Imino hydrogen positions in nucleic acids from density functional theory validated by NMR residual dipolar couplings

Alexander Grishaev,[a] Jinfa Ying,[a] and Ad Bax*

a Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892,

[a] Both authors contributed equally

SUPPORTING INFORMATION

Full reference 12: Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision B.01, *Gaussian, Inc., Wallingford CT,* 2010.

Computational procedure

Initially, the coordinates of isolated A, U, G, and C nucleotides, with their ribose replaced by a methyl group, were energy-minimized using B3LYP/6-31G density functionals, allowing all base atoms to move. The resulting planar bases represent idealized geometries and are taken as frames of reference for the coordinates of the G:N1-H1 and U:N3-H3 vectors. A set of crystal structures solved at resolutions better than 3.0 Å was selected from the RCSB PDB database with PDB codes listed below. For these structures, base pair coordinates were extracted and used as ensembles representative of base pair geometry (Fig. S1). Due to the overwhelming abundance of the G:C and U:A Watson-Crick and G:U wobble base pairs relative to the much less frequently represented U:U and A:U reverse Hoogstein types, only limited subsets of ~200 geometries for Watson-Crick base pairs and ~80 for G-U wobbles were DFT-optimized. In order to facilitate the analysis of the N-H vector deviations and to remove geometry imperfections caused by varying structural resolution, the exact heavy atoms for each nucleotide base in these pairs were replaced by the best-fitted coordinates of the idealized isolated individual bases, mentioned above. While keeping all heavy atoms fixed, the coordinates of the hydrogen atoms in these base pairs were then optimized using B3LYP/6-31G density functionals. The resulting ~600 DFT-optimized geometries of the individual base

pairs were used as input for a non-linear least squares optimization routine, which optimizes the variable parameters in eq 1: coordinates of the virtual point (VP) of attraction, force constants of the stiffness potentials, and the constants that describe attraction to the VP and repulsion with the neighboring H atom(s). In order to match the populations observed in the structural database, the contribution from each base pair type to the discrepancy cost function is weighted to correspond to ratios of 1:2:9:27:82 for A:U reverse Hoogstein / U:U asymmetric / G:U wobble / A:U Watson-Crick / G:C Watson-Crick. Due to the high dimensionality of the optimization space (23 dimensions in total for the base pair types listed above), the starting values for the optimized variables were adjusted in a scheme which gradually increased the dimensionality of the optimization space while simultaneously removing the redundant dimensions. Initially, the potentials for all five base pair types are optimized separately. Such relatively low-dimensional calculations quickly converge and lead to the nucleotide-specific potential parameters, such as stiffnesses of the return potentials and the H-H interaction constants, reaching somewhat different values depending on the particular base pair type. In the next step, joint potentials are formed for G and U nucleotides, by combining G:C Watson-Crick and G:U wobble base pair fits for the G:N1-H1 vectors and U:A Watson-Crick, U:G wobble, U:U asymmetric, and U:A reverse Hoogstein base pair fits for the U:N3-H3 vectors, with the initial values for the fitted variables obtained from the results of the preceding separate base pair optimizations. The results of such calculations, now corresponding to single sets of the stiffness potential constants for a given nucleotide type, are still redundant due to the separate fits of the H-H repulsion potentials, and the parameters describing the U:O2 and U:O4 virtual points of attraction. The locations of these latter VP points were found to be very similar in the separate fits and therefore were defined by a single variable in the final round of fitting, which included all five considered base pair types, starting from the parameters values of the preceding pair of separate U and G potential fits. Definitions of the coordinate frames for the VP points are listed in Supplementary Table S2. The ϕ_{VP} angles in Table 1 of the main text are defined relative to the bisector lines of the heavy atom angle for N atoms (C:N3, A:N1, A:N7) and relative to the C-O bond for carbonyl O atoms (U:O2, U:O4, G:O6).

