RNA Sec. Structure dependence on alphabet size, sequence length and number of hydrogen bonds

Ulf W. Liebal *

March 3, 2007

Structure Neighbourhood Properties around the mfe-Structure

The RNA world theory states that in one stage of the evolution RNA was not only information bearing instrument but also shaped the phenotype directly by active participation as catalyst for reactions. However, catalytic efficiency could be improved by catalysts composed of amino acid monomers. Following this development RNA enzymes were displaced largely by proteins from the phenotype world, surviving in reaction niches at the boundary of RNA-protein world (Ribosomes, tRNA) and in DNA/RNA processing reactions (spliceosome) though.[Smith and Szathmáry, 1995]
The RNA at present, that is capable of a plethora of reactions, consists of four alphabet members A, U, C, G, of which two pairs are possible with an additional wobble GU pair. The question of why there are four bases and not an arbitrary other number was asked and answered several times from differing viewpoints in the past. These studies do not attempt to fully explain or support the hypothetical RNA world theory but instead characterise simplified RNA world subclasses in oder to uncover limits and restrictions the RNA world should be governed by.
Szathmáry [Szathmary, 1992] studied copying fidelity and metabolic efficiency for several alphabet sizes and concludes that the optimal alphabet without mismatch repair is four, but six in the presence of mismatch repair. With a rising number of alphabets the copying fidelity drops sharply since adding more members to the alphabet with keeping the distinction parameters constant (here the base pairing via hydrogen bonds) the more the members will resemble each other with respect to their pairing characteristics. The metabolic efficiency increases with the alphabet size since different monomers add diversity to the strings enabling more efficient catalysis.
Gardner and coworkers [Gardner et al., 2003] studied the evolutionary flexibility of different sized alphabets asking how efficient randomised alphabets could evolve towards a naturally occurring RNA structure. They found that for low copy fidelity regimes the most efficient emulator was the four sized alphabet followed, and outperformed in high fidelity regimes, by the six size alphabets.
Mac Dónaill [Mac Dónaill, 2002] on the other hand investigates for the reason of having two and three hydrogen bonds to distinguish between the alphabet members. He approaches the problem from an information theoretic side and reveals that the existing kind of code is most efficient in error detecting and corresponds to a parity code in digital data transmission.
The goal of this study is to characterise a subset of available structures with different sequence length, $N = \{40(+20)\dots 200\}$, for alphabet sizes $n = \{2, 4, 6\}$, different hydrogen bond characteristics $hb = \{2, 3, 2+3\}$ as well as the GU wobble pairing for four sized alphabet. The alphabet with the characteristics $n = 4$, $hb = 2 + 3$ and GU wobble is called the canonical alphabet. The scrutinised structure subset does not cover all possible random sequences but instead an algorithm was designed such as to evolve from random initialised structures randomly towards a higher secondary structure content. The approach to study general structural properties of alphabets is feasible since the majority of conformational order within evolved structures results not from evolutionary optimisation but from constraints imposed by rules intrinsic to RNA polymer folding [Schultes, 1999].
The sequences are folded with the help of the ViennaRNA package with which it is possible to simulate folding of the canonical alphabet as well as artificial alphabet with arbitrary member size and two and/or three hydrogen bonds between the letters [Hofacker et al., 1994–1998]. Additionally the temperature can be adjusted, GU-wobble pairing can be excluded and the structures around an energy range of the mfe are accessible. The artificial alphabets with six members can have two combinations of hydrogen bonding, here only the constellation of two GC- and one AU-type bonding is considered.

*e-Mail: ulfliebal@web.de

How can a structure be described?
A condition in which nonadjacent monomers of a string interact with each other is called structure. The nonadjacent monomer interaction is governed by the complementary rules of the letters. The occupancy of a structure, i.e. the probability that the structure folds into the structure, is determined by the free energy that is released due to base pairing. The structure that releases the most energy, the minimum free energy (mfe) is the most populated, and characteristic for each sequence. However, not only information about the most populated structure, the mfe structure is important, the relative dominance of the mfe structure against neighbouring low energy structures is interesting as well . Further characterisation may be gathered from the values of the probabilities for pairing between two letters [Schultes, 1999]. The previous measurements describe thermodynamic properties of the structure but there are also measurements needed that describe better the actual shape or topology of the structure. This is performed by the Gardner uniqueness of folding function (F), which is the Frobenius norm of the base pairing probability matrix [Gardner et al., 2003]. The higher the secondary structure content of a structure the higher will F be.

