2210–2214 *Nucleic Acids Research, 2005, Vol. 33, No. 7*
doi:10.1093/nar/gki508

Asymmetry in RNA pseudoknots: observation and theory

Daniel P. Aalberts* and Nathan O. Hodas

Physics Department, Williams College, Williamstown, MA 01267, USA

Received January 16, 2005; Revised March 5, 2005; Accepted March 25, 2005

ABSTRACT

RNA can fold into a topological structure called a pseudoknot, composed of non-nested double-stranded stems connected by single-stranded loops. Our examination of the PseudoBase database of pseudoknotted RNA structures reveals asymmetries in the stem and loop lengths and provocative composition differences between the loops. By taking into account differences between major and minor grooves of the RNA double helix, we explain much of the asymmetry with a simple polymer physics model and statistical mechanical theory, with only one adjustable parameter.

INTRODUCTION

Accurately predicting how biological macromolecules fold is one of the great challenges of our day because 'structure is function'. Encoded in the primary amino acid sequence of proteins are α-helix and β-sheet secondary structures which assemble into final native folds through additional tertiary contacts. The protein folding problem is notoriously difficult because local secondary and non-local tertiary contacts both contribute significantly to the stability of the final fold. In RNA, however, because base-pairing interactions are stronger and more specific typically than tertiary contacts, it is secondary structure which most influences the final fold.

Listing which bases are paired to which other bases uniquely describes the secondary structure of RNA. Base pairs can be annotated with left and right parenthesis pairs; blocks of base pairs, with a letter. In the vast majority of cases, RNA adopts 'nested' secondary structures composed of consecutive helices separated by bulges or by hairpin turns, such as the AABB (((()))))[[[]]] or the ABBA ((([[[[]]]]))) base-pairing patterns. Folding algorithms like MFOLD (1,2) or VIENNARNA (3) restrict themselves to nested structures to benefit from the algorithmic efficiency of dynamic programming. These algorithms ignore the more unusual non-nested structures of pseudoknot folds, such as the ABAB ((([[[[]))))]]]] pattern, depicted in Figure 1.

Pseudoknots have attracted attention as important functional structures of viruses and auto-catalytic RNAs. This class of structures is more highly constrained by non-local base pairs and exhibits particular 3D geometries.

The general pseudoknot problem has been proven to be NP-complete (4) because of non-local contacts. A number of pseudoknot algorithms have been developed recently (4–13), which search for only a subset of pseudoknot structures (14,15). Algorithms which lack such basic biochemical elements as GU wobble base pairs or basic polymer theory are of questionable value. Furthermore, none of these approaches are tested against the ensemble of known pseudoknots.

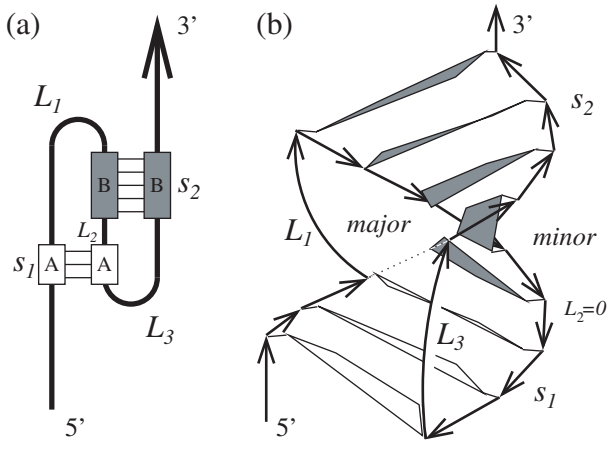

Figure 1. (**a**) An ABAB-pseudoknot is depicted in planar representation. The structure is composed of two double-helical stems (with s_1 and s_2 base pairs) and the three single-stranded loops of lengths L_1, L_2 and L_3 nucleotides. (**b**) The 3D fold of the same knot is depicted. The x, y, z axes point left, out, up. Coaxial stacking interactions between stems 1 and 2 can stabilize the structure, particularly if $L_2 = 0$. Note that loop 1 lies on the major groove side, while loop 3 lies on the minor groove side.

