On the structural repertoire of pools of short, random RNA sequences

Michael Stich, Carlos Briones, and Susanna C. Manrubia*
Centro de Astrobiología (CSIC-INTA),
Instituto Nacional de Técnica Aeroespacial
Ctra. de Ajalvir km. 4
28850 Torrejón de Ardoz, Madrid, Spain

January 9, 2008

* Corresponding author. Email: cuevasms@inta.es; Tel.: +34 915 206 425; Fax: +34 915 206 424

Abstract

A detailed knowledge of the mapping between sequence and structure spaces in populations of RNA molecules is essential to better understand their present-day functional properties, to envisage a plausible early evolution of RNA in a prebiotic chemical environment or to improve the design of *in vitro* evolution experiments, among others. Analysis of natural RNAs, as well as *in vitro* and computational studies, show that certain RNA structural motifs are much more abundant than others, pointing out a complex relation between sequence and structure. Within this framework, we have investigated computationally the structural properties of a large pool (10^8 molecules) of single-stranded, 35 nt-long, random RNA sequences. The secondary structures obtained are ranked and classified into structure families. The number of structures in main families is analytically calculated and compared with the numerical results. This permits a quantification of the fraction of structure space covered by a large pool of sequences. We further show that the number of structural motifs and their frequency is highly unbalanced with respect to

the nucleotide composition: simple structures such as stem-loops and hairpins arise from sequences depleted in G, while more complex structures require an enrichment of G. In general, we observe a strong correlation between subfamilies – characterized by a fixed number of paired nucleotides – and nucleotide composition. Our results are compared to structural repertoire obtained in a second pool where isolated base pairs are prohibited.

Keywords: RNA motif, genotype-phenotype map, RNA folding, structural family, RNA world

1 Introduction

The distribution of RNA structural motifs within pools of random sequences is extremely heterogeneous, as theoretical studies and observation of natural secondary structures demonstrate (Fontana et al., 1993; Schuster et al., 1994). Knowledge of the relationship between sequence and structure space has a theoretical and practical relevance, among other reasons because structural diversity conditions the spectrum of different functionalities present in – and thus selectable from – a random pool of sequences (Lorsch and Szostak, 1994). The role played by parameters such as the sequence length (Sabeti et al., 1997) or the nucleotide composition (Knight et al., 2005; Kim et al., 2007) has been addressed as a way of modifying the functional diversity of random molecular ensembles. Two frequent goals of those studies are to maximize the structural diversity present in the pool and to enhance the presence of certain structures able to perform new functions (Wilson and Szostak, 1999; Gan et al., 2003).

Every RNA sequence can be mapped onto a secondary structure that corresponds to its minimum free energy folded state. The first mathematical studies on this correspondence readily revealed the huge degeneracy existing between the set of all sequences – genotype space, of magnitude 4^n if n denotes the length of the sequences – and the set of their possible secondary structures – a first approximation to the phenotype space (Stein and Waterman, 1978; Waterman, 1978). Calculations based on the compatibility between sequences and structures yield estimates of the average number of sequences that fold into a secondary structure. If isolated pairs are allowed in the secondary structure, there are, on average, about $1.402\ n^{3/2}\ 1.748^n$ sequences of length n folding into each possible secondary structure (Stein and Waterman, 1978). However, this huge number is of little practical relevance in the light of empirical observations and computational results with random pools: the so-called common structures are typically many orders of magnitude more frequent than rare structures (Schuster et al., 1994; Grüner et al., 1996a; Joyce, 2004). While common structures are easily obtained, even in small populations, and do not depend strongly on the nucleotide composition, sequences

folding into rare structures often need to be designed, for instance by means of inverse folding algorithms (Schuster et al., 1994; Hofacker et al., 1994).

A main concern of experimentalists seeking new ribozyme or aptamer activities is how to deviate the structural composition of the initial pools in the *in vitro* experiments from average expectations, thus enhancing for instance the presence of rare structures, or forcing the ensemble to be structurally biased towards specific common structures. One approach has been to maximize the length of the sequences in the starting pool in an attempt to increment the number of different motifs available (Bartel and Szostak, 1993). However, quantitative analyses have shown that long sequences offer little advantage to isolate simple motifs, and their effect might be even inhibitory (Sabeti et al., 1997). More recently, attention has focused on how the probability to obtain a fixed structural motif depends on the nucleotide composition (Knight et al., 2005; Kim et al., 2007). Interestingly, though increases in the size of the initial pool should imply an increase in the amount of different structures present, the dependence of structural diversity on population size has been rarely addressed. Furthermore, computational results indicate that the number of different major topological motifs present in the pool depends very weakly on its size (Gevertz et al., 2005). Modular evolution has been suggested as a plausible way to generate complex structures in a constructive way. This approach can be implemented either through the isolation of simple modules from random populations of short sequences, their directed modification and eventual combination (Sabeti et al., 1997), or as the selective evolution of populations towards specific modules, together with their ligation in suitable environments (Manrubia and Briones, 2007). The latter approach is of particular relevance at prebiotic stages, when emergence of biochemical function appeared in an unsupervised way.

Though our knowledge of the genotype-phenotype map has expanded largely in the last three decades, our understanding still has to be improved in order to comprehend the multiple implications it has on evolution, on setting the conditions for further selection, and on the

dynamic behavior of highly heterogeneous molecular populations. This is the main motivation to undertake the study here presented, where we fold 10^8 random sequences of length $n = 35$ nt and classify the obtained structures into main structure families. We find correlations between the frequency of certain structure families and the nucleotide compositions, and conclude that rare structures are to be found far from the average composition. Hence, they could be enhanced by tuning the fraction of each nucleotide in the sequences. One of our main results concerns the high fraction of sequences folding into topologically simple structure families: most abundant motifs resulting from random polymerization could constitute simple building blocks able to combine into more complex structures.

2 Results

2.1 Distribution and classification of secondary structures with isolated base pairs

In this Section, we describe the results of the folding of 10^8 RNA molecules of length 35 nt consisting of random linear sequences composed of the four types of nucleotides A, C, G, and U. In this first part of the study we allow the presence of isolated base pairs in the secondary structure.

