doi:10.1016/j.jmb.2004.12.012 J. Mol. Biol. (2005) 347, 53–69

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Extracting Stacking Interaction Parameters for RNA from the Data Set of Native Structures

Ruxandra I. Dima[1], Changbong Hyeon[1] and D. Thirumalai[1,2]*

[1]*Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA*

[2]*Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA*

**Corresponding author*

A crucial step in the determination of the three-dimensional native structures of RNA is the prediction of their secondary structures, which are stable independent of the tertiary fold. Accurate prediction of the secondary structure requires context-dependent estimates of the interaction parameters. We have exploited the growing database of natively folded RNA structures in the Protein Data Bank (PDB) to obtain stacking interaction parameters using a knowledge-based approach. Remarkably, the calculated values of the resulting statistical potentials (SPs) are in excellent agreement with the parameters determined using measurements in small oligonucleotides. We validate the SPs by predicting 74% of the base-pairs in a dataset of structures using the ViennaRNA package. Interestingly, this number is similar to that obtained using the measured thermodynamic parameters. We also tested the efficacy of the SP in predicting secondary structure by using gapless threading, which we advocate as an alternative method for rapidly predicting RNA structures. For RNA molecules with less than 700 nucleotides, about 70% of the native base-pairs are correctly predicted. As a further validation of the SPs we calculated *Z*-scores, which measure the relative stability of the native state with respect to a manifold of higher free energy states. The computed *Z*-scores agree with estimates made using calorimetric measurements for a few RNA molecules. Structural analysis was used to rationalize the success and failures of SP and experimentally determined parameters. First, from the near perfect linear relationship between the number of native base-pairs and sequence length, we show that nearly 46% of nucleotides are not in stacks. Second, by analyzing the suboptimal structures that are generated in gapless threading we show that the SPs and experimentally determined parameters are most successful in predicting stacks that end in hairpins. These results show that further improvement in secondary structure prediction requires reliable estimates of interaction parameters for loops, bulges, and stacks that do not end in hairpins.

Keywords: RNA secondary structure; statistical potentials; threading; *Z*-scores; interaction parameters for base-pair stacking

Introduction

Since the pioneering discovery that RNA possesses catalytic activity the list of biological functions involving RNA has continued to expand.[1] Its biological functions include, among others, catalysis of peptide bond formation,[2] translation regulation and intron splicing,[3] and many enzymatic activities.[4] Because there is a direct link between the tertiary structure and the function of biological molecules it is important to determine the native structures of RNA. Unlike in proteins, in which secondary structure stability is coupled to tertiary interactions, secondary structures are more independently stable in RNA. As a result, RNA folding can be separated into two steps, with the first being the formation of secondary structure while the second represents the consolidation of secondary

Abbreviations used: SP, statistical potential; WC, Watson–Crick.

E-mail address of the corresponding author: dt5@umail.umd.edu

structures into compact native states. The two hierarchical levels of organization in RNA can be separated experimentally by adjusting the concentration of counterions (usually Mg^{2+}).

Over the last 20 years, there has been a concerted effort towards developing the tools needed for predicting RNA secondary structures.[5–11] Using these algorithms and experimentally determined thermodynamic data on small oligonucleotides, one can obtain fairly accurately the secondary structures for a large variety of RNA sequences composed of the different elements of secondary structure; namely, stacks and hairpins. Despite the success of algorithms for secondary structures many motifs (bulges, multiloops, pseudoknots) cannot be easily predicted.[12,13] Reliable estimates of thermodynamic interaction parameters are needed for secondary structure predictions to be successful.[9–11] At present, the most accurate parameters for RNA secondary structure come from the pioneering efforts of Turner & Zuker.[9–11] Using a combination of experimental measurements on short oligonucleotides and novel computational approaches, Turner & Zuker and their collaborators have produced interaction parameters that have been remarkably useful. A key finding made by these authors is that context determines the free energy (more precisely the potential of mean force) values. More recently, they have shown that use of experimental constraints in MFOLD greatly improves secondary structure prediction.[11]

In protein folding, the non-redundant structures in the Protein Data Bank (PDB)[14] have been used to obtain estimates of tertiary interactions between amino acid side-chains using methods introduced by Tanaka & Scheraga,[15] and developed further by Miyazawa & Jernigan[16] and others.[17,18] The resulting potentials have been moderately successful in obtaining coarse-grained descriptions of protein structures. To date, similar ideas have not been used to obtain interaction parameters for RNA using the knowledge of their three-dimensional structures. The increase in the number of RNA structures in the PDB allows us to assess the reliability of the knowledge-based parameters for RNA molecules. A key advantage of using PDB for obtaining statistical potentials is that an estimate of the tertiary interactions that stabilize the three-dimensional RNA structures can also be made. The purpose of this work is to utilize the currently available RNA structures to obtain interaction parameters for secondary structure predictions. We find that the calculated stacking interactions compare favorably with the experimentally determined thermodynamic parameters provided a temperature scale is chosen appropriately (see Methods). More importantly, when used in the ViennaRNA package, which does not use experimental constraints, the present statistical potentials and the Turner parameters are equally successful in predicting the secondary structures of RNA molecules.

Results

Stacking potentials

From the dataset of RNA structures (Table 1), we extract the stacking free energy for each combination of base-pairs using equation (3) (Methods), which requires as input the number of stacks of a given type (Table 2). The frequencies of occurrence of the four nucleotides in the sequences in our dataset are: $P_A = 0.23$, $P_C = 0.26$, $P_G = 0.33$ and $P_U = 0.18$. The data in Table 2 and the values of P_A, P_C, P_G, and P_U are used in equation (3) to obtain the statistical stacking energies (Table 3). Comparison with the thermodynamically determined values (shown in parentheses in Table 3) shows that the interaction parameters in the statistical potentials (SPs) agree very well with Turner's stacking free energies. The exceptions are the sign changes in the (UG,UG) and (UG,GU) pairs. There are two possible reasons for the disagreement in the stacking energies for these two pairs between the SP estimates and Turner's rules: (1) the number of intra-chain stacks for these two pairs is relatively small (Table 2). The knowledge-based approach works best only if there are a large number of stacks of a given type in the dataset. Consistent with this reasoning, we find that the largest discrepancy in the trends between the SPs and Turner's values arises whenever the entries in Table 2 are relatively small (<50). (2) Stacking free energies for specific base-pairs are determined by spectroscopic measurements from which the equilibrium constant between the open and closed form of stacked pairs in oligonucleotides is estimated. The experimental conditions in such measurements do not necessarily mimic the physical environment of stable contact formation found in naturally occurring RNA. Indeed, it is known that Turner's parameters have some limitation in predicting the correct native conformation of an RNA sequence when counterion binding plays an important role.[4] In the stacking free energy parameters obtained using the knowledge-based approach, counterion-mediated contacts are included only in a "mean-field" sense.

Gapless threading and fold recognition

We use gapless threading to assess the validity of the computed stacking interaction potentials. For each sequence S in the dataset, with known native state Γ_N, we calculated the native free energy $\Delta G_N(S,\Gamma_N)$ using the PDB structure and the free energies $\{\Delta G(S, \Gamma_D)\}$ when S is mounted on all other decoy structural fragments Γ_D of equal length from longer chains (see Methods). If the calculated free energy parameters are accurate, then the inequality $\Delta G_N(S,\Gamma_N) < \Delta G(S, \Gamma_D)$ should be satisfied for all D. Because the SPs are obtained in a coarse-grained manner using a limited sized RNA database, it is possible that the native state may not correspond to the minimum free energy state for all the test sequences. More importantly, the SPs have not