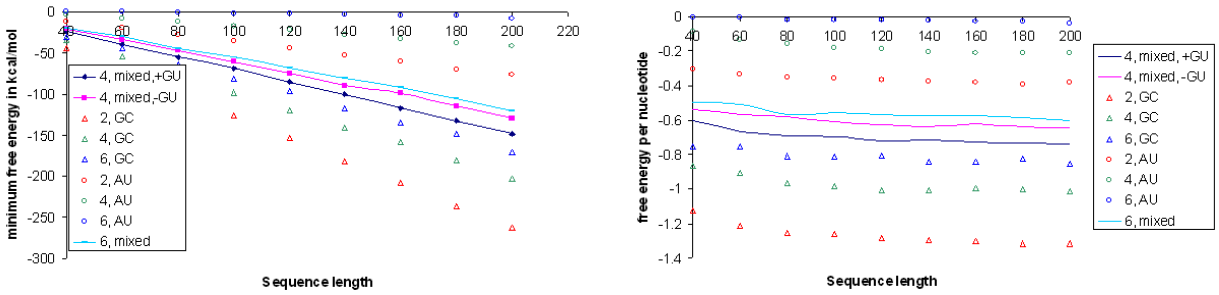
The Variables that are recorded for this study are the minimum free energy (mfe), the base pair probability between any two sequence position through which the Shannon Entropy (Q) and the Gardner uniqueness function of folding (F) is calculated, as well as the number of stable structures (meaning populate-able) and their mean bp distance within an energy range around the mfe and the base pair (bp) distance of the most distinct structure to the mfe structure in the set of neighbouring structures.
Structures were generated by ten cycles of an iterative process. Given a folding structure, a new sequence is generated where bases were chosen at random except that they are constrained to be complimentary where the structure indicates a pairing. This sequence is folded by RNAfold to generate a new structure, which is input to the next iteration. The structure for the first iteration is unfolded, so that the first sequence is unconstrained. This process generates structures which are random (in the sense of being independent of human choice) and highly paired.
The examined mfe indicates the maximum gain of energy for the structure with the highest release of energy. If the mfe is low then many stabilising structures have formed (correspondingly: many hydrogen bonds).
Shannon entropy and Gardner uniqueness of folding function are based on the base pair probability matrix for each sequence. The base pair probability matrix in turn is calculated via thermodynamical likelihoods of base pair formation and is a summary of the many possible alternative formations that compete during the energy minimisation process of polymer folding [McCaskill, 1990]. The Shannon entropy for a structure in this form was developed by Schultes and is calculated by [Schultes, 1999]:

$$Q := -\frac{1}{Q_{max}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij} \log_2 p_{ij} \tag{1}$$

with $Q_{max} = \frac{1}{2} N \log_2 N$, p_{ij} the base pair probability between bases i and j and N the sequence length. The Shannon entropy measures the global uniqueness of folding, that is the uncertainty of the situation, that the sequence is currently occupying the mfe structure. A sequence that acts as catalyst should have a strong tendency to reduce the Shannon entropy as much as possible since a sequence wants to be sure to fold into a specific structure which is able to perform reactions.
Unlike the Shannon entropy, the Gardner uniqueness of folding function will distinguish between a 'well-folded' stable secondary structure and a completely unfolded molecule. It takes the form [Gardner et al., 2003]:

$$F := \sqrt{\frac{1}{N} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}^2 \right)} \tag{2}$$

Sequences that are supposed to catalyse reactions are likely to obey a trend towards a high degree of secondary structures and therefore a high value for F. This is because secondary structures can be more efficiently regulated by evolutionary selection compared to loose and random unstructured stretches. The amount of stable sequences around an energy range of the mfe is included to give insight into the structural plasticity of the alphabets. The structural plasticity describes the capacity of a sequence to assume a variety of energetically favourable shapes by equilibrium among them at constant temperature [Ancel and Fontana, 2000]. Selection leads to a reduction in plasticity to raise the time a sequence spends in a defined state, correspondingly it lowers the Shannon entropy, but the reduction of plasticity causes also a loss of variability. But not only the number of different structures around the mfe is important also the diversity of the structures has to be considered. On that end the distance of the most distant shape



(a) (b)

Figure 1: The dependence of the released energy for folding is plotted versus the sequence length (a) and the free energy per nucleotide for different alphabet sizes with different hydrogen bond properties (triangles GC-type, circles AU-type, red: 2n (alph), green: 4n and blue: 6n). The sequences are randomly generated with a fixed number of alphabet members and subsequently the proportion of secondary structures is increased (see text for explanation). Simulated are 100 sequences at 30° C.

to the mfe structure is recorded. Although only the most distant shape is taken for each sequence, this may nonetheless give insight into the variability of structures around the mfe. Most structures around the mfe structure will be highly similar to the mfe structure differing only by opening of some pairing in the structure, completely different shapes may only provide a fractional deviation from the mean, therefore detecting the extreme may give a better idea about the diversity than the mean alone.

