J. Chem. Sci., Vol. 119, No. 5, September 2007, pp. 525–531.

Base pairing in RNA structures: A computational analysis of structural aspects and interaction energies

PURSHOTAM SHARMA, ABHIJIT MITRA,* SITANSH SHARMA and HARJINDER SINGH
Center for Computational Natural Sciences and Bioinformatics,
International Institute of Information Technology, Hyderabad 500 032
e-mail: abi_chem@iiit.ac.in

MS received 5 May 2007; accepted 6 June 2007

Abstract. The base pairing patterns in RNA structures are more versatile and completely different as compared to DNA. We present here results of *ab-initio* studies of structures and interaction energies of eight selected RNA base pairs reported in literature. Interaction energies, including BSSE correction, of hydrogen added crystal geometries of base pairs have been calculated at the HF/6-31G** level. The structures and interaction energies of the base pairs in the crystal geometry are compared with those obtained after optimization of the base pairs. We find that the base pairs become more planar on full optimization. No change in the hydrogen bonding pattern is seen. It is expected that the inclusion of appropriate considerations of many of these aspects of RNA base pairing would significantly improve the accuracy of RNA secondary structure prediction.

Keywords. Base pairing of RNA; DNA sequence; interaction energy of RNA; RNA structures.

1. Introduction

A major part of molecular biology research has mainly been centered on the role of DNA as the carrier of information in terms of their sequences and on the role of proteins in terms of their functions. For example, structural genomics, the systematic determination of all macromolecular structures represented in a genome, is focused at present exclusively on the proteins. Similarly, though crystal structures of both DNA as well as proteins have been studied widely in the early days of structural biology, proteins have attracted far more attention than DNA. Proteins were considered to be involved with all the biochemical processes and displayed almost infinite variety. DNA on the other hand, was thought to be a fairly stable molecule with a monotonously regular structure.

The structural analysis of RNA molecules was by and large a neglected area possibly for two reasons. One is the prevailing understanding that unlike DNA, RNA molecules have no 'structure' related functional role to play and that in terms of sequence information they have nothing more to offer than what DNA already provides. The other reason is the fact that crystallization of RNA molecules was found to be very difficult. Today, with the determination of high resolution 3D structures of tRNA molecules,[1] Ribozymes,[2] the detailed structures of the 30S and 50S ribosomal fragments[3,4] and even of ribosome[5], the scenario has changed completely on both counts. In recent times the discovery of several ncRNA (non-coding RNA which does not code for proteins but directly performs structural, catalytic or regulatory functions) genes has drawn progressively greater scientific attention to RNA structural studies.

The base pairing patterns in DNA and RNA molecules are completely different. In case of DNA molecule, the hydrogen bonding between the bases takes place through the formation of standard Watson–Crick (WC) base pairs. On the other hand, in case of RNA, apart from standard WC base pairs, the hydrogen bonding between bases can occur through highly versatile non WC base pairing patterns. These non WC base pairs, often called non-canonical base pairs are an important factor in governing the evolution and folding of RNA structures.[6] They are also important in forming tertiary interactions between remote portions of RNA structures, and sometimes they participate in formation of structurally specific and evolutionary conserved regions of RNA structures, called RNA motifs.[7]

*For correspondence

The possible base pairing patterns in RNA structures were studied in detail by Leontis *et al.* They classified the known RNA base pairing types in 12 geometric families[8] on the basis of type of interacting edge and the orientation of the glycosidic bonds. They also elaborated a matrix formulation for these base pairing types, to signify isosteric relationships between these bases.[9]

But apart from classification and characterization of RNA base pairs, an important task is the evaluation of interaction energies of RNA base pairs in 'away from equilibrium' geometries present in actual structural contexts, and to compare these geometries and interaction energies with the fully optimized gas phase geometries of these base pairs. It is true that base pair geometries observed at high resolution in crystal structures of DNA fragments correspond to the minima on potential energy surfaces of isolated DNA base pairs. However, in RNA molecules, base pair geometries are affected by multiple factors, often leading to overall optimization of the complete motif, and hence, many of the geometries of individual base pairs do not correspond even to the local minima on the intrinsic potential energy surface of the interacting subsystems. Further, X-ray crystallography gives us structures, but it does not provide us with any information regarding energy, often resulting in misleading interpretation of observed interactions.[10] Thus, the evaluation of interaction energies for base pairs in the geometry it assumes in the crystal, as well as in fully optimized geometry, is imperative. This would also be helpful in analysing the contribution of base pair hydrogen bonding to the overall stability of the structure.

