Viral RNA and Evolved Mutational Robustness

Andreas Wagner[a,c,*] and Peter F. Stadler[b,c]

[a,*]Department of Biology, University of New Mexico
167A Castetter Hall, Albuquerque, NM 87131-1091, U.S.A.
Phone: +1-505-277-2021; Fax: +1-505-277-0304
wagnera@unm.edu

[b]Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: ++43 1 4277 52737, Fax: ++43 1 4277 52793
studla@tbi.univie.ac.at, stadler@santafe.edu

[c]The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA
Phone: +1-505-984-8800 Ext. 231; Fax: +1-505-982-0565

Abstract
Many properties of organisms show great robustness against mutations. Whether this robustness is an evolved property or intrinsic to genetic systems is by and large unknown. An evolutionary origin of robustness would require a rethinking of key concepts in the field of molecular evolution, such as gene-specific neutral mutation rates, or the context-independence of deleterious mutations. We provide evidence that mutational robustness of the genome of RNA viruses to mutational changes in secondary structure has evolved.

Keywords: Canalization, Epigenetic Landscapes, Neutral Evolution, RNA Folding

1. Introduction

Even the simplest known organisms consist of thousands of molecular components which form a highly integrated network of interactions. An immense number of parameters that specify these interactions have to be finely tuned to ensure survival and reproduction of a cell. With this view in mind, and considering that organisms are constantly subject to mutations, it may seem surprising that multicellular organisms evolved at all. When layer upon layer of suborganismal organization is accumulated in evolution, surely there must be a point at which a system's complexity becomes unsustainable, at which the "house of cards" collapses?

In contrast to this perspective, many subsystems of organisms are quite robust to mutations, as exemplified by the high tolerance of proteins to amino acid changes (Bowie *et al.*, 1990), the resilience of metabolic flux to changes in enzyme activity (Hartl *et al.*, 1985; Kacser and Burns, 1981), and the robustness of developmental pathways to mutations in their constituent genes (Cadigan *et al.*, 1994; Dun and Fraser, 1958; Joyner *et al.*, 1991; Rendel, 1979; Tautz, 1992; Weintraub, 1993). The mechanistic basis for robustness in each of these cases depends on the peculiarities of the system studied, raising the possibility that robustness is an intrinsic characteristic of these systems. A less explored alternative is that observed robustness is a property that has evolved, perhaps because it protects a biological system against deleterious mutations. Here we show that evolution of robustness took place in the functional RNA secondary structures of RNA viral genomes.

In contrast to conventional scenarios of natural selection, where selection acts directly on fitness differences in genetic variants, the selective advantage of a robust system is its decreased susceptibility to deleterious mutations. Because of this indirect, "second order", and thus slow mechanism it is highly unlikely that the evolution of robustness can be directly observed in the laboratory (Wagner, 1999). However, the outcome of this process can be observed: Features of biological systems that are under very strong stabilizing selection, and thus highly conserved, should display higher mutational robustness than non-conserved ones[1]. Anecdotal evidence that can be interpreted in this way derives from a conspicuous lack of variation in many basic *Bauplan* characters, and has led to the concept of *genetic canalization* of these characters, implying robustness in the developmental pathways generating them (Waddington, 1957; Rendel, 1979). However, with the exception of a study on life history traits in *Drosophila* (Stearns and Kawecki, 1994), to our knowledge no systematic effort has been made to test this prediction.

The genome of RNA viruses presents us with an ideal study system in this regard. Parts of the genomic RNA form conserved secondary structures that are crucial for regulating the viral life cycle, e.g., the TAR (Baudin *et al.*, 1993) and the RRE (Dayton *et al.*, 1992) elements of HIV1, the 3' stem-loop structures of Flaviviruses (Brinton *et al.*, 1986), the IRES region of picorna viruses (Jackson and Kaminski, 1995), or the 3'UTR of flaviviruses (Mandl *et al.*, 1998). A large fraction of the genome, on the other hand, encodes proteins. These regions form secondary structures of little or no functional importance but comparable (thermodynamic)

[1] It must be noted that an increase in robustness need not necessarily to increased population mean fitness. This is because mean equilibrium fitness does not depend on the average effect of a mutation, but only on the mutation rate, a population genetic principle often referred to as the Haldane-Muller principle.

stability. A single RNA virus genome thus contains both conserved structural elements, which we expect to display high mutational robustness, and non-conserved elements which serve as the non-robust control.

2. Methods

The number of experimentally documented conserved secondary structure elements is too small to yield anything but anecdotal evidence for our proposition (Huynen *et al.*, 1993). A computational procedure, however, has recently become available that reliably detects conserved secondary structure motifs from a moderate sample of related RNA sequences (Hofacker *et al.*, 1998; Hofacker and Stadler, 1999) by combining thermodynamics-based secondary structure predictions and comparative sequence analysis. It extracts conserved secondary structure elements from moderately sized datasets of related viral sequences using a combination of thermodynamic secondary structure prediction and multiple sequence alignments.