Due to the presence of the H-H repulsion term in eq 1, the predictions of the H^N atom coordinates using the now best-fitted pseudo-potentials are carried out iteratively in three cycles, to ensure self-consistency of the coordinates of the interacting H atoms. This is particularly relevant for proximate pairs of H^N atoms, such as G:H1 and U:H3 for G:U wobble and the two U:H3 atoms in U:U asymmetric base pairs. Moreover, since some of the atoms interacting with the imino H^N of G and U include base NH_2 groups, such as Watson-Crick and U:A reverse Hoogstein, separate potentials of the same form as in eq 1 are derived for those and their coordinates are also optimized until self-consistency is reached by the third cycle of iterations.

We have also evaluated the impact of using idealized geometries for the individual bases with results obtained when using the raw heavy atom coordinates from the crystal structures as input for the DFT calculations. This comparison was carried out for all A:U and G:C Watson-Crick base pairs with observable RDCs in RiboA. The orientations of the N-H vectors obtained from these two sets of input data agree to within $0.4\pm0.2°$. Similarly, there was a negligible impact of using the raw X-ray coordinates versus the idealized best-fit base coordinates as input for the DFT calculations on the quality of the RDC fit (Q = 0.602 vs Q = 0.604). However, in this respect it is worth noting that the X-ray structure of RiboA is of very high crystallographic quality, and that for most RNA structures in the database the use of idealized base geometries becomes more important.

We have also evaluated whether additional terms in the resistance potentials would be beneficial relative to the harmonic form of eq 1. Based on symmetry considerations, we added quartic terms to the out-of-plane and cubic terms to the in-plane stiffness potentials. Reparameterization of the overall potential with these terms included did not result in a noticeable improvement in the accuracy of the predicted (θ, ϕ) values, which indicates that our simple harmonic model for the resistance potentials is sufficient to adequately reproduce the observed angular deviations of the NH vector.

PDB codes corresponding to the RNA structural database used in this study:

1M5K, 2EZ6, 2A43, 1ZHO, 1Z43, 1ZCI, 1ZEV, 1TFW, 1T0D, 1SJ3, 1SJ4, 1S03, 1RC7, 1R3E, 1Q96, 1Q93, 1Q9A, 1NUJ, 1FEU, 1KH6, 1XJR, 1VC7, 2FGP, 2BCZ, 1ZFX, 2GOZ, 1X9C, 2B57, 1Y26, 1U8D, 2BCY, 2CKY, 2D2K, 2GDI, 402D, 405D, 409D, 413D, 420D, 422D, 433D, 434D, 438D, 439D, 466D, 1QCU, 1QC0, 472D, 1CSL, 1D4R, 1DQH, 1DUQ, 1I9X, 1KD3, 1KD5, 1L3Z, 1NLC, 377D, 1RXB, 259D, 255D, 157D, 1SDR, 1RNA, 1F27, 1MWL, 1NTB, 1MSW, 1DRZ, 1CX0, 1B7F, 1QA6, 1DFU, 1C0A, 1DI2, 1DUL, 1EC6, 1F7U, 1FFY, 1G2E, 1F8V, 1HQ1, 1E7X, 1I6U, 1JBR, 1JBT, 1E7K, 1H4S, 1K8W, 1GKW, 1L9A, 1IVS, 1LNG, 1M5O, 1M8V, 1M8W, 1M8X, 1M8Y, 1MJI, 1MZP, 1N78, 1BMV, 2BBV, 6MSF, 1A34, 1ZDJ, 1QTQ, 1SER, 1A9N, 1DK1, 1F7Y, 1G1X, 1EHZ, 1YFG, 1BR3, 1HMH, 429D, 437D, 462D, 483D, 1DUH, 1HR2, 1J8G, 1KXK, 1L2X, 1L8V, 1MSY, 205D, 280D, 361D, 1GID, 301D, 353D, 397D, 1FJG, 1JJ2.

Table S1. Measured and SVD-fitted RDC values in RiboA for the DFT-optimized H coordinates with and without adjacent stacked bases.