An important aspect of RNA research that has attracted a lot of attention for more than two decades is the RNA folding problem. RNA folding seems to be driven principally by hydrogen bonding and base stacking, is replete with complex non-canonical interactions and are highly dependent on environmental factors such as presence of ions and protein co-factors. Recently Meyer *et al* have carried out a detailed theoretical analysis, adequately augmented with experimental evidences from studies on a variety of transcription related processes, highlighting the importance of co-transcriptional folding, and implying that transcription affects folding, in the context of both RNA secondary structure prediction methods as well as for the detection of RNA genes.[11] The most notable assertion that emerges from this and other studies is that most computational RNA secondary structure and folding pathways prediction methods which essentially work around the minimization of free energy of the already synthesized RNA molecule need to be reviewed in the context of the effects of co-transcriptional and protein mediated folding. We feel that availability of different types of base–base interaction energy data will be helpful in developing a better understanding in this regard. Such data would also be helpful in probing into the molecular mechanisms of RNA functions.

In the present work, we carried out comparative *ab-initio* computations of geometries and interaction energies of eight selected RNA non-canonical base pairs in crystal geometry and in fully optimized contexts. The computations will also be a part of the database of interaction energies for different types of base pairings in RNA especially in 'away form equilibrium' situations in keeping with the geometries exhibited in actual structural contexts. Some general features of non-canonical base pairing in RNA structures are expected to emerge from these computations, which will be helpful in framing rules for sequence-structure prediction in RNA.

Interaction energy for canonical RNA base pairs have been calculated by several groups using different methods ranging from HF/6-31G*[12] and MP2/6-31G**[13] to DFT/cc-pVXZ.[14] Not only do all of them display similar trends in results, but geometry optimization of nucleotide bases by MP2/6-31G(2*p*, 2*d*), HF/6-31G(2*p*, 2*d*), HF/6-311G(2*p*, 2*d*), and HF/6-311+G (2*p*, 2*d*) have also been reported[15] to show very similar structural features. It has been confirmed that medium quality *ab-initio* methods provide rather satisfactory estimates of the base pairing energies, sufficiently accurate for most applications, especially regarding the relative stability of the base pairs.[16] We have therefore opted for HF/6-31G** as the method of choice, optimal in terms of speed and reliability in the context of our investigations, for both geometry optimization as well as for computation of interaction energy using Morokuma with BSSE correction.

2. Methods

2.1 *Crystal structure database analysis*

We have run our BPFIND[17] program on 208 good resolution crystal structures of RNA taken from Protein Data Bank (PDB) and computed the occurrence frequency of the base pairs selected for our computations. The PDB ids, base pair ids, and the corre-

Table 1. Classification of RNA base pairs by Leontis and Westhof[8]

No.	Glycosidic bond orientation	Interacting edges	Local strand orientation
1	*cis*	Watson–Crick/Watson–Crick	Anti parallel
2	*trans*	Watson–Crick/Watson–Crick	Parallel
3	*cis*	Watson–Crick/Hoogsteen	Parallel
4	*trans*	Watson–Crick/Hoogsteen	Anti parallel
5	*cis*	Watson–Crick/Sugar Edge	Anti parallel
6	*trans*	Watson–Crick/Sugar Edge	Parallel
7	*cis*	Hoogsteen/Hoogsteen	Anti parallel
8	*trans*	Hoogsteen/Hoogsteen	Parallel
9	*cis*	Hoogsteen/Sugar Edge	Parallel
10	*trans*	Hoogsteen/Sugar Edge	Anti parallel
11	*cis*	Sugar Edge/Sugar Edge	Anti parallel
12	*trans*	Sugar Edge/Sugar Edge	Parallel

sponding percentage occurrence for these selected base pairs are listed in table 1.