2.1.1 Frequency distribution

The 10^8 sequences yielded 5163324 different secondary structures. Since many sequences fold into the same structure, it is of fundamental interest to study how sequences are distributed among structures. These results are summarized in Fig. 1. There are a few hundred structures which are very abundant (with more than 10^4 sequences folding into each of them) and a few million structures that appear only once or very few times. Although for a much smaller pool of random sequences, this has already been described before and had led authors to propose a generalized Zipf's law to describe the curve (Schuster et al., 1994; Grüner et al., 1996a; Schuster and Stadler, 1994; Tacker et al., 1996). Between the two extreme regimes, the curve

in Fig. 1 shows a bump for intermediate ranks (around 3×10^4). Open structures, with rank 0, are not displayed in the figure.

FIGURE 1 NEAR HERE

2.1.2 Classification of structures

In Table 1 we show a classification of the observed 5163324 different secondary structures into 21 families. The classification is obtained according to the number of basic structural motifs found in the structures (see Methods). All structures that are not open consist at least of one stack and one hairpin loop.

TABLE 1 NEAR HERE

The first part of the table contains the structures with one hairpin loop (H=1). From a topological viewpoint, the simplest secondary structure is the stem-loop (SL), composed of exactly one hairpin loop and one stem. Next in complexity comes the hairpin (HP) structure, formed by two stacks, one hairpin loop and one bulge or interior loop. It is important to note the distinction between hairpin structure and the basic structural motif called hairpin loop. Topologically, a hairpin structure can be interpreted as a stem-loop where the stack has been interrupted by a bulge or interior loop. We can define higher-order hairpin structures where two of more interior loops or bulges are present (denoted as HP2 to HP6, i.e., with up to 6 interior loops and/or bulges and 7 stacks). All the structure families with H=1 have one hairpin loop and are therefore topologically linear. We also show in the table the most abundant structure belonging to a given structure family and its frequency. The representative of the stem-loop family is also the most abundant structure found at all.

The next class of structure families is characterized by having two hairpin loops, H=2. In particular, the double stem-loop family (DSL) is characterized by two stems and two hairpin loops. Double stem-loops can be viewed as consisting of two independent stem-loops. Similar to above, we can now allow the stacks being interrupted by interior loops and bulges and in

this way obtain higher-order DSLs. Within the studied sample, we find DSLs up to order 5, although higher-order families turn out to be very sparsely populated. For the classification, we do not distinguish in which of the two stacks the interior loops or bulges are present.

Also formed by two hairpin loops, but with the presence of a multiloop and an additional stack, we find the hammerhead (HH) family. A typical structure of this family has two stem-loops that instead of having ends belonging to different stacks, have the 5' and 3' ends connected to an additional stack. However, sequences folding into HH are relatively rare, and in particular higher-order hammerheads, i.e., HH with interior loops or bulges, are only present in very small numbers.

The next class of families has three hairpin loops, H=3. The case without additional loops corresponds to the triple stem-loop (TSL) family. Also, hammerhead-like structures are found, i.e., structures with a multiloop, which constitute the THH family. For both, TSL and THH, higher-order families with interior loops or bulges are possible. Again, we do not distinguish in which of the stacks the interior loops or bulges are present. In our pool, structures with H=3 are very rare. We also find a structure comprised of four stem-loops, called quadruple stem-loop (QSL). Its appearance is purely anecdotic since there is only one sequence of the sample folding into this H=4 class of structures.

For the SL family, the most frequent structure has 4 base pairs, a hairpin loop of 4 nt, and a dangling end of length 23 nucleotides at the 3' end. The next frequent structures in this family have loops of 3 or 4 nt and stacks of length 4 or 5 bp. In general, the stack of the most abundant structures is located at or very near the 5' end. If we compare the most abundant structure of the SL family with the most abundant structures of the HP and the higher-order HP families, we see that the hairpin loop has always a size of 4 nt and that there is a trend towards shorter stacks, although the total number of paired bases increases. This is due to the fact that smaller loops are unstable and less likely to be minimum free energy structures. Also, higher-order HPs require more stacks. However, short stacks are again relatively unstable and

hence the overall number of base pairs increases. The preference for a hairpin loop of 4 nt is also observed in the most abundant representative of other structure families (see Table 1 and Fig. 2).

2.1.3 Sequence and structure frequencies

The fact that the number of sequences folding into the most abundant structure of a given family decreases dramatically as we go to higher-order families within a class should also be put into the context of the number of possible (and actually realized) structures constituting the family. Theoretical calculations show that this number increases dramatically as the number of basic structural motifs characteristic for the family grows (more on this below). For families with many structures appearing at low frequencies, the sampling may not be sufficiently good and the displayed structure is not necessarily the most abundant structure of the family for larger pools.

TABLE 2 NEAR HERE

In Table 2, we display the main results of the extensive folding carried out. The first line of Table 2 denotes the open structures. All those sequences that do not fold – and that therefore have zero free energy – are formally counted as a single structure. The next line represents the stem-loop structures. We see that 20% of all sequences fold into a SL. Since the SL family has a relatively low number of different structures, many of them are found repeatedly: on average, there are approximately 8600 sequences per SL structure.

The third line gives the respective numbers for the hairpin structures. Around 37% of all sequences fold into a hairpin, revealing that the HP structure is the most probable structure to be formed by a random RNA sequence of length 35 nt. Compared to the 2330 SL structures, there are many more HP structures (182569 in the sample) and therefore we have 204 sequences per HP structure on average. The next family, hairpins of order 2 (HP2), represent the second-largest group with respect to folded sequences: more than 22% of all sequences fold into a

structure of this family. Interestingly, there are ten-fold more structures belonging to this class than to HP, roughly 1.6 million. This makes the HP2 family the second most represented family in terms of found structures. It is only outweighed by the HP3 family, which with 4.6% of the sequences accumulates 34% of the found structures, yielding an average of only 2.6 sequences per structure. The remaining HP structure families are less densely populated and are not discussed in detail here.

The families with two hairpin loops, DSLs and HHs, deserve a short discussion. We see that the double stem-loops are relatively frequent (7.8% of all sequences), while the family encompasses a relatively low number of different structures (82554). Therefore, the ratio of sequences to structures is relatively high, 94. The other DSL families are less densely populated. The hammerhead families are already rare in sequences although they still contribute considerably to structure diversity: 0.5% of the sequences fold into 4.6% of the structures. Sequences folding in structures with three or four hairpin loops (TSL, THH and QSL families) are very rare.