*To whom correspondence should be addressed. Tel: +1 413 597 3520; Fax: +1 413 597 4116; Email: aalberts@williams.edu
Present address:
Nathan O. Hodas, Physics Department, California Institute of Technology, Pasadena, CA 91125, USA

In the following section, we begin by describing the statistics of the pseudoknots in PseudoBase (16). To our knowledge, Ref. (16) is the only online database (http://wwwbio.leidenuniv.nl/~batenburg/pkb.html) focused on pseudoknots. The statistics illuminate key physical characteristics of pseudoknots: (i) the simplest pseudoknots are the most abundant, (ii) these pseudoknots have asymmetric loop and stem lengths and (iii) their loop compositions differ. The asymmetries in the ensemble of pseudoknots have not been characterized previously.

To self-consistently explain the source of these asymmetries, we proceed to develop a polymer physics model and statistical mechanical theory in Section 3. We argue that including the asymmetry of the major versus the minor groove is essential.

CHARACTERIZING PSEUDOBASE AND PSEUDOKNOT ASYMMETRY

PseudoBase is a gold mine of information, allowing us to dig deeply into the properties of pseudoknots. As of January 2005, there are 245 pseudoknots in PseudoBase. After removing duplicate sequences (PKB6 and 9, 25 and 26 and 29, 39 and 40 and 41, 19 and 27, 33 and 34), there are 238 unique pseudoknots. Of these, 230 (97%) are the simple ABAB-pseudoknot variety shown in Figure 1. This most common type of pseudoknot is involved in a number of essential biological processes including RNA self-splicing, translation control and viral frameshifting. It is perhaps not surprising that as the complexity of the knot increases, its likelihood of occurring decreases.

In PseudoBase, there are also six ABACBC kissing hairpin structures (PKB150, 163, 169, 171, 173 and 178), and two more exotic structures (PKB71 and 75). The pseudoknot loops occasionally contain an additional self-contained hairpin loop (e.g. ABACCB) but such nested structures do not change the degree of non-nestedness (e.g. ABAB-class), and in these cases PseudoBase only catalogs the A and B stems.

ABAB-pseudoknots are asymmetric. The distribution of stem lengths s_1 and s_2 are markedly different, as shown in Figure 2a. Excessively long stems are not required for pseudoknot formation; s_1 peaks at 3 bp and s_2 favors 5 or 6 bp.

Loop 2 is often very short (172 of the 230 unique ABAB-pseudoknots, or 75% have $L_2 = 0$; 195 of 230, or 85% have $L_2 \leqslant 1$) resulting in favorable coaxial helix stacking interactions which stabilize the pseudoknot. The Turner rules (17) permit helix stacking for $L_2 \leqslant 1$. In Section 3 we will present a theory for the ABAB class with stacked stems.

In Figure 2b, we also see differences in the distributions of L_1 and L_3 sizes, including multiple peaks. These features may arise because of differences in tertiary interactions between loops and stems.

We observe striking composition biases in the loops of ABAB-pseudoknots. As Table 1 shows, loop 1 is uracil rich while loop 3 tends to be adenine rich, particularly the end of loop 3 which is across from stem 1. These observations are consistent with reports of tertiary contacts (with one to four hydrogen bonds) between loop adenines and the minor grooves of helices, known as A-minor interactions (18–23). Adenine-rich loop 3 is on the minor groove side of stem 1. On the other hand, uracil-rich loop 1 is a more flexible loop (24), and interacts less with the major groove of stem 2 (see Figure 1b).

