J|A|C|S
ARTICLES
Published on Web 01/16/2003

Improving the Accuracy of NMR Structures of RNA by Means of Conformational Database Potentials of Mean Force as Assessed by Complete Dipolar Coupling Cross-Validation

G. Marius Clore*,† and John Kuszewski‡

Contribution from Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0510 and Division of Computational Bioscience, Building 12A, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624

Received September 3, 2002

Abstract: The description of the nonbonded contact terms used in simulated annealing refinement can have a major impact on nucleic acid structures generated from NMR data. Using complete dipolar coupling cross-validation, we demonstrate that substantial improvements in coordinate accuracy of NMR structures of RNA can be obtained by making use of two conformational database potentials of mean force: a nucleic acid torsion angle database potential consisting of various multidimensional torsion angle correlations; and an RNA specific base−base positioning potential that provides a simple geometric, statistically based, description of sequential and nonsequential base-base interactions. The former is based on 416 nucleic acid crystal structures solved at a resolution of ≤2 Å and an *R*-factor ≤25%; the latter is based on 131 RNA crystal structures solved at a resolution of ≤3 Å and an *R*-factor of ≤25%, and includes both the large and small subunits of the ribosome. The application of these two database potentials is illustrated for the structure refinement of an RNA aptamer/theophylline complex for which extensive NOE and residual dipolar coupling data have been measured in solution.

Introduction

NMR structure determination involves seeking the minimum of a target function comprising terms for the experimental NMR restraints, covalent geometry, and nonbonded contacts.[1] The description of the nonbonded contacts can have a significant impact on the accuracy of a NMR structure determination,[2] particularly in the case of nucleic acids where the density of short interproton distances is rather limited.[3] On the basis of the results of cross-validation against independent NMR observables (interproton distance restraints derived from nuclear Overhauser enhancement measurements and dipolar couplings), we recently showed that significant improvements in the accuracy of NMR structures of DNA can be obtained by including both torsion angle and base-base positioning database potentials of mean force in the description of the nonbonded interactions.[3] These statistical potentials, which are derived from high resolution crystal structures, seek to bias sampling during simulated annealing refinement to physically reasonable regions of conformational space within the range of possibilities that are consistent with the experimental NMR restraints.[4,5] Although double stranded DNA can adopt a number of distinct conformations (e.g., A, B, or Z-DNA), interstrand hydrogen-bonding is usually limited to Watson−Crick base pairing, no tertiary interactions are present, and interresidue contacts are generally limited to nucleotides and base pairs adjacent in the linear sequence.[6] Although the local conformation of RNA is typically A-type, RNA can adopt much more complex structures than DNA, including not only a variety of non-Watson−Crick interstrand hydrogen-bonding interactions, but also long-range internucleotide tertiary interactions between nonsequential nucleotides or base pairs.[7−9] As a consequence, the design of the base-base positioning potential employed successfully for DNA[3] in which interactions were limited to linearly sequential intra- and interstrand base-base contacts is not appropriate for RNA. In this paper, we describe a base-base positioning potential of mean force specifically designed for RNA, and demonstrate using complete dipolar coupling cross-validation[10] that the use

* To whom correspondence should be addressed. Phone (301) 496 0782. Fax: (301) 496 0825. E-mail: mariusc@intra.niddk.nih.gov.

† Contribution from Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health.

‡ Division of Computational Bioscience, Building 12A, Center for Information Technology, National Institutes of Health.

(1) Clore, G. M.; Gronenborn, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5891−5898.
(2) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.
(3) Kuszewski, J.; Schwieters, C.; Clore, G. M. *J. Am. Chem. Soc.* **2001**, *123*, 3903−3918.
(4) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *Protein Sci.* **1996**, *5*, 1067−1080.
(5) Clore, G. M.; Schwieters, C. D. *Curr. Op. Struct. Biol.* **2002**, *12*, 146−153.
(6) Berman, H. M.; Gelbin, A.; Westbrook, J. *Prog. Biophys. Mol. Biol.* **1996**, *66*, 255−288
(7) Kallenbach, N. R.; Berman, H. M. *Quart. Rev. Biophys.* **1997**, *10*, 138−236.
(8) Doudna, J. A. *Nature Struct. Biol.* **2000**, *S7*, 954−956.
(9) Mollova, E. T.; Pardi, A. *Curr. Op. Struct. Biol.* **2000**, *10*, 298−302.

10.1021/ja028383j CCC: $25.00

Table 1. Breakdown of Databases Used to Create the Torsion Angle and Base−Base Positioning Potentials of Mean Force

A. nucleic acid torsion angle database potential (resolution ≤2 Å, *R*-factor ≤25%)[a]

	no. of structure
RNA structures	64
A-RNA	24
tRNA	2
protein-RNA	8
drug-RNA	19
unusual RNA	11
DNA structures	332
A-DNA	52
B-DNA	61
Z-DNA	42
drug−DNA	89
protein-DNA	66
unusual DNA	22
DNA/RNA hybrids	20
Total	416

B. RNA specific base-base positioning potential (resolution ≤3 Å, *R*-factor ≤25%)[b]

		no. of valid residue pairs	
	no. of structures	sequential	nonsequential
A-RNA	30	1378	7995
tRNA	9	879	5557
protein-RNA	43	2302	13083
ribosomal protein-RNA	6	8243	64638
drug-RNA	21	461	3075
unusual RNA	22	1664	11002
Total:	131	14927	105350

[a] The torsion angle database potential comprises 26 2D surfaces: α/ϵ_{-1}, α/ζ_{-1}, α/β, α/γ, α/δ, α/ϵ, α/ζ, α/χ, β/ϵ_{-1}, β/ζ_{-1}, β/γ, β/δ, β/ϵ, β/ζ, β/χ, γ/ζ_{-1}, γ/δ, γ/ϵ, γ/ζ, γ/χ, δ/ϵ, δ/ζ, δ/χ, ϵ/ζ, ϵ/χ, and ζ/χ; 8 3D surfaces: $\alpha/\epsilon_{-1}/\zeta_{-1}$, $\zeta/\beta/\zeta_{-1}$, $\alpha/\beta/\gamma$, $\beta/\gamma/\delta$, $\gamma/\delta/\epsilon$, $\delta/\epsilon/\zeta$, $\gamma/\delta/\chi$ and $\delta/\epsilon/\chi$; and 1 4D surface: $\gamma/\delta/\chi/\epsilon$. [b] Includes the 2.4 Å resolution structure of the 50S ribosome subunit[22] and the 3 Å resolution structure of the 30S ribosome subunit.[23]

of these database potential coupled with a torsion angle database potential of mean force leads to a considerable improvement in the accuracy of the NMR structures of an RNA aptamer/theophylline complex for which extensive nuclear Overhauser enhancement (NOE) and residual dipolar coupling data have previously been measured in solution.[11,12]

