Supplementary Material

for

Frequency and Isostericity of RNA Basepairs

Jesse Stombaugh[1,4], Craig L. Zirbel[2,4], Eric Westhof[5], and Neocles B. Leontis[3,4,*]

[1] Department of Biological Sciences, [2] Department of Mathematics and Statistics, [3]Department of Chemistry and [4]Center for Biomolecular Sciences, Bowling Green State University, Bowling Green, OH 43403. [5] Architecture et réactivité de l'ARN, Université Louis Pasteur de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue René Descartes, F-67084 Strasbourg, France.

*To whom correspondence should be addressed
Telephone: (419) 372-8663
Fax: (419) 372-9809
Email: leontis@bgsu.edu

In addition to this pdf file, the supplementary material consists of these six additional files:

1. Stombaugh_et_al_Sup_Mat_S1.fasta
2. Stombaugh_et_al_Sup_Mat_S2.fasta
3. Stombaugh_et_al_Sup_Mat_S3.fasta
4. Stombaugh_et_al_Sup_Mat_S4.pdf
5. Stombaugh_et_al_Sup_Mat_S6_S7_S8.xls
6. Stombaugh_et_al_Sup_Mat_S9.xls

Supplemental Material S1. 5S rRNA sequence alignments. Sequence alignments were downloaded from the Rfam database and parsed to identify the most complete sequence for each species. SeqQR was used to reduce sequence redundancy as described in **Methods**. (See Stombaugh_et_al_Sup_Mat_S1.fasta)

Supplemental Material S2. 16S rRNA sequence alignments. Sequence alignments were downloaded from the European Ribosomal database and parsed to identify the most complete sequence for each species. SeqQR was used to reduce sequence redundancy as described in **Methods**. (See Stombaugh_et_al_Sup_Mat_S2.fasta)

Supplemental Material S3. 23S rRNA sequences alignments. Sequence alignments were downloaded from the European Ribosomal database and parsed to identify the most complete sequence for each species. SeqQR was used to reduce sequence redundancy as described in **Methods**. (See Stombaugh_et_al_Sup_Mat_S3.fasta)

Supplementary Material S4. Reduced redundancy list of NDB/PDB files. This list comprises representative RNA containing PDB files selected as described in **Methods.** (See Stombaugh_et_al_Sup_Mat_S4.pdf)

	Basepairs modeled in Leontis et al. (2002) that have been observed in 3D structures																
	BP Type	PDB	NT1		NT2		Number of Observations	C1'-C1' Distance	1st H-bond: H-bonding atoms		2nd H-bond: H-bonding atoms		3rd H-bond: H-bonding atoms		# of H-Bond Matches	Difference in C1'-C1' Distances	Frequencies from Bacterial rRNA Sequence Alignments
									NT1	NT2	NT1	NT2	NT1	NT2			
Predicted	cWS	-	G	-	A	-	-	8.5	O6	H2	H11	N3	H21	O2'	3/3	0.1	1.4%
Observed		2avy	G	394	A	366	1	8.6	O6	H2	H11	N3	H21	O2'			
Predicted	cWS	-	U	-	C	-	-	6.5	H3	O2	O2	HO2'			0/2	0.6	0.6%
Observed		1lng	U	185	C	233	1	7.1	---	---	---	---					
Predicted	cWS	-	U	-	U	-	-	6.5	H3	O2	O2	HO2'			1/2	0.4	2.2%
Observed		1s72	U	454	U	1362	1	6.1	H3	O2	---	---					
Predicted	tWS	-	A	-	C	-	-	10.0	H62	HO2'	H61	O2			2/2	0.5	1.7%
Observed		1s72	A	843	C	838	2	9.5	H62	HO2'	H61	O2					
Predicted	tWS	-	A	-	U	-	-	10.0	H62	O2'	H61	O2			2/2	0.6	1.3%
Observed		1s72	A	57	U	28	1	9.4	H62	O2'	H61	O2					
Predicted	tWS	-	U	-	U	-	-	8.5	O4	OH2'	H3	O2			1/2	0.3	1.0%
Observed		1grz	U	106	U	258	1	8.2	---	---	H3	O2					
Predicted	cHS	-	G	-	A	-	-	7.6	O6	H2					1/1	0.5	1.6%
Observed		2j01	G	2052	A	2051	1	7.1	O6	H2							
Predicted	cHS	-	U	-	C	-	-	7.0	O2	H5					1/1	0.8	1.2%
Observed		1et4	U	223	C	222	1	6.2	O2	H5							
Predicted	cSs	-	C	-	C	-	-	5.6	O2'	HO2'	HO2'	O2			2/2	0.2	0.1%
Observed		1q86	C	75	C	2542	1	5.8	O2'	HO2'	HO2'	O2					
Predicted	cSs	-	G	-	C	-	-	5.6	O2'	HO2'	HO2'	O3			2/2	0.6	0.1%
Observed		359d	G	50	C	152	13	6.2	O2'	HO2'	HO2'	O3					
Predicted	cSs	-	G	-	U	-	-	5.6	O2'	HO2'	HO2'	O4			2/2	0.3	1.9%
Observed		1s72	G	871	U	866	7	5.9	O2'	HO2'	HO2'	O4					
Predicted	tSs	-	G	-	A	-	-	8.4	N3	H2	H22	N3	H2	O2'	1/3	0.6	1.2%
Observed		2aw4	G	1750	A	2860	2	7.8	---	---	H22	N3	---	---			
	Modeled basepairs not observed in 3D structures																
Predicted	cWH	-	A	-	U	-	-	10.5	H61	O4	N1	H5			-	-	0.2%
Predicted	cWH	-	C	-	U	-	-	10.5	H41	O4	N4	H5			-	-	0.0%
Predicted	tWS	-	C	-	U	-	-	9.4	H42	O2'	H41	O2			-	-	0.3%
Predicted	cSs	-	U	-	U	-	-	5.3	O2'	HO2'	HO2'	O5			-	-	0.2%