Figure 1(a) shows the the increase in the minimum free energy for successively longer sequence length. Obviously for longer strings more hydrogen bonds can form, therefore rising the absolute gain in free energy. The magnitude of the energy gain depends largely on the number of hydrogen bonds that can form between alphabet members, with GC type letters and their three hydrogen bonds releasing more energy than AU-type alphabets with only two hydrogen bonds. The alphabet size effects the energy release in that the probability to find a complementary letter at any position decreases with increasing member choice. The canonical alphabet releases slightly more energy than the canonical alphabet without GU pairing which lies close to the mean of the four GC and AU mfe values. The six letter alphabet with mixed hydrogen bonds is again close to the mean of the two pure hydrogen bondings, which is lower than for the canonical alphabet.
In figure 1(b) the average nucleotide specific gain in free energy is plotted against the sequence length for sequences under investigation. The situation is similar to the graph with absolute mfe in that the main effector is the number of hydrogen bonds and the alphabet size. There is a small increase in the free energy gain per nucleotide for longer sequences. The alphabets with mixed hydrogen bond properties (2 and 3 bonds) are again between the pure bonding patterns.

Figure 2 indicates that the Shannon entropy of the mfe structures for the random generated sequences only marginally correlates with the sequence length. The AU-type alphabets have a higher Shannon entropy, i.e. a higher uncertainty of the actual structure of the sequence, compared to GC-type alphabets implying a higher preference for the mfe structure in GC alphabet. The theoretical reasons for that is either that fewer structures are formed in GC alphabets or that the energy difference between the mfe structure to other stable structures is higher for GC alphabets. Particularly following the latter argument, one could predict that a smaller neighbourhood lies within ΔE of the mfe.
The four sized alphabet with mixed hydrogen bonds obeys a Shannon entropy similar to the four AU-type alphabet. The six mixed alphabet has a lower Entropy compared to the four sized alphabet and is similar to the six GC-type alphabet (six mixed contains two GC-pairs). Surprisingly the wobble GU-pairing reduces the entropy of the four mixed alphabet and transforms it to a similar line like the GC mixed alphabet.

Figure 3 indicates the dependence of the Gardner uniqueness function of folding (F) on the sequence length and the free energy per nucleotide. The Gardner uniqueness function of folding is not influenced strongly by the sequence length as shown in Figure 3 (a). However, the Guff is shaped by the mfe free energy gain per nucleotide in that higher Guff values are reached with rising free energies for a specific alphabet. There seems to be a tendency of the order of the Guff values for the alphabet sizes $2 > 4 > 6$. The reason is because the probability that long complementary stretches in a sequence occur is higher

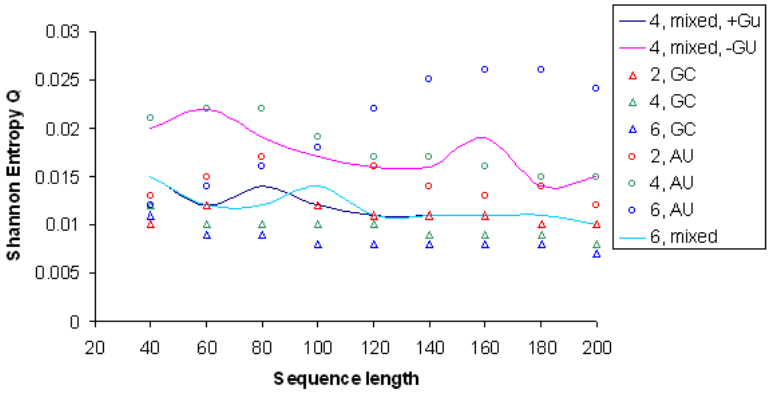
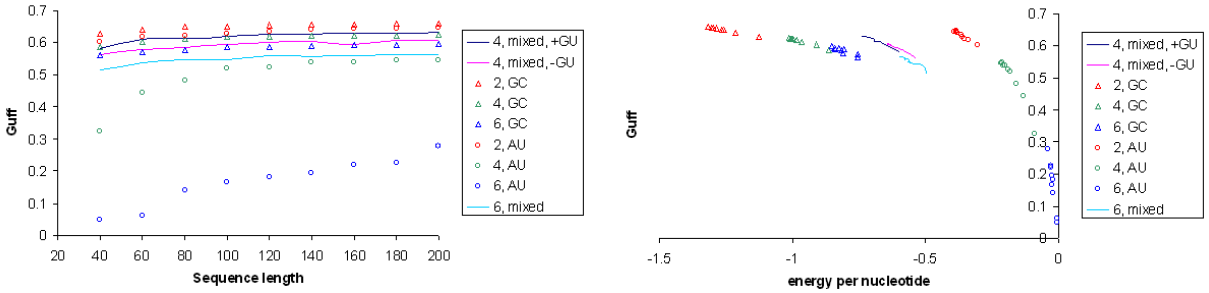