Site in 1Y26	Experimental RDC, Hz	SVD-fitted RDC to DFT-optimized geometries with no stacking	SVD-fitted RDC to DFT-optimized geometries with stacking
U20	18.8	20.0	20.1
U22	20.1	20.5	19.4
U25	11.9	11.6	11.5
U28	17.7	17.0	16.9
U31	4.2	3.9	3.8
U34	20.5	19.3	19.6
G37	18.3	19.6	19.7
G38	16.3	16.5	15.9
U39	8.4	9.6	9.5
U40	7.8	9.7	9.2
U41	11.0	10.9	10.8
G42	20.1	19.1	19.2
G43	17.8	17.1	17.2
G44	12.7	13.1	13.0
G46	18.7	18.4	18.6
G57	15.8	15.4	15.3
G59	9.0	7.4	7.9
U68	10.1	11.5	11.6
U70	19.7	20.7	20.9
U71	15.7	16.2	16.7
G72	4.3	3.5	3.5
U74	11.4	10.7	11.2
U75	18.9	19.0	19.1

U77	15.7	15.7	15.8
U49	3.9	4.2	4.0
U51	14.3	12.9	13.0
U47	16.7	15.3	15.3

Table S2. Definitions of the coordinate frames for the VP locations.[a]

Nucleotide type	H-bond donor atom	A1	OO	A2
C	N3	C2	N3	C4
A	N1	C2	N1	C6
A	N7	C5	N7	C8
U	O2	C2	O2	N3
U	O4	C4	O4	N3
G	O6	C6	O6	N1