Table 1. List of 246 RNA structures used to obtain the statistical potential

Single chain									
17ra	la4t^{P}	**1a51$_{rs}$**	**1a60**	1A9N^{P}	**1afx**	**1aju**	**1AQ3^{P}**	**1arj**	**1ato$_{rz}$**
1atv	laud	**1b36$_{rz}$**	**1BAU**	**1bgz**	1biv^{P}	**1bn0**	**1bvj**	**1byj**	**1c0o$_{rz}$**
1c2w$_{rs}$	1ck5$^{P}_{rs}$	1ck8$^{P}_{rs}$	1cn8^{P}	**1cql**	1cx0$_{rz}$	1d6k$^{P}_{rs}$	1dkl$^{P}_{rs}$	1drz$^{P}_{rs}$	1dul^{P}
1DZS^{P}	**1e4p$_{rz}$**	1E7K^{P}	1e8o$^{P}_{rs}$	**1e95**	**1ebr**	1EC6^{P}	1lekz^{P}	**1esh**	**1esy**
1etf^{P}	1exy^{P}	**1f27^{P}**	1f6u^{P}	**1f7f$_{rz}$**	1f7y$^{P}_{rs}$	**1f84$_{rs}$**	**1f85$_{rs}$**	**1ffz$_{rs,rz}$**	**1fhk**
1fje^{P}	1fjg$^{P}_{rs}$	**1FOQ**	**1fqz$_{rs}$**	1g70^{P}	**1GID$_{rz}$**	**1GRZ$_{rz}$**	1HC8^{P}	1hji^{P}	**1hlx**
1hs2$_{rs}$	**1hwq$_{rz}$**	li3x	1I6U	1i94$^{P}_{rs}$	**1idV$_{rs}$**	**1ie2**	**1ikl**	**1ikd**	1jid^{P}
1JJN	1jo7	**1jP0**	**1jtj**	**1jtw**	**1ju7**	**1jur**	**1k4a**	**1l1w**	**1k5i**
1k6g	**1K9W**	**1kaj**	**1KIS**	**1kks**	**1kos**	**1kp7$_{rs}$**	**1kpz**	**1kxk**	**1l2x**
1l3d	**1L8V$_{rz}$**	1lng	**1lu3$_{rs}$**	**1m5l**	**1m82**	**1mfj**	**1mfk**	**1mfy**	1MJI^{P}
1MMS^{P}	1mnb^{P}	1mnx	1msy	1mt4	1mzp$^{P}_{rs}$	**1n8x**	**1na2**	**1nbr**	**1NBS**
1nc0	**1nz1**	**1o15**	1OOA^{P}	**1oq0**	**1osw**	**1ow9$_{rz}$**	1P5m$_{rs}$	1P5P$_{rs}$	1p9x$_{rs}$
1pbr$_{rs}$	1pjy	1q75	**1qc8**	1qfq^{P}	1qwa$_{rs}$	1qwb	1r2P$_{rz}$	1r7W$_{rs}$	1rfr
1bfn$_{hrz}$	**1rng**	**1rnk**	1roq	1s9s	**1scl**	1sjf$^{P}_{rz}$	**1Slp**	1sy4	1szy
1t4l^{P}	**1tfn**	**1tlr**	**1u2a**	1URN^{P}	**1uuu**	1vbx$^{P}_{rz}$	1vc0$^{P}_{rz}$	**1vop$_{rs}$**	**1wts$_{rs}$**
1zig	**28sp**	**2a91**	**2ldz$_{rz}$**	**2tpk**	**2u2a**	**361D**	**3php**	**430d**	**437d**
480d	**488d$_{rz}$**								
Multiple chain									
157d^{2}(*)	1a34^{2P}(*)	**1a3m^{2}**(*)	**1a4d$^{2'}_{rs}$** (*)	**1ajl$^{2'}$** (*)	**1ajt2$^{'}$** (*)	**1al5^{2}** (*)	1BY4^{2P} (*)	1c04$^{2'}_{rs}$	**1c2x$^{2}_{rs}$** (*)
1c4l^{2}(*)	**1csl^{2}** (*)	1dfu$^{2P}_{rs}$ (*)	**1dqf2$^{'}$** (*)	**1DUQ^{2}** (*)	1dz5^{2P} (*)	**1eka^{2}** (*)	**1ekd^{2}** (*)	**1elh2$^{'}$** (*)	**1f5g^{2}** (*)
1feu$^{2-2'-P}_{rs}$(*)	1ffk$^{2'}_{rs}$	**1fuf^{2}** (*)	**1guc^{2}** (*)	1hys^{2P} (*)	1i6h^{2P} (*)	**1i7j^{2}** (*)	**1i9k^{4}** (*)	**1i9x^{2}** (*)	**1j8g^{4}** (*)
1jbr^{3P}	**1k8s^{2}** (*)	**1kd3^{2}** (*)	**1l3z^{2}** (*)	**1lmv^{2}** (*)	**1lnt^{2}** (*)	1M5K$^{2'P}_{rz}$ (*)	**1mhk^{2}** (*)	**1mis^{2}** (*)	1msw$^{3'P}$ (*)
1muv^{2} (*)	**1mvl^{2}** (*)	**1mv^{6}** (*)	**1mwg^{2}** (*)	1mwl^{2} (*)	**1n35^{2}** (*)	1n53$^{2'}$ (*)	1NUJ$^{2'}$ (*)	**1pbm^{2}** (*)	1pns$^{4'P}_{rs}$
lqc0$^{2-2'}$ (*)	**1qcu^{2}** (*)	**1qes^{2}** (*)	**1qet^{2}** (*)	**1q1n$^{2'P}$** (*)	1qwa$^{2'}_{rs}$	**1rau^{4}** (*)	**1rna^{2}** (*)	**1rxb^{2}** (*)	**1sdr$^{4}_{rs}$** (*)
1TF6^{2P} (*)	**1yfv^{2}** (*)	**205d^{2}** (*)	**259d^{2}** (*)	**280D^{2}** (*)	**2bj2$^{2'}$**	**354d^{2}** (*)	356d^{2} (*)	**357d^{2}** (*)	**359d$^{2'}_{hrz}$**
364d^{2} (*)	**377D^{2}** (*)	**379d$^{2'}_{hrz}$**	**397d^{2}** (*)	**402d^{2}** (*)	**405d^{2}** (*)	**420d^{2}** (*)	**429D$^{2'}_{rz}$** (*)	**433d^{2}** (*)	**438d^{2}** (*)
439d^{2} (*)	**439d^{2}** (*)	**472d^{2}** (*)	**AZOS$^{3'}$**						

The meanings of the superscript and subscript are as follows. *P* signifies RNA–protein complexes. The numerical label 2,3,… indicates the number of chains. The symbol $'$ denotes a complex formed from different chains. *rz* stands for ribozyme structures, while *rs* denotes ribosomal structures. From RNA dimer complexes composed of identical chains (each having a non-zero number of intra-chain contacts) we selected only one subunit of the complex and specified this modification using a capital letter (for example 1GID, 1BY4^{2P}). This was done in order to prevent the introduction of spurious data in the statistical analysis. PDB codes with (*) mark are the non-intra-chain contact structures composed of multiple chains complex that are stabilized only by inter-chain contact. Codes in bold correspond to RNA structures that were selected from the PDB list in January 2003.

Table 2. Number of intra-chain stacks in 172 RNA structures from PDB

CG	GC	GU	UG	AU	UA	
619	850	149	111	273	296	CG
880	740	170	99	313	322	GC
164	226	13	6	66	10	GU
73	116	20	50	40	60	UG
322	357	56	27	106	95	AU
282	262	32	27	80	106	UA

The total number of bases in stacks is 7424.

been optimized to recognize any specific PDB structure. Nevertheless, if the SPs approximately reflect the true energy parameters in RNA structures, then a large percentage of all test sequences should pass the threading test, i.e. the native state should have the lowest free energy.

When we use the statistical parameters for stacking interactions in gapless threading, we find (Figure 1) that about 70% of the RNA secondary structures in our dataset are recognized correctly. This can be improved up to 75% if we include in the success list the predicted structures with backbone RMSD <2 Å.[19] It is remarkable that the prediction rate coincides perfectly with the percentage obtained when using the Turner parameters in gapless threading. This shows that the performance of the SPs is comparable to the well-known Turner interaction parameters for stacking.

To establish the significance of the results obtained when using either the SPs or Turner potentials, it is important to compare these results with predictions generated when using less accurate parameters sets. Mathews *et al.*[11] tested the efficacy of a parameter set (that counts only hydrogen bonds between the canonically base-paired regions) in recognizing the native base-pairs of a set of RNA chains. In this scheme, G–C pairs are assigned three hydrogen bonds while A–U or G–U pairs are given two hydrogen bonds. This parameter set, which acts as a control, allowed Mathews *et al.*[11] to establish that sequence-dependent estimates of thermodynamic interaction parameters greatly enhance the accuracy of RNA secondary structure prediction. Here, we use a different test, in which the stacking interaction parameters are randomly chosen from a Gaussian distribution with the means and variances corresponding to the values in the statistical potential. Because evolutionary pressure has resulted in the observed RNA sequences in nature, we expect that the randomly chosen stacking interaction parameters will not be able to recognize the native secondary structures accurately. Indeed, when used in gapless threading, only 9% of the native secondary structures are predicted correctly (Figure 1). Thus, as surmised by Mathews *et al.*, context-dependent interaction parameters are needed for reliable secondary structure prediction. Apparently, the SPs capture this aspect to the same degree as the Turner parameters.

Success of gapless threading depends on sequence length

We find that gapless threading is less successful for shorter RNA chains than for longer ones (Figure 2). If we restrict ourselves to the length of RNA sequence $L>30$, the percentage of sequences that passes threading increases from 69% to 84%. For a short sequence, it is likely that its native structure has only a few of the basic building blocks of RNA. As a result, a short sequence might be better accommodated by another structure that is part of the native state of a longer RNA. The 1*f*27

Table 3. Stacking energies (in kcal/mol) obtained from statistical analysis

CG	GC	GU	UG	AU	UA	
−2.97	**−3.40**	**−1.79**	**−1.11**	**−3.08**	**−3.04**	CG
(−2.40)	(−3.30)	(−2.10)	(−1.40)	(−2.10)	(−2.10)	
−3.40	**−3.20**	**−2.08**	**−1.31**	**−3.23**	**−3.07**	GC
(−3.30)	(−3.40)	(−2.50)	(−1.50)	(−2.20)	(−2.40)	
−1.78	**−2.08**	**0.82**	**0.82**	**−1.62**	**−0.27**	GU
(−2.10)	(−2.50)	(1.30)	(−0.50)	(−1.40)	(−1.30)	
−1.11	**−1.31**	**0.82**	**−0.88**	**−0.86**	**−1.19**	UG
(−1.40)	(−1.50)	(−0.50)	(0.30)	(−0.60)	(−1.00)	
−3.08	**−3.23**	**−1.62**	**−0.86**	**−2.80**	**−2.56**	AU
(−2.10)	(−2.20)	(−1.40)	(−0.60)	(−1.10)	(−0.90)	
−3.04	**−3.07**	**−0.27**	**−1.19**	**−2.56**	**−2.80**	UA
(−2.10)	(−2.40)	(−1.30)	(−1.00)	(−0.90)	(−1.30)	

The interaction free energies in bold were obtained using the data in Table 2 and equation (3) in Methods. The numbers in parentheses are the experimentally measured values at 37 °C using model oligonucleotides. The average value for the stacking free energy is −1.87 kcal/mol and the dispersion is 1.26 kcal/mol. The corresponding numbers for the Turner parameters are −1.59 kcal/mol and 0.98 kcal/mol, respectively.