The last line of Table 2 summarizes the results for the complete pool, with an average of about 19 sequences per structure. The average free energy is -5.23 kcal/mol (-5.34 kcal/mol if open structures are disregarded). The four nucleotide types are equally distributed in the pool of folded sequences.

From the presented results we can draw some preliminary conclusions. First, the most abundant families in terms of folded sequences are HP, SL, HP2 and DSL: 87.8% of all sequences fold into structures belonging to one of these families. However, these families are not very diverse in terms of number of structures, representing only 36.6%. If we neglect the HP2 family, the difference becomes even more obvious: 65.2% of the sequences fold into 4.2% of the structures. Second, roughly 99.5% of the sequences fold into linear structures according to the classification of Ref. (Gan et al., 2003), in qualitative agreement with the results displayed in Fig. 5 of Ref. (Gevertz et al., 2005). Third, simple structures are the most frequent ones:

every family of order $k+1$ is less abundant in sequences than the preceding one, of order k. The only exception would be hairpins being more frequent than stem-loops, if the former were regarded as "higher-order stem-loops". Fourth, the number of obtained structures within a family – at least for the SL/HP and DSL families, where the statistics is very good – increases to a maximum value as the order increases, and then decreases. This can be explained by the fact that low-order families are composed of relatively few different structural modules and have therefore few representatives. Families with an intermediate number of loops have already many different configurations which are actually found in considerable numbers. However, high-order families are less likely to be realized since they require many short stacks for the molecule length considered.

2.2 Distribution of secondary structures prohibiting isolated base pairs

We carried out a new set of simulations, folding the same 10^8 molecules without permitting isolated base pairs in the secondary structures. It is known that isolated base pairs are unstable with respect to thermodynamic perturbations of the folded structure and do not contribute strongly to the free energy, so this situation might be more suitable to represent an actual experimental pool.

TABLE 3 NEAR HERE

The fundamental result is that instead of more than 5.2 million different structures, we only find 2.3 million. Therefore, in the simulations described above, 2.9 million structures actually contained at least one isolated base pair. The dashed curve in Fig. 1 shows the frequency-rank relation for this case. Since we still have 10^8 sequences, but less structures, some structures are now found more often. Hence, for low ranks, the dashed curve lies above the solid one and then drops. Nevertheless, the main qualitative features of the solid curve are again found: a flat plateau for low ranks, indicating frequent structures, a bump for intermediate ranks, and a steadily falling tail.

FIGURE 2 NEAR HERE

Table 3 gives the main numerical results of the simulation. The first observation is that we have not only considerably fewer structures, but also less structure families. In particular, there are no representatives found for the HP6, DSL5, TSL3, THH2, and QSL families, all high-order families that were populated only by – in total – 50 sequences in the simulations permitting isolated base pairs. Prohibiting isolated base pairs restricts the possible folding results, i.e., the number of possible structures. We find that the number of different structures found has decreased for all families. As an example, Fig. 2 displays the most abundant structures in two representative families of the hairpin class (HP and HP5) in the two situations analysed, that is, permitting and prohibiting isolated base pairs in the structure. With respect to the distribution of the sequences within the families, we observe that the higher-order families are much less populated than before whereas families with simple structures are now more often found: SL, HP and DSL families contain more sequences than before. The strong decrease of the sequences folding into HP2 structures and the simultaneous strong increase of both HP and SL indicate that there is probably a continuous shift from higher-order structures to lower-order ones, i.e., some HP3 now fold into HP2, some HP2 now fold into HP (or directly into SL), and also some HP fold now into SL. The number of open structures must increase, although it does it only slightly, from 2.1% to 2.4%.

Trends in the energetic behavior are less clear: avoiding isolated base pairs turns into an increase of the average free energy for some structure families (SL, HP, HP5, DSL, DSL2, DSL4 and HH) but into a decrease for the rest. If the minimum free energy structure of a given sequence has isolated base pairs, the prohibition of the latter must lead to a structure with a higher (or equal) free energy. However, since sequences generally switch their family as isolated base pairs are prohibited, the average free energy need not necessarily increase for a given family (although it does of course for the total pool). The average free energy is now -5.11 kcal/mol (-5.23 kcal/mol if open structures are disregarded).

2.3 Rank distribution of structure families and nucleotide composition

In this section we describe how the structures belonging to each family are distributed in the frequency-rank diagram. Figure 3(a) shows the corresponding results for structures where isolated base pairs are permitted. The numbers of all structures within a specific rank interval have been summed up. The solid curve contains all sequences and corresponds to the solid curve from Fig. 1.

FIGURE 3 NEAR HERE

We clearly see that the most frequent structures are all stem-loops. Since practically no structures of other families are present for low ranks, the SL curve coincides with the full curve there. For ranks larger than 10^3, stem-loops become rare and the corresponding curve decays rapidly. If we look at the HP curve, we see significant contributions only for intermediate ranks. The curve increases for ranks around 10^3, reaches its maximum, then practically coincides with the full curve for ranks between 4×10^3 and 10^4 and decays smoothly for higher ranks. It is remarkable to observe that the bump of the total curve for intermediate ranks to a good approximation coincides with the maximum of the HP curve and the region where SL and HP families contribute equally to the total number of structures, whereas other families are practically absent: stem-loops are already relatively infrequent while hairpins do not yet dominate. As the rank increases, other families successively appear and start to contribute significantly. In particular, DSL structures are the second-most frequent structures for ranks in the low 10^4. Structures of the HP2 family start to contribute for ranks around 10^4 but become important for ranks 10^5 to low 10^6. Structures of the HP3 family – the most frequent in terms of structures in our sample – start to contribute not before ranks around 10^5, although they dominate the distribution for ranks around 10^6. Finally, hammerhead structures show a maximum in their modest contribution for ranks around 10^6. In summary, the distribution curve is dominated by structures of a single family only within two regions – by SL for ranks

up to 10^3 and by HP for ranks between 4×10^3 and 10^4. For higher ranks, the hairpin families HP2 and HP3 are the main contributors.