RNA secondary structures are computed using a robust and accurate minimum energy folding algorithm (Zuker and Sankoff, 1984), as implemented in the `Vienna RNA Package`, version 1.2 (Hofacker *et al.*, 1994). The energy parameters used by the `Vienna RNA Package` are based on (Freier *et al.*, 1986; Jaeger *et al.*, 1989; He *et al.*, 1991). Complete RNA genomes were folded on CalTech's `Delta` using the message passing version of the minimum folding algorithm described in (Hofacker *et al.*, 1996a,b). The computed structures are in good agreement with experimentally determined secondary structures (Huynen *et al.*, 1997; Schuster *et al.*, 1997; Tacker *et al.*, 1996).

The software package `Clustal W` (Thompson *et al.*, 1994) is used to generate a multiple alignment of the viral *sequences*. The program `alidot` (Hofacker *et al.*, 1998) translates the sequence alignment into an alignment of the *secondary structures*. It does so by generating a list of all base pairs $(i.j)$ occuring in the aligned minimum energy secondary structures of the entire viral data set. This list is then sorted according to the following hierarchical criteria: (1) The more sequences are non-compatible with $(i.j)$, the less supported is the base pair. (2) Symmetric base pairs, i.e., base pairs (i,j) for which i is the most frequent pairing partner of j and j is the most frequently predicted pairing partner of i, are better supported than other base pairs. (3) A base pair with more consistent (compensatory) mutations is better supported. (4) Base pairs that are predicted more frequently are better supported. Scanning this sorted list of base pairs in the order of decreasing support, a base pair is removed if it conflicts with a higher-ranking one that has already been accepted. Finally, base pairs are collected into stems and those stems that are both predicted in a sizeable fraction of the sequences and that are supported by sequence covariations are retained in the final output. A detailed description of this procedure can be found in (Hofacker *et al.*, 1998).

The approach pursued here consists in a comparison of various measures of mutational stability between conserved and non-conserved viral secondary structure elements. Using the method just described, conserved secondary structure elements were extracted from three different groups of single-stranded RNA viruses, namely Dengue virus (Genbank names: DEN2CGA, DEN2CMEMSA, DEN2CMEMSB, DEN2JAMCG, DEN2RCG, DEN3CME, DEN4STRA, DENT1SEQ), Hepatitis C virus (Genbank names: HCU16362, HCU45476, HPCCGAA, HPCCGENOM, HPC-CGS, HPCEGS, HPCHCJ1, HPCJ483, HPCJRNA, HPCJTA, HPCK3A, HPCPP, HPCRNA; these sequences do not contain the 98nt long X-tail at the 3' terminus), and HIV1 (Genbank names: HIVANT70, HIVBCSG3C, HIVCAM1, HIVD31, HIVELI, HIVLAI, HIVMAL, HIVMVP5180, HIVNDK, HIVOYI, HIVRF, HIVU455, HIVZ2Z6). The selected sequences represent a maximally diverse subset of the available completely sequenced genomes for each family. The mean pairwise sequence similarities range from 77 to 81 percent.

A total of 34, 9, and 13 conserved elements were thus isolated from the Dengue, Hepatitis C, and HIV groups, respectively. Further, a sample of 181 non-conserved structural elements was isolated from representatives of each of the three groups (Jamaican genotype of Dengue, GenBank acc.no. M20558, Japanese genotype of Hepatitis C, acc.nos. D14484, D01173; African genotype of HIV, acc.no. L20587). This was done by (i) extracting randomly chosen fragments spanning 300 nucleotides and not overlapping a conserved element from each of these sequence data sets, (ii) computing the secondary structures of these fragments, and (iii) extracting, from each such fragment, a subsequence corresponding to a well-defined secondary structure. A total of 64, 55, and 62 non-conserved elements were thus isolated from Dengue, Hepatitis C, and HIV, respectively.

For each of the extracted conserved and non-conserved elements, we determined two complementary measures of mutational stability. We first analyzed the effects of mutations inside an element by generating all possible single point mutations within all conserved and non-conserved elements ($3n$ mutations for an element of n nucleotides). For each point mutation we then numerically evaluated the minimum free energy secondary structure of the mutated element. We compared the secondary structure of a mutated element to that of the unmutated element using a variety of distance measures for secondary structures, including tree-edit distance and string distance (Hofacker *et al.*, 1994; Hogeweg and Hesper, 1984). The qualitative results obtained were independent of the distance measure used.

Second, we determined the stability of secondary structure elements to changes in sequence context, i.e., to changes in sequences surrounding the element, in two ways: (i) We folded a window of the genomic consensus sequence containing both the conserved element and flanking sequences of different lengths, ranging from one nucleotide to the length of the element. We folded each of these windows, and compared the secondary structure of the segment encompassing the element of interest to its "wild-type" structure. (ii) To emulate the effect of changes in sequence context, we attached random RNA sequences of varying length to a conserved element of interest, folded the resulting sequence, and compared the resulting structure to the "wild-type" as in (i). Ideally, one would of course wish to introduce mutations anywhere in the viral genome, and to evaluate their effect on any particular element by folding the entire mutated genome. This is computationally infeasible. Computing the secondary structures of long RNAs sequences is a time-consuming task since CPU requirements scale as $\mathcal{O}(n^3)$ and memory requirements scale as $\mathcal{O}(n^2)$ with the chain length n. The computation for a single HIV-1 sequence, $n \approx 9200$, for instance, takes several hours on a massively parallel computer and requires more than 1GByte of memory. These considerations motivated the more limited approach we pursued here.

