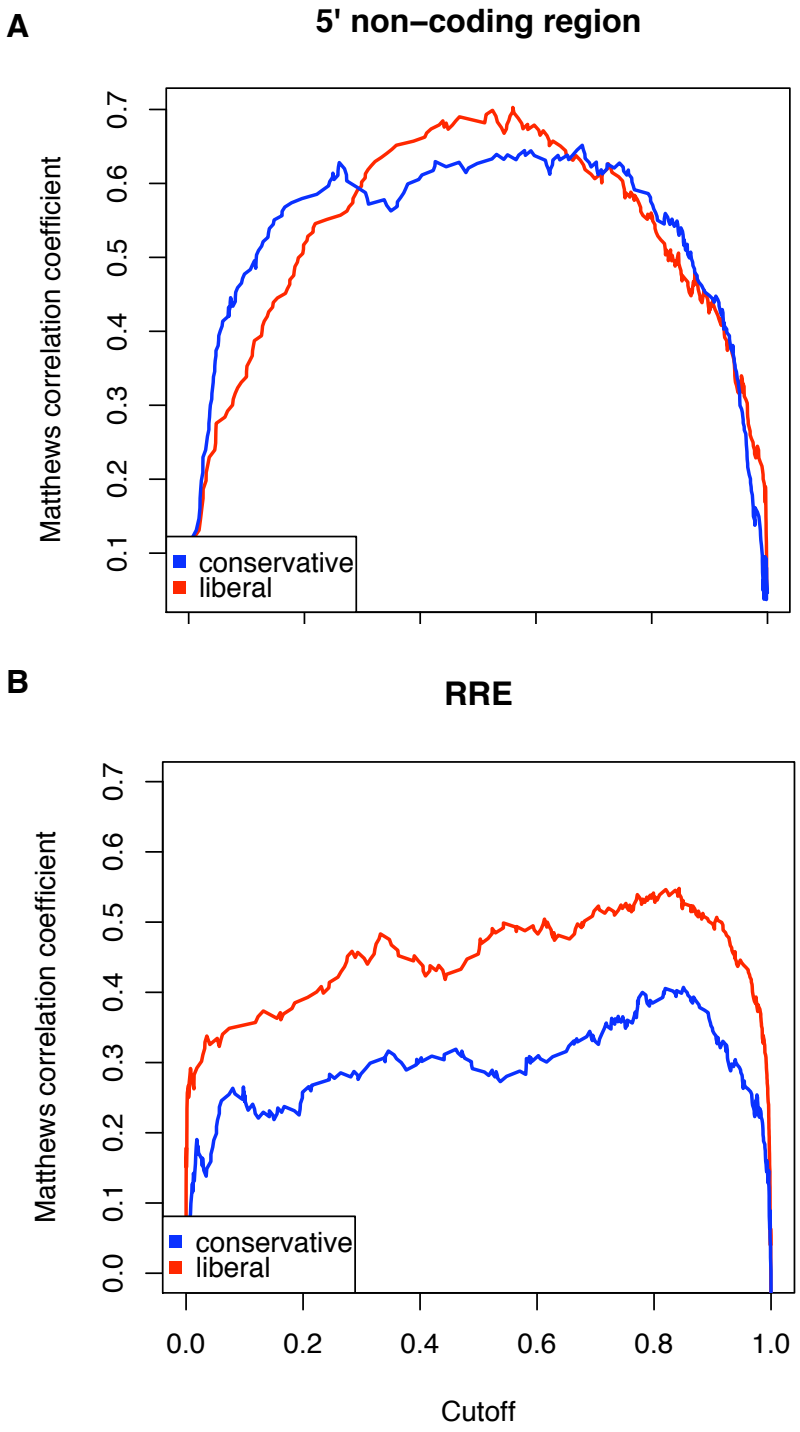


Figure 8: Performance of the Matthews correlation coefficient over various pairing probability cutoffs. The correlation between the structural labels of chemical-thermodynamic and RNA-Decoder is shown for the two large, known structural regions in HIV-1, (a) the RRE (positions 7709-8061 in HXB2) and (b) the 5' non-coding region (466-797 HXB2). The cutoff indicates the minimum value of pairing probability required for a position to be considered a stem.

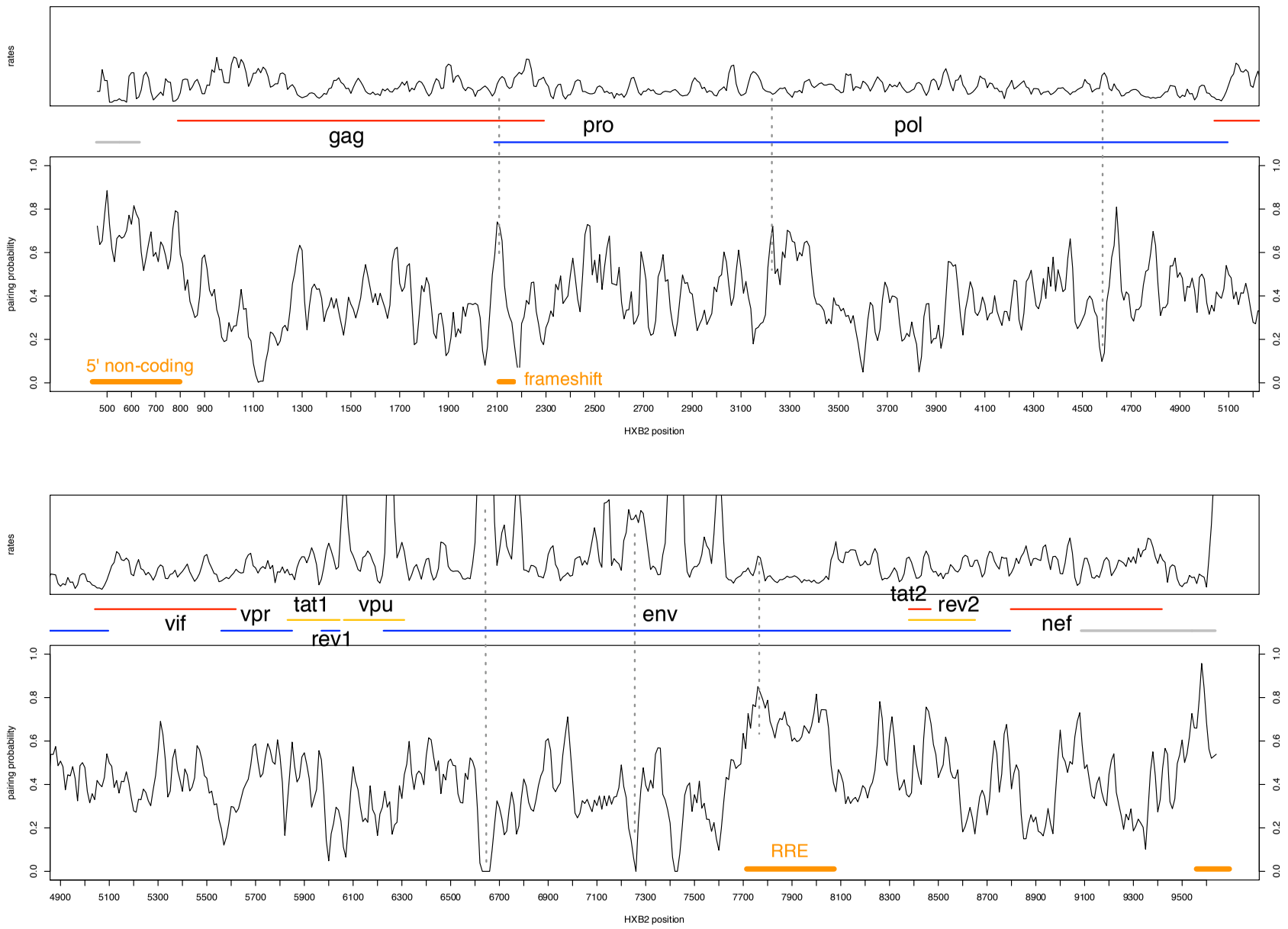

rates
gag
pro
pol
pairing probability
5' non-coding
frameshift
HXB2 position
rates
vif
vpr
tat1
rev1
vpu
env
tat2
rev2
nef
pairing probability
RRE
HXB2 position

