# Concurrent neutral evolution of mRNA secondary structures and encoded proteins

**Jaromir Konecny*, Michael Schöniger*, Ivo L. Hofacker†, Marc-Denis Weitze*, G. Ludwig Hofacker***

*Theoretical Chemistry, Technical University Munich, Lichtenbergstr. 4, D 85747 Garching, Germany

†Theoretical Chemistry, University Vienna, Währingerstraße 17, A-1090 Wien, Austria

Correspondence to: G. Ludwig Hofacker, hofacker@theochem.tu-muenchen.de

## Abstract

Messenger RNA sequences often have to preserve functional secondary structure elements in addition to coding for proteins. We present a statistical analysis of retroviral mRNA which supports the hypothesis that the natural genetic code is adapted to such complementary coding. These sequences are still able to efficiently explore the space of possible proteins by point mutations. This is borne out by the observation that in stem regions of retroviral mRNA foldings silent mutations on one strand are preferentially accompanied by conservative mutations on the other. Distances between amino acids based on physicochemical properties are used to quantify the conservation of protein function under the constraint of maintained RNA secondary structure. We find that preservation of RNA secondary structure by compensatory mutations is evolutionary compatible with the efficient search for new variants on the protein level.

## Introduction

The genetic code is known to create a high acceptance rate of amino acid replacements in proteins by random point mutations. Messenger RNAs of existing species, as far as they are functionally selected, then have to compromise in their evolution between their own and the encoded proteins function (King and Jukes 1969; White et al. 1972; Min Jou et al. 1972; Ball 1973; Fitch 1974; Hasegawa et al. 1979; Zama 1990; Huynen et al. 1992; Konings 1992). Twenty seven years ago White III et al. (1972) argued that the redundancies in the triplet code allowed for mRNA sequences satisfying structural requirements for both protein and RNA. They concluded that King and Jukes (1969) claim for selective neutrality of mutations in redundant bases (synonymous mutations) had to be weakened. Ball (1973) enriched the discussion by three alternative hypotheses concerning the relationship between secondary structures of mRNAs and the amino acid sequences encoded by them. His reasoning, in the interpretation by Fitch (1974), was as follows:

1. Secondary structure is a natural consequence of the physical chemistry and there is no optimization.
2. Optimization of secondary structure occurs only within the degeneracy of the genetic code.
3. Optimization of RNA secondary structure imposes limitations in the nucleotide sequence and therefore in the amino acid sequence as well.

While Fitch felt that hypothesis 3 was biologically plausible he discarded it (in contrast to Ball) on account of his interpretation of MS2 viral coat protein data. Here we will here present evidence for hypothesis 3 by using an extended computational analysis, and relate this to other (prebiotic) aspects of the genetic codes determination.

The structure of the universal genetic code is such that it encodes 20 amino acids by 61 codons (43 of which 3 are stop codons). Note that the properties of an amino acid are essentially determined by the second position of its codon, where U encodes large hydrophobic amino acids, C uncharged, slightly hydrophilic ones, A hydrophilic, mainly charged ones and G determines mainly uncharged ones. Third positions are often degenerate; i.e. do not determine the amino acid and a mutation of a such position is called silent. The mutation of a first codon position causes mainly a change to a physico-chemically related amino acid (a conservative mutation).

The genetic code exhibits, beyond the well-known neutral replacement patterns of amino acid substitutions, optimal properties by favoring the simultaneous quasi-neutral evolution of proteins encoded in two complementary strands of DNA/RNA (Konecny et al. 1993; Konecny et al. 1995). The principle of complementary coding of two proteins on one double-strand must also hold if one protein is coded by a folded, single stranded RNA. Secondary, and therefore tertiary, structure depends crucially on the stems of paired nucleotides. Thus, if the secondary structure of this RNA is under selective pressure, most of the mutations in the stems have to be rejected or compensated by corresponding replacements of paired nucleotides. Such double (or compensatory) mutations will most likely preserve the secondary structure of the functional RNA while single, uncompensated, point mutations can give rise to large conformational changes. Compensatory mutations are indeed frequently observed in phylogenetic data (Dixon and Hillis 1993).