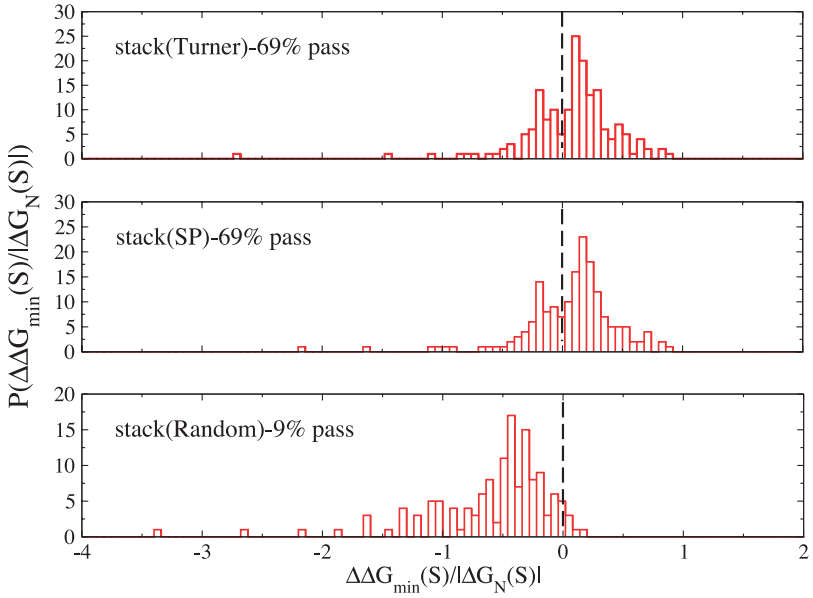

Figure 1. Distribution of relative energy gaps for the 172 chains in the dataset. The energy gap $\Delta\Delta G_{min}(S)$ of a test sequence S depends on the results of threading. If threading fails ($\Delta G_{min}(S, \Gamma_D) < \Delta G_N(S, \Gamma_N)$ for some D) $\Delta\Delta G_{min}(S) = \Delta G^{(-)}_{min}(S) - \Delta G_N(S, \Gamma_N)$ where $\Delta G^{(-)}_{min}(S)$ is the lowest free energy obtained when sequence S is mounted on fragment set $\{\Gamma_D\}$. If threading passes ($\Delta G_N(S, \Gamma_N) < \Delta G(S, \Gamma_D)$ for all D) $\Delta\Delta G_{min}(S) = \Delta G^{(+)}_{min}(S) - \Delta G_N(S, \Gamma_N)$ where $\Delta G^{(+)}_{min}$ is the free energy of the first excited state (Methods) from the native energy $\Delta G_N(S, \Gamma_N)$. Here, Γ_N refers to the structure of the native state of the sequence S and Γ_D is a decoy structure.

composed of 19 nucleotides is a typical example that illustrates this situation (Figure 3). Comparison of the native free energy $\Delta G_N(1f27, \Gamma_N)$ and $\Delta G(1f27; \{\Gamma_D\})$ shows that a structural fragment Γ^* satisfies $\Delta G(1f27,\Gamma^*) < G_N(1f27,\Gamma_N)$ computed using the SP parameter set. The free energy of Γ^* fragment from positions 2 to 20 from the long structure 1ffz is the lowest in gapless threading, i.e. $\Delta G(1f27, \Gamma^*) < \Delta G_N(1f27,\Gamma_D)$ for all D including the native state. The secondary structure map shows that the octaloop in the native state of 1*f*27 can further lower its free energy by forming a tetraloop with a bulge (Figure 3). The formation of an additional base-pair stabilizes 1*f*27. This particular example shows that tests of the performance of our SPs would be more reliable if they can be performed on datasets containing a larger number of longer chains. At present this drawback cannot be avoided, given the paucity of RNA tertiary structures for long chains.

The inability of gapless threading to predict the structure of the 19 nucleotide hairpin correctly might be because the loop region forms a double strand with another RNA strand in 1*f*27 (Figure 3). To rule out this possibility, we obtained the structure for this sequence using the ViennaRNA package, which includes energetic interactions arising from loops. The predicted structure coincides with that in the PDB file and suggests that one of the crucial assumptions in threading, namely, that the substructure of a short sequence is independent of the rest of the structure, may fail when L is small. This is because short chains have greater structural flexibility than when L is large. The finding that the ViennaRNA package predicts the structures correctly shows that loop interaction

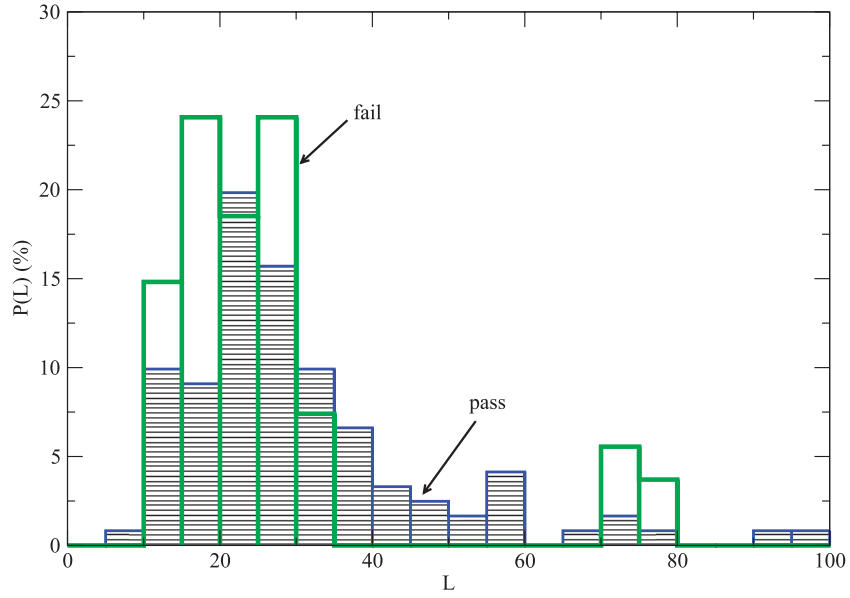

Figure 2. Histogram showing the results for recognition of native RNA structures using gapless threading as a function of sequence length. The results were obtained using the SP stacking parameters. The distribution in green represents structures that fail in gapless threading and those in blue correspond to successful recognition of the native state.

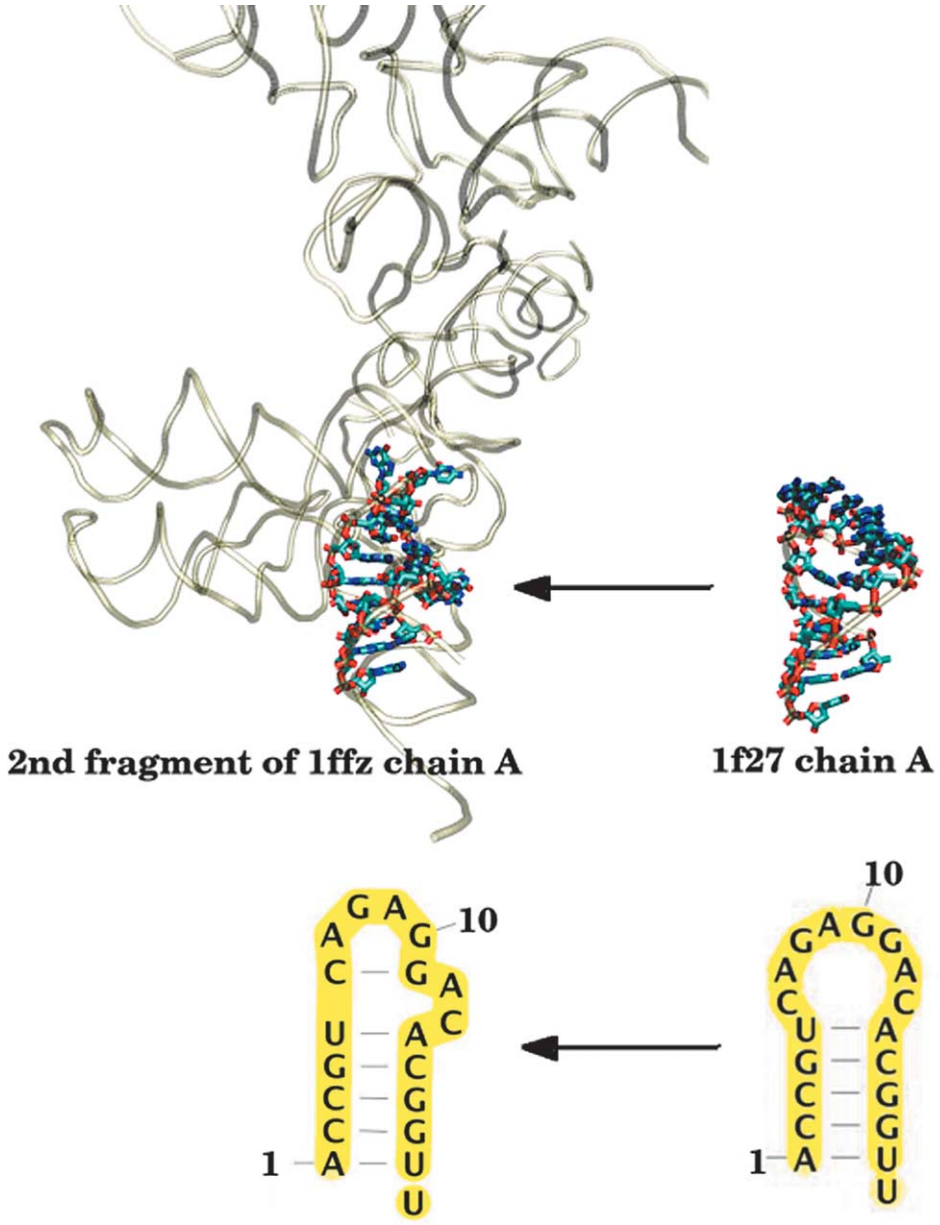

Figure 3. Illustration of the failure of gapless threading for a short sequence. The lowest free energy structure for 1*f*27 is not the native one (top right corner) but coincides with the second fragment of 1ffz (top left corner). The lowering of the free energy is achieved by gaining a 6C-11G interaction in the loop region (see the bottom of the Figure) that is not present in the native state of 1*f*27.

energies, which are not included in threading, rather than formation of a double strand with the rest of the RNA, leads to an incorrect structure in this case.