3. Results

The genetic perturbations that may affect the secondary structure of a short sequence element within a larger genome fall into two classes: those that take place within the element, and those that take place in the remainder of the genome. We compared the effect of point mutations in conserved and non-conserved elements by exhaustively mutating all positions within each isolated element in the way described above. Figures 1a and 1b show extreme examples of a non-conserved and a conserved HIV-1 secondary structure element, respectively, along with histograms of the distribution of structure distances for all single point mutations. As judged by the average structure distance, the conserved element is six-fold more stable than the non-conserved element.

Figure 2a summarizes analogous results for all conserved and non-conserved structure elements, using two different structure distance measures. It shows that conserved elements are consistently more stable than non-conserved elements, although the stability increase of conserved elements is in general less than 50 percent. Figure 2b considers only the fraction of single point mutants whose secondary structure does not differ at all from that of the unmutated element. Here, conserved elements longer than 70 base pairs show a higher frequency of neutral mutations with respect to secondary structure than non-conserved elements.

The second class of mutations is located outside the secondary structure elements of interest. Due to the nonlocal nature of secondary structure formation they may nevertheless affect the sub-structure of interest. A distant position that underwent mutation may now interact preferably with the structural motif under consideration (in which case the structure of interest is disrupted), or with a position in its close vicinity, thereby changing the structures in its vicinity. Formerly unpaired regions may thus become part of a secondary structure motif. Conversely, sequences that were part of a helix before the mutation, may become available for interaction with the element of interest. Mutations that affect this "genomic context" of a sequence element are by no means rare (Fontana *et al.*, 1993b).

Because it is infeasible to compute the secondary structure of a complete genome for a large number of mutants, we assess the sensitivity of a structural element to changes in its genomic context by adding either genomic or random flanking sequences of varying length as described in "Methods". Adding flanking *genomic* sequences of varying size emulates the effect of structural changes elsewhere in the molecule, whereas adding *random* sequences models the effect of changes in sequence context. The latter is motivated by the observation that many RNA viruses have regions with enormously high sequence variability. Insertions and deletions have an especially profound influence on this variability. Their abundance is exemplified by our sample of HIV1 sequences, where approximately 7 insertions or deletions per 1000nt are observed. Another impressive example is the hypervariable part of the 3' noncoding region in flaviviruses (Wallner *et al.*, 1995) which is located adjacent to a particularly well-conserved piece of secondary structure (Mandl *et al.*, 1998) but which itself has been completely randomized.

The effects of genomic and random context differ greatly: The grayscale in Figure 3a,e corresponds to the similarity in secondary structure: light (dark) points correspond to fragments with attain secondary structures that are similar (dissimilar) to the fragment in Figure 3d. String distance, that is, the Hamming distance of