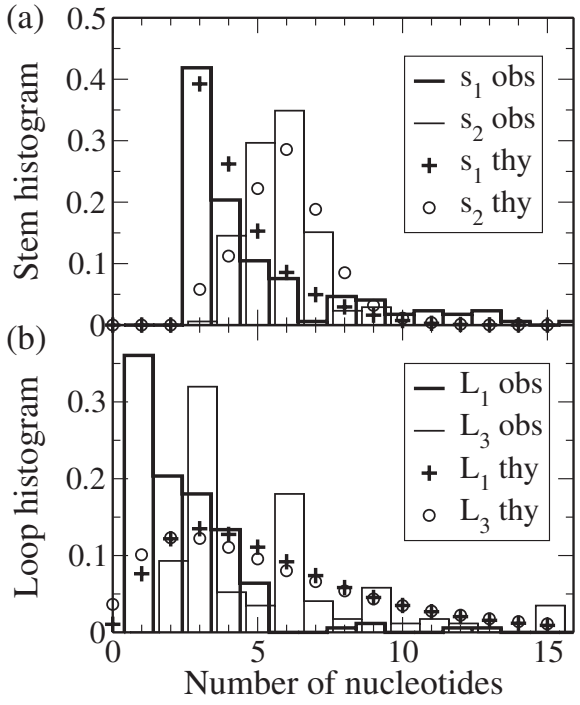

Figure 2. The statistics of ABAB-pseudoknots in PseudoBase (obs) with $L_2 = 0$ is compared with our theory (thy). (**a**) Stems favor different numbers of base pairs s_1 and s_2. (**b**) Loop lengths L_1 and L_3 are also asymmetric.

Table 1. The overall base composition of loops 1 and 3 differs

	A	C	G	U
Loop 1	27.0	15.6	17.9	39.5
Loop 3	46.1	14.3	11.1	28.5
Loop 3 (last)	63.9	11.4	4.4	20.3
Loop 3 (first)	35.1	9.4	11.4	44.1
Stem 1	18.0	27.7	32.1	22.1
Stem 2	19.6	28.2	30.5	21.8

Loop 3 has a high percentage of adenines which makes it prone to A-minor stacking interactions with stem 1. Loop 1 has a high percentage of uracils, making it a more flexible loop and more interaction neutral. The adenines in loop 3 are strongly biased toward the 3′ end of the loop. The large fraction of uracils at the start (5′ side) of loop 3 enhances loop flexibility in the turn.

The asymmetries in the populations of stem and loop lengths have not been explained by previous pseudoknot algorithms and models (4–13). The algorithms in Refs (5–8,13) are all symmetric with respect to stem and loop lengths (i.e. transforming an ABAB-into a BABA-pseudoknot by interchanging stems 1 $\leftrightarrow$ 2 and loops 1 $\leftrightarrow$ 3). The phenomenological estimates of Gultyaev and co-workers (25) do provide different free energies for loops 1 and 3 but result from *ad hoc* assumptions rather than polymer physics.

We assert that the differences in stem and loop sizes arise primarily from major/minor groove asymmetries and use this fact to reproduce the population of pseudoknots observed in PseudoBase.

ABAB PSEUDOKNOT MODEL

The dominant contributions to the free energy of ABAB-pseudoknots are (i) base-pairing of stems and (ii) entropy of the loops. The overall free energy of the complex is then

$$\Delta G = \Delta G_{s_1} + \Delta G_{s_2} - TS(s_1, s_2, L_1, L_2, L_3), \quad \mathbf{1}$$

where ΔG_{s_j} is the free energy of helix j and $S(s_1, s_2, L_1, L_2, L_3)$ is the entropy of the loops.

Stems, RNA duplex

Step one is to describe the base-paired stems. The cartesian coordinates of complementary bases in double-helical A-form RNA are approximately:

$$\mathbf{r}_W = \left\{ r\cos\left(\frac{2\pi s}{N_t}\right), r\sin\left(\frac{2\pi s}{N_t}\right), hs \right\},$$
$$\mathbf{r}_C = \left\{ r\cos\left(\frac{2\pi s}{N_t} + \phi\right), r\sin\left(\frac{2\pi s}{N_t} + \phi\right), hs + H_{\text{off}} \right\}. \quad \mathbf{2}$$

Here s indexes both the nucleotide on the Watson strand and its complement on the Crick strand. The coordinates of the 4′ carbon from the six double-helical RNA structures which appear in the Protein Data Bank (26) (1AL5, 1RNA, 1RRR, 1RXB, 1SDR and 433D) were incorporated in a least-squares fit to obtain values for the model parameters: the number of base pairs per helical turn $N_t = 11.2 \pm 0.3$, the radius of the 4′ carbon $r = 9.9 \pm 0.2$ Å, the height per stack is $h = 2.7 \pm 0.2$ Å, the phase angle between complementary strands $\phi = 1.6 \pm 0.1$ rad $= 93 \pm 4°$ and the vertical offset between complementary strands $H_{\text{off}} = -4.2 \pm 1.4$ Å.