Experimental and numerical *Z*-scores

The success of the potentials in structure prediction is often assessed using *Z*-scores (see Methods). The distribution of Z_G, which depends on the sequence length, for the structures for which threading is successful shows $0<|Z_G|<60$ (Figure 4(a)). The relatively large Z_G values suggest that the parameters of the SP are fairly accurate. An advantage of using Z_G scores is that they can, in principle, be estimated using calorimetric measurements. If we ignore conformational fluctuations of RNA in the folded state, Z_G can be approximated as:

$$Z_G \cong \frac{\Delta G}{\sqrt{k_B T^2 C_P}} \quad (1)$$

where ΔG is the counterion-dependent free energy of stability of the native state and C_P is the specific heat. From equation (6) (Methods) it follows that:

$$\sigma(\Delta G(S,\Gamma)) = \sqrt{\langle \Delta G^2 \rangle - \langle \Delta G \rangle^2}$$

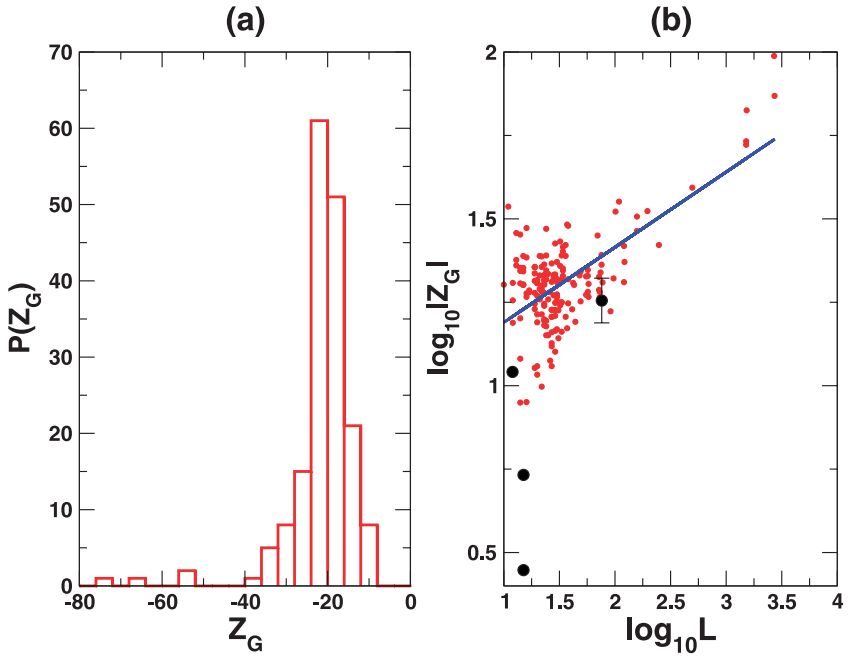

Figure 4. (a) Distribution of Z_G scores using gapless threading (equation (6) in Methods). We calculated Z_G using the stacking SP. (b) Plot of $|Z_G|$ as a function of sequence length L. A fit of Z_G to L^{α} yields $a = 0.23$ with a correlation coefficient of 0.65. Estimates of Z_G for selected RNA (Table 4) using experimental calorimetric data are shown as filled circles. The good agreement between calculated and experimental estimates is another measure of the quality of the SP.

which from fluctuation formula is given by:

$$\sqrt{k_B T^2 C_P}$$

The numerator in equation (6) is roughly given by ΔG, and hence equation (1) is an approximation to equation (6). The experimental estimate of Z_G includes contributions from both secondary and tertiary interactions that stabilize the native state. Our calculations include contributions only from secondary structure free energy parameters. Because these dominate in RNA, the comparison between computed experimental estimates of Z_G is appropriate. Given that ΔG and C_P are extensive variables, it follows that $|Z_G| \propto \sqrt{L}$ provided L, the sequence length, obeys $L \gg 1$. To verify that the predicted range is meaningful, we have estimated Z_G using equation (1) from calorimetric data for a few RNA molecules. We find that (see Table 4) for RNA with $L \leq 80$, Z_G is in the range $-6 < Z_G < -21$. These values are in accord with the computed numbers using the SPs (Figure 4(a)). The favorable comparison between the computed and experimentally estimated Z_G values lends further credence to the procedure used to obtain the stacking interaction parameters.

Although the dataset of RNA structures is small and not varied enough to obtain a reliable fit, we find that $|Z_G|$ increases with L, which implies that the SPs stabilize the native state of longer chains to a greater extent than those of shorter chains. The linear fit of the $\log_{10}|Z_G|$ *versus* $\log_{10}$L (Figure 4(b)) gives a slope of 0.23 (with a correlation coefficient of 0.65). There are two possible reasons for the deviation of the exponent α in $Z_G \propto L^{\alpha}$ from 0.5. (i) The range of L is not large enough. (ii) The composition of structural motifs in the dataset is very limited such that the decoy structures for the majority of the threaded RNA chains do not represent the full spectrum of their excited states. Nevertheless, the overall increase of Z_G with L is in accord with the prediction from equation (1).

Assessing accuracy

As a further test of the efficacy of the statistical potentials in predicting secondary structures we obtained, following Mathews *et al.*,[11] the total number of base-pairs predicted N_T^V and the number of correctly predicted base-pairs N_C^V. These quantities, for the SPs and the Turner parameters, were calculated using the readily available ViennaRNA package. Comparison of the results for all RNA structures and different RNA families (see Table 5) shows that the performance of the SPs is

Table 4. The estimates of Z_G from the experiment data

RNA	L (length)	$-\Delta H$ (kcal/mol)	T (K)	C_p (kcal/mol K^{-1})	Z_G	Reference
$tRNA^{PHE}$, $tRNA^{ASP}$	~76	~310	~300	$1.2 \leq C_p \leq 2.4$	$-21 \leq Z_G \leq -15$	34
Segment of ribosomal RNA						35
(i) 0.1 mM Mg^{2+} buffer	~<20	11.2	~300	~0.024	~−5.4	
(ii) 3 mM Mg^{2+} buffer	~<20	18.2	~300	~0.24	~−2.7	
GGXCGAAAGYCC (X,Y=C,G)	12	48.2	~300	~0.1	~−11.5	36

Table 5. Prediction of base-pairs using ViennaRNA package (170 structures)

RNA[a]				Turner			SP	
	N_{nucl}[b]	N_{nat}[c]	N_T^{Vd}	N_C^{Ve}	%[f]	N_T^V	N_C^V	%
[g]RNA–protein complex	2031	604	621	350	**57.9**	643	379	**62.7**
Ribosome	1893	558	580	343	**61.5**	598	375	**67.2**
Ribozyme	1894	549	607	342	62.3	614	334	60.8
Pseudoknot[h]	657	195	184	40	20.5	189	40	20.5
Simple RNA[i]	4298	1350	1384	1114	82.5	1412	1097	81.3
Total (neighbor, *i* with *j*, $j\pm1$)[j]	6641	2045	2075	1520	74.3	2127	1532	74.9
Total(*i* with *j*)	6641	2045	2075	1478	72.2	2127	1506	73.6

[a] The list of 170 structures used to produce the Table is available upon request. The largest chain used in the calculation is 1ffz ($L=495$). We removed structures with $L>700$ (1ffk chain 0, 1p9x chain 0, 1fjg chain A, 1i94 chain A, 1pns chain A) to produce results comparable to that in the Table 1 of Mathews *et al*.[11] If the number of native base-pairs in these structures is included, the numbers in N_{nat} in Table 6 are recovered.
[b] Total number of nucleotide in the dataset.
[c] Total number of native base-pairs.
[d] Total number of predicted base-pairs.
[e] Total number of correctly predicted native base-pairs.
[f] Ratio of e and c.
[g] There is an overlap between the structures in RNA–protein complex and Ribosome.
[h] In this category we included only the structures that are labeled as pseudoknots in the PDB header.
[i] These are structures that excluded pseudoknots, hammerhead ribozymes, and RNA–protein complex.
[j] Following Mathews *et al*.,[11] we considered a prediction to be successful if base *i* is paired with either *j* or $j\pm1$.

comparable to that of the Turner potentials. In particular, the percentage of the native base-pairs for SP is 74%, whereas for the Turner potential it is 72% (Table 5). Because of the differences in the set of query sequences used in the calculation, the value of 72% for the Turner's potential differs slightly from that reported in Table 1 by Mathews *et al*.[11]

From Table 5 it also follows that interaction parameters for secondary structures are not accurate enough to predict pseudoknots or even RNA structures that form complexes with proteins. We note that the SP parameter set is slightly better than the Turner potential in predicting RNA structures that are found in complexes with protein or in ribosome. The reason for the improvement may be due to the ability of SP to capture the effect of counterion-induced environment beyond the canonical Watson–Crick pairing. In the next section we show that this is indeed the case in the secondary structure prediction of the P5abc domain of group I intron.

We have also used suboptimal structures (generated by threading) to calculate N_{DG^0} and N_{MAX} (see Methods) using SP. We were motivated to compute these quantities because threading is an alternative rapid way to predict RNA secondary structures. In addition to the interaction parameters the efficacy of threading depends on the number and diversity of structures in PDB. By comparing N_{MAX} (more precisely N_{DG^0}) the confidence level of threading can be assessed. From Table 6, we learn that if the native structure is present in the data set, the prediction of native base-pairs is excellent. This is hardly surprising, because threading is successful in correctly predicting the secondary structure in nearly 70% of the cases (Figure 1). Moreover, the predictions from threading are best for long sequences (Figure 2), which dominate N_{MAX}.

A more stringent and meaningful test of the usefulness of threading is to compute N_{MAX} or N_{DG^0} by excluding the native structure from the data. When this is done we find (Table 6) that, on average, slightly less than 50% of the native base-pairs are predicted correctly. When only the family of "Simple RNA" structures are considered (Table 6) we find that threading succeeds in nearly 70% of the cases when the lowest free energy is picked. This percentage increases to 82% if the best suboptimal structure is chosen. These numbers are similar to those in Table 1 of Mathews *et al*.,[11] for group I intron which in our classification is a Ribozyme or Simple RNA.

Comparison between the predictions of threading and the secondary structure package has implications for obtaining secondary structures for a query sequence. If the predictions of threading and MFOLD (or ViennaRNA package) are similar, then in all likelihood the resulting structure is native-like. From Table 6 it follows that this scenario is most likely for simple RNAs as long as the RNA is neither too short (<20) or too great (at best 300). In light of these results, we propose that results obtained using a combination of the two methods (threading and dynamic programming) can increase the confidence of secondary structure prediction.