Figure 9 *(preceding page)*: Conserved structure predictions for HIV-1 group M genomes. The smoothed pairing probability is shown in the bottom of each 3-part panel. High values indicate a high probability of being in a stem position in any possible structure assumed by the sequence. The smoothed values reported are the average pairing probability across 30-nucleotide windows. Small regions of predicted structure are visible as isolated peaks, while larger regions have several consecutive points at high values. Regions of known structure are indicated by orange horizontal lines, and include the 5' non-coding region (part of which is repeated at the 3' end of the genome), the *gag-pro-pol* frameshift, and the RRE. The upper part of the panels shows the relative per-position rate of evolution. Dashed gray lines are drawn as guides in select places to show the correspondence of the peaks in the top and bottom panels. The middle part of the panel shows the coding regions of the HIV-1 genome, color-coded according to reading frame: frame 1 (*gag, vif, nef) – red; frame 2 (tat1, vpu) – yellow; frame 3 (pro, pol, vpr, rev1, env) – blue. The LTR regions are shown in gray (R-U5 at the 5' end and U3-R at the 3' end).*

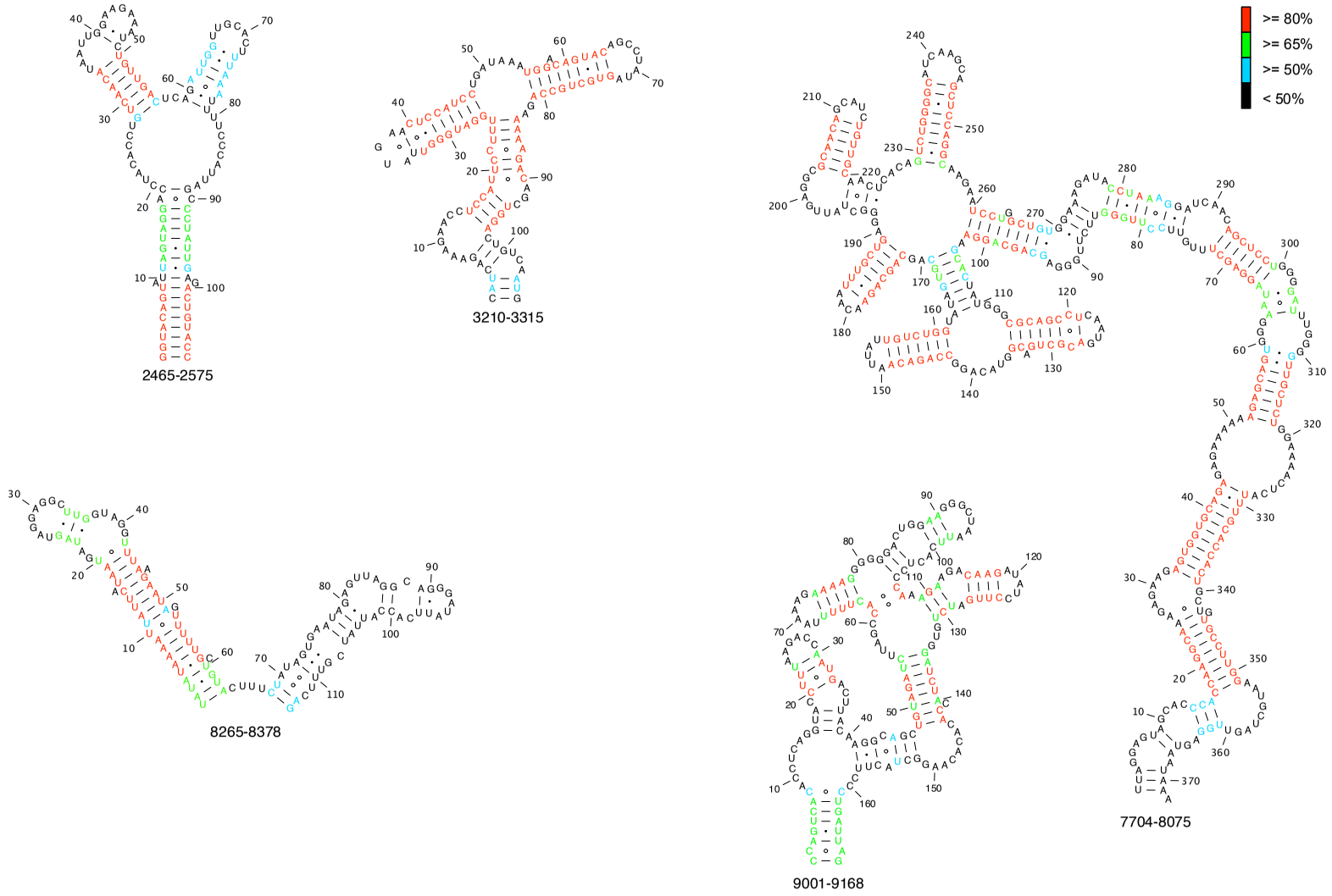


Figure 10: Selected structures predicted by RNA-Decoder.Depicted are the highest-probability folds for selected structure regions in HIV-1. The structures are depicted using the HXB2 reference sequence, and the 5' and 3' positions refer to HXB2 numbering. Base-paired regions are color-coded according to the reliability (i.e. posterior probability) of the prediction. Structures were drawn using xRNA (`http://rna.ucsc.edu/rnacenter/xrna/xrna.html`).

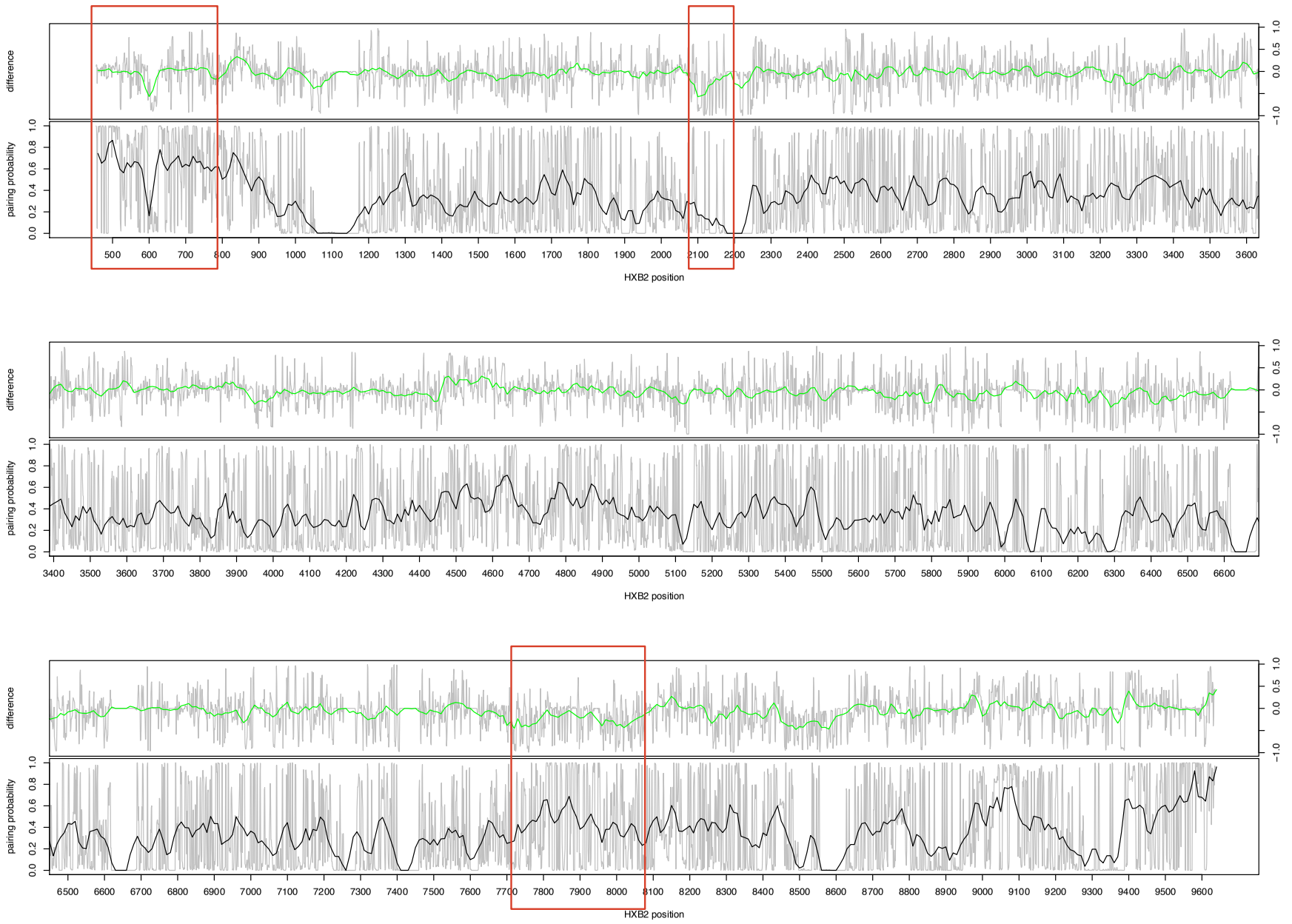

difference
pairing probability
HXB2 position
difference
pairing probability
HXB2 position
difference
pairing probability
HXB2 position