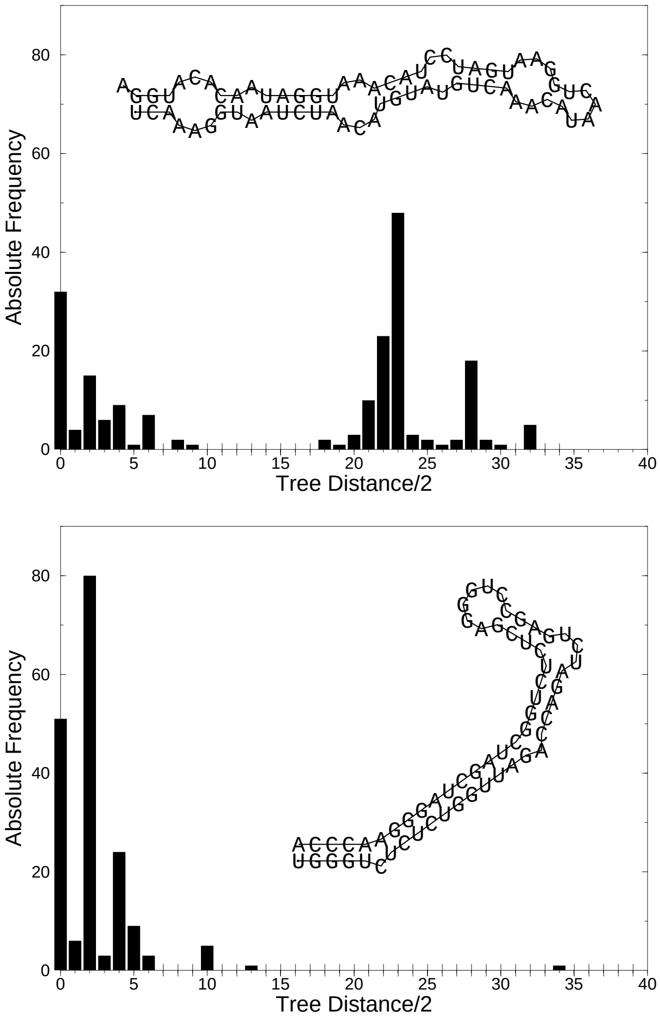

Figure 1: Conserved and non-conserved elements differ in robustness.
Above: Absolute frequency of structure distances for all single point mutations of the *non-conserved* HIV-1 structural element shown in the inset (position 7089-7155 of the HIV-1 sequence HIVANT70, Genbank acc. no. M31171). The tree edit distance plotted here is one of several distance measures for RNA secondary structures (Fontana *et al.*, 1993a). Mean/standard error of structure distance normalized by length of element: $4.7 \times 10^{-1}/2.3 \times 10^{-2}$.
Below: Absolute frequency of structure distances for all single point mutations of a HIV-1 *conserved* secondary structure element. Mean/standard error of normalized distance: $7.7 \times 10^{-2}/7.8 \times 10^{-3}$.

the dot-parenthesis encoding of the secondary structure, see Hogeweg and Hesper (1984), was used as a structure distance measure. Despite the striking differences in the appearance of the two grid plots for genomic and random context, the overall distance (gray-level) increases in both as more nucleotides are added.

Figure 3b,c,f,g shows that conserved secondary structure elements display a higher degree of robustness to addition of genomic or random sequences than non-conserved

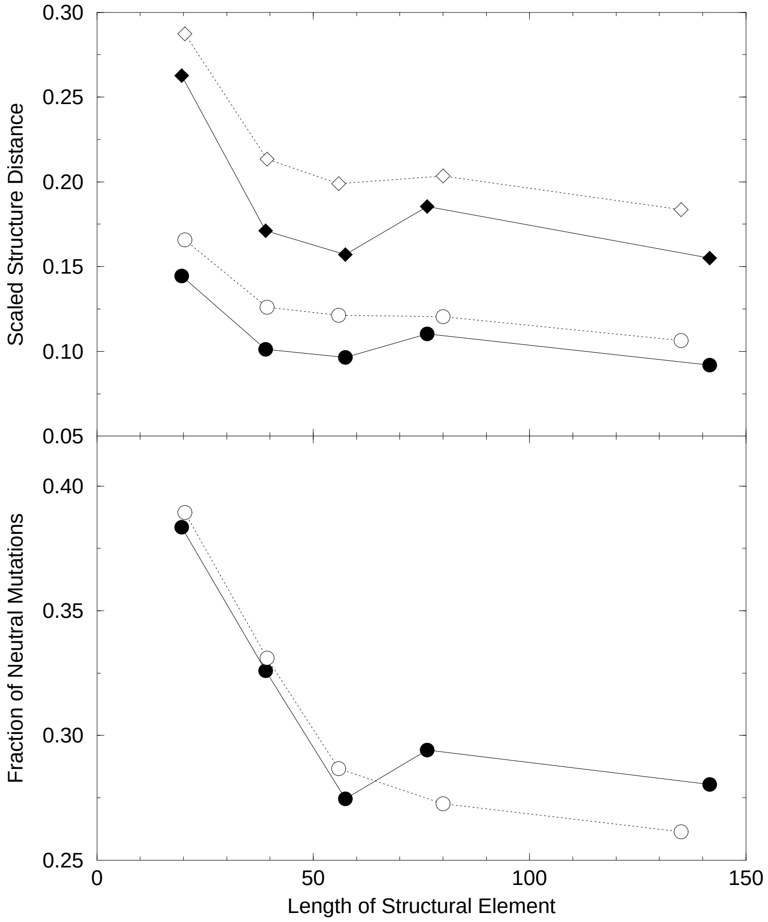

Figure 2: Robustness against point mutations.
Above: Mean normalized tree edit distance (diamonds) and string distance (circles) for all single point mutations of non-conserved (open symbols) and conserved (black symbols) secondary structure elements, as a function of the primary sequence length. Elements were binned into length intervals (9-30 nucleotides: 19 conserved/73 non-conserved elements; 30-50: 19/45; 50-70: 9/28; 70-90: 3/15; 90-192: 6/20), for which the mean lengths are plotted on the abscissa.
Below: Mean fraction of point mutations that cause no deviation from the secondary structure of the native element for non-conserved (○) and conserved (●) structure elements. Differences in this fraction of ***structurally*** neutral mutations are only apparent for elements longer than 70 nucleotides.

elements, with the possible exception of sequences between 70 and 90 nucleotides. The number of sequences in this interval, however, is very small in our data set, and thus highly susceptible to statistical outliers.