Specific applications

To further test the accuracy of the stacking free energy parameters in the SP we have predicted the secondary structures of a few selected sequences. Such applications are meaningful because the

Table 6. Base-pair prediction by gapless threading

	N_{nat}[a]	$N_{\Delta G^\circ,N}$[b] (%)	$N_{\Delta G^\circ}$[c] (%)	$N_{max,N}$[d] (%)	N_{MAX}[e] (%)	N_{struct}[f]
RNA–protein complex	2659	2648 (99.6)	945 (35.5)	2659 (100)	1087 (40.9)	51
Ribosome	3276	3267 (99.7)	948 (28.9)	3276 (100)	1102 (33.6)	38
Ribozyme	549	542 (98.7)	390 (71.0)	549 (100)	445 (81.1)	23
Simple RNA (<700)[g]	1350	1327 (98.2)	949 (70.3)	1350 (100)	1109 (82.1)	114
Total (<700)	2045	2011 (98.3)	1422 (69.5)	2045 (100)	1706 (83.4)	170
Total	4763	4729 (99.3)	1973 (41.4)	4763 (100)	2327 (48.9)	175

[a] Total number of native base-pairs.
[b] Total number of base-pairs in the lowest free energy structures including the native structure.
[c] Total number of base-pairs found in the lowest free energy structures excluding the native structure.
[d] Total number of base-pairs in the best suboptimal structure including the native structure.
[e] Total number of base-pairs found in the best suboptimal structure excluding the native structure.
[f] Number of structures in each family.
[g] Only simple RNA structures with $L<700$ are considered.

extracted parameters have not been optimized to recognize any specific structure. We substituted the stacking interaction parameters either with Turner's estimates[11] or with our SPs, while retaining the parameters for loops, bulges, and mismatches as in the ViennaRNA package. The predicted secondary structures from each parameter set are compared with the PDB structures.

Hammerhead ribozyme (1 rmn)

The 49 nucleotide hammerhead ribozyme is composed of two hairpin loops (Figure 5). The secondary structures predicted using both Turner potential and SP coincide with that from the crystal structure, and we found 57 similar examples out of 168 test sequences. The structures are usually simple without pseudoknot or mismatches and all the hydrogen bonds are found as expected from Watson–Crick pairs.

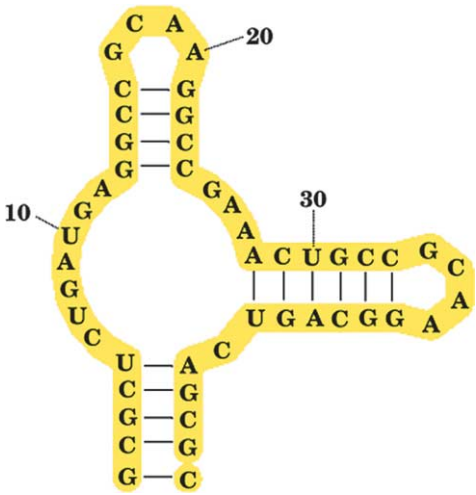

Figure 5. Predictions of secondary structure of hammerhead ribozyme (1rmn) using the SP. In this case the structure obtained using SP and the Turner parameters is identical and coincides with the structure in the PDB.

Ribonucleic acid (1 cql)

The secondary structure in the crystal structure is shown in Figure 6(a) while the predictions using the Turner's parameters and the SPs are shown in Figure 6(b) and (c), respectively. The predicted structures deviate from Figure 6(a). Both the interaction parameters predict base-pairing 7U-37A, which is absent from the crystal structure. With the SPs, the pairs 5U–39G and 6U–38G are predicted to be unbound, which contradicts the experimental finding (compare Figure 6(a) and (c)). This failure can be understood by noting that the stacking interaction involving UG and GU pairs is found to be slightly repulsive (Table 3). In this example, the use of Turner interaction parameters leads to a more accurate structure than the prediction using the SP. In particular, the Turner parameters predict correctly that 5U–39G and 6U–38G are paired (Figure 6(b)).

P5abc domain of Tetrahymena thermophila *group I intron ribozyme*

Wu & Tinoco[21] showed that the P5abc domain of the group I intron ribozyme locally changes its secondary structure upon folding to its native tertiary conformation upon addition of excess Mg^{2+}. The NMR secondary structure at low [Mg^{2+}] differs from that found in the crystal structure (Figure 7(a)) of the 56nt–P5abc domain. In the crystal structure of the P4–P6 domain, five Mg^{2+} are coordinated to the A-rich bulge, P5c–L5c stem–loop region and a three helix junction (rich in GA base-pairs).[20] It is likely that the formation of the unusual metal ion core triggers a local rearrangement of the secondary structure as Mg^{2+} concentration is increased.[21,22] If discrete Mg^{2+} play a crucial role in consolidating tertiary interactions by coordinating specifically with certain type of base-pairs then their role has to be taken into account if the crystal secondary structure is to be predicted correctly. The experimentally determined parameters are from studies of oligonucleotides in the absence of specific ion binding. As a result, we do not expect that the predictions using the Turner parameters to coincide with the structure in

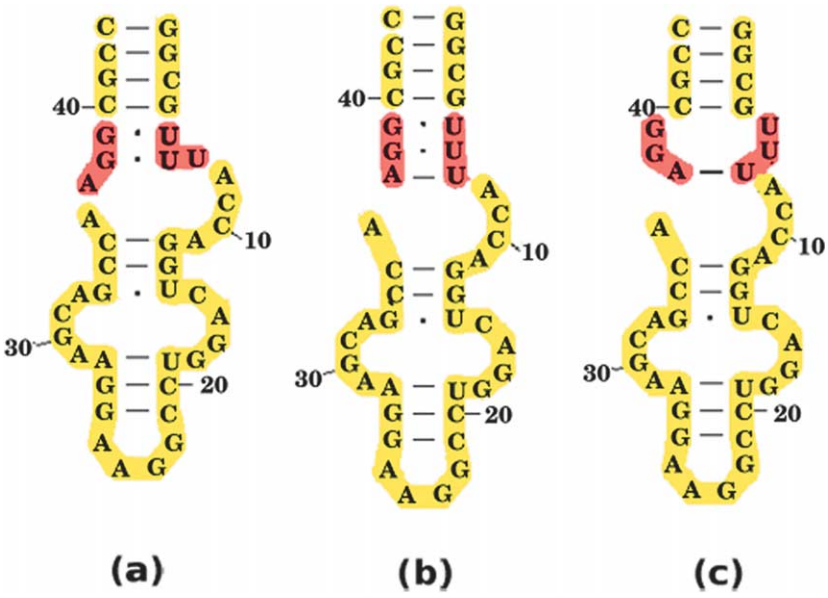

Figure 6. (a) Secondary structure of 1cql from the crystal state. The predictions of the structures using the Turner and SP parameters are given in (b) and (c), respectively. The region in which the predicted structures deviate from the native state are colored in red. Comparison of (b) and (c) with (a) shows that the Turner's parameters predict a more accurate structure than the SP.

Figure 7(a). This is indeed the case, as can be seen in Figure 7(b), which shows that the structure coincides with that obtained using NMR at low [Mg^{2+}].

The structure obtained using the SP shows (Figure 7(c)) that P5c-L5c region is predicted correctly, as in the crystal structure. Because the SP is obtained from native structures, the Mg^{2+}-dependent stacking interactions are approximately taken into account. It is interesting that the differences between these structures in the P5c–L5c region obtained with the two parameters are reminiscent of the secondary structure rearrangement in the transition from the solution structure to crystal structure of P5abc domain.[21] Although the predicted structure of the A-rich bulge region deviates from the crystal structure, it is encouraging that the SP can predict correctly local regions whose structures are influenced by specific metal ion binding.

The helix junction formed by three stacked regions is not predicted correctly either by Turner or by SP parameters. This is mainly due to the absence of an appropriate interaction parameter

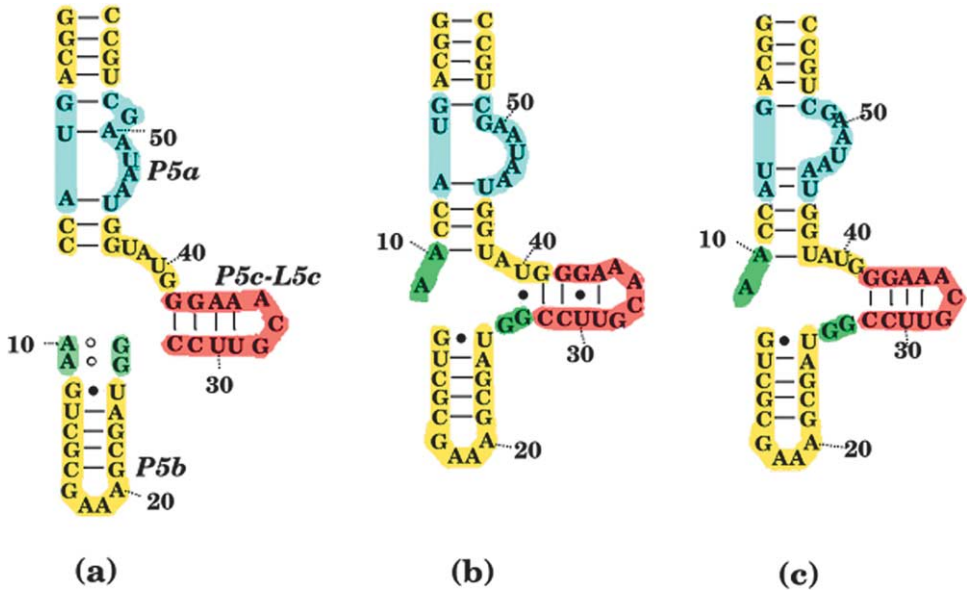

Figure 7. (a) The crystal secondary structure of P5abc domain in *Tetrahymena thermophila* group I intron ribozyme. Structures with minimum free energy produced by the ViennaRNA package using Turner parameters and the SP are displayed in (b) and (c), respectively. The structure corresponding to the statistical stacking potentials ((c)) coincides with the native structure in the P5c–L5c region (compare the red parts in (a) and (c)). In the same region the predictions using the Turner parameters deviate from the structure in (a). The structure in (b) in the P5c–L5c region resembles that seen at low [Mg^{2+}]. Neither potential predicts the GA pairs (shown in green) in the three helix junction region correctly.

for GA mismatch pairs. To assess if the prediction near the three helix junction can be improved by eliminating GA base-pairs, we replaced 10A and 11A by U. The prediction for secondary structure using the ViennaRNA package shows that 11U–26G is paired as in Figure 7(a). However, the expected pairing 10U–27G is not predicted correctly.

tRNA

As a final example we used the SP and Turner parameters to predict the secondary structures of tRNA. For the SP, this test is meaningful because the tRNA structures were not included in our dataset (see the caveat in the caption to Figure 8). We replaced the modified nucleotides such as D, Ψ, and U_m with U. In this case the use of from SP and Turner parameters yields identical secondary structure (Figure 8(b)) that coincides with the crystal structure (Figure 8(a)).