Figure 11 *(preceding page)*: HIV-1 M and SIVcpz comparison plot. In each two-part panel, the bottom part shows the pairing probability resulting from an alignment of HIV-1 groups M, N and O and SIVcpz (gray). A smooth of the data is shown in black (the average across 30-nucleotide windows). The top part of each panel shows the difference between the pairing probability for this alignment and for the HIV-1 group M alignment (gray). Values greater than zero are places where this alignment predicts a higher pairing probability than the group M sequences only, while values less than zero are positions where the group M sequences give a stronger pairing prediction than this alignment. A smooth of this data is shown in green. Red boxes are drawn to indicate the positions of the 5' non-coding region, the frameshift, and the RRE. All positions are according to the HXB2 reference sequence.

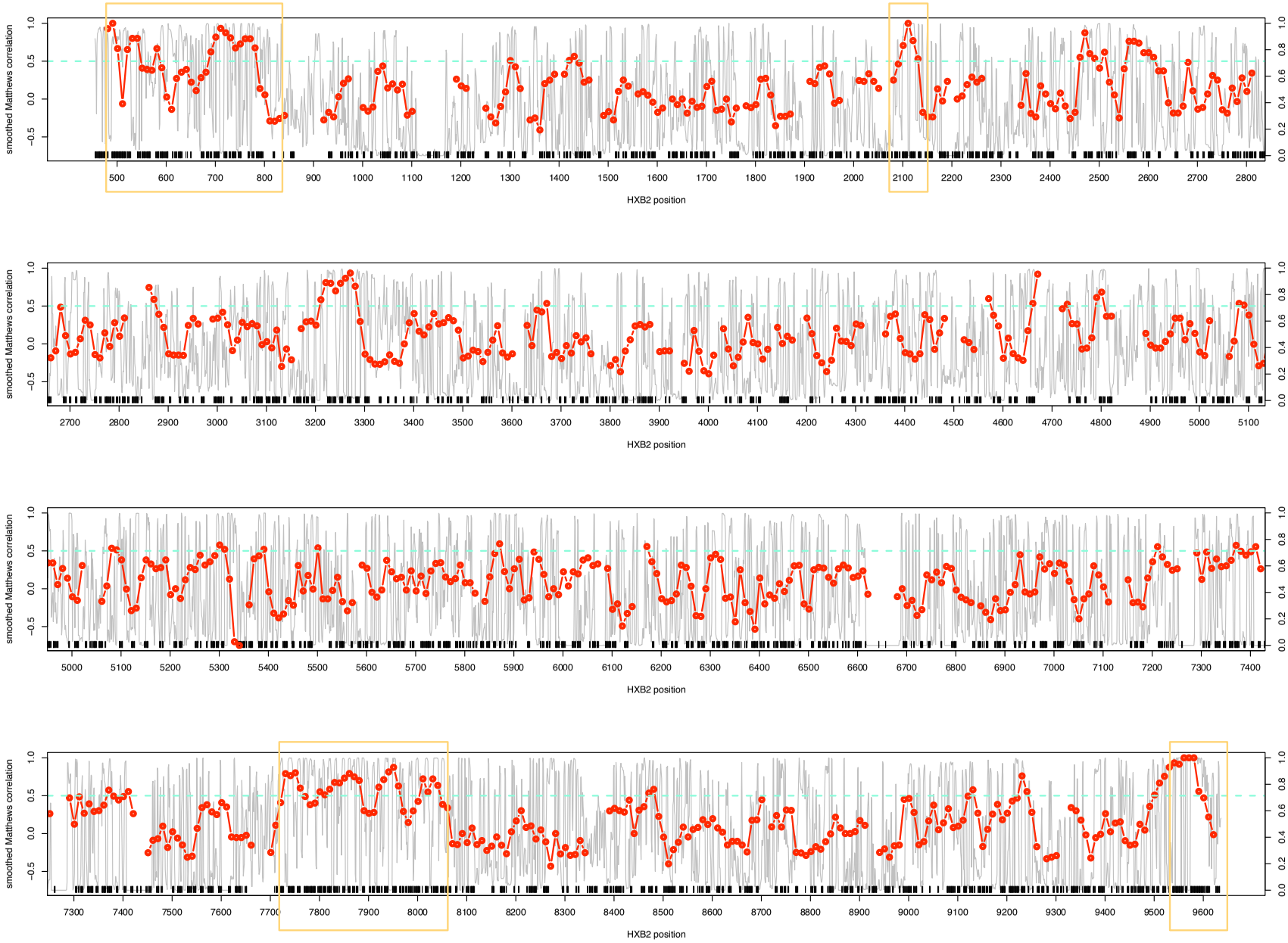

smoothed Matthews correlation
HXB2 position
500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800
2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000 5100
5000 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000 7100 7200 7300 7400
7300 7400 7500 7600 7700 7800 7900 8000 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000 9100 9200 9300 9400 9500 9600

Figure 12 *(preceding page)*: Agreement of chemical-thermodynamic method and RNA-Decoder across the HIV-1 genome. The Matthews correlation coefficient (left y-axis) is shown in red for overlapping 30-nucleotide windows across the HIV-1 genome. Values below zero indicate worse than random correlation between the two methods, or regions of strong disagreement between the methods. Missing values occur when the correlation is undefined; typically these are regions where either method predicts no stems in the entire window and thus the correlation is uninformative. Our threshold for selecting areas with strong joint predictions is shown by the blue dashed line. Regions with more than one consecutive point falling at or above this line are listed in Table 1. The actual pairing probability values reported by RNA-Decoder are shown in gray (values are according to the right y-axis) and the stem predictions by chemical-thermodynamic are shown as short vertical black lines on the x-axis. Regions of known structure are indicated by orange boxes, and include the 5' non-coding region (part of which is repeated at the 3' end of the genome), the *gag-pro-pol* frameshift, and the RRE.