It is therefore worth investigating whether a mutation not affecting the function of the encoded protein would also preserve the mRNA secondary structure with its biological function. It should be born in mind that this likely effect of complementary evolutionary compatibility differs greatly from the assumption that the redundancy of the genetic code is used for optimal folding of mRNAs, especially for thermodynamic stability, as suggested by Fitch. The latter regarded pairing of second and third codon positions (2-3 pairing) as optimal, i.e. pairing of strong amino acid determining positions with most degenerate ones. However, from the viewpoint of evolutionary compatibility pairing of fixed codon positions with flexible ones (2-3 pairing) is highly disadvantageous. Given the inflexibility of the second codon position silent mutations in the third codon position of the complementary strand can not be compensated. In contrast to that, 1-3 pairing should be most favorable in evolutionary terms because silent (acceptable) mutations would be accompanied by conservative (often acceptable) ones. If it is true that RNA secondary structures are subjected to selection, then base pairing of such kind is advantageous allowing for the least restricted evolution of complementary sites with respect to protein evolution.

## The rev responsive element

To investigate further the interdependence of primary and secondary structures of mRNA one needs to find examples of well conserved foldings with biological function. Such multiple coding systems were observed, for example, in some lentiviruses, including the rev responsive element (RRE; an RNA secondary structure involved in regulating the transport of unspliced mRNA to the cytoplasm) which is located within the coding region of the env gene (Konings 1992). The essential structure encompasses about 200 nucleotides and its structure is predicted to be a multibranched stem-loop (Fig. 1).

**Secondary structure calculation.** RNA secondary structures were computed using the RNAfold program from the Vienna RNA Package (Hofacker et al. 1994). The program implements the calculation of minimum free energy structures and base pairing probabilities using the algorithms of Zuker and Stiegler (1981) and McCaskill (1990). The energy parameters used in the current version 1.2 are identical to those in Zuker's mfold 2.3 and are summarized in Walter et al. (1994).

**Computational analysis and simulation.** To assess the likely changes in functionality of a protein due to point mutations in its mRNA we define the *antisense distances* (ASD), derived from neutral-evolution replacement frequencies (Dayhoff 1978). Average ASDs for mRNAs were determined by the algorithm sketched in Fig. 2: Each possible silent mutation in stems of the RRE structure was compensated on the opposite strand under the assumption that base pairing is preserved. Usually, this compensatory mutation causes an amino

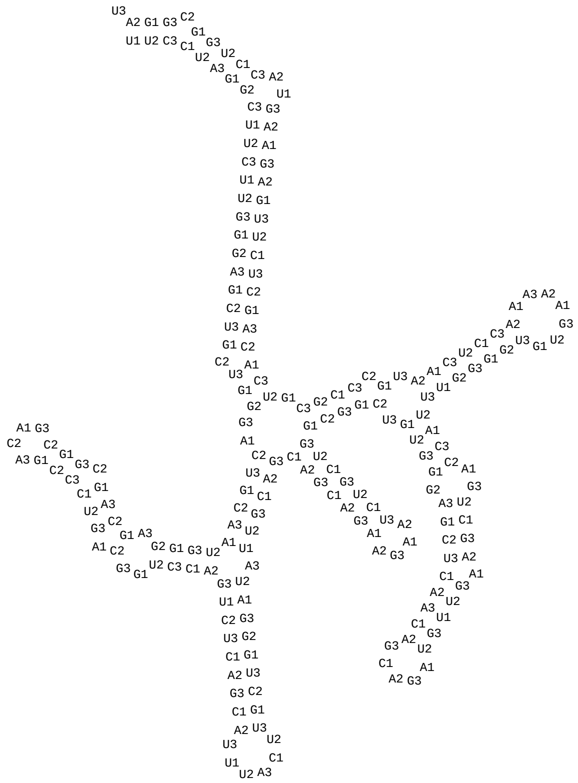

Figure 1. Secondary structure of the simian SIVAGM155 RRE sequence computed with RNAfold. Numbers indicate first, second and third codon positions.

acid replacement in the corresponding protein. The evolutionary distance between these two residues is quantified using a distance matrix based on the physico-chemical properties of the amino acids (Borstnik et al. 1987). If a silent mutation is compensated by a silent one on the opposite strand we define the distance as zero. G-U pairs were not included in our analysis since they are much rarer than A-U and G-C pairs and would by force introduce an artificial weighting as they contribute much less to the stability of the folded RNA structure. After averaging over all mutation distances caused by all possible silent mutations on the opposite strand of the given mRNA structure we have a measure for the adaptation of the native sequence to evolutionary compatible coding.