It is worth comparing these findings with curves describing the relative content of the different types of nucleotides, shown in Fig. 3(b). We observe that all frequent structures, for ranks up to 10^3 and hence dominated by SL structures, arise from sequences depleted in G, with a mean density around 21%. The shortage of G is balanced by a significant increase in U and A, around 26% and 27%, respectively. Although possibly increased slightly compared to 25%, the fraction of C remains rather constant. However, as HP structures appear in ranks around 10^3, the curve for G increases strongly, while the curve for A decreases. Also C and U decrease, but only slightly.

2.4 Analytical results

Given a structure family as defined above, it is interesting to calculate exactly the total number of different possible structures within. To this end, one should know the structural constraints or elements involved, such as hairpin loops, internal loops or stacks, and calculate the different ways in which the n nucleotides in the linear sequence can combine to yield those elements. Our analysis of the structures obtained reveals that the number l of pairs in a structure is strongly correlated with the composition of a sequence. This is the reason why we calculate the number of different structures corresponding to sequences of length n forming structures with l base pairs, irrespectively of the size of loops (as long as they are formed by at least three nucleotides) and length of open ends. This combinatorial calculation does not consider energetic restrictions and can neither estimate the probability that a random sequence folds into each of the possible secondary structures. We have carried out this analysis for the four most abundant families obtained in the computational studies: SL, HP, HP2, and DSL.

All stem-loop structures have a terminal hairpin loop and a compact stack of size l. The most abundant SL structure, as represented in Table 1, has $l = 4$. In general, the number

$S^{\mathrm{SL}}(n,l)$ of stem-loop structures that can be formed with sequences of length n and stacks of length l, with a size of the hairpin loop equal or larger than 3, is

$$S^{\mathrm{SL}}(n,l) = \frac{1}{2}(n-2l-1)(n-2l-2), \tag{1}$$

with $l \geq 1$. To calculate the corresponding number for the HP family, it is necessary to take into account that at least one unpaired nucleotide interrupts the stack. In general,

$$S^{\mathrm{HP}}(n,l) = A^{\mathrm{HP}}(l)(n-2l+4)(n-2l-1)(n-2l-2)(n-2l-3), \tag{2}$$

and $l \geq 2$. The coefficient $A^{\mathrm{HP}}(l)$ depends on the minimum number of pairs required to form a stack. In the present case, we accept that stacks can be formed by a single pair, and get $A^{\mathrm{HP}}(l) = (l-1)/24$. When restrictions are imposed on the length of stacks, this is the only coefficient that is modified in the expression for the total number of possible structures with l pairs (see below). A similar calculation yields the number of structures belonging to the double stem-loop family DSL, formed by two ligated stacks that enclose hairpin loops,

$$S^{\mathrm{DSL}}(n,l) = A^{\mathrm{DSL}}(l)(n-2l-2)(n-2l-3)(n-2l-4)(n-2l-5), \tag{3}$$

with $A^{\mathrm{DSL}}(l) = (l-1)/24$ and $l \geq 2$. Finally, the calculation for the family HP2 is slightly more involved, since a second internal loop or bulge has to be taken into account. The final result is

$$S^{\mathrm{HP2}}(n,l) = A^{\mathrm{HP2}}(l)(n-2l+10)(n-2l+3)(n-2l-1)(n-2l-2)(n-2l-3)(n-2l-4), \tag{4}$$

with $A^{\mathrm{HP2}}(l) = (l-1)(l-2)/1440$ and $l \geq 3$.

Just for comparison, consider the total amount of possible structures with eight pairs in each of those families, that is $n = 35$ and $l = 8$. While there are only $S^{\mathrm{SL}}(35,8) = 153$ stem-loops with these characteristics, one finds $S^{\mathrm{HP}}(35,8) = 32844$ simple hairpins, $S^{\mathrm{DSL}}(35,8) = 16660$ structures of type double stem-loop, and a much larger amount $S^{\mathrm{HP2}}(35,8) = 1366596$ of hairpins with two internal loops (or bulges). In our simulation, all of the 153 stem-loops

are found, 25426 structures in the HP family, 11101 of type DSL, and 349193 structures in the HP2 group.

The ratio between the number of different structures found in our computations and the number of possible structures as obtained from the previous equations is displayed in Fig. 4 for each of those four families as a function of the number of pairs in the structure. We see that structures with a small number of base pairs (low l) are rare in the random pool, due to their small folding energy. The space of structures is better sampled for intermediate values of l, where the degeneracy between genotype and phenotype is higher: there are more sequences corresponding to each structure in that range than for extreme values of l. In fact, the curves decrease at high l despite the fact that structures with many base pairs are more stable. For large l, the sampling becomes insufficient as we consider families with more structural modules (see Table 1) as reflected in the ordering of the curves.

FIGURE 4 NEAR HERE

2.5 Relationship between structure, minimum free energy, and sequence composition

We have observed a strong correlation between the appearance of certain structural elements, notably the number of base pairs in a secondary structure, and two quantities of experimental relevance: the composition of nucleotides in the sequence and the minimum free energy of the folded molecule. Consider as an example the four different subfamilies (with $l = 3$, 5, 7, and 9 base pairs) within the SL family represented in Fig. 5. Figure 5(a) illustrates the correlation between the proportion of each type of nucleotide and the number of pairs formed, while Fig. 5(b) demonstrates the correlation between the energy of each secondary structure and the average composition of the sequences folding in that structure. Each point in the plots corresponds to the properties of one structure in the subfamily, and composition and energy are averaged over all the sequences that have that secondary structure as minimum free energy structure. The dispersion observed is due to other structural elements (loops at the

end of the stack and dangling ends) that differ among structures and contribute in different ways to determine the energy and the composition of the folded state. The stack of the most abundant SL structures is placed, independently of its length, very close to the 5'-end of the sequence. This possibly reflects the asymmetry of the energy contributions of dangling ends in the folded structure (Hofacker et al., 1994). Long dangling ends have actually a stabilizing effect that has been experimentally observed in RNA-RNA duplexes, where they increase the stability of the structure due to the cooperative stacking interactions among the unpaired nucleotides (Ohmichi et al., 2002). In general, the strong correlations here observed with the number of base pairs weaken as more structural elements appear in the structure.