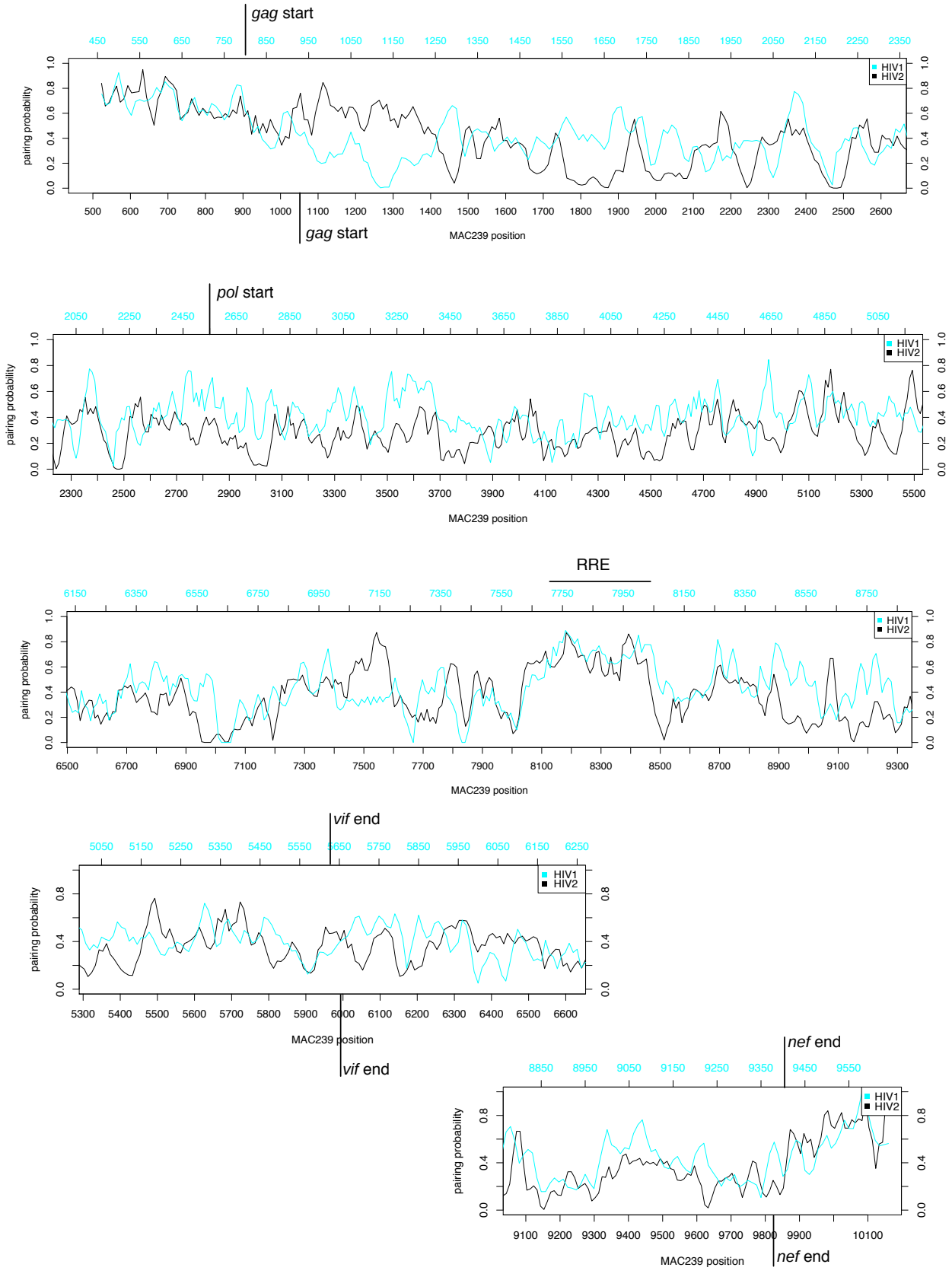

gag start
pol start
RRE
vif end
nef end
HIV1
HIV2
pairing probability
MAC239 position

Figure 13 *(preceding page)*: Similarities and differences in predicted structures for HIV-1 and HIV-2. Smoothed pairing probabilities for HIV-1 and HIV-2 are shown compared over genome regions: (a) the 5' non-coding region and *gag*, (b) *pro* and *pol*, (c) *env*, (d) all genes between *pol* and *env* (including *vif*), and (e) *nef* and the 3' non-coding region. In each panel, the signals are plotted according to their own x-axis numbering, with HIV-2 in black (axis below the plot) and HIV-1 in blue (axis above the plot). Data for the two genomes is not aligned, but it does begin and end at homologous positions (i.e. the start and end of genes). There are length differences between the HIV-1 and HIV-2 signals for the regions shown.

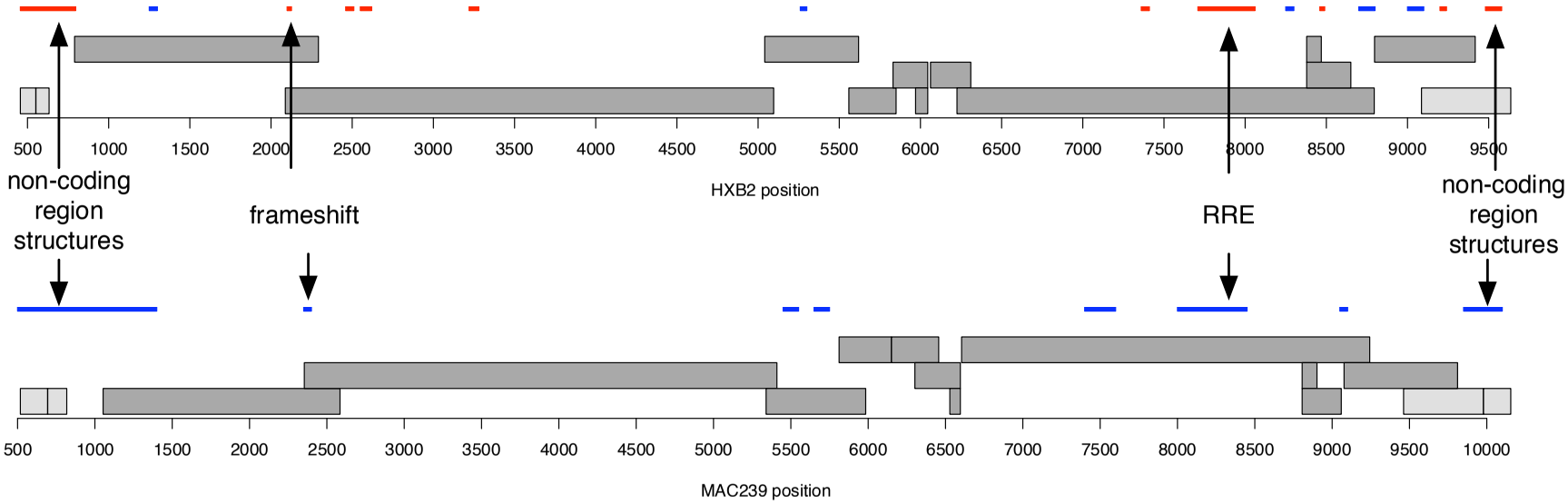


Figure 14: Regions of predicted structure in HIV-1 and HIV-2. Best-supported regions of predicted structure for HIV-1 and HIV-2 are shown in blue and red above the gene layout for the respective genomes. Structures supported by data from a chemical-thermodynamic analysis of the HIV-1 genome are shown in red. Structures predicted only by a high pairing probability are shown in blue. Dark gray rectangles represent the genes, which are arranged vertically according to their reading frame. Light gray rectangles represent portions of the LTR.