Kendall's τ rank correlation coefficient between stability measures for 53 and 117 conserved and non-conserved structure elements, respectively, are listed in Table 1. Two different measures for stability to the sequence context (added 5' and 3' nucleotides) are considered, one in which stability to all contexts with less than 100 nucleotides is averaged, ("context length ≤ 100", see also Figure 3), and a second

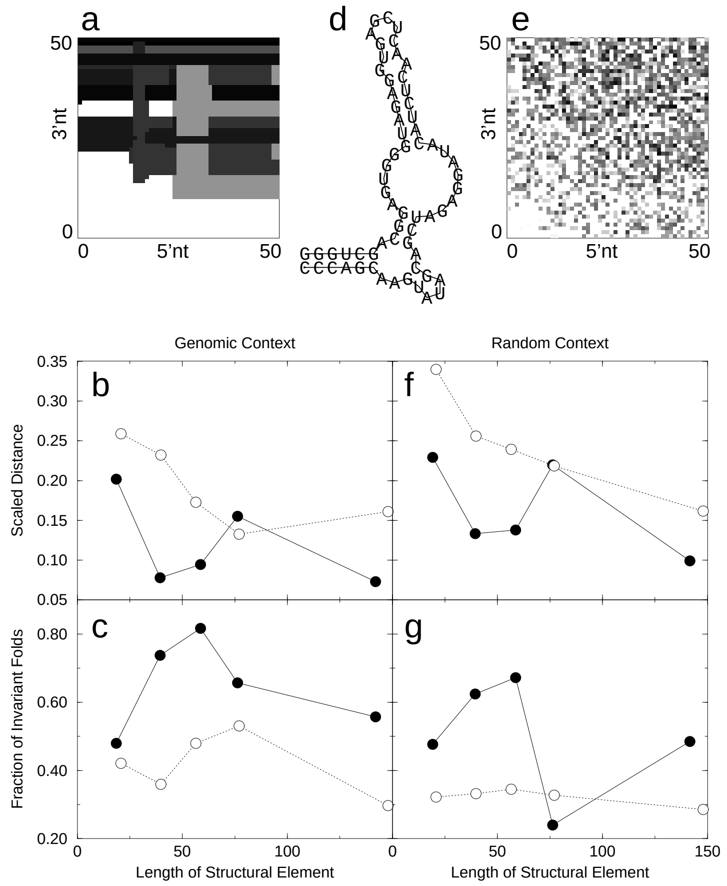

Figure 3: Robustness of secondary structure elements to changes of the genomic context. We consider flanking sequences of different lengths at the 3'- and the 5'-end of a secondary structure element (**d**, taken from the Dengue secondary structure alignment, positions 5399-5462) that are either part of the genomic consensus sequence, (**a-c**), or randomly generated sequences (**e-g**). The 2D grid plots (**a**, **e**) display the structure distance of the mutants to the original structure (**d**) as a function of the number of nucleotides added to the 5' and 3' ends (x- and y-axis, respectively). The mean structure distance (**b**) and the fraction of identical structures (**c**) for all fragments with up to 100 nucleotides of *genomic* sequence added to the fragment is plotted for 53 conserved ($\circ$) and 117 non-conserved ($\bullet$) structure elements. Analogous data are shown in (**f**, **g**) for the addition of up to 50 *random* nucleotides.

measure where the number of nuclotides added is equal to a constant fraction, $n/2$, of the length n of an element. ("context length $= n/2$"). All three stability measures, stability to point mutations, stability to varying genomic context lengths, and stability to random contexts, are significantly correlated. Significance tests were performed against the null-hypothesis that the correlation coefficients are equal to zero (Sokal and Rohlf, 1981). All correlation coefficients are significant at $\mathbf{P} < 0.05$. Pearson product-moment correlation coefficients (not shown) have values similar to those given here. There is a conspicuous increase in all three pairwise correlations

Table 1: Correlations (Kendall's τ rank correlation coefficient) between robustness measures for point mutations (p.m.), genomic context (g.c) and random context (r.c.) for two different context lengths ℓ. See text for details.

	$\ell \leq 100$		$\ell = n/2$	
	τ	**P**	τ	**P**
Conserved Structure Elements				
p.m./g.c.	0.44	4×10^{-6}	0.31	1×10^{-3}
p.m./r.c.	0.59	6×10^{-10}	0.35	2×10^{-4}
g.c./r.c.	0.51	7×10^{-8}	0.46	1×10^{-6}
Non-Conserved Structure Elements				
p.m./g.c.	0.26	5×10^{-5}	0.15	2×10^{-2}
p.m./r.c.	0.57	$< 10^{-10}$	0.30	3×10^{-6}
g.c./r.c.	0.36	2×10^{-8}	0.41	2×10^{-10}

from non-conserved to conserved fragments, although its significance can not be evaluated, given the non-normality of the underlying distribution.

In summary, our results indicate that conserved secondary structure elements from three groups of RNA viruses are more robust to genetic perturbations than their non-conserved counterparts, regardless of whether the perturbations occur inside or outside the element. This indicates that robustness in RNA secondary structures has evolved in RNA viruses.