Consider the typical ABAB-pseudoknot, with $L_2 = 0$ and helices 1 and 2 stacked. In this configuration loop 1 must traverse the distance from the junction between the helices to the other end of stem 2 across the *major* groove. This distance is

$$D_{L_1} = |\mathbf{r}_W(s) - \mathbf{r}_C(s + s_2)|, = \sqrt{2(1 - \cos\theta_{s_2})r^2 + (hs_2 + H_{\text{off}})^2}, \quad \mathbf{3}$$

where $\theta_{s_2} = (2\pi s_2/N_t + \phi)$, is the phase angle between the strands.

The other loop must traverse the distance from the junction between the helices to the other end of stem 1 across the *minor* groove. This distance is

$$D_{L_3} = |\mathbf{r}_C(s) - \mathbf{r}_W(s + s_1)|, = \sqrt{2(1 - \cos\theta_{s_1})r^2 + (hs_1 - H_{\text{off}})^2}, \quad \mathbf{4}$$

with $\theta_{s_1} = (2\pi s_1/N_t - \phi)$. The sign difference in θ and in the H_{off} term arises from the major/minor groove asymmetry. In Figure 3, we show how the distances differ in the two cases.

Loops

Step two is to estimate the loop entropy. In the standard Gaussian approximation, a chain of N links of length a has end-to-end separation distance between D and $D + d$ with probability

$$p_G(D, N) = 4\pi D^2 d\left(\frac{3}{2\pi N a^2}\right)^{3/2} \exp\left\{\frac{-3D^2}{2Na^2}\right\}, \quad \mathbf{5}$$

where $a = 6.2$ Å and $d = 0.1$ Å is our model's one free parameter. Other polymer physics models, such as the worm-like chain, self-avoiding chain (13), freely-jointed chain models, could also be used in place of Equation (5). The entropic contribution of loop 1 to Equation (1) can be obtained from Equation (5), taking $D = D_{L_1}$ and $N = (L_1 + 1)$ links for L_1 nucleotides. To be explicit, the total entropy is $S = R\ln[p_G(D_{L_1}, L_1 + 1)p_G(D_{L_3}, L_3 + 1)]$.

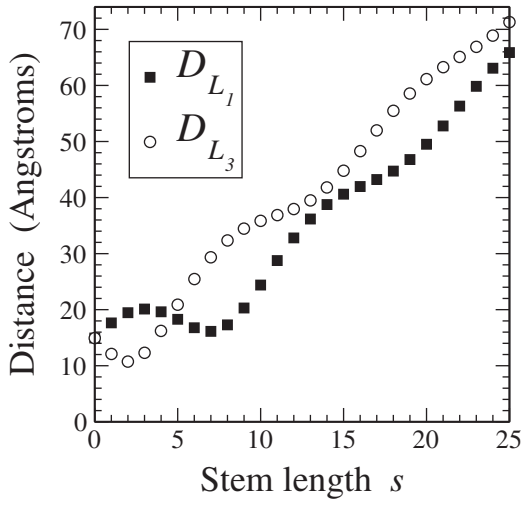

Figure 3. The distances D_{L_1} and D_{L_3} across the major or minor groove as a function of the number of bases s in the associated stem. The differences are due to the geometries of major- and minor grooves.

ABAB probability

The probability of an ABAB-pseudoknot with lengths $\{s_1, s_2, L_1, L_2 = 0, L_3\}$ is the product of a degeneracy factor for the ABAB pattern and the likelihood of that pattern resulting in a pseudoknot.

The degeneracy of the ABAB pattern is $4^{s_1+s_2+L_1+L_3}$, out of all patterns $4^{2s_1+2s_2+L_1+L_3}$, because of the required complementarity. For the sake of simplicity, we ignore bulge loops in stems (which occur in ~30% of structures) at this stage.