Discussion

Hierarchical nature of RNA folding and success of statistical potentials

Taking advantage of the steadily growing number of structures for RNA molecules in the PDB in the last five years we used one of the simplest, but quite powerful, methods to obtain reliable estimates of stacking interaction parameters. Surprisingly, our estimates of the stacking interaction parameters compare well with Turner's values, which are obtained using thermodynamic data in short oligonucleotides. The success of our estimates for the stacking energies in predicting secondary structures using gapless threading is comparable to that obtained with Turner's values in similar tests. Even more interestingly, our potentials are able to capture some of the influence of tertiary fold formation on the secondary structure as seen in the secondary structure prediction of P5abc (Figure 7).

The set of stacking interaction potentials performs extremely well in threading tests, as they are able to recognize the real native state conformation of 69% of the chains in our dataset. If we restrict ourselves to the longer chain with $L>30$ then the percentage of folds recognized increases. The value of 69% (or 84% for $L>30$) is best appreciated if we compare it with the performance of SPs in proteins. For example, the Miyazawa–Jernigan pairwise potentials[16] lead to approximately 83% success in gapless threading. When secondary structure information is included in these potentials the success rate in gapless threading native state recognition increases.[23] Our procedure in RNA potential is similar to this latest approach in proteins and, as a result, this can serve as an explanation of the rate of success of our statistical potentials in predicting RNA folds. In general, it is easier to predict secondary structures that are independently stable, as in RNA, than those whose stability is dictated by tertiary interactions, as in proteins.

Accuracy in estimates of stacking interaction parameters improves with the growth in database

The knowledge-based approach used to

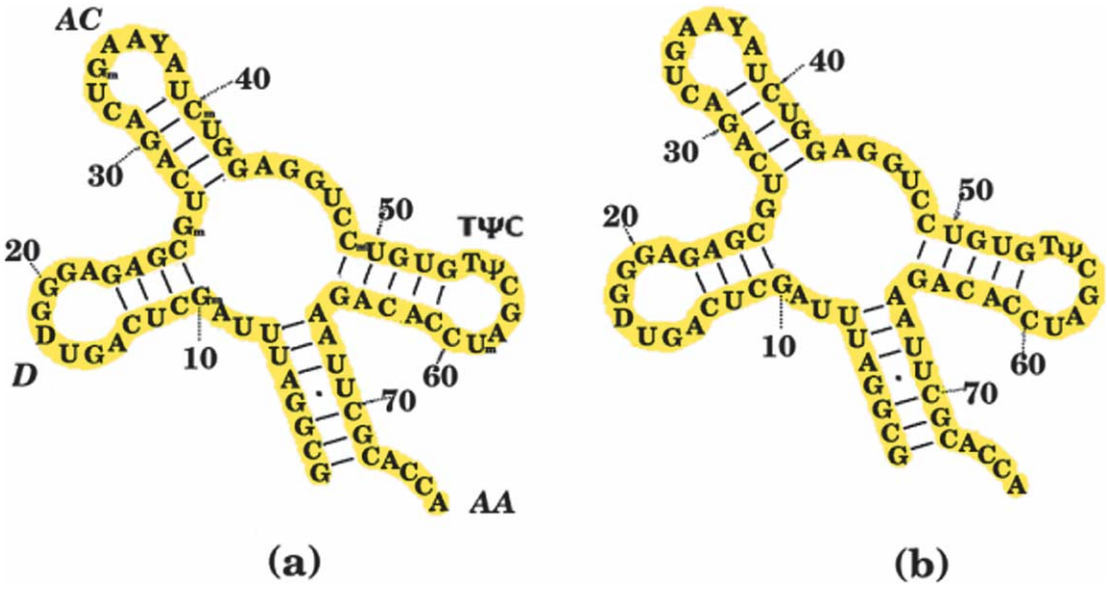

Figure 8. The identical prediction of the secondary structure of tRNA using the SP and the Turner parameters coincide with that in the crystal structure (compare (a) and (b)). This is illuminating because tRNA structures were not used to obtain the SP parameters. However, we should note that structures of certain regions of the ribosome and tRNA are identical. In particular, A-tRNA and P-tRNA are embedded in the small subunit of the 70 S ribosome (see Figure 1(b) of Vila-Sanjurjo *et al.*[32]).

Table 7. Stacking energies (in kcal/mol) obtained from statistical analysis of 109 structures

CG	GCI	GU	UG	AU	UA	
(111)[a]	(144)	(35)	(10)	(96)	(80)	
−2.80[b]	−3.33	−1.65	−0.63	−3.40	−3.40	CG
(189)	(156)	(45)	(13)	(81)	(82)	
−3.33	−3.25	−1.90	−0.68	−3.31	−3.37	GC
(29)	(33)	(5)	(5)	(20)	(3)	
−1.65	−1.91	0.29	0.75	−1.71	−0.21	GU
(19)	(17)	(2)	(12)	(5)	(8)	
−0.63	−0.68	0.75	−0.83	−1.10	−1.23	UG
(72)	(76)	(12)	(15)	(19)	(29)	
−3.40	−3.31	−1.71	−1.10	−2.44	−3.15	AU
(88)	(82)	(7)	(12)	(20)	(22)	
−3.40	−3.37	−0.21	−1.23	−3.15	−2.98	UA

List of structures used from January 2003 release of PDB are given in bold in Table 1.
[a] The numbers in parentheses represent the number of intra-chain stacks for the 109 structures.
[b] Stacking interaction energies using equation (3) in Methods.

extract interaction parameters is not without limitations.[24,25] The assumption that the ensemble of interactions observed in the PDB structures is in "equilibrium" may not be accurate. We believe that this problem is not critical for RNA, in which the secondary structures are inherently stable. As long as the number of contacts or base-pairs of a given type is sufficiently large, reliable estimates of the stacking and presumably tertiary interactions can be made. To validate this assertion we obtained the stacking parameters using the dataset of 109 structures that were available in January 2003 in PDB. Comparison of the resulting interaction parameters (Table 3) and Table 7 shows that as the number and variety of RNA structures grows there is closer agreement between the SP values and the experimentally determined thermodynamic parameters. Thus, the quality of the SP can be greatly improved by increasing the number and diversity of structures in the dataset. However, parameters for the rare base-pair stacking interactions (like GU-UG) cannot be estimated reliably using the available PDB structures. Even with the increase in the available RNA structures in PDB from January 2003 until present, the number of UG–GU stacking interactions increases by only 1. For these rare, but important interactions, experimental estimates are needed. Nevertheless, it is encouraging that for the more preponderant set of interactions the estimates of the SP appear to converge to the experimental values as the database of structures increases and diversifies.

Stacks that end in hairpins are most easily recognized

It is clear that when reliable values for the base-stacking interactions are available then the currently available algorithms can predict RNA secondary structures fairly accurately provided pseudoknots are excluded. This is illustrated in the present study by the accurate predictions of tRNA and hammerhead ribozyme structures. More generally, we found that stacks that end in hairpins are easiest to predict. To illustrate this, we computed the frequency of each native base-pair recognized by structural fragments other than the native state during threading. In Figure 9 we represent the distribution of frequencies, in the dataset of suboptimal structures, of each of the native base-pairs in 1GID, 1NBS and AZOS. To understand the significance of these distributions, we also calculated the histogram of frequencies based on random considerations. If the number of native base-pairs recognized by a fragment Γ_i is n_i ($i=1,2,\ldots,N_\Gamma$), the random probability p_R of finding the particular base-pair in the set of fragments is $p_R = \sum_{i=1,2,\ldots,N_\Gamma} n_i/N_{BP}^{NAT}N_\Gamma$. This quantity, which is independent of the base-pair type or index, is represented by the green line in Figure 9. In contrast to p_R, Figure 9 shows that certain clusters of base-pairs are formed at values that are much larger than the random value (i.e. are found above the green line), which implies that sequence and structural context determine the interaction parameters. We mapped these clusters to the corresponding secondary structures and discovered that they all correspond to base-pairs that form stacks ending in hairpins (indicated as blue in the secondary structure of Figure 9). This finding indicates that secondary structure prediction procedures (either threading or dynamic programming algorithms) can easily recognize these structural motifs. Interestingly, the number of peaks in Figure 9 is identical with the number of stacks ending in hairpins. The peak heights and the width depend on the structure, which implies that sequence context plays an important role in determining the local structure. Regions that occur with low probability, like those from base-pair index 10 to 14 in AZOS corresponding to pseudoknot structures, are very difficult to predict. The results in Figure 9 can also be used as a measure of the specificity of the RNA fold: the large probability regions are generic, while the low-frequency regions identify the characteristic architecture of the fold. As an example, in 1NBS the middle region consisting of two internal multiloops is the distinct structural element corresponding to the low-frequency portion of the histogram. In this example, accurate prediction of