Figure 2: Shannon Entropy (Q) of the mfe Structures are plotted against the sequence length for sequences with different alphabet properties (triangles GC-type, circles AU-type, red: 2n, green: 4n and blue: 6n). The sequences are randomly generated with a fixed number of alphabet members and subsequently the proportion of secondary structures is increased (see text for explanation). Simulated are 100 sequences at 30° C.

for low sized alphabets, thereby increasing the amount of highly structured shapes, which is measured by F.

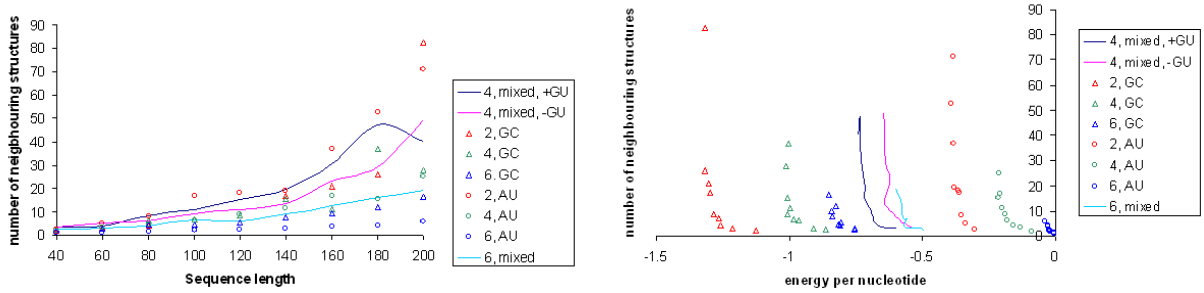


(a) (b)

Figure 3: Plotting of the Gardner uniqueness function of folding (Guff) versus the sequence length (a), free energy per nucleotide (b) (triangles GC-type, circles AU-type, red: 2n, green: 4n and blue: 6n). The sequences are randomly generated with a fixed number of alphabet members and subsequently the proportion of secondary structures is increased (see text for explanation). Simulated are 100 sequences at 30° C.

The number of structures for each sequence that lie within energy distance $\Delta E = 1\frac{kcal}{mol}$ of the mfe structure increases rapidly with the released energy per nucleotide for specific alphabet properties as visualised in figure 4. AU-type alphabets generate a slightly lower number of neighbouring structures compared to GC-type alphabets due to the low energy gain per nucleotide which weakens the tendency of buildup of hydrogen bonds over short complementary stretches. The major influence however, is given by the alphabet size, where the effects can be explained, as in the observations of the Gardner uniqueness of folding function, that the probability to find complementary regions in a string of a low member alphabet is higher than for a larger sized alphabet, and therefore there exist more possible states. In this graph clear differences between the canonical alphabet and the six sized mixed alphabet can be observed, in that the number of adjacent structures is much lower for the six sized mixed alphabet. The GC-type alphabets have alphabet dependent a higher number of neighbouring structures compared to AU, however the Shannon entropy of the GC-type structures is lower compared to AU-type structures. One might conclude in concordance with the arguments presented for figure 2 that the final number of structures for GC-type alphabets is smaller than for AU. This is because most structures for GC alphabets are very stable, whereas for AU-type alphabets a higher number of weaker structures exist.