4. Discussion

Conventional methods of multiple sequence alignment can be used to identify highly conserved sequence elements. In analogy, a multiple alignment of secondary structures can be used to identify highly conserved structural elements. The sequence information, or more precisely, the presence or absence of compensatory mutations is utilized to verify the existence of a conserved element (Hofacker *et al.*, 1998). Thus structural elements are identified as conserved not because of their thermodynamic stability or because of sequence conservation but because there is a consistenly predicted structure in the presence of sequence variation. This comparative approach compensates for the occasional inaccuracies in individually predicted secondary structures. The method has been applied to a variety of virus families and readily produces the known conserved elements such as TAR and RRE of HIV-1, the panhandle structure of Hanta virus, and the 3'UTR of flavivirus (Hofacker *et al.*, 1998; Hofacker and Stadler, 1999).

Conservation of secondary structure features in sequences with about 80% pairwise identity (as in our sample) is likely to be a consequence of functional importance of the secondary structure. Statistical surveys (Fontana *et al.*, 1993a; Bonhoeffer *et al.*, 1993) show that even a small number of mutations is sufficient to alter secondary structure completely, and at a sequence divergence of only 10 percent the overwhelming majority of sequences will fold into structures with at most vague similarities. Conservation of secondary structure among sequences with only 90 percent similarity thus should be seen as a consequence of stabilizing selection, indicating the functional importance of the structure rather than as a consequence of sequence similarity.

The importance of the *in silico* approach taken here is emphasized by the fact that the large numbers of perturbations analyzed here can not be easily studied in laboratory experiments (e.g., more than 20,000 point mutants were generated and folded). The computational approach, however, is not free of problems. While the similarity among the genome sequences used in this study is low enough to allow us to assume that conserved secondary structures have conserved function (Schuster *et al.*, 1994; Schuster and Stadler, 1998), we do not know what this function is. Consequently, we do not know how much deviation in secondary structure is admissible for an element to be functional, and thus for a mutation to be functionally neutral, as opposed to structurally neutral. We take the difference in the measures of *structural* robustness as an indication for a similar difference in the fraction of *functionally* neutral mutations for conserved and non-conserved RNA sequences.

In theory, very simple differences among conserved and non-conserved sequences could be responsible for their differential mutational robustness. These might include differences in GC content and/or the number of paired bases. However, in our data set we did not find any obvious correlations between GC content and the number of paired bases on one hand, and mutational robustness on the other hand. Mutational stability might also be correlated with thermodynamic stability (W. Fontana, pers. comm.). If functionally important secondary structure elements have to be thermodynamically exceptionally stable, then high mutational robustness could be an indirect consequence of selection for thermodynamic stability. However, many viral secondary structure motifs need to be sufficiently flexible to unfold for processes such as replication or translation. Furthermore, it has been shown experimentally that mutations leading to increased thermodynamic stability of several important viral secondary structure motifs have deleterious effects (Olsthoorn *et al.*, 1994; Honda *et al.*, 1996).

RNA secondary structures have critical functions not only in RNA viruses, but also in cells. Examples abound and include mRNAs, rRNAs, tRNAs, RNAse P, and tmRNA (10Sa RNA of *Escherichia coli*) (Felden *et al.*, 1997; Aziz and Munro, 1987; Graves *et al.*, 1987; Birchmeier *et al.*, 1983; Easterwood and Harvey, 1997). However, they represent only one of many forms of suborganismal organization. A number of recent results suggest that the evolution of robustness may be a general phenomenon also on higher levels of biological organization, such as that of genetic networks. (Nowak *et al.*, 1997; Stearns and Kawecki, 1994; Wagner *et al.*, 1997; Wagner, 1999). This raises a question about the general preconditions under which mutational robustness might evolve in biological systems. There are at least three such preconditions.

First, the system under consideration must be under strong stabilizing selection. Otherwise, robustness would incur no selective advantage. Second, there must be alternative ways to achieve the same biological function with different degrees of robustness. Possible examples run the gamut of biological organization. They include different amino acid sequences folding into polypeptides of the same function which vary only in their sensitivity to amino acid replacements, as well as alternative interaction patterns among the gene products in a developmental pathway, where each pathway architecture leads to the same patterning process, but has differential susceptibility to mutations. Beyond the observation that polygenic systems with non-additive gene interactions allow such alternative organizations (Wagner *et al.*, 1997; Wagner, 1996), little is known about this prerequisite for the evolution of robustness.

Third, the amount of variation in robustness generated by mutations has to be sufficiently large for selection to be effective. The reasons lie in the fact that the evolution of robustness is a second order phenomenon: it is not the robust system itself, but its offspring that is better adapted, on average. This requires high mutation rates μ or large (effective) population sizes N_e; for the evolution of robustness via redundant gene functions, for instance, $N_e\mu \gg 1$ may be necessary (Wagner, 1999). We note that RNA viruses may be an ideal study system in this regard, because of their enormously high mutation rates of up to 10^{-3} per position and replication (Domingo *et al.*, 1995; Drake, 1993).

Evolution of robustness, if widely occurring, has profound implications for our understanding of molecular evolution. From a pragmatic prespective, it provides a good candidate mechanism for an evolving fraction of neutral vs. non-neutral mutations. Everything else being equal, this implies an evolving neutral mutation *rate*, and might necessitate an even more careful choice of molecules for molecular systematics studies.