We used the data set of five human and simian RRE sequences that were analyzed by Konings (1992). Those sequences were shown by Konings (1992) to be rather distinct and therefore represent a sample as independent as possible out of the 60 RRE sequences that were available then. We restrict ourselves to this limited number of sequences to keep our results comparable to the study by Konings (1992). In order to judge how well the five native RRE sequences are adapted to evolutionary compatible coding we conducted Monte

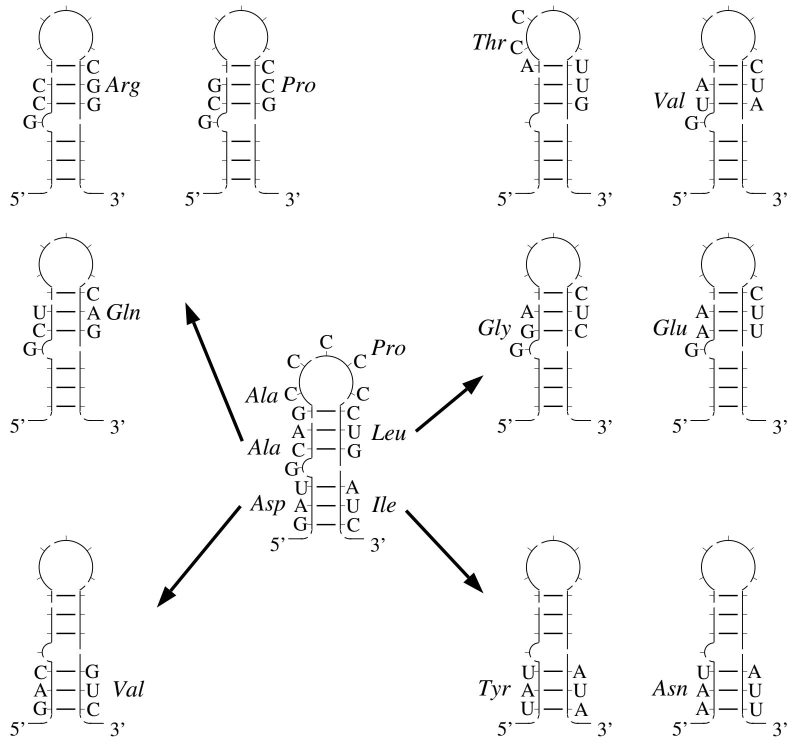


FIGURE 2. Algorithm to calculate the average antisense distance (ASD). Each folded mRNA sequence (center) was subjected to all possible silent mutations in order to determine ASD. The first codon GAU (Asp) allows for one silent mutation to GAC which corresponds to the conservative amino acid interchange from Ile (AUC) to Val (GUC) on the opposite strand (lower left). For the second codon GCA (Ala) three silent mutations are possible (upper left) yielding amino acid replacements of Leu (CUG) by Arg (CGG), Pro (CCG) and Gln (CAG) on the opposite strand. The Leu codon CUG can silently mutate to four others (upper right), whereas the Ile codon AUC allows for two silent mutations (lower right).

Carlo simulations. Artificial sequences folding into the RRE secondary structure were constructed by "inverse folding" using a version of the RNAinverse program from the Vienna Package, modified to yield only sequences without stop codons. The inverse folding procedure constructs sequences folding into a predefined target structure by an optimization procedure starting from a random sequence. In that way we generated essentially unrelated sequences with the same secondary structure. The efficiency of this process depends strongly on the target structure. In particular, the existence of isolated base pairs, i.e. helices of length one, can make the search very slow. To obtain a sample of 200 sequences for each native structure without excessive computational effort, we have used not only sequences that fold exactly into the target RRE structure,