FIGURES 5 AND 6 NEAR HERE

This fact can be already observed when we compare two families in our classification, take SL and HP2 as an example. Consider the average values of the composition and the energy as a function of the number of pairs as represented in Fig. 6. In Fig. 6(a) the average energy of structures in each family decreases as the number of pairs in the structure increases, as expected. However, the decrease is stronger for the SL family, since the number of unpaired elements that contribute positively to the total energy (one hairpin loop and the open ends) is smaller than the number contributing to hairpins of the class HP2 (one hairpin loop, two internal loops or bulges, and the open ends). Correlations between differences in composition and the number of pairs also weaken: while the ratio [A]/[G] varies between 1.7 and 0.6 for the SL family, the interval shrinks to 1 - 0.65 for the HP2 family. The differences that we observe here are thus relevant for small molecules or motifs, and disappear as the length of sequences increases and more structural elements can contribute to structures with a fixed number of pairs. In Fig. 6(b) we display how the mean free energy decreases as the number of base pairs increases.

2.6 Other structural constraints and variations in the sequence length

The analysis above has been also carried out under the condition that isolated base pairs are forbidden, so that a stack is formed at least by two consecutive pairs. As discussed, this is a more biologically meaningful – though topologically less rich – situation: it enormously reduces the total number of possible secondary structures and eliminates in particular rare structures with a large folding energy that are not found in practice. The qualitative results obtained do not strongly depend on this structural constraint, though the space of possible structures is better sampled in this case (see Fig. 7). One can repeat the calculation of the number of possible structures in each of the four most abundant families and obtain the new coefficient $A(l)$ for each family, the rest of the expression remaining unchanged. The calculation yields $A^{\mathrm{HP}}(l) = A^{\mathrm{DSL}}(l) = (l-3)/24$ (now with $l \geq 4$, since there are two stacks of length two at least) and $A^{\mathrm{HP2}}(l) = (l-4)(l-5)/1440$ (with $l \geq 6$).

FIGURE 7 NEAR HERE

There are theoretical results that yield the total exact number of possible compatible secondary structures for sequences of length n when different structural constraints are considered (Stein and Waterman, 1978; Hofacker et al., 1998; Liao and Wang, 2004). Those values can be compared with the number of structures measured in our simulations to better comprehend the degree of sampling of the structure space. For length $n = 35$ nt, the total number of compatible secondary structures is 1.214×10^{10} if single pairs are allowed, and this quantity decreases to 1.572×10^{7} under the condition that a stack is formed at least by two consecutive pairs. The difference is huge and reveals that almost 99.9% of structures in the first case are unstable from a thermodynamic viewpoint due to the presence of isolated pairs. In the pool where we have prohibited isolated base pairs, we have identified about 15% of all possible structures. Taking into account that we have only sampled the tiny fraction of 10^{-13} of the sequence space, this is a remarkably high content. On the other hand, this again talks about

the difference between common and rare secondary structures: actually, if all the structures without isolated base pairs would be equally possible given a random sequence, our pool of 10^8 sequences should cover practically all the structure space, with about ten sequences per structure. This is however not the case: common structures become dominant as the length of the sequence grows, and are the only structures found asymptotically (as $n \to \infty$); rare structures are obtained from a fraction of sequences that tends to zero as the length of the sequence diverges (Grüner et al., 1996a). As a consequence, the repertoire of structures obtained in a large enough pool of sequences is very likely a good representative of structures with a functional and evolutionary significance.

3 Discussion

The structural space of RNA is vastly smaller than its sequence space. In order to deepen into the features of such degeneracy, we have folded in silico two pools of 10^8 random RNA sequences of length 35 nt and classified the obtained secondary structures – about 10^6 – in roughly 20 structure families (see Methods and Table 1). We found that hairpins are the most probable structures formed by a random RNA sequence of this length. In fact, more than half of the sequences in our pools fold in structures belonging to HP and HP2 families. The third most probable structural conformation is the stem-loop. All together, about 80% of the sequences fold into one of these preferred structure families, irrespectively of whether isolated base pairs are allowed or prohibited (Tables 2 and 3). Therefore, our results support the hypothesis that hairpins and stem-loops are likely preferred building blocks of functional RNA structures.

The fact that there exist common and rare structures in a pool of random sequences is well-known and is described by the frequency distribution of structures. The frequency-rank diagram (Fig. 1) qualitatively recovers the curves obtained by others before (Schuster et al., 1994; Grüner et al., 1996a; Schuster and Stadler, 1994; Tacker et al., 1996) with the

characteristic bump for intermediate ranks. However, our simulations represent a much larger pool and reveal that the shape of the curve is directly related to the presence of different structure families (Fig. 3(a)). In particular, we document that the intermediate bump is due to the switch from stem-loops to simple hairpins as most abundant structure family.

From an evolutionary point of view it is plausible that, at early stages of the RNA world, the repertoire of structures yielded by pools of the kind here analyzed contained an abundance of common structures that constituted the raw modular material for further biochemical evolution (Manrubia and Briones, 2007). Currently, certain simple structures are found in RNA molecules with a probability above random expectation, presumably revealing that their presence was enhanced through evolution and selection. This is the case with the two most abundant families as here identified. Hairpin motifs are very abundant in nature (Hendrix et al., 2005) and have been described as essential secondary structures of RNA that guide its folding process in vivo, modulate gene expression in both RNA and DNA genomes, protect messenger RNA from degradation, serve as a recognition motif for certain RNA binding proteins or act as a substrate for enzymatic reactions (Svoboda and Di Cara, 2006). In turn, the stem-loop is a basic motif present in all the higher-order RNA structures (Gan et al., 2003), and it has been described to be more abundant and stable in non-coding regions of prokaryotic genomes than expected by chance (Petrillo et al., 2006). Recently, genome-wide surveys for non-coding RNAs in long vertebrate genomes have proven that a large fraction of the small ncRNAs fold in structures belonging to the HP and SL structure families (Pedersen et al., 2006; Backofen et al., 2007).