[a]The x axis of these frames coincides with the sum of the normalized vectors A1-OO and A2-OO for the N atoms and with the vector A1-OO for the O atoms. The z axis is defined as the vector product of (A1-OO)*(A2-OO). See also Fig. S5.
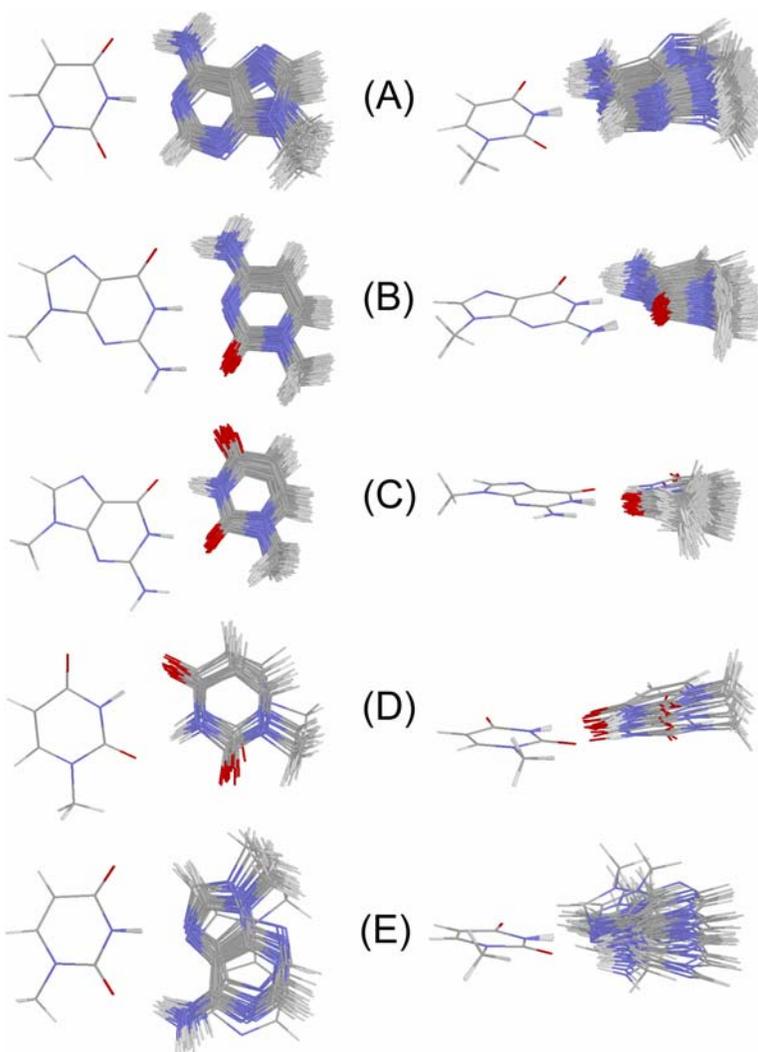
Figure S1. Base pair geometries from the structural database with H positions optimized by the DFT calculations, used as input for parameterization of eq 1. For the five different types of base pairing shown, coordinates of one of the two nucleotides were superimposed, such as to highlight the variation in their relative orientations and translations. Top and side views are shown. (A) U:A Watson-Crick; (B) G:C Watson-Crick; (C) G:U wobble; (D) U-U asymmetric; (E) U:A reverse Hoogstein.
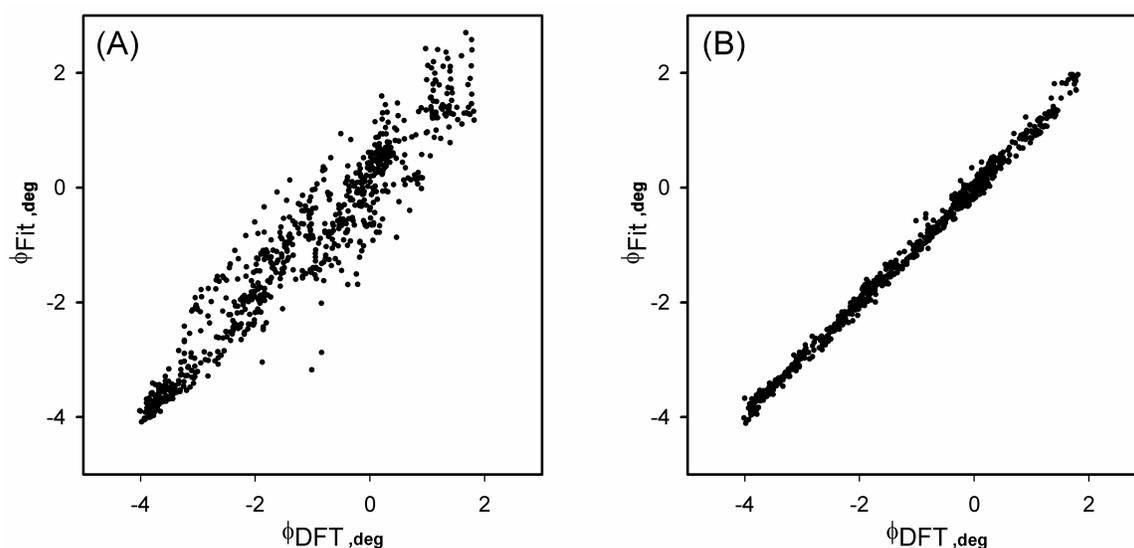