2.2 *Geometry of the systems*

The initial structures of base pairs were built by extracting the coordinates of base pairs from respective PDB files using RASMOL[18] software. Hydrogen atoms were added to these base pairs using MOLDEN[19] software. The sugar portions attached to base pairs in RNA structures were removed, and C1′ atoms were respectively replaced by hydrogen atoms during model building. The change in base pair geometry on relaxed geometry optimization was estimated by superposing one of the bases of the 'relaxed' pair with the corresponding base of the 'unrelaxed' or 'crystal geometry' pair and calculating the respective RMSD[20] values.

2.3 *Computational details*

The GAMESS-US[21] package was used for all calculations. We have optimized the structures of base pairs using the HF/6-31G (*d*, *p*) basis set. We have calculated basis set superposition error (BSSE) corrected interaction energies between the bases of base pair by Morokuma[22] method using the HF/6-31G (*d*, *p*) basis set.

Two different sets of total interaction energies were evaluated. In the first approach, we optimize positions of all atoms in the base pair, and the BSSE corrected interaction energy was calculated relative to the fully optimized and isolated bases. Then we separately correct for the deformation energy in such a way that, using the monomer basis sets, we evaluate the monomer energies in the deformed (complex) and optimized isolated monomer geometries, i.e.

$$E^A_{\text{def}} = E^A_{\text{Dimer}} - E^A_{\text{Monomer}}, \tag{1}$$

where *A* stands for individual monomer base. Thus the interaction energy of the base pair is defined in the following way:

$$\Delta E^{A\ldots B} = E^{A\ldots B} - (E^A + E^B) + E^A_{\text{def}} + E^B_{\text{def}}, \tag{2}$$

In the second approach, we optimize the position of hydrogen atoms only, while all the heavy atoms were frozen as in the X-ray structure, using the IFREEZ option of GAMESS. The interaction energies were calculated relative to the unrelaxed isolated bases using a rigid body approximation, i.e. base pair was rigidly fragmented into the bases, which implies that deformation energy is not defined here. Thus the interaction energy in this case is defined as

$$\Delta E^{A\ldots B} = E^{A\ldots B} - (E^A + E^B), \tag{3}$$

Such an approach has already been applied in literature.[23]

3. Results and discussions

Optimized geometries and relative interaction energies were computed for the base pairs reported in table 2. The interaction energies were also computed for these base pairs in crystal geometry after adding

Table 2. PDB ids and base pair ids of systems used for computation. The percentage occurrence non-canonical base pairs is calculated.

System number	Base pair	Edge interactions	Source structure PDB ID	Base pair ID	Percentage occurrence
1	AA	HH *trans*	1QVG	A2691(O)–A2703(O)	1·51
2	UU	WW *cis*	1FFK	U26(O)–U517(O)	1·21
3	AA	WH *trans*	1FFK	A460(O)–A455(O)	0·57
4	GG	WH *trans*	1QVG	G868(O)–G775(O)	0·20
5	GU	SW *trans*	1J2X	G592(A)–U460(A)	0·03
6	AG	W + H *cis*	1NJP	A1061(O)–G2731(O)	0·03
7	GC	WH + *trans*	1DRZ	G161(B)–C141(B)	0·05
8	AU	WW *cis*	1ASY	A607–U666	Canonical

Table 3. Relevant geometrical parameters of inter base contacts for the hydrogen optimized crystal geometries of the base pairs.