Supplemental Table S5. Comparison of basepairs predicted by Leontis et al. (2002) with instances found using FR3D in 3D structures published since 2002. Only four predicted basepairs have still not been observed (lower part of table).

Supplementary Material S6. Isosteric Matrices (IM) for each geometric family showing IDI values for all base combinations that occur in that geometric basepair family. In some otherwise symmetric families, namely cWW, cHH, and tHH, the AA, CC, GG, and UU basepairs are not symmetric and so appear in the IDI table twice. Because in tWW, the AA, CC, GG, UU basepairs are perfectly symmetric, there is no need to list them twice. Finally, with cSS and tSS, the basepairs are not symmetric at all, and so all basepairs appear twice. The columns and rows of this Excel file may need to be re-sized for readability. (See "Sup_Mat_S6" worksheet in Stombaugh_et_al_Sup_Mat_S6_S7_S8.xls)

Supplementary Material S7. IDI values calculated for all base combinations within each geometric basepair family. The columns and rows of this Excel file may need to be re-sized for readability. (See "Sup_Mat_S7" worksheet in Stombaugh_et_al_Sup_Mat_S6_S7_S8.xls)

Supplementary Material S8. Basepair frequencies from 5S, 16S and 23S Bacterial sequence alignments, *E.c.* and *T.th.* 3D structures, and reduced redundancy set of 3D structures. Frequencies from the Bacterial sequence alignments were calculated using the Bacterial "conserved core" for each molecule (5S, 16S and 23S), determined from the 3D structural alignments (see **Supplemental Table S9**). The columns and rows of this Excel file may need to be re-sized for readability. (See "Sup_Mat_S8" worksheet in Stombaugh_et_al_Sup_Mat_S6_S7_S8.xls)

Supplementary Material S9. 3D structural alignments for 5S, 16S and 23S rRNAs of *E.c.* and *T.th.* For the 5S and 23S rRNA alignments the *H.m.* structures are included. The 3D alignment for each rRNA appears as a separate worksheet. The IDI values are color-coded as in previous IDI tables. IDIs were calculated between aligned basepairs of *E.c.* and *T.th.* using basepair exemplars for the basepair detected by FR3D (column "M") or using the coordinates from the 3D structures (column "N"). When the IDI values were greater than 3.3, the basepair positions were analyzed manually and the reason for the high IDI was identified (column "O") as one of four possibilities: (1) Same base combination and same basepair family, but difference in 3D modeling leading to higher IDI than calculated from exemplars; (2) Isosteric or near isosteric base combination and same basepair family, but difference in 3D modeling leading to higher IDI than calculated from exemplars; (3) Non-isosteric base combination and same basepair family; (4) Basepair is adjacent to a variable Motif. The columns and rows of this Excel file may need to be re-sized for readability. (See Stombaugh_et_al_Sup_Mat_S9.xls)

Supplementary Material References

1. Schuwirth, B.S., Borovinskaya, M.A., Hau, C.W., Zhang, W., Vila-Sanjurjo, A., Holton, J.M. and Cate, J.H. (2005) Structures of the bacterial ribosome at 3.5 A resolution. *Science*, **310**, 827-834.
2. Selmer, M., Dunham, C.M., Murphy, F.V.t., Weixlbaumer, A., Petry, S., Kelley, A.C., Weir, J.R. and Ramakrishnan, V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313**, 1935-1942.
3. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T. and Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327-339.
4. Klein, D.J., Moore, P.B. and Steitz, T.A. (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *Journal of molecular biology*, **340**, 141-177.