To estimate the free energy of the stems, we compose random strings with $s_1 + s_2$ consecutive complementary base pairs bookended with mismatch pairs, then calculate their binding free energy using BINDIGO (27), finding:

$$G_{\text{stem}}(s_1 + s_2) = (-2.14 \text{ kcal/mol})(s_1 + s_2 - 4.88).$$

For the loop entropy, we use the Gaussian approximation, Equation (5), assuming the loops must traverse the distances given by Equations (3) and (4). Thus,

$$p_{ABAB} = \exp\{-\beta G_{\text{stem}}(s_1 + s_2)\} \times p_G(D_{L_1}, L_1 + 1)p_G(D_{L_3}, L_3 + 1), \quad \mathbf{6}$$

is the Boltzmann factor for ABAB-pseudoknots, with $\beta^{-1} = RT_{37^\circ} = 0.62$ kcal/mol.

We estimate the free energy of the optimal nested fold of an ensemble of $N = 2s_1 + 2s_2 + L_1 + L_3$ randomly selected nucleotides using MFOLD (2), finding

$$p_{\text{nest}} = \exp\{-\beta(-0.286 \text{ kcal/mol})(N-17)\}, \quad \mathbf{7}$$

for the Boltzmann factor for nested folds.

Combining the degeneracy factors and the Boltzmann likelihoods, the probability of a pseudoknot is thus

$$p_\psi = \frac{1}{4^{s_1+s_2}} \frac{p_{ABAB}}{p_{ABAB} + p_{\text{nest}} + 1}. \quad \mathbf{8}$$

The 1 in the denominator includes the Boltzmann factor for an open polymer configuration ($G_{\text{open}} = 0$).

To compare Equation (8) with the histograms of Figure 2, we simply sum the other degrees of freedom. For example, to obtain the s_1 distribution, we compute

$$\sum_{s_2, L_1, L_3} p_\psi, \quad \mathbf{9}$$

and analogously for the other sub-ensembles. The agreement of theory and observation is excellent. Studying the properties of the ensemble can reveal insights into the folding problem that individual cases may not.

ABAB-pseudoknots form because of their low energy, with about three-quarters of nucleotides base paired, versus about half of bases paired in nested structures. However, because pseudoknots require many base pairs constrained to the ABAB pattern, they remain unlikely in sequence space.

CONCLUSIONS

Pseudoknots are rare compared with conventional nested secondary structures but their structure gives them biological importance. Of the pseudoknots that appear, the ABAB-type are by far the most common. The structures of these ABAB-pseudoknots are asymmetric. We have argued that this asymmetry is due to structural differences between the major and minor groove. Our simple model is consistent with the observed asymmetry of s_1 and s_2. The statistical mechanical theory Equation (8) provides remarkable agreement with experiment as seen in Figure 2. This suggests that PseudoBase is a representative sample of ABAB-pseudoknot characteristics in nature and that we can now compute pseudoknot abundances in aggregate. Using free energies specific to a given sequence, we can also use the Boltzmann factors to calculate the likelihood of forming a particular pseudoknot.

Models which ignore major/minor groove asymmetry will predict the same free energies for an ABAB-pseudoknot and its BABA counterpart. For example, the symmetry of the theory in Ref. (8) arises because those authors effectively take $H_{\text{off}} = 0$ and $\phi = \pi$ rad $= 180^\circ$, in disagreement with the actual A-form structural asymmetry. We predict that the differences between loops 1 and 3 will destabilize many of the BABA version pseudoknots due to the differences between Equations (3) and (4), the decrease in A-minor interactions and the increased rigidity of the major groove loop.

The rarity of more complicated folds makes comparisons with observed distributions infeasible. Nevertheless, other pseudoknot types like kissing hairpins can be treated with methods similar to those presented. In addition, our simple theory could be extended to permit the possibility of secondary strucure within the loops (e.g. an ABACCB structure) and to permit flexibility between the stems when $L_2 > 0$.