		Hydrogen optimized crystal geometry			Fully optimized geometry		
Base pair	Edge interactions	Donor atom (*X*)	Acceptor atom (*Y*)	*X*–*Y* distance	Donor atom (*X*)	Acceptor atom (*Y*)	*X*–*Y* distance
AA	HH *trans*	N6(A1)	N7(A2)	3·21	N6(A1)	N7(A2)	3·19
		N6(A2)	N7(A1)	2·81	N6(A2)	N7(A1)	3·19
UU	WW *cis*	O4(U1)	N3(U2)	3·03	O4(U1)	N3(U2)	2·98
		O2(U2)	N3(U1)	2·93	O2(U2)	N3(U1)	2·97
AA	WH *trans*	N6(A2)	N1(A1)	3·09	N6(A2)	N1(A1)	3·18
		N6(A1)	N7(A2)	2·97	N6(A1)	N7(A2)	3·16
GG	WH *trans*	N1(G2)	N7(G1)	2·78	N1(G2)	N7(G1)	2·96
		N2(G2)	O6(G1)	3·05	N2(G2)	O6(G1)	3·28
GU	SW *trans*	N2(G)	O2(U)	3·11	N2(G)	O2(U)	2·98
		N3(U)	N3(G)	3·39	N3(U)	N3(G)	3·15
AG	W + H *cis*	N1(A)	N7(G)	2·97	N1(A)	N7(G)	2·99
		N6(A)	O6(G)	3·15	N6(A)	O6(G)	2·85
GC	HW + *trans*	N4(C)	N7(G)	3·15	N4(C)	N7(G)	3·01
		N3(C)	O6(G)	2·92	N3(C)	O6(G)	2·79
AU	WW *cis*	N2(U)	N1(A)	2·94	N2(U)	N1(A)	2·99
		N6(A)	O4(U)	2·85	N6(A)	O4(U)	3·08

the hydrogen atoms and optimizing their position after freezing the heavy atoms. The optimized geometries are the ideal geometries that would be obtained in the gas phase in the absence of any other effects, whereas the hydrogen optimized crystal geometries of the base pairs mimic the geometries of the base pairs in the actual crystal structure contexts.

The base pairs were selected from different RNA base pair families shown in table 1. Two of the base pairs having a reasonably high occurrence frequency in the 208 analysed RNA crystal structures (systems 1 and 2 in table 2), as detected by the BPFIND program, were selected for our computations. Two base pairs (systems 3 and 4) having relatively smaller occurrence frequencies, and one base pair having very low occurrence frequency (system 5) was selected. Two protonated base pairs (systems 6 and 7) from different RNA base pair families were selected for our computations. Apart from these, one canonical base pair (system 8) was also selected for our computations. The hydrogen bonds of the base pairs in the hydrogen optimized crystal geometries and the fully optimized geometry are reported in table 3. The superposition of the base pairs in optimized and the crystal geometry are shown in figure 1. The root mean square deviation values (RMSD) values are reported in table 4.

The interaction energies of the base pairs in the crystal geometries and fully optimized geometries are reported in table 5. The interaction energies for