Figure S2. Impact of inclusion of the repulsive H-H term in eq 1 on the fits of the in-plane N-H angle, ϕ, to values obtained by DFT for 390 G:U wobble base pairs which include both points from the database and systematic samples over the grid of inter-nucleotide rotations and translations. Results shown include both G:N1-H1 and U:N3-H3 vectors. (A) Fit without inclusion of the repulsive term. (B) Fit with inclusion of the repulsive term.
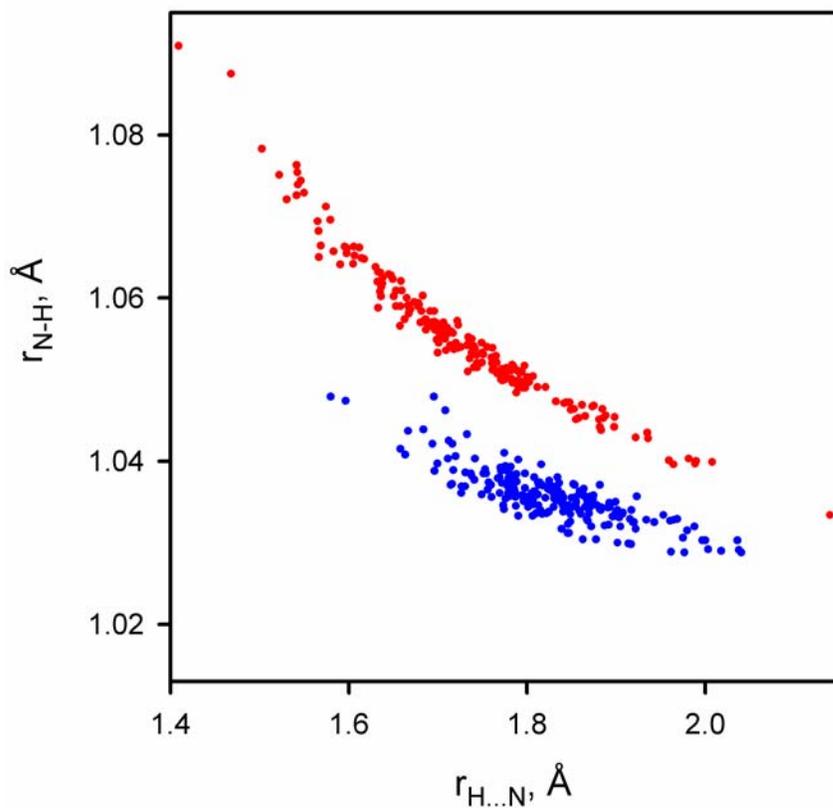
Figure S3. Variation of the imino N-H bond length, r_{N-H}, of the G and U nucleotides within G:C (blue) and U:A (red) Watson-Crick base pairs as a function of the H-bonding distance $r_{H...N}$ (G:H1-C:N3 and U:H3-A:N1), for relative base orientations taken from the database. The lowest value of the vertical axis on this graph (1.0125 Å) corresponds to the DFT-derived N-H bond length in isolated U and G nucleotides.
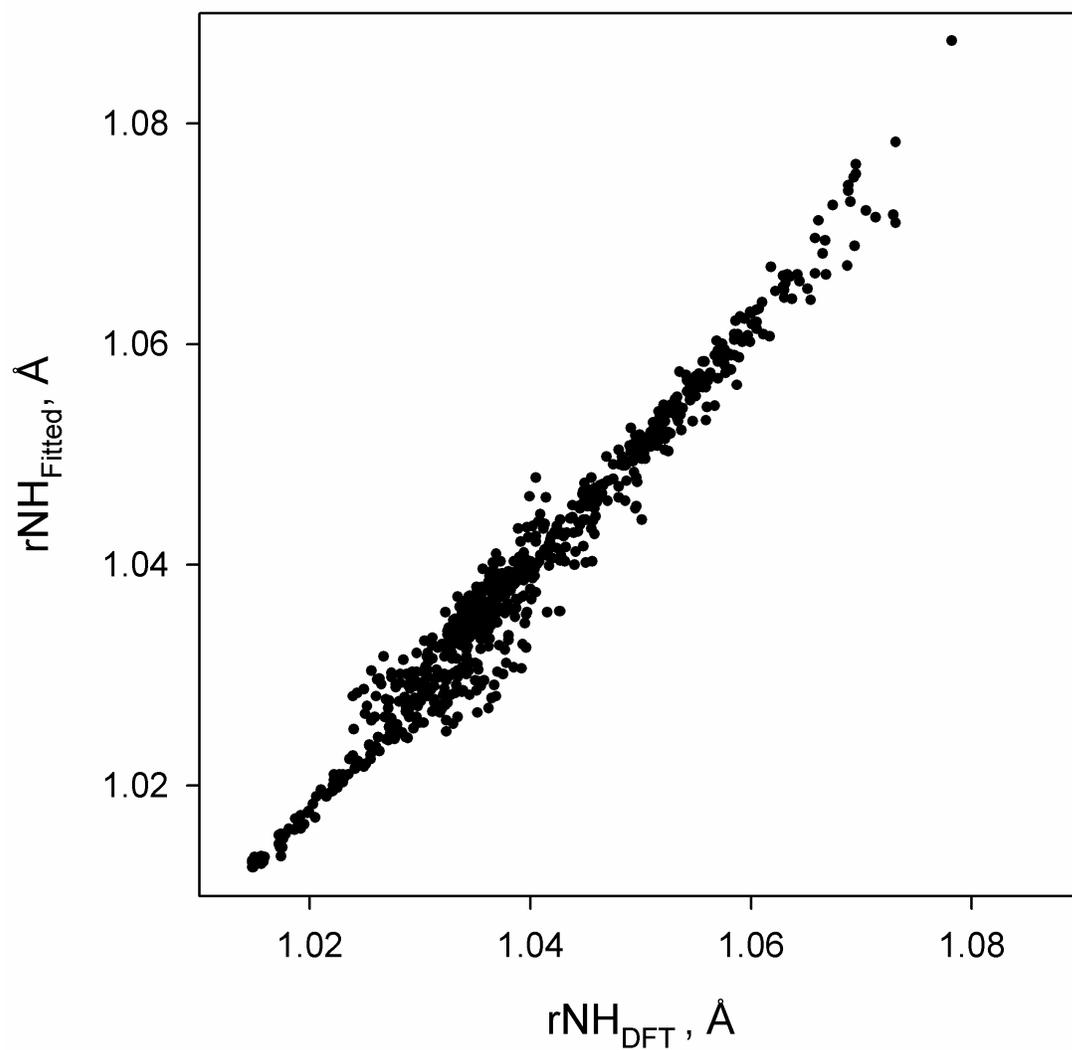
Figure S4. Comparison of the DFT–optimized N-H bond lengths with values obtained from eq 1. Data for all five considered base pair types are shown.