The mean bp distance of the neighbouring structures to the mfe structure increases fairly linear with the sequence length (not shown) which means that with the increasing number of neighbourhood structures (indicated in figure 4) of longer sequences a higher amount of differing stable structures is accessible. Figure 5 visualises the mean maximum bp distance and obeys an unusual behaviour for the canonical and six mixed letter alphabet. The maximum bp distance is determined by the alphabet size such that the four and six letter alphabets have structures that that are at a constant distance, whereas the



(a) (b)

Figure 4: The number of neighbouring structures within distance $\Delta E = 1\frac{kcal}{mol}$ is compared to the sequence length and the free energy per nucleotide (triangles GC-type, circles AU-type, red: 2n, green: 4n and blue: 6n). The sequences are randomly generated with a fixed number of alphabet members and subsequently the proportion of secondary structures is increased (see text for explanation). Simulated are 100 sequences at 30° C.

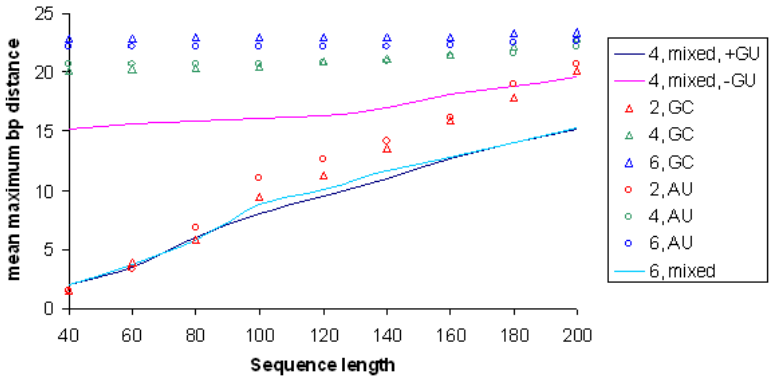


Figure 5: illustrates the mean maximum bp distance versus the sequence length for different alphabets, generated for structures within $\Delta E = 1\frac{kcal}{mol}$ of the mfe. See inset legend for details. The sequences are randomly generated with a fixed number of alphabet members and subsequently the proportion of secondary structures is increased (see text for explanation). Simulated are 100 sequences at 30° C.

two sized alphabets start with maximal distant structures that are quite similar and increasing the bp distance linearly with increasing length (and increasing amount of adjacent structures). The reason for the differential properties of the maximum bp distance for different alphabet sizes is that the probability to find nearly equally stable but dissimilar structures compared to the mfe structures is low for two letter alphabets. However, interesting in this case is the threshold behaviour such that the shape of the two letter alphabet differs so markedly from the four and six alphabet with even hydrogen bonds. Even more striking is the behaviour of the mixed alphabets, particularly canonical and six mixed, which seem to emulate the characteristics of the two sized alphabets. The canonical alphabet without GU wobble (4, mixed) has a lower maximum bp distance structure than four letter alphabets with either two or three hydrogen bonds. Expressed as extreme the hydrogen bond mixing seems to reduce the structural diversity at the bottom of the folding funnel. An additional surprising property is that the GU-wobble pairing further reduces the "structural diversity" around the mfe, to arrive at maximum distances very similar to six mixed alphabet.

Properties of AU- and GC-type alphabets Alphabets with AU-type bonding (two hydrogen bonds) have a much less stabilised mfe compared to GC-type alphabets (three hydrogen bonds)(figure 1(a)). As a result of that the folding funnel is less deep resulting in a slightly higher entropy for mfe structures of AU-type compared to GC type (figure 2).
The alphabet size choice determines the shallowness of the folding tunnel, that is the number of neighbouring sequences in a given energy interval and the structural diversity of this neighbouring structures. The less the heterogeneity of the alphabet the more shallow is the folding funnel and more structures are available for occupancy, which can be followed for the high structure number for two sized alphabets leading to a low neighbourhood for six member alphabets in figure 4. Similarly the structural diversity around the mfe structure is interestingly higher for four and six letter alphabets. This structural diversity around the mfe must not be mixed with the overall structural diversity which was not measured in this study. Concerning the propensity of complex structure formation, which could be thought of as approx-