Methods

Database Potentials. The database potentials are derived from the structures present in the Nucleic Acid Database[13] as of March 2001. The torsion angle and base-base positioning potentials are distributed with Xplor-NIH.[14]

The DELPHIC torsion angle database potential of mean force, E_{deltor}, consists of a set of multidimensional potential surfaces derived from high-resolution crystal structures describing various torsion angle correlations in two-, three-, and four-dimensions (Table 1A).[15] The raw potentials are fitted by sums of multidimensional quartic bell functions as described previously,[15] and E_{deltor} is given by[15a]

$$E_{\text{deltor}} = k_{\text{deltor}} \sum_{i=1}^{i=N} E_{\text{deltor}}(i) \tag{1}$$

where k_{deltor} is a unitless force constant, N is the number of DELPHIC torsion restraints (i.e., the number of torsion angle angle potential surfaces), and $E_{\text{deltor}}(i)$ is the sum of the quartic bell functions fitted to the potential surface appropriate for a particular set of torsion angle correlations. In the case of a two-dimensional surface correlating torsion angles α and β, for example, $E_{\text{deltor}}(i)$ is of the form

$$E_{\text{deltor}}(i) = \sum_{j=1}^{j=Q} \text{torsionQuart}(i,j) \tag{2a}$$

where Q is the number of quartic bells used to fit the raw DELPHIC torsion angle potential of mean force, and

$$\text{torsionQuart}(i,j) = \text{height}(j)\cdot\alpha\text{Frac}(i,j)^2\cdot\beta\text{Frac}(i,j)^2 \tag{2b}$$

where height(j) is the height of a particular fitted quartic bell function j (in kcal·mol^{-1}), and

$$\alpha\text{Frac}(i,j) = \begin{cases} [\alpha_{\text{width}}(j)^2 - \Delta\alpha(i,j)^2]/\alpha_{\text{width}}(j)^2 & \text{if } |\Delta\alpha(i,j)| < \alpha_{\text{width}}(j) \\ 0 & \text{if } |\Delta\alpha(i,j)| \geq \alpha_{\text{width}}(j) \end{cases} \tag{2c}$$

$$\beta\text{Frac}(i,j) = \begin{cases} [\beta_{\text{width}}(j)^2 - \Delta\beta(i,j)^2]/\beta_{\text{width}}(j)^2 & \text{if } |\Delta\beta(i,j)| < \beta_{\text{width}}(j) \\ 0 & \text{if } |\Delta\beta(i,j)| \geq \beta_{\text{width}}(j) \end{cases} \tag{2d}$$

where $\alpha_{\text{width}}(j)$ and $\beta_{\text{width}}(j)$ are the widths of the fitted quartic bell function j along the α and β axes, and $\Delta\alpha(i,j)$ and $\Delta\beta(i,j)$ are the minimal angular distances from the center of the fitted quartic bell function j (along each axis) to the current values of the torsion angles from DELPHIC torsion restraint i.

The raw DELPHIC base-base positioning potentials of mean force are likewise fitted by sums of multidimensional quartic bell functions, and the energy for the DELPHIC base-base positioning potential, E_{delpos}, is given by[3]

$$E_{\text{delpos}} = k_{\text{delpos}} \sum_{i=1}^{i=N} E_{\text{delpos}}(i) \tag{3}$$

where k_{delpos} is a unitless force constant, N is the number of DELPHIC positional restraints (i.e., the number of base-base positional potential surfaces) and $E_{\text{deltor}}(i)$ is the sum of the quartic bell functions fitted to the potential surface type appropriate for the four orienting atoms of restraint i

$$E_{\text{delpos}}(i) = \sum_{i=1}^{j=Q} \text{positionQuart}(i,j) \tag{4a}$$

where Q is the number of quartic bells used to fit the raw DELPHIC positioning potential of mean force, and

$$\text{positionQuart}(i,j) = \text{height}(j)\cdot x\text{Frac}(i,j)^2\cdot y\text{Frac}(i,j)^2\cdot z\text{Frac}(i,j)^2 \tag{4b}$$

where height(j) is the height of a particular fitted quartic bell function j (in kcal·mol^{-1}), and

(10) Clore, G. M.; Garrett, D. S. 1999 *J. Am. Chem. Soc.* **1999**, *121*, 9008−9012.
(11) Zimmerman, G. R.; Jenison, R. D.; Wick, C. L.; Simorre, J.-P.; Pardi, A. *Nature Struct. Biol.* **1997**, *4*, 644−649.
(12) Sibille, N.; Pardi, A.; Simorre, J.-P.; Blackledge, M. *J. Am. Chem. Soc.* **2001**, *123*, 12 135−12 146.
(13) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.-H.; Srinivasan, A. R.; Schneider, B. *Biophys. J.* **1992**, *63*, 751−759; Available on-line at http://www.ndbserver.rutgers.edu/NDB.
(14) Schwieters, C. D.; Kuszewski, J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160*, 66−74; Executables and source code for Xplor-NIH are available on-line at http://nmr.cit.nih.gov/xplor-nih.
(15) (a) Clore, G. M.; Kuszewski, J. *J. Am. Chem. Soc.* **2002**, *124*, 2866−2867. (b) Clore, G. M.; Kuszewski, J. *J. Magn. Reson.* **2000**, *146*, 249−254.