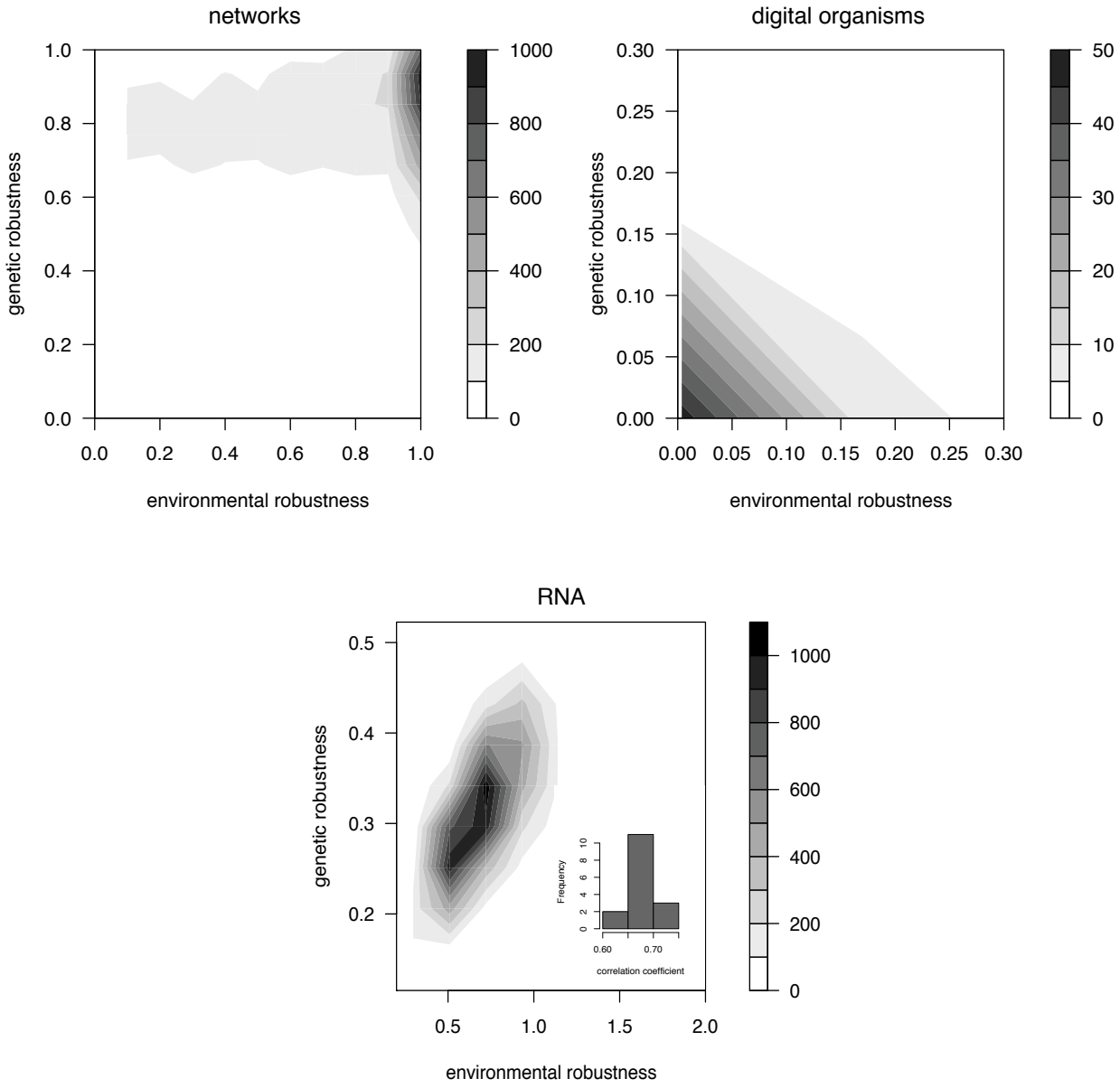


Figure 15: Environmental and genetic robustness are correlated in all three models. For each model is shown the distribution of genetic robustness versus environmental robustness among individuals chosen without regard for their robustness. For regulatory networks, Kendall's τ = 0.14, p-value = 2.2x10-16. For digital organisms, Kendall's τ = 0.06, p-value = 0.4. For RNA secondary structures, a representative distribution for a single collection of sequences is shown, with Pearson's correlation coefficient = 0.68, p-value = 0. This distribution of values is typical among the 17 collections we surveyed (inset = correlation coefficients for all 17 collections, which were all statistically significant at alpha = 5%).

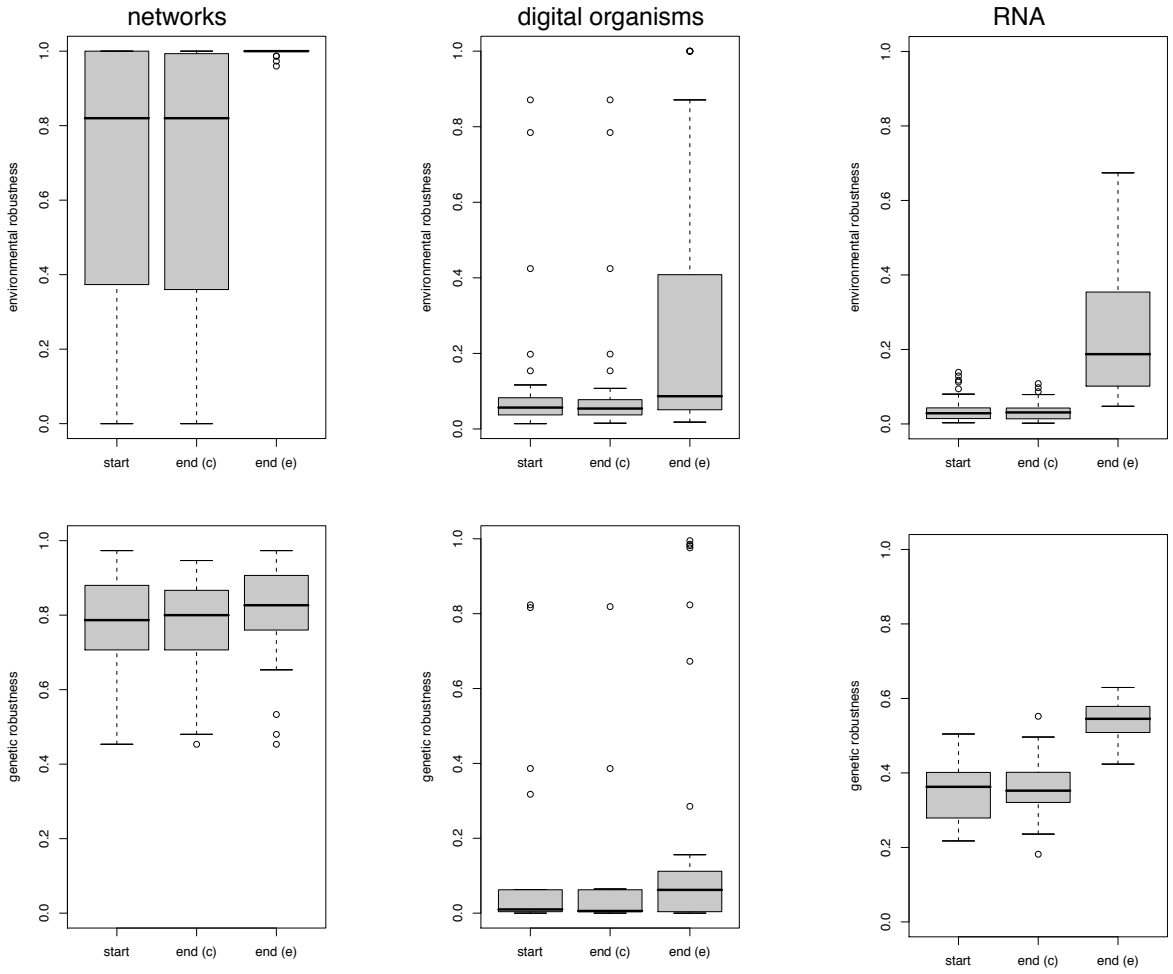


Figure 16: Genetic robustness (relative fitness following a mutation) increases under selection for environmental robustness. Boxplot distributions of median robustness values across populations are shown at the start and end of the experiment for each model. Ending distributions are shown for both the control populations (indicated by (c)), which lacked selection for environmental robustness, and the experimental populations (indicated by (e)). Starting distributions for control and experimental populations were identical for all models. Number of generations elapsed during the experiment varied: 400 for networks, 15,000 for digital organisms, 250,000 for RNA. Whisker lines are drawn at the most extreme data point that is no more than 1.5 times the interquartile range from the box, and points beyond are shown as circles. Endpoint values for experimental populations only are significant by a Wilcoxon sign-rank test (networks, digital organisms) or a t test (RNA), with a significance value less than 0.01.