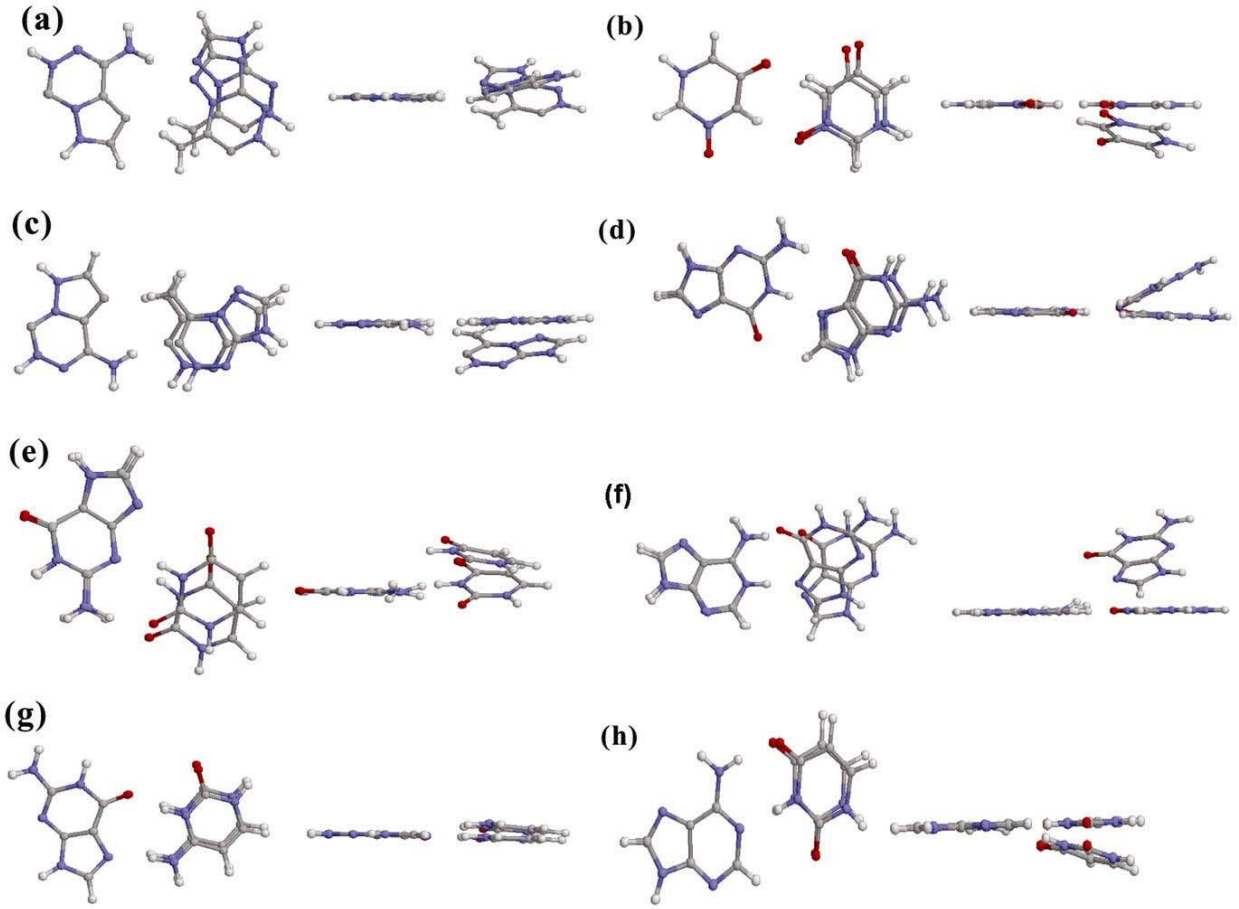

Figure 1. Superposed structures of crystal geometry and optimized geometry of base pairs. (**a**) System 1 (**b**) System 2 (**c**) System 3 (**d**) System 4 (**e**) System 5 (**f**) System 6 (**g**) System 7 and (**h**) System 8 (see table 1 for complete description). For each system, the figures on the left side are the top views and on the right side are the side views of the superposed systems respectively.

the fully optimized geometries are corrected for the deformation energy term as described in the methods section. It is seen that in general, the interaction energy in the crystal geometry is relatively smaller than in case of fully optimized geometry. But, in case of systems 4, 6, 7 and 8 the interaction energy is slightly smaller in case of optimized geometry as compared to the crystal geometry. The interaction energy of the protonated base pairs is higher than the non-protonated base pairs. This is expected, as the presence of charged species increases the electrostatic component of the interaction energy to a greater extent which increases the value of total interaction energy.

It may be noted that there is not much difference in the interaction energy of system 8, which is the standard WC geometry, from the interaction energy of other non-canonical base–base pairs which are presented in this work. Traditionally, non-canonical base pairs have been considered as perturbations to the overall secondary structure of RNA (leading to functional motifs), which is mainly determined by strong canonical interactions and G.U Wobble base pairs. But, it seems that the importance of non-canonical base pairs is comparable, if not more, than the canonical base pairs in terms of extent of base pair interaction and stability. It thus appears that the detailed studies of the interaction energies of these base pairs will be helpful in developing a better framework of RNA secondary structure prediction.

In all these cases, the hydrogen bonding pattern is same in case of crystal geometry and the fully opti-

Table 4. Interaction energies for the hydrogen optimized as well as fully optimized geometry (RMSD values are for crystal structure with hydrogen atoms added and geometry optimized and fully optimized geometry).