$$x\mathrm{Frac}(i,j) = \begin{cases} [x_{\mathrm{width}}(j)^2 - \Delta x(i,j)^2]/x_{\mathrm{width}}(j)^2 & \text{if } |\Delta x(i,j)| < x_{\mathrm{width}}(j) \\ 0 & \text{if } |\Delta x(i,j)| \geq x_{\mathrm{width}}(j) \end{cases} \quad (4c)$$

$$y\mathrm{Frac}(i,j) = \begin{cases} [y_{\mathrm{width}}(j)^2 - \Delta y(i,j)^2]/y_{\mathrm{width}}(j)^2 & \text{if } |\Delta y(i,j)| < y_{\mathrm{width}}(j) \\ 0 & \text{if } |\Delta y(i,j)| \geq y_{\mathrm{width}}(j) \end{cases} \quad (4d)$$

$$z\mathrm{Frac}(i,j) = \begin{cases} [z_{\mathrm{width}}(j)^2 - \Delta z(i,j)^2]/z_{\mathrm{width}}(j)^2 & \text{if } |\Delta z(i,j)| < z_{\mathrm{width}}(j) \\ 0 & \text{if } |\Delta z(i,j)| \geq z_{\mathrm{width}}(j) \end{cases} \quad (4e)$$

where $x_{\mathrm{width}}(j)$, $y_{\mathrm{width}}(j)$, and $z_{\mathrm{width}}(j)$ are the widths (in Å) of the fitted quartic bell function j along the local x, y, and z axes, respectively, and

$$\Delta x(i,j) = x\mathrm{Pos}(i) - x\mathrm{Cen}(j) \quad (4f)$$

$$\Delta y(i,j) = y\mathrm{Pos}(i) - y\mathrm{Cen}(j)$$

$$\Delta z(i,j) = z\mathrm{Pos}(i) - z\mathrm{Cen}(j)$$

where xCen(j), yCen(j), and zCen(j) are the coordinates of the center of the quartic bell function j, and xPos(i), yPos(i), and zPos(i) are the local, standardized coordinates of the oriented atom I′ of restraint i, which are defined using the global coordinates of the orienting atoms I, J, K (of the first base), and the oriented atom I′ (of the second base) of DELPHIC position restraint i, as described in ref 3.

Simulated Annealing. All simulated annealing calculations were carried out in torsion angle space[16] using the NMR molecular structure determination package Xplor-NIH.[14] In addition to terms for the nonbonded interactions, the target function comprises quadratic square-well potentials for the distance and torsion angle restraints,[17] a harmonic potential for the dipolar couplings,[18] and a harmonic potential for Watson−Crick base pair planarity restraints to prevent undue buckling while allowing propeller twisting to occur[3]. Three main sets of calculations were carried out using three different descriptions of the nonbonded interactions: (i) a quartic van der Waals repulsion term;[17] (ii) A 6−12 Lennard−Jones van der Waals and electrostatic term from the all-hydrogen CHARMM nucleic acid empirical energy function;[19] (iii) A quartic van der Waals repulsion term together with the torsion angle[15] and base−base positioning[3] database potentials of mean force designed for nucleic acids and RNA, respectively. The resulting structures are referred to as ⟨R⟩ , ⟨LJ⟩ , and ⟨R + Db⟩, respectively. In addition, a fourth set of calculations, yielding structures ⟨LJ + Db⟩, was also carried out in which the 6−12 Lennard−Jones and electrostatic potentials were combined with the torsion angle and base-base positioning database potentials.

The quartic van der Waals repulsion term, E_{rep}, is given by[17]

$$E_{\mathrm{rep}} = \begin{cases} 0 & \text{if } r \geq s_{\mathrm{vdw}}.r_{\mathrm{min}} \\ k_{\mathrm{vdw}}(s_{\mathrm{vdw}}.r_{\mathrm{min}}{}^2 - r^2)^2 & \text{if } r < s_{\mathrm{vdw}}.r_{\mathrm{min}} \end{cases} \quad (5)$$

where k_{vdw} is a force constant; r the distance between a pair of atoms; r_{min}, the corresponding sum of the van der Waals radii between the two atoms of the pair; and s_{vdw} a van der Waals radius scale factor (whose optimal value is 0.78) to account for the absence of an attractive component to the potential.

The simulated annealing protocol is similar to that previously described for DNA[3] and comprises three steps: (i) 10 ps of dynamics at 3000 K, in which all nonbonded interactions involving either the quartic van der Waals repulsion term or the Lennard-Jones and electrostatic terms are turned off with the exception of those between C1′ atoms; (ii) 119 cycles of 0.2ps each in which all nonbonded interactions are turned on, the temperature is slowly reduced from 3000 to 25 K, and the force constants for the various terms in the target function are gradually increased to their final values; and (iii) a few cycles of torsion angle minimization. The final values of the force constants are as follows: 1 kcal·mol^{-1}·Hz^{-2} for the dipolar couplings, 30 kcal·mol^{-1}.Å^{-2} for the distance restraints, 200 kcal·mol^{-1}·rad^{-2} for the torsion angle restraints, 20 kcal·mol^{-1}.Å^{-2} for the planarity restraints, except for the end base-pair (G1−C33) where a force constant of 80 kcal·mol^{-1}·Å^{-2} was used; 4 kcal·mol^{-1}·Å^{-4} for the quartic van der Waals repulsion term with a scale factor of 0.78 for the van der Waals radii; 1 for the torsion angle database potential; and 0.3 for the base-base positional database potential. In the case of the calculations with the Lennard−Jones and electrostatic terms, the parameters for these two potentials are left unchanged during the entire course of the calculation. (Note that a $1/r$ screening function is employed for the electrostatics,[20] and the net charge on the phosphate group is reduced to −0.32 e;[21] nonbonded interactions are switched off between 9.5 and 10.5 Å using a cubic switching function, and pairs up to 11.5 Å are included in the nonbonded list).

Results and Discussion

Torsion Angle Database Potential. The torsion angle database potential of mean force comprises a set of multidimensional potential surfaces (26 2D, 8 3D, and 1 4D) describing various torsion angle correlations (see footnote a to Table 1). The raw multidimensional potential surfaces are derived from 416 crystal structures of nucleic acids (64 RNA and 332 DNA) solved at ≤2 Å resolution with an R-factor of ≤25%. The breakdown of structures is shown in Table 1A. The raw potential surfaces, each of which comprise an average of 3207 ± 386 examples, are then fitted by a sum of multidimensional quartic functions,[15a] and these fitted functions are incorporated as a pseudo-potential into the target function for refinement.[15b] There are more DNA structures than RNA ones, but this does not pose a problem since there are numerous representatives in the DNA database whose local backbone structure is similar to RNA. Only structures solved at a resolution of ≤2 Å resolution were employed, since the sugar−phosphate backbone torsion angles, sugar pucker and glycosidic bond torsion angles can only be determined accurately from high-resolution crystallographic data.