Our interest in this paper has been to estimate properties of the ensemble of ABAB-pseudoknots and compare those with observed pseudoknots. To study a particular pseudoknot, values specific to its sequence should be used in place of the general G_{stem} and G_{nest} average values given.

ACKNOWLEDGEMENTS

The authors thank Jesse Dill for noticing the duplications in PseudoBase and assisting with Table 1. This work was supported by National Institutes of Health grant GM068485. Funding to pay the Open Access publication charges for this article was provided by Williams College.

Conflict of interest statement. None declared.

REFERENCES

1. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
2. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
3. Hofacker,I.L., Fontana,W., Standler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
4. Lyngsø,R.B. and Pedersen,C.N. (2000) RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, **7**, 409–427.
5. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
6. Rivas,E. and Eddy,S.R. (1999) The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, **16**, 334–340.
7. Orland,H. and Zee,A. (2002) RNA folding and large N matrix theory. *Nucl. Phys. B.*, **620**, 456–476.
8. Isambert,H. and Siggia,E.D. (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl Acad. Sci. USA*, **97**, 6515–6520.
9. Akutsu,T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, **104**, 45–62.
10. Dirks,R.M. and Pierce,N.A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, **24**, 1664–1677.
11. Lyngsø,R.B. and Pedersen,C.N. (2000) Pseudoknots in RNA secondary structures. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB '00)*, 8–11 April, Tokyo, Japan, pp. 201–209.
12. Uemura,Y., Hasegawa,A., Kobayashi,S. and Yokomori,T. (1999) Tree adjoining grammars for RNA structure prediction. *Theor. Comp. Sci.*, **210**, 277–303.
13. Lucas,A. and Dill,K.A. (2003) Statistical mechanics of pseudoknot polymers. *J. Chem. Phys.*, **119**, 2414–2421.
14. Haslinger,C. (2001) Predition algorithms for restricted RNA pseudoknots. Ph.D. dissertation, Universität Wien, Vienna, Austria.
15. Condon,A., Davy,B., Rastegari,B., Zhao,S. and Tarrant,F. (2004) Classifying RNA pseudoknotted structures. *Theor. Comp. Sci.*, **320**, 35–50.

16. van Batenburg,F.H.D., Gultyaev,A.P., Pleij,C.W.A., Ng,J. and Oliehoek,J. (2000) Pseudobase: a database with RNA pseudoknots. *Nucleic Acids Res*., **28**, 201–204.
17. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*., **288**, 911–930.
18. Michiels,P.J.A., Versleijen,A.A.M., Verlaan,P.W., Pleij,C.W.A., Hilbers,C.W. and Heus,H.A. (2001) Solution structure of the pseudoknot of SRV-1 RNA, involved in ribosomal frameshifting. *J. Mol. Biol*., **310**, 1109–1123.
19. Nixon,P.L. and Giedroc,D.P. (2000) Energetics of a strongly pH dependent RNA tertiary structure in a frameshifting pseudoknot. *J. Mol. Biol*., **296**, 659–671.
20. Egli,M., Minasov,G., Su,L. and Rich,A. (2002) Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. *Proc. Natl Acad. Sci. USA*, **99**, 4302–4307.
21. Battle,D.J. and Doudna,J.A. (2002) Specificity of RNA–RNA helix recognition. *Proc. Natl Acad. Sci. USA*, **99**, 11676–11681.
22. Strobel,S.A. (2002) Biochemical identification of A-minor motifs within RNA tertiary structure by interference analysis. *Biochem. Soc. Trans*., **30**, 1126–1131.
23. Nissen,P., Ippolito,J.A., Ban,N., Moore,P.B. and Steitz,T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA*, **98**, 4899–4903.
24. Aalberts,D.P., Parman,J.M. and Goddard,N.L. (2003) Single-strand stacking free energy from DNA beacon kinetics. *Biophys J*., **84**, 3212–3217.
25. Gultyaev,A.P., van Batenburg,F.H.D. and Pleij,C.W.A. (1999) An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, **5**, 609–617.
26. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
27. Hodas,N.O. and Aalberts,D.P. (2004) Efficient computation of optimal oligo–RNA binding. *Nucleic Acids Res*., **32**, 6636–6642.