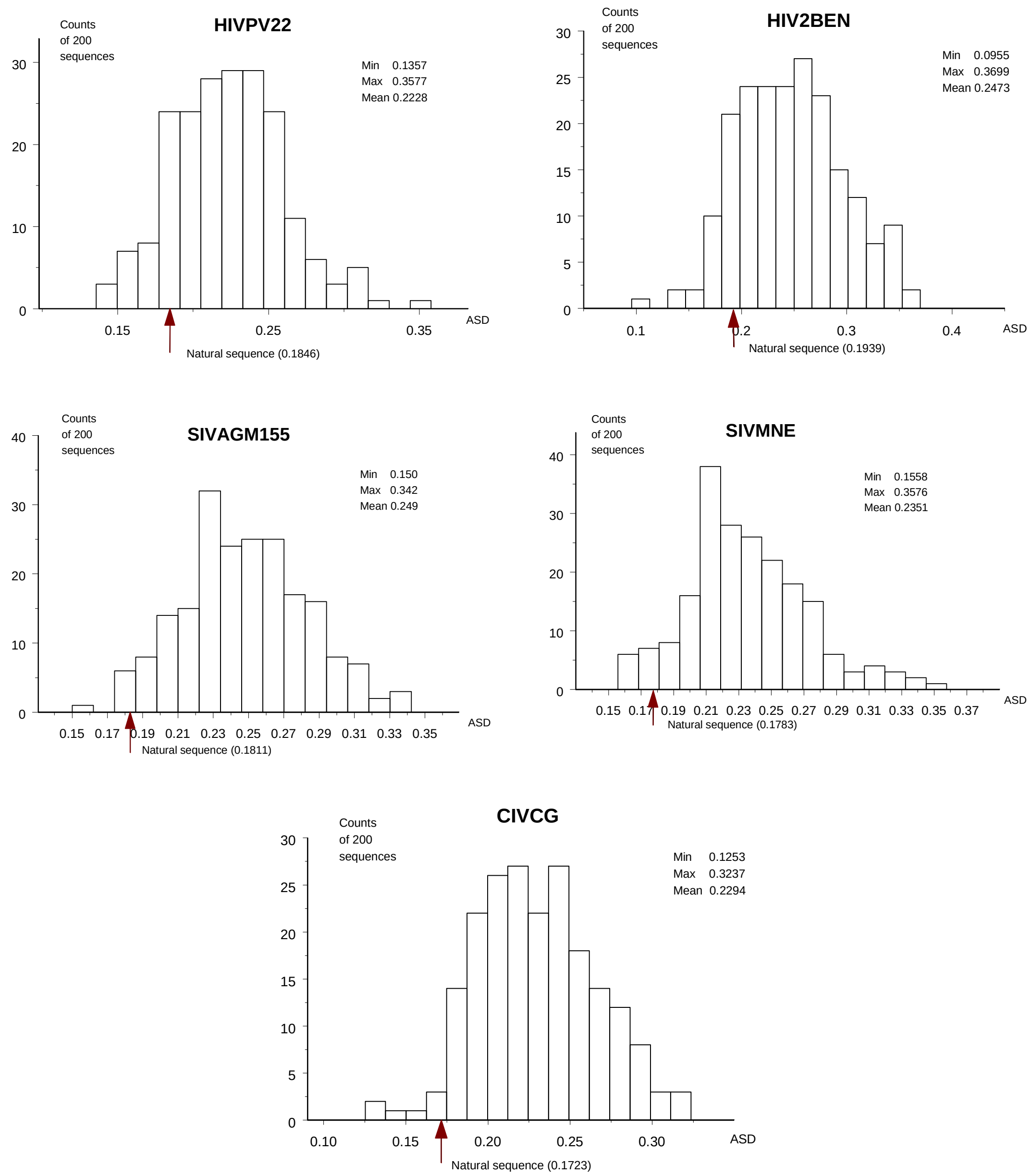


FIGURE 3. Average antisense distance distribution of 200 randomized sequences of same or very similar secondary structure. Arrows indicate the value obtained for the natural sequence. The data set comprises one representative sequence from each of four major phylogenetic groups: human immunodeficiency virus groups 1 (A, HIVPV22) and 2 (B, HIV2BEN), African green monkey virus (C, SIVAGM155) and macaque virus (D, SIVMNE), as well as one single species, the chimpanzee immunodeficiency virus (E, CIVCG).

but included sequences whose minimum free energy structure differs by a single base pair from the target structure.