RNA Base−Base Positioning Potential. The base-base positioning potential of mean force is derived from 131 RNA crystal structures solved at a resolution of ≤3 Å and an R-factor ≤25% (Table 1B). These database includes both the 2.4 Å resolution structure of the large 50S ribosomal subunit[22] and the 3 Å resolution structure of the small 30S ribosomal subunit,[23] which make up ∼39% and ∼18%, respectively, of all the base-base interactions in the database. Because the bases comprise large rigid planar groups, their positions can still be relatively accurately determined even at comparatively low resolution. The overall position of each base is defined by the coordinates of three orienting atoms (I, J, K) that have been translated and

(16) Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* ***2001****, 152,* 288−302.
(17) Nilges, M.; Gronenborn, A. M.; Brünger, A. T.; Clore, G. M. *Prot. Eng.* **1998**, *2*, 27−38.
(18) Clore, G. M.; Gronenborn, A. M.; Tjandra, N. *J. Magn. Reson.* **1998**, *131*, 159−162.
(19) Nillson, L.; Karplus, M. *J. Comput. Chem.* **1986**, *7*, 691−716.
(20) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1993**, *74*, 187−217.
(21) Nillson, L.; Clore, G. M.; Gronenborn, A. M.; Brünger, A. T.; Karplus, M. *J. Mol. Biol.* **1986**, *188*, 455−475.
(22) Ban, N.; Nissen, P.; Hansen, J.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 905−920.

rotated into a standard geometry;[3] the relative geometry of a second base with respect to the first base is defined by the Cartesian coordinates of its three oriented atoms (I′, J′, K′) to which the same rotations and translations have been applied.[3] (The orienting atoms I, J, and K are N7, N6/O6, and N3 for A/G bases; and C6, N4/O4, and O2 for C/U bases). Thus, the orientation of the second base relative to the first is described by three separate 3D surfaces. The RNA base-base positional potential comprises two components: a set of 96 $[(3 + 3) \times 4^2]$ 3D surfaces representing sequential (i, i ± 1) base−base interactions, and 48 (3×4^2) 3D surfaces representing all nonsequential intra- and interstrand base−base interactions. The standard coordinate space over which the base−base positioning potentials are calculated comprises a 20 Å per side cube with atom J of the first base at the origin, atom I along the negative *x* axis and atom K in the *xy* plane.[3] The average number of examples per 3D surface is 445 ± 157 for the sequential database, and 4732 ± 1608 for the nonsequential one. As in the case of the torsion angle database potential, the raw 3D surfaces are fitted by a sum of three-dimensional quartic functions that are then used in the target function for refinement.

Both the torsion and the base−base positioning potentials deal solely with interactions that are close in space. Consequently, the effects of crystal packing on the global structure of nucleic acids[6] do not in any way decrease the utility of these database potentials in NMR structure determination because the databases are sufficiently large to include all conformations that are likely to exist in solution.

Description of RNA System used to Assess the Impact of the Database Potentials. To assess the impact of the torsion angle and base−base positioning potential on the accuracy of RNA structures determined by NMR, we made use of previously acquired experimental NMR data on an RNA aptamer/theophylline complex solved by Pardi and colleagues.[11,12] This NMR structure has several features that make it ideally suited for the present study. First, the RNA/theophylline complex represents one of the few RNA structures that have been solved on the basis of both extensive NOE-derived interproton distance restraints[11] and $^{13}C-^{1}H$ residual dipolar couplings,[12] thereby permitting the use of dipolar coupling cross-validation as an independent means of assessing accuracy.[10] Second the RNA/theophylline complex contains a range of RNA structural motifs which provide a rigorous test of the database potentials. In addition to the presence of regular A-RNA type stems, the RNA/theophylline complex features non-Watson−Crick base-pairing, the presence of three base triples, a base-zipper, 1−3−2 and interstrand stacking motifs, and an S-turn in the backbone containing a reversed sugar.[11]

Because the "true" solution structure is unknown, accuracy can only be judged by indirect means. The simplest approach, which has been extensively employed in work on proteins, is to compare the calculated NMR structures to an existing high-resolution crystal structure.[1] The agreement between observed and calculated values of NMR observables (such as dipolar couplings, chemical shift anisotropy, chemical shifts and J couplings) is typically excellent for high resolution protein crystal structures, and usually significantly better than for the corresponding NMR structures refined in the absence of these observables.[1] One can therefore conclude that, in general, structures of proteins in the crystal and in solution are very similar, and hence protein crystal structures usually provide a good reference point for judging accuracy.[1] For nucleic acids, however, the situation is far more complex, since it is well-known that crystal packing forces can have a significant impact on global structure.[6] For example, the palyndromic Dickerson DNA dodecamer is asymmetric and kinked in the crystal,[24,25] but symmetric and essentially straight in solution.[3,26] Moreover, in the case of RNA, there are no examples for which both a high-resolution crystal structure has been determined and extensive NMR measurements, including residual dipolar couplings, are available. An alternative approach using cross-validation to assess accuracy must therefore be employed.[10,27,28]

Complete Dipolar Coupling Cross-Validation. Cross-validation is a statistical method in which the structure calculation is carried out omitting a subset of the data (the test set) while refining against the remaining data (the working set).[10,27,28] The quality of the fit and, consequently, the accuracy of the calculated structures are cross-validated by the agreement between the structures and the test set. Thus, cross-validation allows one to determine how well the data in the test set are predicted by structures calculated on the basis of the working data set, and a more accurate structure will predict the test data set better than a less accurate one. The cross-validated free-*R* factor is routinely employed in macromolecular crystallography[27a] and is directly correlated with a model's phase accuracy.[27b] Residual dipolar couplings measured in dilute liquid crystalline media are ideally suited for cross-validation: they provide both local and global orientational information;[29,30] they can be accurately measured with known experimental error;[29] and a dipolar coupling *R*-factor, R_{dip}, which scales between 0% and 100% can be readily calculated (0% representing a perfect fit, and 100% a random orientation of internuclear vectors).[10]