Roughly 2% of the sequences in our pools remain as open structures (Tables 2 and 3). Regarding its nucleotide composition, these sequences have an [A]/[G] ratio higher than 2.2, far above the value obtained for any of the folded structure families. The preference for A-rich or G-depleted sequences in unfolded regions of RNA has been described in nature, in poly-A tracks of mRNAs, in ribosomal RNAs, and in locally unfolded RNA regions prone to

interact with other intra- or intermolecular RNA regions as well as with some RNA-binding proteins (Khanam et al., 2006; Gutell et al., 2000; Hackermüller et al., 2005; Hiller et al., 2006). Also of interest, although not surprising due to the basic thermodynamic parameters used in `RNAfold` (see Methods), our results show the clear preference for hairpin loops 4 nt long in all structure families. In fact, the most abundant structure of each family shows one or more of these ubiquitous tetraloops (see Table 1), whose stability and abundance has been described in extant RNA molecules (Hendrix et al., 2005; Jaeger et al., 1989; Moore, 1999; Zorn et al., 2004).

Apart from their relation with RNA topologies currently found in nature, our results can be also useful for the design and optimization of RNA *in vitro* evolution experiments. Results from different experimental approaches show that evolved aptamers (Carothers et al., 2004; Lee et al., 2004) or small catalytic RNAs (Puerta-Fernández et al., 2003; Lilley, 2005) tend to have simple topologies, with linear or low branched motifs analogous to those preferentially obtained here. Our results also show that the amount and frequency of structural motifs is highly unbalanced with respect to the abundance of the four nucleotides in the pool (Fig. 3(b)). In particular, all frequent structures (practically being simple stem-loops) are formed by sequences with – on average – a significant low content in G.

The dependence between structural complexity and composition can be exploited in order to force the appearance of motifs with certain structural – and thus functional – properties. In particular, we observe a correlation between the number of base pairs in a secondary structure and the composition of a typical sequence folding into that structure. Rare structures are characterized in our pool by having a larger number of structural elements. This is only possible in structures with several stacks (implying a relatively low number of base pairs in the stacks and an abundance of loops) that thus have to be highly stable: this is achieved by lowering the content in U, and increasing the content in G, to improve stability. Actually, other authors have analysed how a bias in composition enhance or impair the presence of

structural motifs of interest (Fontana et al., 1993; Kim et al., 2007; Gevertz et al., 2005), with the goal of optimizing the chance of finding specific functions (Knight et al., 2005; Knight and Yarus, 2003).

We also have studied, for the stem-loop structure family, how the total length of the stacks correlate with the composition of the sequence (Fig. 5). We see that, for fixed A, the more G is present in the sequence, the more base pairs are formed. Conversely, for a fixed amout of G, the more A is present, the less base pairs are formed. Also, we observe that – beyond the fact that the number of base pairs determines to a large extent the folding energy (Fig. 6(b)) – the more G is present, the lower the energy for a given number of base pairs. The variation of the energy according to the G content may be of the same order as the total folding energy. Subsequently, we have investigated how the ratio [A]/[G] depends on the number of base pairs (Fig. 6(a)). We see for the most common structure families (representing almost 90% of the sequences) that the smaller this ratio, the more base pairs are formed.

Our results are in agreement with the common knowledge about the contributions of different types of nucleotides, and the dependence of the number of base pairs to the folding energy, as derived from experiments and reflected in the parameter values used in `RNAfold`. Nevertheless, the results shown here go beyond a qualitative validitation of the folding program since they represent statistically significant data obtained from large-scale simulations of random sequences. Also, the use of the classification in structure families introduced here, helps to sharpen the notion of common vs rare structures, or rather common vs rare structure families. Much work performed in the 1990s on the statistics of RNA secondary structures and the sequence-structure map focuses on interesting aspects not discussed here, e.g., the dependence on the length of the molecule, structure of neutral networks, shape space covering, landscapes and dependence on folding algorithms (Schuster et al., 1994; Grüner et al., 1996a; Schuster and Stadler, 1994; Tacker et al., 1996; Grüner et al., 1996b).

We have complemented our computational results by analytic calculations of the number of

possible structures within a given family. It turns out that for the common structure families a lot of structures are actually found in the compuational folding (Figs. 4 and 7). This is particular true if isolated base pairs are prohibited, where the quality of the sampling of the sequence space for common structure families is very good.

Our findings offer additional evidence to improve the design of the initial pools for *in vitro* RNA evolution experiments, and indicate that structural heterogeneity would increase with a mixture of random sequences originating from different pools that differ in their average nucleotide composition. Actually, rare structures departing largely from the average composition of a unique pool might be impossible to access in large sequences where fluctuations in composition are limited by their length. Examples in our study are families QSL, HP6, HH3, THH and THH2, which strongly differ from the average composition (see Tables 2 and 3). Large departures from average composition are easier to obtain with short sequences. Several independent *in vitro* evolution experiments have documented that the catalytic core of certain ribozymes can be trimmed to only 13 to 30 nt in length (reviewed in (Joyce, 2004; Puerta-Fernández et al., 2003)). In parallel, recent results also indicate that the length of ligand binding aptamer motifs can be easily reduced to 25-30 nt and in some cases to even smaller molecules with as few as 12-13 nt (Majerfeld and Yarus, 2005; Anderson and Mecozzi, 2006). Hence, pools of random sequences of length up to 35 nt cover a relatively larger part of their structure space in comparison to longer molecules. The use of short sequences in theoretical and experimental studies might suffice to yield the structural modules required to obtain fully functional RNA molecules.

4 Methods

4.1 Programming and computational resources

Simulations have been carried out at the Itanium II cluster of INTA (Instituto Nacional de Técnica Aeroespacial, Spain). For random number generation, we relied on the Mersenne

Twister and Ziff's FSR4 algorithms as provided by GNU Scientific Library (GSL), Version 1.7 (see http://www.gnu.org/software/gsl). For secondary structure folding (minimum free energy), we use the routine `fold()` from the Program `RNAfold` of the Vienna RNA package (Hofacker et al., 1994), version 1.5, with the energy parameter set based on Ref. (Mathews et al., 1999). It must be noticed that `RNAfold`, as most folding programs, does not allow for pseudoknots or other kind of tertiary structures. However, and in particular for the relatively short molecules considered here, secondary structures are a very good approximation of the tertiary structures since a major part of the folding energy corresponds to the secondary structure formation. No search for suboptimal structures was performed in this study.