Perhaps more profound is the emphasis it puts on the role of the genetic context in which mutations occur. Similar to the effect of a point mutation in an RNA secondary structure element, which depends on the local (robust or non-robust) structural context, the effect of a mutation in a developmental gene will depend on other genes in the developmental pathway (the genetic "background"). While there are clear examples of mutations that will be deleterious in almost any context, the fitness effects of most mutations may be context-dependent. If context is indeed of the essence, then we will have to rethink time-honored and gene-centered notions in population genetics, such as that of a "genetic load"(Crow and Kimura, 1970), or the very notion of a "deleterious mutation".

Acknowledgments. We would like to thank Walter Fontana and Martijn Huynen for invaluable discussions. Financial support of the Santa Fe Institute, the Austrian *Fond zur Förderung der Wissenschaftlichen Forschung*, P12591-INF, and computational support by the Albuquerque High Performance Computing Center (AHPCC) is gratefully acknowledged.

References

Aziz, N. and Munro, H. N. Iron regulates ferritin mRNA translation through a segment of its 5' untranslated region. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8478–8482 (1987).

Baudin, F., Marquet, R., Isel, C., Darlix, J. L., Ehresmann, B. and Ehresmann, C. Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *J. Mol. Biol.* **229**, 382–397 (1993).

Birchmeier, C., Folk, W. and Birnstiel, M. L. The terminal RNA stem-loop structure and 80 bp of spacer DNA are required for the formation of 3' termini of sea urchin H2A mRNA. *Cell* **35**, 433–440 (1983).

Bonhoeffer, S., McCaskill, J. S., Stadler, P. F. and Schuster, P. RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur. Biophys. J.* **22**, 13–24 (1993).

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. and Sauer, R. T. Deciphering the message in protein sequences: tolerance to amino acid substitution. *Nature* **247**, 1306–1310 (1990).

Brinton, M. A., Fernandez, A. V. and Dispoto, J. H. The 3'-nucleotides of flavivirus genomic RNA form a conserved secondary structure. *Virology* **153**, 113–121 (1986).

Cadigan, K. M., Grossniklaus, U. and Gehring, W. J. Functional redundancy: the respective roles of the two sloppy paired genes in drosophila segmentation. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6324–6328 (1994).

Crow, J. F. and Kimura, M. *An introduction to population genetics theory*. New York: Harper & Row (1970).

Dayton, E. T., Konings, D. A., Powell, D. M., Shapiro, B. A., Butini, L., Maizel, J. V. and Dayton, A. I. Extensive sequence-specific information throughout the CAR/RRE, the target sequence of the human immunodeficiency virus type 1 Rev protein. *J. Virol.* **66**, 1139–1151 (1992).

Domingo, E., Holland, J. J., Biebricher, C. K. and Eigen, M. Quasispecies: The concept and the word. In: *Molecular Basis of Virus Evolution* (Gibbs, A., Calisher, C. H. and García-Arenal, G., eds.), pp. 171–180. Cambridge UK: Cambridge University Press (1995).

Drake, J. W. Rates of spontaneous mutations among RNA viruses. *Proc. Natl. Acad. Sci., USA* **90**, 4171–4175 (1993).

Dun, R. B. and Fraser, A. S. Selection for an invariant character – "vibrissa number" – in the house mouse. *Nature* pp. 1018–1019 (1958).

Easterwood, T. R. and Harvey, S. C. Ribonuclease P RNA: models of the 15/16 bulge from Escherichia coli and the P15 stem loop of Bacillus subtilis. *RNA* **3**, 577–585 (1997).

Felden, B., Himeno, H., Muto, A., McCutcheon, J. P., Atkins, J. F. and Gesteland, R. F. Probing the structure of *escherichian coli* 10Sa RNA (tmRNA). *RNA* **3**, 89–103 (1997).

Fontana, W., Konings, D. A. M., Stadler, P. F. and Schuster, P. Statistics of RNA secondary structures. *Biopolymers* **33**, 1389–1404 (1993a).

Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D. and Schuster, P. RNA folding and combinatory landscapes. *Phys. Rev. E* **47**, 2083–2099 (1993b).

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. and Turner, D. H. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA* **83**, 9373–9377 (1986).

Graves, R. A., Pandey, N. B., Chodchoy, N. and Marzluff, W. F. Translation is required for regulation of histone mRNA degradation. *Cell* **48**, 615–626 (1987).

Hartl, D. L., Dykhuizen, D. E. and Dean, A. M. Limits of adaptation: the evolution of selective neutrality. *Genetics* **11**, 655–674 (1985).

He, L., Kierzek, R., SantaLucia, J., Walter, A. E. and Turner, D. H. Nearest-neighbor parameters for GU mismatches. *Biochemistry* **30** (1991).

Hofacker, I. L., Fekete, M., Flamm, C., Huynen, M. A., Rauscher, S., Stolorz, P. E. and Stadler, P. F. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.* **26**, 3825–3836 (1998).

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M. and Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188 (1994).