In the case of the RNA/theophylline complex it has been shown that the NOE-derived data is not sufficient to define the overall orientation of the two stems and that this can only be achieved by the incorporation of residual dipolar couplings.[12] As a consequence, one cannot simply calculate a set of structures based solely on NOE data and expect the conformational database potentials to produce any significant improvement in the overall agreement between calculated and observed dipolar couplings. In addition, because each dipolar coupling only contains information relating to an individual interatomic vector, it is insufficient to carry out a set of calculations using only a single working set and test set, as is done in crystallography where each reflection contains information on the entire molecule. One therefore has to resort to complete dipolar coupling cross-validation[10] to assess the impact of the various nonbonded terms on structure accuracy. To this end the dipolar couplings were divided into 10 pairs of working and test data sets chosen at random, and comprising 70% and 30%, respec-

(23) Wimberley, B. T.; Brodersen, D. E.; Clemons, W. M.; Morgan-Warren, R. J.; Carter, A. P.; Vonrhein, C.; Hartsch, T.; Ramakrishnan, V. *Nature* **2000**, *407*, 327−339.
(24) Dickerson, R. E.; Drew, H. R. *J. Mol. Biol.* **1981**, *149*, 761−786.
(25) Shui, X.; McFail-Isom, L.; Hu, G. G.; Williams, L. D. *Biochemistry* **1998**, *37*, 8341−8355.
(26) Tjandra, N.; Tate, S.; Ono, A.; Kainosho, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 6190−6200.
(27) (a) Brünger, A. T. *Nature* **1992**, *355*, 472−475. (b) Brünger, A. T. *Acta Crystallogr.* **1993**, *D49*, 24−36.
(28) Brünger, A. T.; Clore, G. M.; Gronenborn, A. M.; Saffrich, R.; Nilges, M. *Science* **1993**, *261*, 328−331.
(29) Bax, A.; Kontaxis, G.; Tjandra, N. *Methods Enzymol.* **2001**, *339*, 127−174.
(30) Prestegard, J. H.; Kishore, A. I. *Curr. Op. Chem. Biol.* **2001**, *5*, 584−590.