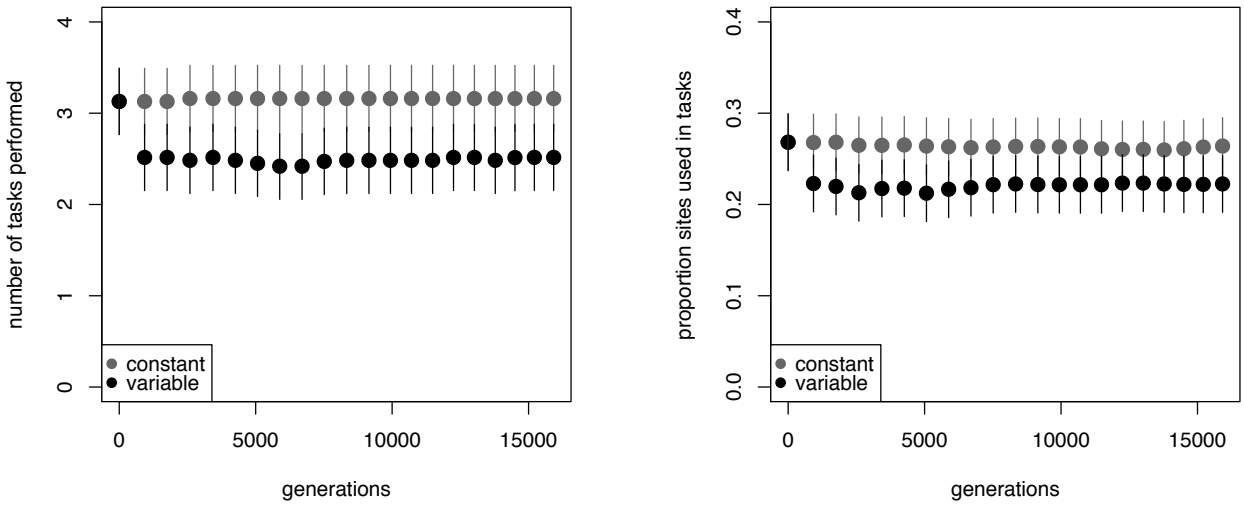


Figure 17: Proximate mechanisms for increasing genetic robustness as a correlated response in the digital organisms model. AVIDANs achieve robustness by reducing the proportion of their genomes that is susceptible to deleterious mutations. Shown at left is the average across all populations of the number of logic functions ("tasks") performed over the course of the experiment. Shown at right is the average across all populations of the number of instructions (sites) used to perform logic functions. The total number of instructions in the genome was 70.

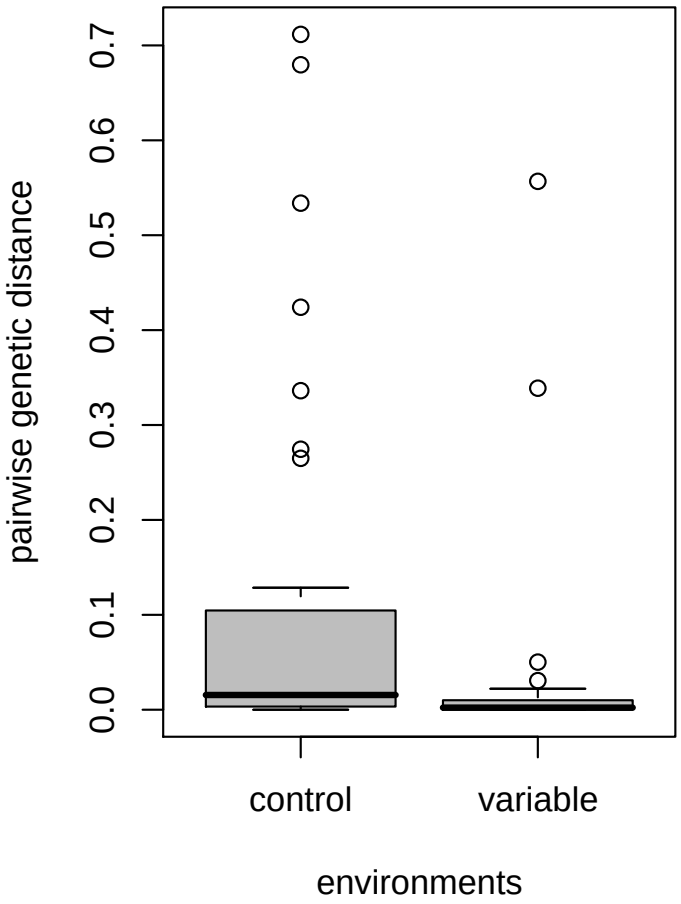


Figure 18: Pairwise genetic diversity in digital organism populations is reduced in variable environments. Shown is the boxplot distribution of the Hamming distance for all populations at the end of the evolution experiment. Populations evolved in the constant environment are shown at left, those evolved in the variable environment are shown at right. These distributions are significantly different by a Wilcoxon signed rank test (p-value = 0.004). Whisker lines are drawn at the most extreme data point that is no more than 1.5 times the interquartile range from the box, and points beyond are shown as circles.

Bibliography

[1] AB Abecasis, P Lemey, N Vidal, T de Oliveira, M Peeters, R Camacho, B Shapiro, A Rambaut, and AM Vandamme. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype g is a circulating recombinant form. *J Virol*, 81(16):8543–8551, 2007.

[2] L. W. Ancel and W. Fontana. Plasticity, evolvability, and modularity in rna. *J Exp Zool*, 288(3):242–283, 2000.

[3] RB Azevedo, R Lohaus, S Srinivasan, KK Dang, and CL Burch. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature*, 440(7080): 87–90, 2006.

[4] CS Badorrek and KM Weeks. Rna flexibility in the dimerization domain of a gamma retrovirus. *Nat Chem Biol*, 1(2):104–111, 2005.

[5] R Belshaw, OG Pybus, and A Rambaut. The evolution of genome compression and genomic novelty in rna viruses. *Genome Res*, 17(10):1496–1504, 2007.

[6] B Berkhout. Structural features in tar rna of human and simian immunodeficiency viruses: a phylogenetic analysis. *Nucleic Acids Res*, 20(1):27–31, 1992.

[7] B Berkhout. The primer binding site on the rna genome of human and simian immunodeficiency viruses is flanked by an upstream hairpin structure. *Nucleic Acids Res*, 25(20): 4013–4017, 1997.

[8] B Berkhout and I Schoneveld. Secondary structure of the hiv-2 leader rna comprising the trna-primer binding site. *Nucleic Acids Res*, 21(5):1171–1178, 1993.

[9] Elhanan Borenstein and Eytan Ruppin. Direct evolution of genetic robustness in microrna. *Proc Natl Acad Sci U S A*, 103(17):6593–6598, 2006.

[10] RR Breaker. Complex riboswitches. *Science*, 319(5871):1795–1797, 2008.

[11] CL Burch and L Chao. Epistasis and its relationship to canalization in the rna virus phi 6. *Genetics*, 167(2):559–567, 2004.

[12] JV Chamary and LD Hurst. Evidence for selection on synonymous mutations affecting stability of mrna secondary structure in mammals. *Genome Biol*, 6(9):R75, 2005.

[13] JH Chen, SY Le, and JV Maizel. Prediction of common secondary structures of rnas: a genetic algorithm approach. *Nucleic Acids Res*, 28(4):991–999, 2000.

[14] Y Chen, DB Carlini, JF Baines, J Parsch, JM Braverman, S Tanda, and W Stephan. Rna secondary structure and compensatory evolution. *Genes Genet Syst*, 74(6):271–286, 1999.

[15] P Clote. Introduction to special issue on rna. *J Math Biol*, 56(1-2):3–13, 2008.

[16] K Clyde and E Harris. Rna secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. *J Virol*, 80(5):2170–2182, 2006.

[17] VS Cooper and RE Lenski. The population genetics of ecological specialization in evolving escherichia coli populations. *Nature*, 407(6805):736–739, 2000.

[18] MC Cowperthwaite, JJ Bull, and LA Meyers. Distributions of beneficial fitness effects in rna. *Genetics*, 170(4):1449–1457, 2005.

[19] MC Cowperthwaite, JJ Bull, and LA Meyers. From bad to good: Fitness reversals and the ascent of deleterious mutations. *PLoS Comput Biol*, 2(10), 2006.

[20] BR Cullen. Role and mechanism of action of the apobec3 family of antiretroviral resistance factors. *J Virol*, 80(3):1067–1076, 2006.

[21] CK Damgaard, ES Andersen, B Knudsen, J Gorodkin, and J Kjems. Rna interactions in the 5' region of the hiv-1 genome. *J Mol Biol*, 336(2):369–379, 2004.

[22] J. A. de Visser, J. Hermisson, G. P. Wagner, L. Ancel Meyers, H. Bagheri-Chaichian, J. L. Blanchard, L. Chao, J. M. Cheverud, S. F. Elena, W. Fontana, G. Gibson, T. F. Hansen, D. Krakauer, R. C. Lewontin, C. Ofria, S. H. Rice, G. von Dassow, A. Wagner, and M. C. Whitlock. Perspective: Evolution and detection of genetic robustness. *Evolution Int J Org Evolution*, 57(9):1959–1972, 2003.