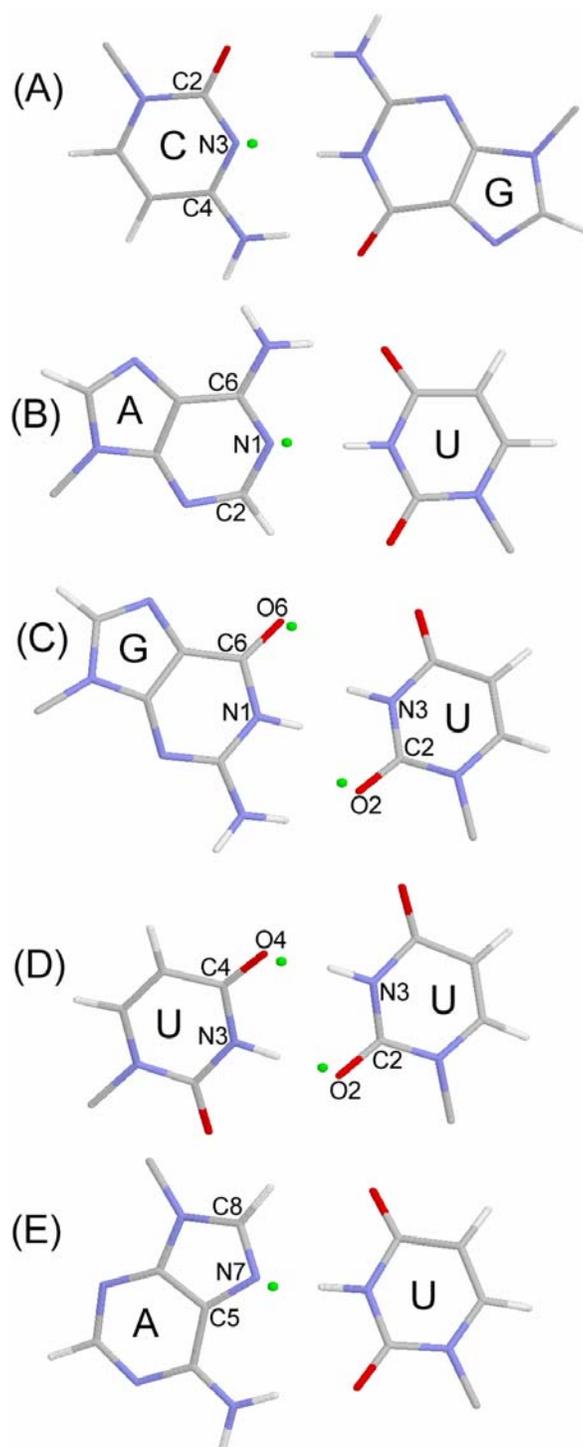
Figure S5. Locations of the virtual attractor points, as obtained from eq 1, are shown as green spheres for the five base pair types considered. The atoms forming the coordinate systems listed in Table S2 are marked. (A) G:C Watson-Crick, (B) U:A Watson-Crick, (C) G:U wobble, (D) U:U asymmetric, (E) A:U reverse Hoogstein.
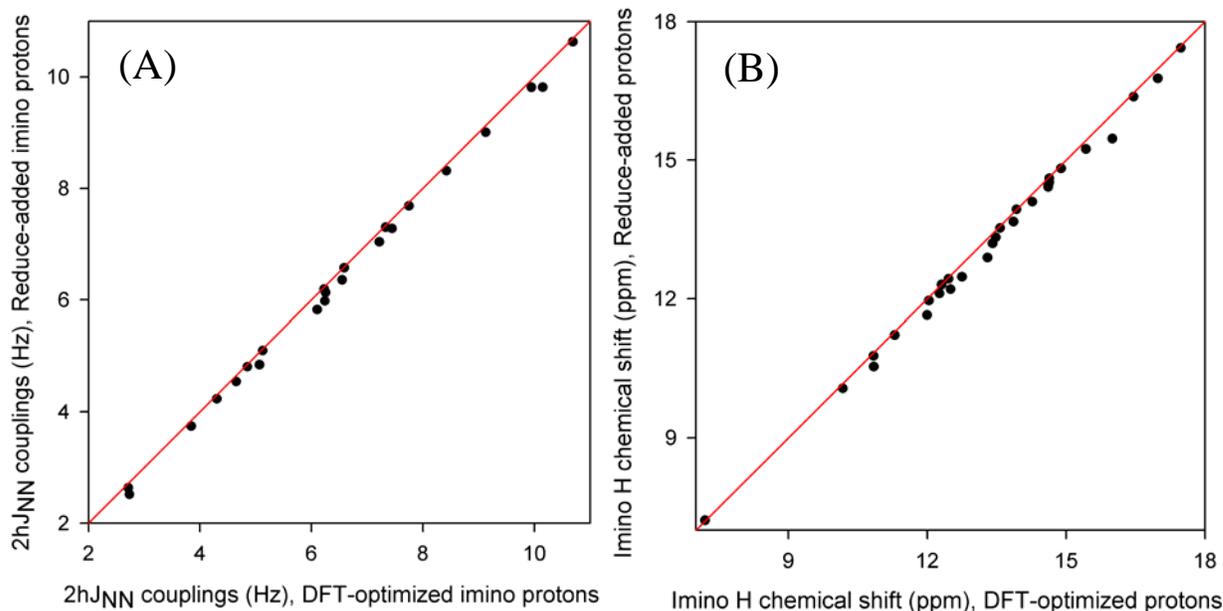
Figure S6. Impact of DFT optimization of imino N-H bond orientation on $^{2h}J_{NN}$ and imino ^{1}H chemical shifts. Correlation plot of (A) DFT-derived values for through-H-bond scalar coupling $^{2h}J_{NN}$ and (B) imino ^{1}H chemical shifts, generated for base-paired nucleotides in RiboA, with protons added either in their idealized positions using the program REDUCE, or in DFT-derived N-H orientations, but with their bond length scaled to the REDUCE value of 1.00 Å.   All the coupling and shielding calculations were carried out using DFT at the B3LYP/6-311++G** level in Gaussian09. The calculated coupling constant is the sum of the Fermi contact, spin-dipolar, paramagnetic spin-orbit, and diamagnetic spin-orbit contributions (Helgaker et al., J. Chem. Phys. 2000, 113, 9402-9409; Sychrovský et al., J. Chem. Phys. 2000, 113, 3530-3547). For the shielding computation, the Gauge-Independent Atomic Orbital (GIAO) method (Ditchfield, R. Mol. Phys. 1974, 27, 789-807) was applied. The ^{1}H chemical shifts reported here are indirectly referenced to the isotropic shielding (31.93 ppm) of tetramethylsilane (TMS), inferred from the calculated isotropic shielding (31.80 ppm) at the same level of theory and the experimental shift (0.13 ppm) of methane. See Barfield et al. J. Am. Chem. Soc. 2001, 123, 4014-4022 and references therein for more details on the procedure of referencing calculated ^{1}H chemical shifts.