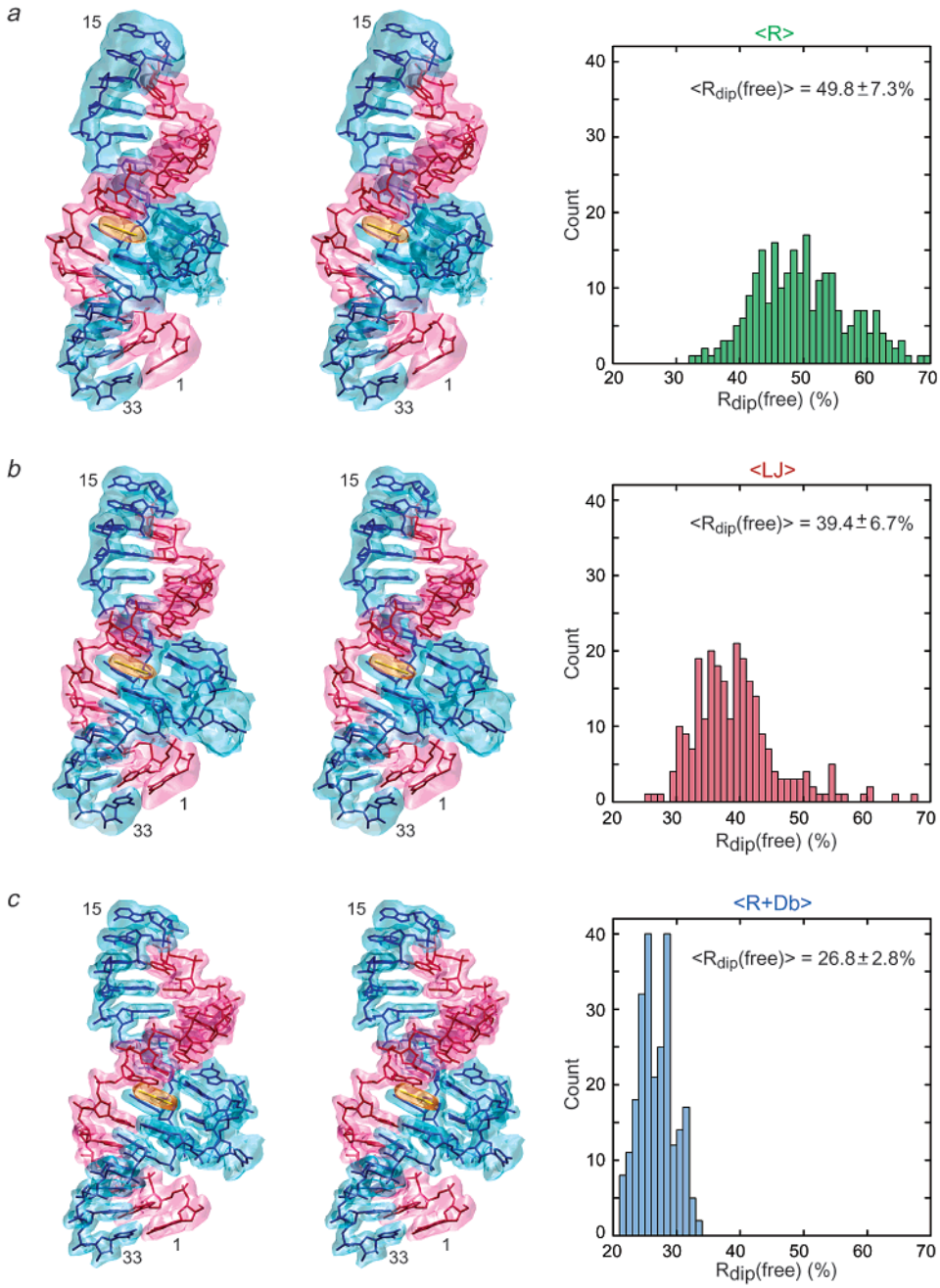

Figure 1. Comparison of the three different descriptions of the nonbonded contacts on the calculated NMR structures and dipolar coupling free *R*-factors, R_{dip}(free), for an RNA aptamer/theophylline complex. (a) Quartic van der Waals repulsion term; (b) Lennard−Jones van der Waals and electrostatic potential terms; and (c) quartic van der Waals repulsion term together with torsion angle and base-base positioning database potentials of mean force. The right-hand panels show histograms of R_{dip}(free) for the three ensembles of 250 simulated annealing structures calculated with complete dipolar coupling cross-validation. The left-hand panels show stereoviews of the corresponding regularized mean coordinates together with an isosurface of the reweighted atomic density map[31] drawn at a value of 15% of maximum calculated from the ensembles of 250 simulated annealing structures using a constant atomic radius of 0.8 Å for all atoms. Nucleotides 1−14 are shown in red, nucleotides 15−33 in blue and the theophylline in yellow. For complete cross-validation, the dipolar couplings were divided into 10 pairs of working and test data sets chosen at random and partitioned in a ratio of 70% (working) to 30% (test). 25 simulated annealing structures were calculated for each working/test pair of dipolar couplings yielding a total of 250 structures. R_{dip}(free) represents the agreement between observed and calculated dipolar couplings for the test set which are *not* used in the structure calculation.

tively, of the data. The working sets are used in refinement, whereas the corresponding test sets are employed only for cross-validation, that is to calculate a dipolar coupling free *R*-factor, R_{dip}(free). 25 simulated annealing structures were calculated for each pair, resulting in a total of 250 structures per calculation.

Results of Structure Calculations. The results of the three sets of simulated annealing calculations are summarized in Figure 1 and Table 2. The agreement with the experimental restraints included in the target function (that is the distance and torsion angle restraints and the working set of dipolar couplings) is broadly comparable for all three ensembles of structures and is consistent with experimental error (Table 2). The ⟨R + Db⟩ structures satisfy the torsion angle restraints somewhat better than the other structures which probably reflects a smoother path to the global minimum region of the target function as a consequence of the introduction of the torsion angle database potential. On the other hand, the dipolar coupling working *R*-factor, R_{dip}(work), is smallest for the ⟨R⟩ structures

Table 2. Structural Statistics[a]

	⟨R⟩	⟨LJ⟩	⟨R + Db⟩
dipolar coupling R-factors (101)[b]			
R_{dip}(free) (%)	50.3 ± 7.4	39.2 ± 6.6	26.8 ± 2.8
R_{dip}(work) (%)	5.4 ± 0.6	6.2 ± 0.6	8.6 ± 0.5
r.m.s. deviation from other experimental restraints			
distances (275) (Å)[c]	0.095 ± 0.010	0.076 ± 0.008	0.089 ± 0.003
torsion angles (110) (°)[d]	0.18 ± 0.11	0.09 ± 0.12	0.01 ± 0.04
coordinate precision (Å)[e]			
all residues	1.81 ± 0.34	1.37 ± 0.26	0.65 ± 0.18
excluding C27	1.69 ± 0.35	1.26 ± 0.25	0.59 ± 0.18
measures of end-to-end length			
R_{gyr} (Å)	15.56 ± 0.37	14.63 ± 0.27	14.95 ± 0.23
$r_{C1'(1)-C1'(15)}$ (Å)	47.2 ± 2.1	43.4 ± 1.6	44.7 ± 1.0
$r_{C1'(33)-C1'(15)}$ (Å)	51.8 ± 1.9	48.1 ± 1.2	47.7 ± 0.9

[a] The notation of the structures is as follows: ⟨*x*⟩ is an ensemble of 250 simulated annealing structures calculated with complete dipolar coupling cross-validation (see footnote b). $\bar{x}$ are the average coordinates derived from each ensemble; $(x)_r$ are the restrained regularized mean coordinates. The nonbonded terms for the three sets of structures are as follows: ⟨R⟩, structures calculated with a quartic van der Waals quartic repulsion term; ⟨LJ⟩ , structures calculated with the Lennard−Jones van der Waals and electrostatic potentials using the all-hydrogen CHARMM TOPNAH1ER1 nucleic acid parameters.[19] ⟨R + Db⟩, structures calculated with the quartic van der Waals repulsion term together with the torsion angle and base-base positioning database potentials of mean force. The number of terms for the various experimental restraints are given in parentheses. [b] There are a total of 101 experimentally measured $^{13}C-^{1}H$ dipolar couplings, comprising 55 dipolar couplings within the sugars (C1′−H1′, C2′−H2′, and C3′−H3′) and 46 within the bases (C8−H8, C6−H6, C5−H5, C2−H2).[12] The dipolar couplings were divided into 10 pairs of working and test data sets chosen at random and partitioned in a ratio of 70% (working) to 30% (test). 25 simulated annealing structures were calculated for each pair, and the results represent the averages obtained for all 250 calculated structures. Note the structures are only refined against the working set of dipolar couplings. The dipolar coupling R-factor is defined as the ratio of the rms deviation between observed and calculated values to the expected rms deviation if the vectors were randomly oriented. The latter is given by $\{2D_a^2[4+3\eta^2]/5\}^{1/2}$, where D_a is the magnitude of the axial component of the alignment tensor and η the rhombicity.[10] The values of D_a and η, obtained from the distribution of dipolar couplings,[33] are −18.7 Hz and 0.15, respectively. [c] There are 223 NOE-derived interproton distance restraints comprising 30 intraresidue, and 86 sequential ($|i-j|=1$), 17 medium ($1<|i-j|<5$) and 90 long ($|i-j|\geq 5$) range interresidue restraints.[11] In addition, there are 52 distance restraints for eight Watson−Crick base pairs and one G−U wobble pair[3]. None of the structures exhibit NOE restraint violations greater than 1 Å. The average number of violations between 0.5 and 1.0 Å is 1.13 ± 0.73 for the ⟨R⟩ structures, 0.77 ± 0.66 for the ⟨LJ⟩ structures, and 0.13 ± 0.33 for the ⟨R + Db⟩ structures. [d] There are 110 loose torsion angle restraints: 31 δ torsion angle restraints derived from ^{3}J coupling constant measurements (with 4 residues, A7, C22, U23, and G26, restrained to a 2′-endo sugar pucker with $\delta = 145 \pm 20°$; and 27 residues restrained to a 3′-endo sugar pucker with $\delta = 80 \pm 20°$; the sugar puckers for U24 and C27 were allowed to float);[11] 33 χ angle restraints (−150 ± 90°) to restrain the glycosidic bond torsion angles to the anti-range;[11] 9 α (−160 ± 50°), 9 β (−70 ± 50°), 10 γ (60 ± 30°), 9 ϵ (−60 ± 40°), and 9 ζ (180 ± 50°) restraints for stem 1 (residues 1−5/29−33). None of the structures exhibit torsion angle violations greater than 5°. [e] The coordinate precision is defined as the average rms difference between the 250 individual simulated annealing structures and the mean coordinates (obtained by averaging the coordinates of the 250 simulated annealing structures best-fitted to each other).