[23] K Deforche, R Camacho, KV Laethem, B Shapiro, Y Moreau, A Rambaut, AM Vandamme, and P Lemey. Estimating the relative contribution of dntp pool imbalance and apobec3g/3f editing to hiv evolution in vivo. *J Comput Biol*, 14(8):1105–1114, 2007.

[24] Richard Durbin and Sean Eddy. *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK : New York, 1998.

[25] I Dworkin and G Gibson. Epidermal growth factor receptor and transforming growth factor-beta signaling contributes to variation for wing shape in drosophila melanogaster. *Genetics*, 173(3):1417–1431, 2006.

[26] Mario A. Fares, Mario X. Ruiz-Gonzalez, Andres Moya, Santiago F. Elena, and Eladio Barrio. Endosymbiotic bacteria: Groel buffers against deleterious mutations. *Nature*, 417 (6887):398–398, 2002.

[27] J Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[28] MT Ferris, P Joyce, and CL Burch. High frequency of mutations that expand the host range of an rna virus. *Genetics*, 176(2):1013–1022, 2007.

[29] T Flatt. The evolutionary genetics of canalization. *Quarterly Review of Biology*, 80(3): 287–316, 2005.

[30] W Fontana and P Schuster. A computer model of evolutionary optimization. *Biophys Chem*, 26(2-3):123–147, 1987.

[31] W Fontana and P Schuster. Continuity in evolution: on the nature of transitions. *Science*, 280(5368):1451–1455, 1998.

[32] PP Gardner and R Giegerich. A comprehensive comparison of comparative rna structure prediction approaches. *BMC Bioinformatics*, 5:140, 2004.

[33] G. Gibson and G. Wagner. Canalization in evolutionary genetics: a stabilizing theory? *Bioessays*, 22(4):372–380, 2000.

[34] RS Harris and MT Liddament. Retroviral restriction by apobec proteins. *Nat Rev Immunol*, 4(11):868–877, 2004.

[35] M Hasegawa, H Kishino, and T Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22(2):160–174, 1985.

[36] J Hermisson and GP Wagner. The population genetic theory of hidden variation and genetic robustness. *Genetics*, 168(4):2271–2284, 2004.

[37] O Hobert. Gene regulation by transcription factors and microrna s. *Science*, 319(5871): 1785–1786, 2008.

[38] IL Hofacker, M Fekete, C Flamm, MA Huynen, S Rauscher, PE Stolorz, and PF Stadler. Automatic detection of conserved rna structure elements in complete rna virus genomes. *Nucl. Acids Res.*, 26(16):3825–3836, 1998.

[39] I.L. Hofacker, P.F. Stadler, and R.R. Stocsits. Conserved rna secondary structures in viral genomes: a survey. *Bioinformatics*, 20(10):1495–1499, 2004.

[40] P Hogeweg and B Hesper. Evolutionary dynamics and the coding structure of sequences: Multiple coding as a consequence of crossover and high mutation rates. *Computers & Chemistry*, 16(2):171–182, 1992.

[41] EC Holmes. Error thresholds and the constraints to rna virus evolution. *Trends Microbiol*, 11(12):543–546, 2003.

[42] MA Huynen, DA Konings, and P Hogeweg. Multiple coding and the evolutionary properties of rna secondary structure. *J Theor Biol*, 165(2):251–267, 1993.

[43] MA Huynen, PF Stadler, and W Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci U S A*, 93(1):397–401, 1996.

[44] H Innan and W Stephan. Selection intensity against deleterious mutations in rna secondary structures and rate of compensatory nucleotide substitutions. *Genetics*, 159(1): 389–399, 2001.

[45] K Katoh, K Misawa, K Kuma, and T Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, 30(14):3059–3066, 2002.

[46] L Katz and CB Burge. Widespread selection for local rna secondary structure in coding regions of bacterial genes. *Genome Res*, 13(9):2042–2051, 2003.

[47] TL Kieffer, P Kwon, RE Nettles, Y Han, SC Ray, and RF Siliciano. G→a hypermutation in protease and reverse transcriptase regions of human immunodeficiency virus type 1 residing in resting cd4+ t cells in vivo. *J Virol*, 79(3):1975–1980, 2005.

[48] M Kimura. The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, 64:7–19, 1985.

[49] H Kitano. Biological robustness. *Nat Rev Genet*, 5(11):826–837, 2004.

[50] PS Klosterman, AV Uzilov, YR Bendana, RK Bradley, S Chao, C Kosiol, N Goldman, and I Holmes. Xrate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, 7:428, 2006.

[51] JL Knies, KK Dang, TJ Vision, NG Hoffman, R Swanstrom, and CL Burch. Compensatory evolution in rna secondary structures increases substitution rate variation among sites. *Mol Biol Evol*, In Review, 2008.

[52] B Knudsen and J Hein. Rna secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, 1999.

[53] B Knudsen and J Hein. Pfold: Rna secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, 2003.

[54] S Kosakovsky Pond and SV Muse. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*, 22(12):2375–2385, 2005.

[55] SL Kosakovsky Pond, SD Frost, and SV Muse. Hyphy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679, 2005.

[56] JM Lanchy and JS Lodmell. An extended stem-loop 1 is necessary for human immunodeficiency virus type 2 replication and affects genomic rna encapsidation. *J Virol*, 81(7): 3285–3292, 2007.

[57] JM Lanchy, CA Rentz, JD Ivanovitch, and JS Lodmell. Elements located upstream and downstream of the major splice donor site influence the ability of hiv-2 leader rna to dimerize in vitro. *Biochemistry*, 42(9):2634–2642, 2003.

[58] SY Le, JH Chen, MJ Braun, MA Gonda, and JV Maizel. Stability of rna stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (hiv-i). *Nucleic Acids Res*, 16(11):5153–5168, 1988.

[59] SY Le, MH Malim, BR Cullen, and JV Maizel. A highly conserved rna folding region coincident with the rev response element of primate immunodeficiency viruses. *Nucleic Acids Res*, 18(6):1613–1623, 1990.

[60] T Leitner, B Korber, M Daniels, C Calef, and B Foley. *HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences, 2005*. Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, 2005.

[61] RE Lenski, C Ofria, TC Collier, and C Adami. Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400(6745):661–664, 1999.

[62] RE Lenski, C Ofria, RT Pennock, and C Adami. The evolutionary origin of complex features. *Nature*, 423(6936):139–144, 2003.

[63] EA Lesnik, R Sampath, and DJ Ecker. Rev response elements (rre) in lentiviruses: an rnamotif algorithm-based strategy for rre prediction. *Med Res Rev*, 22(6):617–636, 2002.

[64] H Liang and LF Landweber. A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome Res*, 16(2):190–196, 2006.

[65] R Luck, G Steger, and D Riesner. Thermodynamic prediction of conserved secondary structure: application to the rre element of hiv, the trna-like element of cmv and the mrna of prion protein. *J Mol Biol*, 258(5):813–826, 1996.

[66] DH Mathews, J Sabina, M Zuker, and DH Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *J Mol Biol*, 288(5):911–940, 1999.

[67] DH Mathews, MD Disney, JL Childs, SJ Schroeder, M Zuker, and DH Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proc Natl Acad Sci U S A*, 101(19):7287–7292, 2004.

[68] JS McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[69] LK McMullan, A Grakoui, MJ Evans, K Mihalik, M Puig, AD Branch, SM Feinstone, and CM Rice. Evidence for a functional rna element in the hepatitis c virus core gene. *Proc Natl Acad Sci U S A*, 104(8):2879–2884, 2007.

[70] CD Meiklejohn and DL Hartl. A single mode of canalization. *Trends in Ecology & Evolution*, 17(10):468–473, 2002.

[71] EJ Merino, KA Wilkinson, JL Coughlan, and KM Weeks. Rna structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (shape). *J Am Chem Soc*, 127(12):4223–4231, 2005.

[72] IM Meyer. A practical guide to the art of rna gene prediction. *Brief Bioinform*, 8(6): 396–414, 2007.