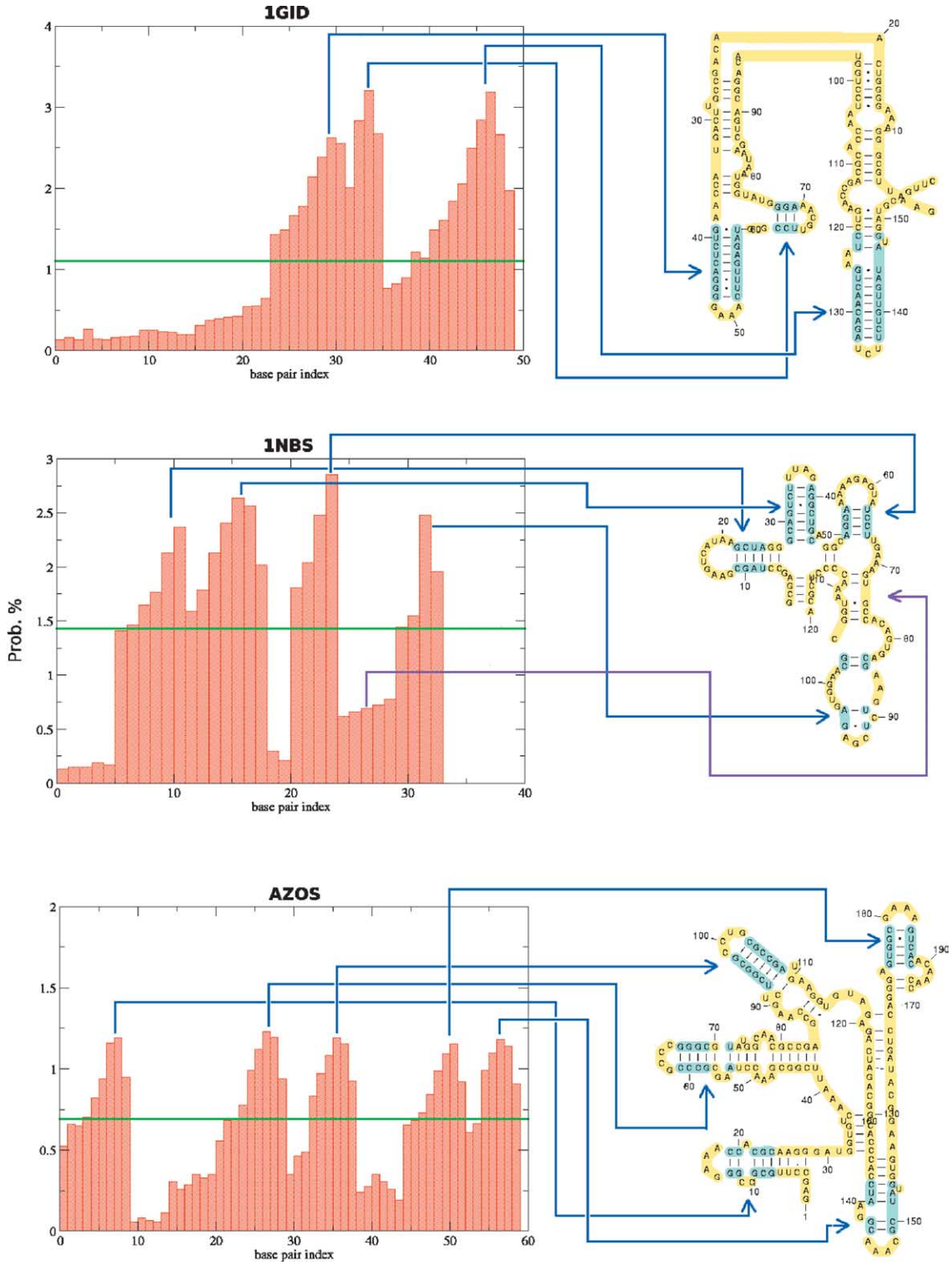

Figure 9. Probability distribution of occurrence of specific base-pairs in three representative structures. The distribution for each structure was constructed from an ensemble of structures that is generated using gapless threading. The ensemble includes only suboptimal structures and excludes the native states. For 1GID and 1NBS the native structures are in the PDB and for AZOS we use the model structure reported by Rangan *et al*.[33] The three parts of the Figure show that the base-pairs with high probability precisely map onto (see arrows) the most easily predicted regions of the secondary structures. They correspond to stacks (blue region of the structures) that end in hairpins.

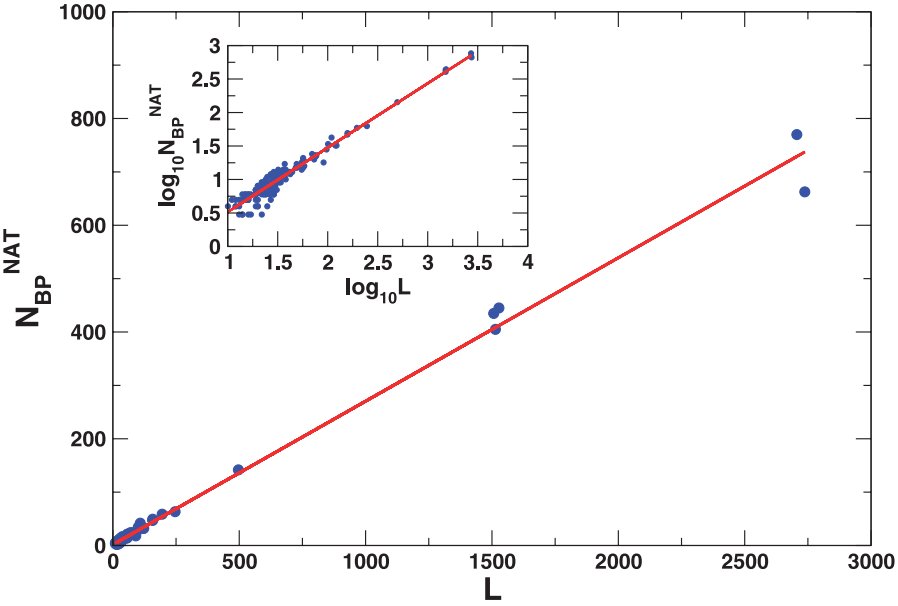

Figure 10. The dependence of the number of base-pairs in the native state of a given structure as a function of sequence length, L. The red line is a linear fit to the data, i.e. $N_{BP}^{NAT} = 0.27L$ with a correlation coefficient of 1.0. The inset, in which $\log_{10} N_{BP}^{NAT}$ as a function of $\log_{10} L$ is plotted, shows that the largest deviation from linearity occurs at small L values. Using the relation $N_{BP}^{NAT} \approx (L - x)/2$ (x is the number of nucleotides that are not base-paired) we estimate that for $L > 30$ the average percentage of a sequence in non-base-paired region (bulges, loops, dangling ends, etc.) is 46%.

the structure requires interaction parameters that can recognize these low-frequency regions.

Inherent difficulties in accurate secondary structure prediction

One of the greatest difficulties in increasing the accuracy of the secondary structure prediction is that a substantial fraction of nucleotides do not engage in base-pairing. Using the data set of structures we found that the number of base-pairs N grows linearly with L the sequence length (Figure 10). The linear growth of N with L (with a slope of less than 0.5, which is the expected value if all the nucleotides are engaged in Watson–Crick base-pairing) is satisfied with great accuracy. However, the slope is only 0.27, which implies that only 54% of a RNA chain is engaged in base-pairing. The remaining 46%, which is a large percentage of nucleotides, remain as single strands forming bulges, loops, or dangling ends. The free energy of forming such motifs is difficult to determine experimentally. Substantial improvements in secondary structure prediction can be achieved only if the free energy cost of formation of the single-stranded region can be calculated or measured.

Conclusions

Even though there are intrinsic limitations of the knowledge-based approach for obtaining SPs, our results show that the major drawback in our procedure comes from the limited number and variety of RNA structures corresponding to short sequences (~40 nucleotides long). The most serious obstacle in improving the quality of the statistical potential is the lack of structures of long sequences (100 nucleotides and more) and for sequences with varied composition. We are confident that an improvement of the quality and size of RNA tertiary structure set can produce better potentials. Considering the excellent performance of statistical stacking potentials, we believe that with increasing number of structures the quality of extracted interaction free energies will continue to improve. As the database of structures involving RNA continues to grow it will be useful to see how interaction parameters extracted from various complexes (RNA–proteins, RNA–DNA structures) vary. These calculation will give insights into the way context modifies the interaction of RNA with other biomolecules.

Methods

Extraction of interaction energies between nucleotides

Following the pioneering work by Tanaka & Scheraga,[15] we use a "knowledge-based" statistical approach to extract interaction potentials for nucleotides from RNA structures. For clarity we describe the strategy for extracting pair interaction potentials. The computation of the statistical potentials (SP) involves the following steps: (1) let N_{ij} be the number of pairs of nucleotides of type i and j that are in contact in the given 3D structures. (2) The

probability of finding the (*ij*) pair in contact is $P_{ij} = N_{ij}/\sum_{i,j} N_{ij}$. (3) The probability that such a pair would be in contact by random chance is $P_R = P_i P_j$, where P_i is the probability of appearance of nucleotide i (i=A, C, G, U) in the ensemble of sequences. The state representing the random occurrence of nucleotides is taken to be the reference state. (4) If we assume that Boltzmann statistics apply to the ensemble of contacts in RNA (both at the secondary and tertiary structure levels), then the interaction "free energy" is:

$$\Delta G_{ij} = -\lambda k_B T \ln\left(\frac{P_{ij}}{P_R}\right) \quad (2)$$

where λ is a scale factor, k_B is the Boltzmann's constant, and T is the temperature. The interacting pairs are taken to be at "equilibrium" at an effective temperature λT. Despite obvious problems with this approach,[24,26,27] the knowledge-based method has had moderate success in estimating interaction energies in proteins.[15–17,28–30]

Estimates of stacking energies from RNA structures

Given that sequence context determines the free energy (more precisely the potential of mean force) value[9–11] raises the question of whether the free energy parameters, even at the secondary structure level, vary if tertiary interactions are explicitly taken into account. The growing database of RNA structures permits an evaluation of this issue.

One of the major difficulties in using experimental and computational methods to unambiguously extract thermodynamic parameters is the presence of a variety of secondary structure elements (stacks, bulges, internal loops, internal multiloops, and hairpins). It is generally assumed that the free energy of a specific RNA secondary structure can be written as the sum of the various elements (therefore neglecting any cross-interaction contributions). The stacking of consecutive hydrogen bonded pairs of nucleotides makes the largest contribution to the stabilization of secondary structure. For this reason, we represent the secondary structure contribution to the free-energy of the RNA structure, ΔG, only as the sum of the stacking interaction energies.