				Full optimized geometry		
Base pair	Interacting edges and orientation	RMSD	Hydrogen optimized geometry (kcal/mol)	Interaction energy (kcal/mol)	Deformation energy (kcal/mol)	Total (kcal/mol)
AA	HH *trans*	0·92	–1·64	–7·04	0·65	–6·39
UU	WW *cis*	0·72	–8·32	–9·09	0·63	–8·46
AA	WH *trans*	0·84	–6·49	–8·03	0·23	–7·80
GG	WH *trans*	1·14	–16·17	–16·71	1·16	–15·55
GU	SW *trans*	1·10	–5·81	–9·21	0·96	–8·25
A·G	W + H *cis*	0·84	–31·99	–31·42	1·79	–29·63
GC	HW + *trans*	0·39	–37·56	–37·32	2·00	–35·32
AU	WW *cis*	0·62	–11·32	–10·42	0·66	–9·76

Table 5. Dipole moments of the base pair complex and the resultants of dipole moments of individual bases for the hydrogen optimized crystal geometry and fully optimized geometry. All values are in Debye.

	Hydrogen optimized geometry		Full optimized geometry	
Base pair	Dipole moment (D) of base pair complex	Resultant of the individual base dipole moments (D)	Dipole moment (D) of the base pair complex	Resultant of the individual base dipole moments (D)
AA	0·35	0·43	1·89	1·79
UU	3·28	3·36	2·38	2·39
AA	4·47	4·44	4·84	4·74
GG	11·33	9·37	10·05	8·36
GU	5·20	5·37	4·58	5·00
AG	1·73	3·21	1·11	11·68
GC	6·31	9·03	5·48	205·78
AU	3·98	3·20	2·89	2·36

mized geometry, except that the constrained and deformed hydrogen bonds in the crystal geometry becomes more linear in the optimized geometry.

The dipole moments of these base pair complexes, as well as the dipole moments obtained by the vector addition of the dipole moments of the individual monomers, have been reported in table 5, both for base pairs in the crystal geometry as well as in the fully optimized geometry respectively. It is seen that except for the systems 1 and 3, where two adenines are interacting with each other, the dipole moment of the crystal geometry is larger than in the fully optimized geometry. The dipole moment of system 1 in optimized geometry, systems 3 and 4 both in crystal and fully optimized geometry, and system 5 in crystal geometry, is greater than the resultant dipole moment obtained by the vector addition of two monomer dipole moments. The resultant dipole moment of the monomers is greater than the dipole moment of the dimer complex in case of system 1 in crystal geometry, system 5 in fully optimized geometry, and systems 2, 6, 7 and 8, both in crystal and fully optimized geometry. But the difference in magnitudes of the resultant dipole of the monomers and the dipole moment of the complex is not great, except in the case of systems 6 and 7, which demands special comment. In case of system 6, the orientation of the bases with respect to each other in the optimized geometry of the base pair is such that the dipole moment arising from the polar amino group of adenine and carbonyl group of guanine cancel each other, resulting in a relatively lower value of dipole moment in the optimized geometry. In case of system 7, the large dipole moments of the monomers C and G get cancelled with each other in the base pair complex geometry. The amino groups of C and G, and the carbonyl groups of the monomers point in opposite directions. This results in a less polarized charge distribution and the overall dipole moment in the complex geometry is reduced.

4. Conclusions

The stability and interaction energies of non-canonical base pairs are found to be comparable to the standard canonical base pairs of RNA. The dipole moments of these base pairs are different from the resultant dipole moment of the monomers, which indicates significant charge re-distribution and interaction on complexation. There is thus a serious reason to believe that there are facets of RNA base pairs chemistry other than the canonical interactions, which are energetically feasible and hence possible. Thus the role of RNA is not merely that of a transcriptor and translator; it is likely to involve far greater complexity than conceived earlier. Our observations lend further support to the possibility that RNA world could be a significant contributor to the evolutionary process and it may even have been a precursor to the DNA based evolution.

The detailed study of occurrence frequency, structural and electronic properties, and dipole moments of RNA base pairs will be helpful in developing a better framework for RNA secondary structure prediction. Maintenance of data of different types of base–base interaction energy for RNA base pairs will be helpful for understanding the effects of co-transcriptional and protein mediated folding and also to probe into the molecular mechanisms of RNA functions.