The routine `fold()` is called with the default parameters, i.e., it allows Watson-Crick and G-U base pairing and the temperature is set to 37°C. No special stabilizing energy contributions for tetraloops are assumed. Dangling end energies are assigned only to unpaired bases adjacent to stacks in free ends and multiloops. A base cannot participate simultaneously in two dangling ends. For the simulations where we prohibit single base pairs, we call the routine `fold()` with option `noLonelyPairs` (Hofacker et al., 1994).

Secondary structures are obtained in the standard bracket notation being the default output of the routine `fold()`. For the classification using the number of basic RNA structural elements and counting of length of loops and stacks, we use the Shapiro notation provided by the function `b2Shapiro` from the Vienna RNA package.

4.2 Classification of secondary structures

Structures are classified according to the number of their basic elements: (i) stacks, i.e., regions formed by base pairs, (ii) loops, i.e., unpaired regions embraced by stacks, and (iii) external elements, i.e., unpaired nucleotides which are not part of a loop. Loops can be classified into hairpin loops (loops at the end of a stack), interior loops (loops connecting two stacks), bulges (unpaired bases within the chain whose neighbors are paired with nucleotides that are direct

neighbors in the complementary strand), and multiloops (loops with more than two adjacent stacks). The first criterion for classification is the number of hairpin loops, denoted by H and varying from 1 to 4 for the sample. For a given structure, the number of stacks, S, is either equal to H or larger (if there are other loops present) and represents the second level of classification. The third level of classification is then given by the sum of interior loops and bulges, I+B, together with M, the number of multiloops. The number of external elements is not used for classification (see Table 1).

Acknowledgements

The authors wish to acknowledge the technical assistance of Ruth Lobo and Pilar Viñado with the computations carried out at the Itanium II cluster of INTA, and the support of Isidro Cano and Hewlett-Packard within the project *OriGenes*.

Author contributions. All authors conceived and designed the study. MS performed the numeric calculations, SCM the analytic calculations. All authors analyzed the data and wrote the paper.

Funding. This work was supported by Ministerio de Educación y Ciencia (FIS2004-06414), INTA, INSA, EU and CAM.

References

P. C. Anderson and S. Mecozzi. Minimum sequence requirements for selective RNA-ligand binding: A molecular mechanics algorithm using molecular dynamics and free-energy techniques. *J. Comp. Chem.*, 27:1631–1640, 2006.

R. Backofen, S. H. Bernhart, C. Flamm, C. Fried, G. Fritzsch, J. Hackermüller, J. Hertel, I. L. Hofacker, K. Missal, A. Mosig, S. J. Prohaska, D. Rose, P. F. Stadler, A. Tanzer, S. Washietl, and S. Will. RNAs everywhere: Genome-wide annotation of structured RNAs. *J. Exp. Zool. (Mol. Dev. Evol.)*, 308B:1–25, 2007.

D. P. Bartel and J. W. Szostak. Isolation of new ribozymes from a large pool of random sequences. *Science*, 261:1411–1418, 1993.

J. M. Carothers, S. C. Oestreich, J. H. Davis, and J. W. Szostak. Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.*, 126:5130–5137, 2004.

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

H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res.*, 31:2926–2943, 2003.

J. Gevertz, H. H. Gan, and T. Schlick. In vitro RNA random pools are not structurally diverse: a computational analysis. *RNA*, 11:853–863, 2005.

W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monatsh. Chem.*, 127:355–374, 1996a.

W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II.

Structures of neutral networks and shape space covering. *Monatsh. Chem.*, 127:375–389, 1996b.

R. R. Gutell, J. J. Cannone, Z. Shang, Y. Du, and M. J. Serra. A story: unpaired adenosine bases in ribosomal RNAs. *J. Mol. Biol.*, 304:335–354, 2000.

J. Hackermüller, N.-C. Meisner, M. Auer, M. Jaritz, and P. F. Stadler. The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: a quantitative model. *Gene*, 345:3–12, 2005.

D. K. Hendrix, S. E. Brenner, and S. R. Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Quart. Rev. Biophys.*, 38:221–243, 2005.

M. Hiller, R. Pudimat, A. Busch, and R. Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, 34:e117, 2006.

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

I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 88:207–237, 1998.

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

G. F. Joyce. Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.*, 73:791–836, 2004.

T. Khanam, R. S. Muddashetty, A. Kahvejian, N. Sonenberg, and J. Brosius. Poly(a)-binding protein binds to a-rich sequences via RNA-binding domains 1+2 and 3+4. *RNA Biol.*, 3: 170–177, 2006.

N. Kim, H. H. Gan, and T. Schlick. A computational proposal for designing structured RNA pools for in vitro selection of RNAs. *RNA*, 13:478–492, 2007.

R. Knight and M. Yarus. Finding specific RNA motifs: Function in a zeptomole world? *RNA*, 9:218–230, 2003.

R. Knight, H. De Sterck, R. Markel, S. Smit, A. Oshmyansky, and M. Yarus. Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids. *Nucleic Acids Res.*, 33:5924–5935, 2005.

J. F. Lee, J. R. Hesselberth, L. A. Meyers, and A. D. Ellington. Aptamer database. *Nucleic Acids Res.*, 32:D95–D100, 2004.

B. Liao and T. Wang. General combinatorics of RNA secondary structure. *Math. Biosci.*, 191:69–81, 2004.

D. M. J. Lilley. Structure, folding and mechanisms of ribozymes. *Curr. Opin. Struct. Biol.*, 15:313–323, 2005.

J. R. Lorsch and J. W. Szostak. *In vitro* evolution of new ribozymes with polynucleotide kinase activity. *Nature*, 371:31–36, 1994.

I. Majerfeld and M. Yarus. A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res.*, 33:5482–5493, 2005.

S. C. Manrubia and C. Briones. Modular evolution and increase of functional complexity in replicating RNA molecules. *RNA*, 13:97–107, 2007.

D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

P. B. Moore. Structural motifs in RNA. *Annu. Rev. Biochem.*, 68:287–300, 1999.

T. Ohmichi, S.-I. Nakano, D. Miyoshi, and N. Sugimoto. Long RNA dangling end has large energetic contribution to duplex stability. *J. Am. Chem. Soc.*, 124:10367–10372, 2002.

J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, 2:e33, 2006.