We obtain the stacking interaction energies directly from the RNA structures deposited in the PDB[14] using a modification of equation (2). We identify the pairs of nucleotides that are hydrogen bonded in an RNA structure. Two nucleotides are hydrogen bonded if at least one of their possible pairs of donor-acceptor heavy atoms is within a cut-off distance of 4 Å. For example, we identify a Watson–Crick (WC) A–U pair if N6 from A and O4 from U or N1 from A and N3 from U are within 4 Å of each other. The 4 Å cut-off, which is larger than the typical 2.8–3.0 Å distance in an ideal hydrogen bonded base-pair, is chosen to account for possible non-ideal bonds and possible imperfections in the RNA structures due to the X-ray resolution or to the NMR NOESY signals. In the list of hydrogen bonded pairs we include all the WC and reverse WC, Hoogsteen and reverse Hoogsteen, wobble and reverse wobble pairs for A–U, U–A, C–G, G–C, G–U, U–G. Based on the list of hydrogen bonded pairs, we determine the stacks as pairs of consecutive base-pairs or base-pairs separated by a one nucleotide long bulge. Using this procedure on all RNA structures in our dataset we calculate the number of stacks of each of the 36 possible types, $N_{stack}(ij, kl)$. We also calculate the frequency of appearance of each type of nucleotide in the dataset of RNA sequences, P_i (i={A,C,G,U}). In analogy with equation (2), the statistical free energy for each type of stack is:

$$\Delta G_{ST}(ij,kl) = -\lambda k_B T \ln\left(\frac{P_{ST}(ij,kl)}{P_{ST}^{(rand)}(ij,kl)}\right) \quad (3)$$

where:

$$P_{ST}(ij,kl) = \frac{N_{ST}(ij,kl) + N_{ST}(kl,ij)}{N_{ST}} \quad (4)$$

and:

$$P_{ST}^{(rand)}(ij,kl) = (2 - \delta_{ij,kl}) P_i P_j P_k P_l \quad (5)$$

and i,j,k,l=A,C,G,U. Following Turner,[10] we assumed that the matrix ΔG_{ST} (ij, kl) is symmetric, i.e. $\Delta G_{ST}(ij, kl) = \Delta G_{ST}$ (kl,ij). Because in obtaining the above quantities we use RNA structures determined under a variety of conditions (X-ray, NMR, presence or absence of ions, pH and temperature), it is difficult to obtain the scale factor λ (equation (3)). Since this is just an overall multiplicative factor, which does not affect the distribution of values of stacking energies, we relied on the Turner rules for estimating λ. We equated the largest negative value from our list of ΔG_{ST} (ij,kl) values with the largest negative value from the Turner set, $\Delta G_{ST}^{(Turner)}(ij,kl)$ and we assigned $\lambda = \min\{\Delta G_{ST}^{(Turner)}(ij,kl)\}/\min\{\Delta G_{ST}(pq,rs)\}$.

Threading of RNA chains: *Z*-scores

The use of the known crystal or NMR structures as folding templates in the context of homology modeling is often used to determine structures of proteins and RNA. The alignment of a probe sequence into a given structure is called threading.[28,31] To measure the degree of fitness between the threaded sequence and the structure an energy function must be used. The key principles of threading are (i) the number of basic folds is limited, and (ii) that the nucleotide preferences for different structural motifs provide sufficient information to choose among folds. In gapless threading a sequence of length L is threaded through all the fragments of the same length from a structure of length L_{str} (with the condition $L \le L_{str}$). For each of the resulting ($L_{str} - L + 1$) structural fragments, we evaluate the contacts using the original sequence and then the sequence to be threaded is mounted on

it and the interactions are scored based on the nucleotides from the latter sequence. Then the scores from the various fragments are compared and the fragment with the best score is selected as the most likely candidate for the native structure of the threaded sequence because it places the nucleotides into preferred structural environments and near other preferred nucleotide types.

To test the performance of the statistically determined potentials in recognizing the native state of RNA chains, we resorted to gapless threading. We mounted each chain in our dataset on each fragment of equal length taken from RNA structures that are at least as long as the test chain. For each fragment, Γ, we first produced the list of stacks and two-body contacts based on its own sequence. We calculated $\Delta G\,(S, \Gamma_D)$ of the test chain, S, on the fragment Γ_D where $\Delta G = \sum \Delta G_{ST}(ij, kl)$. Here $\Delta G_{ST}\ (ij,\ kl) = \Delta G_{ST}\ (S(p_1(\Gamma))S(p_2(\Gamma)),\ S(p_3(\Gamma))S(p_4(\Gamma)))$ with $S(l)$ being the nucleotide at position l in chain S and $p_k(\Gamma)$ $(k=1,2,3,4)$ is the position along the structure Γ for nucleotide number k from the given stack. Threading is considered to be successful if $\Delta G_N(S, \Gamma_N)$, corresponding to mounting a sequence on its native structure Γ_N, is the lowest among all ΔG values obtained for the sequence. If threading is successful, we computed the Z-score, which is:

$$Z_G(S) = \frac{\Delta G_N(S, \Gamma_N) - \langle \Delta G(S, \Gamma) \rangle}{\sigma(\Delta G(S, \Gamma))} \qquad (6)$$

where $\langle \Delta G_R(S) \rangle$ and $\sigma(\Delta G_R(S))$ are the average and the corresponding standard deviation of the ΔG values obtained when mounting the sequence S on all possible structures Γ excluding Γ_N. Given that $Z_G(S)$ measures the free energy separation between the native state and all other modified structures, it follows that a large negative $Z_G(S)$ value implies enhanced native state stability. Because Z_G scores can be estimated from experiments it is a useful measure for assessing the efficacy of the statistical potentials for RNA.

Extraction of number of base-pairs using threading of RNA chains

In addition to the number of native base-pairs N_{nat} from the suboptimal or decoy structures that are generated as threading fragments $\{\Gamma_D\}$ we also determined the number of native base-pairs in the lowest free energy structure among the threading fragments for each RNA sequence. We also computed the total number of native base-pairs found in the fragments that has the largest number of base-pairs in common with the native structure, N_{MAX}. The values of these and similar quantities have been used by Mathews *et al*.[11] to assess the accuracy of measured thermodynamics parameters. A note of caution is in order regarding the meaning of these two quantities. Threading is a tool used in structure prediction, i.e. when one wants to establish what is the most likely conformation adopted by a chain based solely on the sequence and the available structures. Therefore, threading is needed only when the native structure of an RNA sequence is not known. With this in mind, we report two numbers: one obtained when including the native structure of the RNA chain among the threading fragments and the second when the native structure is excluded from the threading set. Because it is desirable to have a procedure that discriminates between the native state and the decoy structures, the second set of numbers is more physically relevant.

Acknowledgements

This work was supported, in part, by a grant (CHE02-09340) from the National Science Foundation.

References

1. Doudna, J. & Cech, T. (2002). The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
2. Piccirilli, J., McConnell, T., Zaug, A., Noller, H. & Cech, T. (1992). Aminoacyl esterase activity of the *Tetrahymena* ribozyme. *Science*, **256**, 1420–1424.
3. Giege, R., Frugier, M. & Rudinger, J. (1998). tRNA mimics. *Curr. Opin. Struct. Biol.* **8**, 286–293.
4. Tinoco, I., Jr & Bustamante, C. (1999). How RNA folds. *J. Mol. Biol.* **293**, 271–281.
5. Zuker, M. & Stiegler, P. (1981). Optimal computer folding of larger RNA sequences using thermodynamics and auxilliary information. *Nucl. Acids Res.* **9**, 133–148.
6. McCaskill, J. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
7. Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188.
8. Hofacker, I. (2003). Vienna RNA secondary structure server. *Nucl. Acids. Res.* **31**, 3429–3431.
9. Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H. & Muller, P. (1994). Coaxial stacking of helices enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
10. Zuker, M., Mathews, D. & Turner, D. (1999). *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology*, Kluwer Academic Publishers, Dordecht.
11. Mathews, D., Sabina, J., Zuker, M. & Turner, D. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
12. Rivas, E. & Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285**, 2053–2068.
13. Orland, H. & Zee, A. (2002). RNA folding large N matrix theory. *Nucl. Phys. B*, **620**, 456–476.

14. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Databank. *Nucl. Acids Res.* **28**, 235–242.
15. Tanaka, S. & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three dimensional structures of proteins. *Macromolecules*, **9**, 945–950.
16. Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.
17. Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **213**, 859–883.
18. Skolnick, J., Jaroszewksi, L., Kolinski, A. & Godzik, A. (1997). Derivation and testing of pair potentials for protein folding When is the quasichemical approximation correct? *Protein Sci.* **6**, 676–688.
19. Dima, R., Settanni, G., Micheletti, C., Banavar, J. & Maritan, A. (2000). Extraction of interaction potentials between amino acids and between amino acids and solvent molecules from native protein structures. *J. Chem.Phys.* **112**, 9151–9163.
20. Cate, J., Gooding, A., Podell, E., Zhou, K., Golden, B., Kundrot, C. *et al*. (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, **273**, 1678–1685.
21. Wu, M. & Tinoco, I., Jr (1998). RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci. USA*, **95**, 11555–11560.
22. Thirumalai, D. (1998). Native secondary structure formation in RNA may be a slave to tertiary folding. *Proc. Natl Acad. Sci. USA*, **95**, 11506–11508.
23. Miyazawa, S. & Jernigan, R. (1999). An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins: Struct. Funct. Genet.* **36**, 357–369.
24. Thomas, D. P. & Dill, K. A. (1996). An iterative method for extracting energy-like quantities from protein structures. *J. Mol. Biol.* **257**, 457–469.
25. Betancourt, M. & Thirumalai, D. (1999). Pair potentials for protein folding: choice for reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **8**, 361–369.
26. Godzik, A., Kolinski, A. & Skolnick, J. (1995). Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* **4**, 2107–2117.
27. Dima, R., Banavar, J. & Maritan, A. (2000). Scoring functions in protein folding and design. *Protein Sci.* **9**, 812–819.
28. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
29. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–89.
30. Kolinski, A., Godzik, A. & Skolnick, J. (1993). A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: application to designed helical proteins. *J. Chem. Phys.* **98**, 7420–7433.
31. Bowie, J., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
32. Vila-Sanjurjo, A., Ridgeway, W., Seymaner, V., Zhang, W., Santoso, S., Yu, K. & Cate, J. H. D. (2003). X-ray crystal structures of the WT and a hyper-accurate ribosome from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, 8682–8687.
33. Rangan, P., Masquida, B., Westhof, E. & Woodson, S. A. (2003). Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proc. Natl Acad. Sci. USA*, **100**, 1574–1579.
34. Privalov, P. L. & Filimonov, V. V. (1978). Thermodynamic analysis of transfer RNA unfolding. *J. Mol. Biol.* **122**, 447–464.
35. Laing, L. G. & Draper, D. E. (1994). Thermodynamics of RNA folding in a conserved ribosomal RNA domain. *J. Mol. Biol.* **237**, 560–576.
36. Horton, T. E., Maderia, M. & DeRose, V. J. (2000). Impact of phosphorothioate substitutions on the thermodynamic stability of an RNA GAAA tetraloop: an unexpected stabilization. *Biochemistry,* **39**, 8201–8207.

Edited by B. Honig

(Received 14 September 2004; received in revised form 1 November 2004; accepted 6 December 2004)