Table 3. Atomic rms Differences between the Regularized Mean Coordinates[a]

	atomic rms difference (Å)[b]		
	$(R+Db)_r$	$(LJ)_r$	$(R)_r$
$(R+Db)_r$		2.27	1.92
$(LJ)_r$	1.89		2.05
$(R)_r$	2.10	1.66	

[a] The regularized mean coordinates are derived from the average coordinates of the 250 simulated annealing structures by restrained regularized minimization and include all the dipolar couplings. [b] Values above the diagonal are for all residues, and below the diagonal exclude C27 which is poorly determined by the experimental NMR restraints.

and largest for the ⟨R + Db⟩ structures. However, because these R_{dip}(work) values correspond to rms differences between observed and calculated dipolar couplings of 1−2 Hz which are comparable to the experimental error, and the measured dipolar couplings span a range of −36 to +26 Hz,[12] the small differences in R_{dip}(work) between the three sets of structures cannot be regarded as significant.

The relative accuracy of the three ensembles of structures is readily assessed by comparison of the dipolar coupling free *R*-factors, R_{dip}(free). It is evident from the distribution of R_{dip}(free) for each ensemble of structures (Figure 1, right-hand panels) that the description of the nonbonded contacts has a dramatic effect on accuracy. R_{dip}(free) has an average value of 49.8 ± 7.3% for the ⟨R⟩ structures, 39.4 ± 6.7% for the ⟨LJ⟩ structures, and 26.8 ± 2.8% for the ⟨R + Db⟩ structures. Thus, the Lennard−Jones and electrostatic terms reduce ⟨R_{dip}(free)⟩ by a factor of ∼1.3 relative to a quartic van der Waals repulsion term alone. The inclusion, however, of the two database potentials in combination with the quartic van der Waals repulsion term reduces ⟨R_{dip}(free)⟩ much further: by a factor of ∼1.5 relative to the Lennard−Jones and electrostatic terms, and ∼2 relative to the quartic van der Waals repulsion term alone. Equally importantly, the distribution of R_{dip}(free) is ∼2.5 to 3 times narrower for the ⟨R + Db⟩ structures than for the ⟨R⟩ and ⟨LJ⟩ structures (Figure 1). Thus R_{dip}(free) values range from 21 to 34% for the ⟨R+Db⟩ structures, from 26 to 67% for the ⟨LJ⟩ structures, and from 33 to 69% for the ⟨R⟩ structures. One can therefore conclude that the inclusion of the torsion angle and base−base positioning database potentials result in a substantial increase in accuracy, as measured by complete dipolar coupling cross-validation.

What does an ⟨R_{dip}(free)⟩ value of 26.8 ± 2.8% observed for the ⟨R+Db⟩ structures relate to in terms of structure quality? The simplest means of providing a qualitative answer to this question is to survey a variety of protein crystal structures for which N−H backbone dipolar couplings have been measured in our laboratory (G. M. C., unpublished data). The measurement error for normalized $^{15}N-^{1}H$ and $^{13}C-^{1}H$ dipolar couplings are comparable, so that the values of R_{dip}^{CH} for the RNA/theophylline complex can be directly compared to those of R_{dip}^{NH} for proteins. R_{dip}^{NH} is found to be correlated to crystallographic resolution, and ranges from ∼15% to ∼27% for protein structures solved at resolutions of 1.5 to 2.5 Å. This suggests, that the ensemble of ⟨R + Db⟩ structures calculated with the torsion angle and base-base positioning database potentials is