Hofacker, I. L., Huynen, M. A., Stadler, P. F. and Stolorz, P. E. Knowledge discovery in RNA sequence families of HIV using scalable computers. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, pp. 20–25. Portland, OR: AAAI Press (1996a).

Hofacker, I. L., Huynen, M. A., Stadler, P. F. and Stolorz, P. E. *RNA Folding and Parallel Computers: The Minimum Free Energy Structures of Complete HIV Genomes*. Tech. rep., SFI, Santa Fe, New Mexico (1996b). #95-10-089.

Hofacker, I. L. and Stadler, P. F. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.* (1999). In press, Santa Fe Institute Preprint 98-06-58.

Hogeweg, P. and Hesper, B. Energy directed folding of RNA sequences. *Nucl. Acids Res.* **12**, 67–74 (1984).

Honda, M., Brown, E. A. and Lemon, S. M. Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA. *RNA* **2**, 955–968 (1996).

Huynen, M. A., Gutell, R. and Konings, D. A. M. Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.* **265**, 1104–1112 (1997).

Huynen, M. A., Konings, D. A. M. and Hogeweg, P. Multiple coding and the evolutionary properties of RNA secondary structure. *J. Theor. Biol.* **16**, 251–267 (1993).

Jackson, R. J. and Kaminski, A. Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. *RNA* **1**, 985–1000 (1995).

Jaeger, J. A., Turner, D. H. and Zuker, M. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci., USA* **86**, 7706–7710 (1989).

Joyner, A. L., Herrup, K., Auerbach, B. A., Davis, C. A. and Rossant, J. Subtle cerebellar phenotype in mice homozygous for a targeted deletion of the en-2 homeobox. *Science* **25**, 1239–1243 (1991).

Kacser, H. and Burns, J. A. The molecular basis of dominance. *Genetics* **97**, 639–666 (1981).

Mandl, C. W., Holzmann, H., Meixner, T., Rauscher, S., Stadler, P. F., Allison, S. L. and Heinz, F. X. Spontaneous and engineered deletions in the 3'-noncoding region of tick-borne encephalitis virus: Construction of highly attenuated mutants of flavivirus. *J. Virology* **72**, 2132–2140 (1998).

Nowak, M. A., Boerlijst, M. C., Cooke, J. and Maynard-Smith, J. Evolution of genetic redundancy. *Nature* **38**, 167–171 (1997).

Olsthoorn, R. C., Licis, N. and vanDuin, J. Leeway and constraints in the forced evolutions of a regulatory RNA helix. *EMBO J.* **13**, 1994 (1994).

Rendel, J. M. Canalization and selection. In: *Quantitative genetic variation* (Thompson Jr., J. N. and Thoday, J. M., eds.), pp. 139–156. New York: Academic Press (1979).

Schuster, P., Fontana, W., Stadler, P. F. and Hofacker, I. L. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Royal Soc. (London) B* **255**, 279–284 (1994).

Schuster, P. and Stadler, P. F. Sequence redundancy in biopolymers: A study on RNA and protein structures. In: *Viral Regulatory Structures* (Myers, G., ed.), vol. XXVIII of *Santa Fe Institute Studies in the Sciences of Complexity*, pp. 163–186. Reading MA: Addison-Wesley (1998). Santa Fe Institute Preprint 97-07-67.

Schuster, P., Stadler, P. F. and Renner, A. RNA Structure and folding. From conventional to new issues in structure predictions. *Curr. Opinion Struct. Biol.* **7** (1997). 229–235.

Sokal, R. R. and Rohlf, F. J. *Biometry*. New York: Freeman (1981).

Stearns, S. C. and Kawecki, T. J. Fitness sensitivity and the canalization of life history traits. *Evolution* **48**, 1438–1450 (1994).

Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L. and Schuster, P. Algorithm independent properties of RNA secondary structure prediction. *Eur. Biophys. J.* **25**, 115–130 (1996).

Tautz, D. Redundancies, development and the flow of information. *BioEssays* **14**, 263–266 (1992).

Thompson, J. D., Higgs, D. G. and Gibson, T. J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680 (1994).

Waddington, C. H. *The strategy of the genes*. New York: Macmillan (1957).

Wagner, A. Does evolutionary plasticity evolve? *Evolution* **50**, 1008–1023 (1996).

Wagner, A. Redundant gene functions and natural selection. *J. Evol. Biol.* **12**, 1–16 (1999).

Wagner, G. P., Booth, G. and Bagheri-Chaichian, H. Towards a population genetic theory of canalization. *Evolution* **51**, 329–347 (1997).

Wallner, G., Mandl, C. W., Kunz, C. and Heinz, F. X. The flavivirus 3'-noncoding region: Extensive size heterogeneity independent of evolutionary relationships among strains of tick-borne encephalitis virus. *Virology* **213**, 169–178 (1995).

Weintraub, H. The MyoD family and myogenesis: redundancy, networks, and thresholds. *Cell* **75**, 1241–1244 (1993).

Zuker, M. and Sankoff, D. RNA secondary structures and their prediction. *Bull.Math.Biol.* **46**, 591–621 (1984).