[73] IM Meyer and I Miklos. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mrnas and pre-mrnas. *Nucleic Acids Res*, 33(19):6338–6348, 2005.

[74] LA Meyers, JF Lee, M Cowperthwaite, and AD Ellington. The robustness of naturally and artificially selected nucleic acid secondary structures. *J Mol Evol*, 58(6):681–691, 2004.

[75] R Montville, R Froissart, SK Remold, O Tenaillon, and PE Turner. Evolution of mutational robustness in an rna virus. *PLoS Biology*, 3(11):1939–1945, 2005.

[76] S.V. Muse. Evolutionary analyses of dna sequences subject to constraints on secondary structure. *Genetics*, 139(3):1429–1439, 1995.

[77] EA Ostrowski, DE Rozen, and RE Lenski. Pleiotropic effects of beneficial mutations in escherichia coli. *Evolution Int J Org Evolution*, 59(11):2343–2352, 2005.

[78] C Pace, J Keller, D Nolan, I James, S Gaudieri, C Moore, and S Mallal. Population level analysis of human immunodeficiency virus type 1 hypermutation and its relationship with apobec3g and vif genetic variation. *J Virol*, 80(18):9259–9269, 2006.

[79] K Pachulska-Wieczorek, KJ Purzycka, and RW Adamiak. New, extended hairpin form of the tar-2 rna domain points to the structural polymorphism at the 5' end of the hiv-2 leader rna. *Nucleic Acids Res*, 34(10):2984–2997, 2006.

[80] JC Paillart, M Dettenhofer, XF Yu, C Ehresmann, B Ehresmann, and R Marquet. First snapshots of the hiv-1 rna structure in infected cells and in virions. *J Biol Chem*, 279 (46):48397–48403, 2004.

[81] J Parsch, JM Braverman, and W Stephan. Comparative sequence analysis and patterns of covariation in rna secondary structures. *Genetics*, 154(2):909–921, 2000.

[82] JS Pedersen, R Forsberg, IM Meyer, and J Hein. An evolutionary model for protein-coding regions with conserved rna structure. *Mol Biol Evol*, 21(10):1913–1922, 2004.

[83] JS Pedersen, IM Meyer, R Forsberg, P Simmonds, and J Hein. A comparative method for finding and folding rna secondary structures within protein-coding regions. *Nucleic Acids Res*, 32(16):4925–4936, 2004.

[84] JS Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, ES Lander, J Kent, W Miller, and D Haussler. Identification and classification of conserved rna secondary structures in the human genome. *PLoS Comput Biol*, 2(4):e33, 2006.

[85] M Peeters and V Courgnaud. *Overview of Primate Lentiviruses and their Evolution in Non-human Primates in Africa*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, 2002.

[86] O Peleg, EN Trifonov, and A Bolshoy. Hidden messages in the nef gene of human immunodeficiency virus type 1 suggest a novel rna secondary structure. *Nucleic Acids Res*, 31(14):4192–4200, 2003.

[87] O Peleg, V Kirzhner, E Trifonov, and A Bolshoy. Overlapping messages and survivability. *J Mol Evol*, 59(4):520–527, 2004.

[88] SR Proulx and PC Phillips. The opportunity for canalization and the evolution of genetic networks. *American Naturalist*, 165(2):147–162, 2005.

[89] SA Rifkin, D Houle, J Kim, and KP White. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature*, 438(7065):220–223, 2005.

[90] E Rivas and SR Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding rnas. *Bioinformatics*, 16(7):583–605, 2000.

[91] SL Rutherford and S Lindquist. Hsp90 as a capacitor for morphological evolution. *Nature*, 396(6709):336–342, 1998.

[92] HA Schmidt, K Strimmer, M Vingron, and A von Haeseler. Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–504, 2002.

[93] P Schuster, W Fontana, PF Stadler, and IL Hofacker. From sequences to shapes and back: a case study in rna secondary structures. *Proc Biol Sci*, 255(1344):279–284, 1994.

[94] ML Siegal and A Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proc Natl Acad Sci U S A*, 99(16):10528–10532, 2002.

[95] P. Simmonds and D. B. Smith. Structural constraints on rna virus evolution. *J. Virol.*, 73(7):5787–5794, 1999.

[96] P Simmonds, A Tuplin, and DJ Evans. Detection of genome-scale ordered rna structure (gors) in genomes of positive-stranded rna viruses: Implications for virus evolution and host persistence. *RNA*, 10(9):1337–1351, 2004.

[97] T Sing, O Sander, N Beerenwinkel, and T Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, 2005.

[98] JE Stajich, D Block, K Boulez, SE Brenner, SA Chervitz, C Dagdigian, G Fuellen, JG Gilbert, I Korf, H Lapp, H Lehvaslaiho, C Matsalla, CJ Mungall, BI Osborne, MR Pocock, P Schattner, M Senger, LD Stein, E Stupka, MD Wilkinson, and E Birney. The bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618, 2002.

[99] SC Stearns, M Kaiser, and TJ Kawechki. The differential genetic and environmental canalization of fitness components in drosophila melanogaster. *J Evol Biol*, 8(5):539–557, 1995.

[100] W Stephan. The rate of compensatory evolution. *Genetics*, 144(1):419–426, 1996.

[101] R Suspene, C Rusniok, JP Vartanian, and S Wain-Hobson. Twin gradients in apobec3 edited hiv-1 dna reflect the dynamics of lentiviral replication. *Nucleic Acids Res*, 34(17): 4677–4684, 2006.

[102] Caroline Thurner, Christina Witwer, Ivo L. Hofacker, and Peter F. Stadler. Conserved rna secondary structures in flaviviridae genomes. *J Gen Virol*, 85(5):1113–1124, 2004.

[103] A Tuplin, J Wood, DJ Evans, AH Patel, and P Simmonds. Thermodynamic and phylogenetic prediction of rna secondary structures in the coding region of hepatitis c virus. *RNA*, 8(6):824–841, 2002.

[104] A Tuplin, DJ Evans, and P Simmonds. 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(Pt 10):3037–3047, 2004.

[105] E van Nimwegen, JP Crutchfield, and M Huynen. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A*, 96(17):9716–9720, 1999.

[106] A Wagner. Does evolutionary plasticity evolve? *Evolution*, 50(3):1008–1023, 1996.

[107] A Wagner. *Robustness and Evolvability in Living Systems*. Princeton University Press, Princeton, 2005.

[108] A. Wagner and P. F. Stadler. Viral rna and evolved mutational robustness. *J Exp Zool*, 285(2):119–127, 1999.

[109] GP Wagner, G Booth, and H Bagheri-Chaichian. A population genetic theory of canalization. *Evolution*, 51(2):329–347, 1997.

[110] John Wakeley. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution*, 11 (4):158–162, 1996.

[111] Stefan Washietl and L. Hofacker, Ivo. Consensus folding of aligned sequences as a new measure for the detection of functional rnas by comparative genomics. *Journal of Molecular Biology*, 342(1):19–30, 2004.

[112] J Watts and KM Weeks. personal communication. 2008.

[113] CO Wilke, JL Wang, C Ofria, RE Lenski, and C Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, 2001.

[114] Christina Witwer, Susanne Rauscher, Ivo L. Hofacker, and Peter F. Stadler. Conserved rna secondary structures in picornaviridae genomes. *Nucl. Acids Res.*, 29(24):5079–5089, 2001.

[115] S Wuchty, W Fontana, IL Hofacker, and P Schuster. Complete suboptimal folding of rna and the stability of secondary structures. *Biopolymers*, 49(2):145–165, 1999.

[116] Z Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol*, 11(9):367–372, 1996.

[117] Q Yu, R Konig, S Pillai, K Chiles, M Kearney, S Palmer, D Richman, JM Coffin, and NR Landau. Single-strand specificity of apobec3g accounts for minus-strand deamination of the hiv genome. *Nat Struct Mol Biol*, 11(5):435–442, 2004.

[118] AJ Zaug and TR Cech. The intervening sequence rna of tetrahymena is an enzyme. *Science*, 231(4737):470–475, 1986.

[119] M Zuker. On finding all suboptimal foldings of an rna molecule. *Science*, 244(4900): 48–52, 1989.

[120] M Zuker and P Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, 1981.