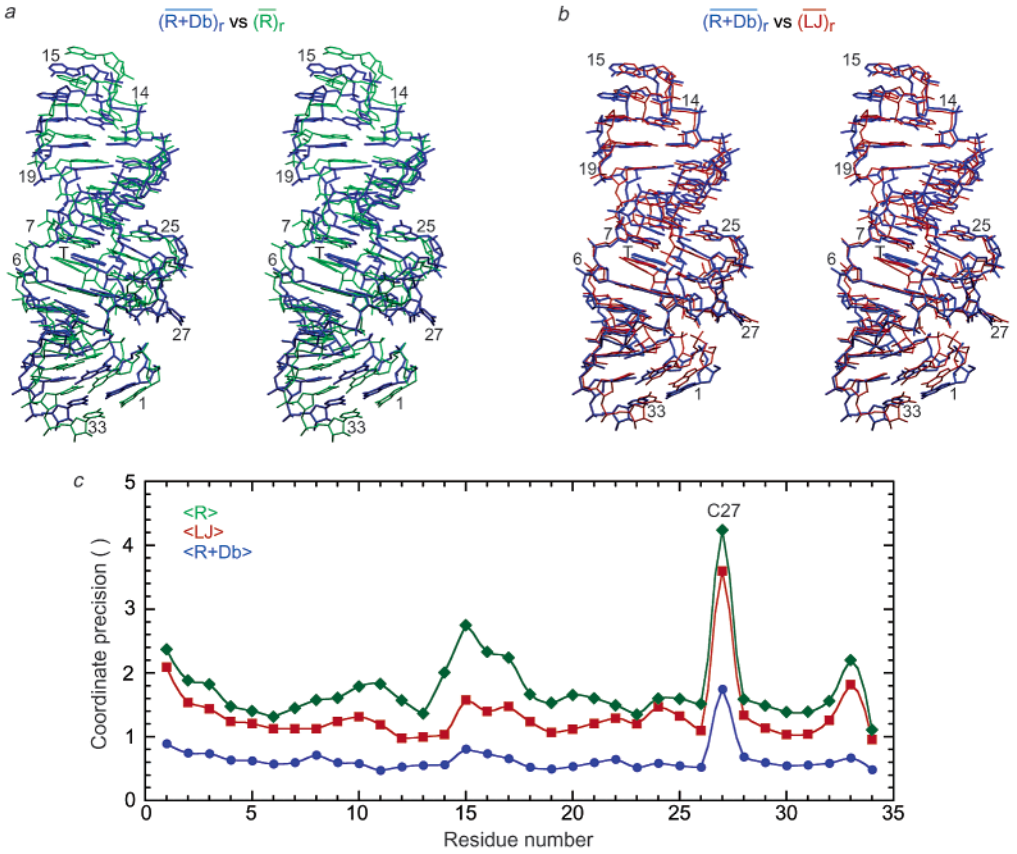

Figure 2. Best-fit superpositions of the restrained regularized mean coordinates: (a) $(R + Db)_r$ versus $(R)_r$; (b) $(R + Db)_r$ versus $(LJ)_r$. (c) Plots of coordinate precision versus residue number for the three ensembles of 250 simulated annealing structures each. Residue 34 is theophylline. The color coding is as follows: blue, $(R + Db)_r$; red, $(LJ)_r$; and green, $(R)_r$.

approximately equivalent in accuracy to a 2.5 Å resolution protein crystal structure.

The left-hand panels of Figure 1 illustrate the conformational space sampled in the three ensembles of structures using an atomic density probability map representation.[31] Best-fit superpositions of the three restrained regularized mean structures are shown in Figure 2a and b, and plots of coordinate precision as a function of residue number are displayed in Figure 2c. The overall topology and RNA fold of the three ensembles of structures are clearly the same. However, excluding C27 which is poorly determined by the experimental NMR restraints, the pairwise atomic rms difference between the $(R)_r$, $(LJ)_r$ and $(R + Db)_r$ restrained regularized mean structures ranges from ~1.7 to 2.1 Å. Thus, there are significant structural differences, both global and local, between the three ensembles of structures. This is also reflected in the overall dimensions of the structures, as measured by both the radius of gyration (R_{gyr}) and the two end-to-end distances, $\langle r_{C1'(1)-C1'(15)} \rangle$ and $\langle r_{C1'(33)-C1'(15)} \rangle$: the ⟨R⟩ ensemble is expanded and the ⟨LJ⟩ one slightly compressed relative to ⟨R + Db⟩. It is also worth noting that, in this instance, the precision of the coordinates (Table 2 and Figure 2c) is directly correlated to R_{dip}(free) (Table 2 and Figure 1). The overall coordinate precision of the ⟨LJ⟩ structures (1.4 ± 0.3 Å) is comparable to that reported for the structures calculated by Sibille et al.[12] using all the dipolar couplings and the Lennard-Jones potential from the AMBER4[32] force field (1.5 ± 0.2 Å). The overall precision of the ⟨R⟩ ensemble is somewhat lower (1.7 ± 0.4 Å), whereas that of the ⟨R + Db⟩ one is significantly higher (0.7 ± 0.2 Å).

A fourth set of calculations was also carried out combining the 6−12 Lennard−Jones and electrostatic potentials with the torsion angle and base−base potentials of mean force. The R_{dip}(free) and coordinate precision of the resulting ensemble of structures, ⟨*LJ* + *Db*⟩, have values of 27.2 ± 3.1% (with a range of 21−34%) and 0.6 ± 0.1 Å, respectively, which are almost identical to the corresponding values for the ⟨R + Db⟩ structures (Table 2). In addition, the atomic rms difference between the $(LJ + Db)_r$ and $(R + Db)_r$ mean coordinates is 0.6 Å which is comparable to the precision of both sets of coordinates. One can therefore conclude that the ⟨*LJ* + *Db*⟩ and ⟨*R* + *Db*⟩ ensembles are essentially the same within coordinate error. Thus, the introduction of the torsion angle and base-base positioning potentials of mean force removes artifactual and systematic distortions arising from conventional descriptions of the non-bonded interactions, either in terms of a simple repulsive potential to prevent atomic overlap or empirical 6−12 Lennard−Jones and electrostatic potentials.

Concluding Remarks

We have shown using complete dipolar coupling cross-validation that, even for an RNA data set comprising quite

(31) Schwieters, C. D.; Clore, G. M. *J. Biomol. NMR.* **2002**, *23*, 221−225.

(32) Pearlman, D. A.; Case, D. A.; Caldwell, J. C.; Seibel, G. L.; Singh, U. C.; Weiner, P.; Kollman, P. A. *AMBER 4.0*, **1991**, University of California, San Francisco.

(33) Clore, G. M.; Gronenborn, A. M.; Bax, A. *J. Magn. Reson.* **1998**, *133*, 216−221.

extensive NOE-derived interproton distance restraints and dipolar couplings, the description of the nonbonded contacts used in the target function for simulated annealing has a large impact on both coordinate accuracy and precision, and local and global structure. A purely repulsive van der Waals term leads to expanded structures of lower precision and accuracy, because on entropic grounds, there are more expanded than compacted configurations that satisfy the experimental restraints. Lennard−Jones van der Waals and electrostatic terms result in some improvement in accuracy, relative to a purely repulsive van der Waals term, but tend to lead to structural compression, presumably because of the attractive component in the Lennard−Jones term. The addition of both torsion angle and base-base positioning potentials of mean force to the description of the nonbonded contacts (either van der Waals repulsion or Lennard−Jones plus electrostatics), however, leads to very substantial improvements in accuracy, as judged by a large decrease in R_{dip}(free) which is accompanied by a concomitant increase in precision. Concomitantly, the introduction of the database potentials obviates systematic distortions associated with particular empirical descriptions of the nonbonded interactions. We therefore conclude that the routine use of the torsion angle and base-base positioning potentials should lead to significant improvements in the accuracy and quality of RNA structures generated from NMR data. In addition, these two database potentials may also be helpful in the refinement of low resolution (>3 Å) crystal structures, in modeling of RNA structures, and possibly in molecular dynamics studies of RNA as well.

Acknowledgment. This work was supported in part by the Intramural AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health (G.M.C.). We thank Art Pardi for kindly providing us with the list of experimental NOE-derived interproton distance and torsion angle restraints for the RNA/theophylline complex that were used in the present study. The coordinates of the restrained regularized mean structure $(R + Db)_r$ have been deposited in the RCSB protein data bank (PDB accession code 1O15).

JA028383J