For each randomized sequence the ASD was calculated by the algorithm described above. The obtained distributions of Fig. 3 show that the native RRE sequences are well adapted to evolutionary compatible coding with statistical significance in 3 of 5 cases.

## Discussion and conclusion

In our view, Fig. 3 indicates that evolutionary compatible coding plays a significant role in mRNAs with biologically relevant, and therefore selected, secondary structures. Vice versa, the mutational tolerance of those structures reveals also the adaptation of the universal genetic code to such multiple and complementary coding. In addition, it underlines the importance of the complementary relationships of the genetic code.

RRE is known to be a unique example for a relatively large unit of mRNA that has a strongly conserved secondary structure with important biological function. Our calculations based on the assumption of complementary coding could not be performed with less obvious examples than RRE. To the best of our knowledge there is no other such striking instance available in the mRNA databases. Nevertheless, with the upcoming wealth of new sequence data we expect to find independent mRNA sequences supporting our hypothesis.

Our results enable us to gain better insights into the simultaneous evolution of folded mRNA molecules and the proteins encoded by them. They shed new light on the emergence of the genetic code in a prebiotic RNA world (Gilbert 1986). The first isolated RNAs had to possess very compact and preserved structures to maintain their function. It is a distinct possibility that the universal genetic code was optimized in this RNA world in accordance with the hypothesis of compatible coding (Konecny et al. 1993; Konecny et al. 1995).

Possible applications of this concept include the optimization of currently used RNA folding programs and, even more promising, the search for biologically important mRNA structure elements in large sequence databases. We conjecture that those elements should adapt particularly well to evolutionary compatible coding indicated by small average antisense distances.

## References

Ball LA (1973) Secondary structure and coding potential of the coat protein gene of bacteriophage MS2. Nature New Biol 242:44-45

Borstnik B, Pumpernik D, Hofacker GL (1987) Point mutations as an optimal search process in biological evolution. J Theor Biol 125:249-268

Dayhoff MO (1978) Atlas of protein sequence and structure. Vol. 5. National Biomedical Research Foundation. Washington DC

Dixon MT Hillis DM (1993) Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. Mol Biol Evol 10:256-267

Fitch WM (1974) The large extent of putative secondary nucleic acid structure in random nucleotide sequences or amino acid derived messenger-RNA. J Mol Evol 3:279-291

Gilbert W (1986) The RNA world. Nature 319:618

Hasegawa M, Yasunaga T, Miyata T (1979) Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res 7:2073-2079

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem 125:167-188. http://www.tbi.univie.ac.at/~ivo/RNA/

Huynen MA, Konings DAM, Hogeweg P (1992) Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. J Mol Evol 34:280-291

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788

Konecny J, Eckert M, Schniger M, Hofacker GL (1993) Neutral adaptation of the genetic code to double-strand coding. J Mol Evol 36:407-416

Konecny J, Schniger M, Hofacker GL (1995) Complementary coding conforms to the primeval comma-less code. J Theor Biol 173:263-270

Konings DAM (1992) On the coexistence of multiple codes in messenger RNA molecules. Comp Chem 16:153-163

McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29:1105-1119

Min Jou W, Haegeman G, Ysebaert M, Fiers W (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 237:82-88

Walter AE, Turner DH, Kim J, Lyttle MH, Muller P, Mathews DH, Zuker M (1994) Co-axial stacking of helices enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc Nat Acad Sci USA. 91:9218-9222

White HB III, Laux BE, Dennis D (1972) Messenger RNA structure: Compatibility of hairpin loops with protein sequence. Science 175:1264-1266

Zama M (1990) Codon usage pattern in a2(I) chain domain of chicken type I collagen and its implications for the secondary structure of the mRNA and the synthesis pauses of the collagen. Biochem Biophys Res Commun 167:772-776

Zuker M, Stiegler P (1981) Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucl Acids Res 9:133-148