M. Petrillo, G. Silvestro, P.-P. Di Nocera, A. Boccia, and G. Paolella. Stem-loop structures in prokaryotic genomes. *BMC Genomics*, 7:170, 2006.

E. Puerta-Fernández, C. Romero-López, A. Barroso-delJesús, and A. Berzal-Herranz. Ribozymes: recent advances in the development of RNA tools. *FEMS Microbiol. Rev.*, 27: 75–97, 2003.

P. C. Sabeti, P. J. Unrau, and D. P. Bartel. Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool. *Chem. Biol.*, 4: 767–774, 1997.

P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computers Chem.*, 18:295–324, 1994.

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

P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Discr. Math.*, 26:261–272, 1978.

P. Svoboda and A. Di Cara. Hairpin RNA: a secondary structure of primary importance. *Cell. Mol. Life Sci.*, 63:901–918, 2006.

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

M. S. Waterman. Secondary structure of single-stranded nucleic acids. In *Studies in Foundation and Combinatorics*, volume 1 of *Advances in Mathematics Supplementary Studies*, pages 167–212. Academic Press, 1978.

D. S. Wilson and J. W. Szostak. *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.*, 68:611–647, 1999.

J. Zorn, H. H. Gan, N. Shiffeldrim, and T. Schlick. Structural motifs in ribosomal RNAs: Implications for RNA design and genomics. *Biopolymers*, 73:340–347, 2004.

TABLE CAPTIONS

Table 1.

Classification of secondary structures. We show a classification of the observed secondary structures into 21 structure families. The classification is obtained according to the number of basic elements (see Methods). Besides these numbers we show the most abundant structure of the family and its frequency. The first part of the table contains the structures with one hairpin loop, H=1, which comprise the families stem-loop (SL), hairpin (HP), and the higher-order hairpin structures (HP2 to HP6). The next class of families is characterized by having two hairpin loops, H=2, and therefore at least two stacks. This class of families encompasses the double stem-loop (DSL) families. The hammerhead (HH) families are present up to HH3. The next class of families has three hairpin loops, H=3, and at least three stacks. The case without any further loops corresponds to the triple stem-loop (TSL) family. TSL families are found up to TSL3. Also, hammerhead-type of structures are found, i.e., structures with a multiloop, denoted as THH and present up to THH2.

Table 2.

Sequence and structure frequencies: results for regular folding permitting isolated base pairs. We show, for each structure family, the number of sequences folding into a structure belonging to it, its relative contribution (number of sequences divided by 10^8, in percent), the number of structures belonging to the given family and its relative contribution (number of computationally obtained structures divided by total number of structures, 5163324, in percent). The next column gives the average number of sequences per structure. Then, we give the average free energy of a sequence folding into a structure of the family (in kcal/mol); in paranthesis for the total value we give the value without considering open structures. In the next columns the relative content (in percent) of the different nucleotides for the sequences is given. The last

column gives the dispersion $1/\sqrt{nN_f}$ in points of percent of the relative nucleotide content, where N_f is the number of sequences in the corresponding family and $n = 35$. This value indicates whether deviations from the overall mean, i.e., 25%, are significant.

Table 3.

Sequence and structure frequencies: results for folding prohibiting isolated base pairs. The meaning of the columns is the same as in Table 2. Main differences are that the number of total different structures is now 2300308, and several higher-order families are absent (HP6, DSL5, HH3, TSL3, THH2, QSL).

FIGURE CAPTIONS

Figure 1.

Frequency distribution. We fold 10^8 RNA random sequences of length 35 nt. For regular folding, the 10^8 sequences fold into 5163324 different secondary structures. We order the structures according to their abundance and display the frequency of a secondary structure as a function of its rank (solid curve). For folding prohibiting isolated base pairs, we find 2300308 structures (dashed curve). Rank 0, representing the open structure, is not displayed.

Figure 2.

Most abundant structures in two representative families of the hairpin class. We show the six most abundant structures in the HP (above) and HP5 (below) families in our pool when single base pairs are permitted (left) and prohibited (right). Small numbers on the dangling ends of HP structures stand for the number of unpaired nucleotides at that end. The numbers below each structure indicate how many sequences folded into it. As can be seen, prohibiting single pairs in stacks increases the abundance of structures in the HP family, while the HP5 family becomes depleted.

Figure 3.

Frequency distribution according to (a) family and (b) nucleotide composition. (a) We have binned in boxes of powers of 2 the numbers of all structures belonging to the interval and have determined the absolute frequency of the corresponding sequences for each family in the bin. Note the double-logarithmic scale of the diagram. The solid curve corresponds to all sequences (cf. solid curve from Fig. 1). The other curves correspond to the sequences folding into different family structures, as shown in the legend. For low ranks, the SL curve coincides with the total curve. Around the bump of the total curve, SL become less frequent and HP

structures appear. See main text for further explanations. (b) We have summed up the numbers of all structures belonging to a specific rank interval and have determined the average nucleotide density for the corresponding sequences for a given bin. Again, the main frequency distribution (solid curve) is given for comparison.

Figure 4.

Sampling quality. Ratio between the number of structures found in the pool of 10^8 sequences and the number of possible structures. Each curve corresponds to one structure family (SL, HP, DSL, HP2) and each point gives the ratio for structures with a fixed number of pairs.

Figure 5.

Correlation between nucleotide content, energy, and number of pairs in the secondary structures of the SL family. (a) Total number of A nucleotides as a function of the total number of G nucleotides for stem-loop structures with stacks of length $l = 3$, 5, 7, and 9. (b) Energy of stem-loop structures as a function of their G content, stack length as in (a). There are well-defined average values of the ratio [A]/[G] and of the energy for each fixed value of l.

Figure 6.

Average value of the [A]/[G] content and of the energy for the four major structure families as a function of the number of pairs in the structure. (a) The represented value of [A]/[G] has been obtained through the interpolation of a straight line through least squares to data analogous to that of Fig. 5(a). Error bars correspond to the error of the obtained slope. (b) Mean free energy of same set of structures.

Figure 7.

Sampling quality. Ratio between the number of structures found in the pool of 10^8 sequences

and the number of possible structures. Same as Fig. 4 under the condition that no isolated base pairs appear in the structure. The space of structures is better sampled under this structural restriction.