Acknowledgements

This work was partially supported by the grant from the Department of Biotechnology. We thank the Centre for development in advanced computing (CDAC) for computational support. P S and S S thank the Council of Scientific and Industrial Research (CSIR), New Delhi for Junior research fellowships.

References

1. (a) Kim S H, Suddath F L, Quigley G J, McPherson A, Sussman J L, Wang A H, Seeman N C and Rich A 1974 *Science* **185** 435; (b) Robertus J D, Ladner J E, Finch J T, Rhodes D, Brown R S, Clark B F and Klug A 1974 *Nature* (*London*) **250** 546
2. (a) Scott W G, Finch J T and Klug A 1995 *Cell* **81** 991; (b) Ferré-D'Amaré A R, Zhou K and Doudna J A 1998 *Nature* (*London*) **395** 567; (c) Golden B L, Gooding A R, Podell E R and Cech T R 1998 *Science* **282** 259; (d) Wedekind J E and McKay D B 1999 *Nat. Struct. Biol*. **6** 261
3. Wimberley B T, Guymon R, McCutcheon J P, White S W and Ramakrishnan V 1999 *Cell* **97** 491
4. Ban N, Nissen P, Hansen J, Moore P B and Steitz T A 2000 *Science* **289** 905
5. Yusupov M M, Yusupova G Z, Baucom A, Lieberman K, Earnest T N, Cate J H D and Noller H F 2001 *Science* **292** 883
6. Šponer J E, Špačková N, Leszczynski J and Šponer J 2005 *J. Phys. Chem*. **B109** 11399
7. (a) Leontis N B and Westhof E Q 1998 *Q. Rev. Biophys*. **31** 399; (b) Leontis N B, Lesocute A and Westhof E 2006 *Curr. Opin. Struct. Biol*. **16** 279; (c) Cate J H, Gooding A R, Podell E, Zhou, K H, Golden B L. Kundrot C E, Cech T R and Doudna J A 1996 *Science* **272** 1678; (d) Scott W G, Murray J B, Arnold J R P, Stoddard B L and Klung A 1996 *Science* **274** 2065
8. Leontis N B and Westhof E 2001 *RNA* **7** 499
9. Leontis N B, Stombaugh J and Westhof E 2002 *Nucleic Acids Res*. **30** 3497
10. Šponer J and Hobza P 2003 *Collect. Czech. Chem. Commun*. **68** 2231
11. Meyer I M and Miklos I 2004 *BMC Mol. Biol* **5** 10
12. Gould I R and Kollman P A 1994 *J. Am. Chem. Soc*. **116** 2493
13. Danilov V I and Anisimov V M 2005 *J. Biomol. Struct. Dyn*. **22** 471
14. Hesselmann A, Jansewn G and Schütz M 2006 *J. Amer. Chem. Soc*. **128** 11730
15. Mukherjee S, Majumdar S and Bhattacharyya D 2005 *J. Phys. Chem*. **B109** 10484
16. Šponer J, Jurečka P and Hobza P 2004 *J. Am. Chem. Soc*. **126** 10142
17. Das J, Mukherjee S, Mitra A and Bhattacharyya D 2006 *J. Biomol. Struct. Dyn.* **24** 149
18. (a) Roger Sayle and James Milner-White E 1995 *Trends Biochem. Sci*. **20** 374; (b) Herbert J Bernstein 2000 *Trends Biochem. Sci*. **9** 453
19. Schaftenaar G and Noordik J H 2000 *J. Comput. Aided Mol. Design* **14** 123
20. Bhattacharyya D, Koripella S C, Mitra A, Rajendran V B and Sinha B 2006 *IIT Hyderabad Technical Report TR No. IIIT/TR/2006/25* (http://www.iiit.net/techreports/reports.html)
21. Schmidt M W, Baldridge K K, Boatz J A, Elbert S T, Gordon M S, Jensen J, Koseky S, Matsunaga N, Nguyen K A, Su S J, Windus T L, Dupuis M and Montgomery J A 1993 *J. Comput. Chem*. **14** 1347
22. Kitaura K and Morokuma K 1976 *Int. J. Quantum Chem*. **10** 325
23. Oliva R, Cavallo L and Tramontano A 2006 *Nucl. Acids Res*. **34** 865