imated by the Gardner uniqueness function of folding, low letter alphabets have a higher probability of complex structures compared to high letter alphabets. Which is the effect now for the mixing of different amount of hydrgenbonds? For the four letter alphabet this yields a medium mfe gain between GC and AU free energies, but not the mean of the Shannon entropy, where the mixed alphabet is unordered like the AU-type alphabet. Examination of the Guff reveals that the mixed four letter alphabet behaves rather like a four GC- type alphabet. The mixed alphabet results in more neighbouring structures than the even type alphabets.
The six alphabet is energetically stabilised comparable to the four mixed alphabet, however, it obeys a lower Shannon entropy. The propensity of complex structures (Guff) is lower, just because the reduced probability of long complementary stretches. Because of the steeper folding funnel, due to a smaller number of structures which are close the mfe, the number of adjacent structures is higher compared to AU-type alphabets, additionally the diversity of neighbouring structures is very low, figure 5. The effect of the GU-wobble pairing of the canonical alphabet is most pronounced in terms of the Shannon entropy which is reduced to levels of the six mixed alphabet (figure 2), as well as the diversity around the mfe structure where it mirrors six letter alphabet (figure 5), suggesting that the additional pairing option acts like a new letter pair. The nature invented wobble pairing to get the benefits of a six letter alphabet without wasting metabolic energy of building a new letter pair.

Future perspectives

To describe adequately the collective structure charachteristics of the sets of studyied RNA sequences properties of the folding funnel should be deeper investigated. Here, the number of structures within an energy range was recorded which gives information about the accessible structures by thermal movement. However, interesting information could be gathered through the proportion of structures that ly within this energy range compared to all available structures for the sequence. This information is already partly used to compute the Shannon entropy but can be more directly calculated via the base pairing probability matrix.
The variations of the shown variables are not shown in this report, however this information is valuable as well and it will be included in future.
The energy range used to calculate neighbouring structures was arbitrarily choosen to be $1\frac{kcal}{mol}$ a more reasonable value would be $5kT$ which corresponds to the loss of two G·C/C·G stacking interactions and amounts to $3\frac{kcal}{mol}$ at 37° C [Ancel and Fontana, 2000].
To lay the study onto a statistic solid ground the number of simulated sequences for each sequence length will be increased from currently 100 to 1000.

References

L.W. Ancel and W. Fontana. Plasticity, evolvability, and modularity in RNA. *Journal of Experimental Zoology*, 288(3):242–283, 2000.

S. Bonhoeffer, JS. McCaskill, PF. Stadler, and P. Schuster. RNA multi-structure landscapes. *European Biophysics Journal*, 22(1):13–24, 1993.

KA. Dill. Principles of protein folding–A perspective from simple exact models. 4(4):561–602, 1995.

W. Fontana, DAM. Konings, PF. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33(9):1389–1404, 1993a.

W. Fontana, PF. Stadler, EG. Bornberg-Bauer, T. Griesmacher, IL. Hofacker, M. Tacker, P. Tarazona, ED. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47(3): 2083–2099, Mar 1993b. doi: 10.1103/PhysRevE.47.2083.

PP. Gardner, BR. Holland, V. Moulton, M. Hendy, and D. Penny. Optimal alphabets for an RNA world. *Proceedings: Biological Sciences*, 270(1520):1177–1182, 2003.

DJ. Hill, MJ. Mio, RB. Prince, TS. Hughes, and JS. Moore. A field guide to foldamers. *Chem Rev*, 101 (12):3893–4012, 2001.

IL. Hofacker, W. Fontana, PF. Stadler, and P Schuster. Vienna RNA Package, 1994–1998. free software.

D.A. Mac Dónaill. A parity code interpretation of nucleotide alphabet composition. *Chem. Commun*, 18:2062–2063, 2002.

JS McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.

JS Reader and GF Joyce. A ribozyme composed of only two different nucleotides. *Nature*, 420(6917): 841–4, 2002.

C. Reidys, C.V. Forst, and P. Schuster. Replication and mutation on neutral networks. *Bulletin of Mathematical Biology*, 63(1):57–94, 2001.

J. Rogers and G.F. Joyce. The effect of cytidine on the structure and function of an RNA ligase ribozyme. *RNA*, 7(03):395–404, 2001.

E.A. Schultes. Estimating the Contributions of Selection and Self-Organization in RNA Secondary Structure. *Journal of Molecular Evolution*, 49(1):76–83, 1999.

P. Schuster. RNA based evolutionary optimization. *Origins of Life and Evolution of the Biosphere*, 23 (5-6):373–391, 1993.

J.M. Smith and E. Szathmáry. *Major Transitions in Evolution.* Oxford Univ Press, 1995.

E. Szathmary. What is the Optimum Size for the Genetic Alphabet? *Proceedings of the National Academy of Sciences*, 89(7):2614–2618, 1992.