A Physical Picture of Atomic Motions within the Dickerson DNA Dodecamer in Solution Derived from Joint Ensemble Refinement against NMR and Large-Angle X-ray Scattering Data†

Charles D. Schwieters*,‡ and G. Marius Clore*,§

Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Building 12A, Bethesda, Maryland 20892-5624, and Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Building 5, Bethesda, Maryland 20892-0520

Received September 19, 2006; Revised Manuscript Received November 27, 2006

ABSTRACT: The structure and dynamics of the Dickerson DNA dodecamer [5′d(CGCGAATTCGCG)$_2$] in solution have been investigated by joint simulated annealing refinement against NMR and large-angle X-ray scattering data (extending from 0.25 to 3 Å^{-1}). The NMR data comprise an extensive set of hetero- and homonuclear residual dipolar coupling and ^{31}P chemical shift anisotropy restraints in two alignment media, supplemented by NOE and ^3J coupling data. The NMR and X-ray scattering data cannot be fully ascribed to a single structure representation, indicating the presence of anisotropic motions that impact the experimental observables in different ways. Refinement with ensemble sizes (N_e) of ≥2 to represent the atomic motions reconciles all the experimental data within measurement error. Cross validation against both the dipolar coupling and X-ray scattering data suggests that the optimal ensemble size required to account for the current data is 4. The resulting ensembles permit one to obtain a detailed view of the conformational space sampled by the dodecamer in solution and permit one to analyze fluctuations in helicoidal parameters, sugar puckers, and BI−BII backbone transitions and to obtain quantitative metrics of atomic motion such as generalized order parameters and thermal *B* factors. The calculated order parameters are in good agreement with experimental order parameters obtained from ^{13}C relaxation measurements. Although DNA behaves as a relatively rigid rod with a persistence length of ∼150 bp, dynamic conformational heterogeneity at the base pair level is functionally important since it readily permits optimization of intermolecular protein−DNA interactions.

Although DNA is one of the stiffest polymers known with a persistence length of ∼50 nm, equivalent to ∼150 bp of B-DNA (*1*, *2*), it is well-known that the intrinsic deformability and local flexibility of DNA are integral components of its function (*3*). For example, DNA interacts with numerous transcription factors. To achieve optimal intermolecular contacts, the conformation of the specific DNA target site must be able to adapt itself to the shape of the interaction surface on the interacting protein. In addition to causing localized conformational changes in the DNA with no long-range structural impact (*4*), binding of a protein to DNA may induce both relatively minor degrees of DNA bending (10−20°) as seen with numerous major groove binding proteins and extensive DNA bending (>60°) and/or kinking observed with numerous minor groove architectural binding proteins (e.g., HMG-box proteins, TATA binding protein, and integration host factor) (*5*). In addition, DNA is tightly packaged and supercoiled within chromatin (*6*).

Structural studies of DNA have focused principally on single-crystal X-ray diffraction and solution NMR spectroscopy. Crystallography has not only elucidated the structures of different conformational forms of DNA (B, A, and Z) at the atomic level but also provided evidence of conformational heterogeneity within a single form of DNA (*7*, *8*). The picture afforded by crystallography, however, is static in nature. Solution NMR methods have also provided structural insights into DNA (*5*, *9*, *10*). One approach has sought to derive very precise single structures (precision of ≤0.2 Å) using extensive NMR data, including not only traditional NOE-derived short (<5 Å) interproton distance restraints but also residual dipolar coupling (RDC)[1] and ^{31}P chemical shift anisotropy (CSA) restraints (*11*−*15*). The most recent of these studies suggested that although a very precise structure could be generated that satisfied the experimental data reasonably well, there was evidence of conformational heterogeneity at the level of the deoxyribose sugars since the sugar RDCs could not be fully accounted for within experimental

† This work was supported by the CIT and NIDDK Intramural Research Programs of the National Institutes of Health.

* To whom correspondence should be addressed. C.D.S.: e-mail, charles.schwieters@nih.gov; telephone, (301) 402-4914. G.M.C.: e-mail, mariusc@mail.nih.gov; telephone, (301) 496-0782; fax, (301) 496-0825.

‡ Center for Information Technology.

§ National Institute of Diabetes and Digestive and Kidney Diseases.

[1] Abbreviations: RDC, residual dipolar coupling; CSA, chemical shift anisotropy; SAXS, small-angle X-ray scattering; LAXS, large-angle X-ray scattering; pos-db, base−base positional database potential of mean force; rms, root-mean-square.

10.1021/bi061943x
Published on Web 01/18/2007

error (*15*). Another approach has sought to represent traditional NOE data by an ensemble of structures to account for the presence of internally inconsistent restraints (*16*−*19*). Unfortunately, NOE refinement does not cross-validate beyond an ensemble size of 2 (*20*, *21*).

NOE-derived interproton distance and *J*-coupling-derived torsion-angle restraints provide purely local, short-range (≤5 Å), semiquantitative, structural information (*9*, *22*). RDC data for fixed distance vectors (e.g., bonds) and CSA measurements provide quantitative orientational information relative to the alignment tensor of the orienting medium (*23*, *24*). RDC data for variable distance vectors (e.g., $^{31}P-^{1}H$, as well as $^{1}H-^{1}H$ separated by more than two bonds) yield both distance and orientational information (*23*−*25*). Complementary structural information in solution can also be extracted from solution X-ray scattering intensities, which are sensitive to larger-scale distances (*26*, *27*). Moreover, via measurement of X-ray scattering intensities at large scattering angles, distances as small as the spacing between adjacent base pairs can be reliably probed (*28*, *29*).

Recent work on proteins has shown that a physical picture of backbone atomic motions in solution can be obtained by ensemble refinement against RDCs in multiple alignment media and that the amplitudes of the derived motions are consistent with those obtained from relaxation measurements (*30*−*32*). In this paper, we seek to derive a physical picture of motions in DNA using a similar approach based upon ensemble refinement against NOE, *J* coupling, RDC and CSA NMR data, and solution large-angle X-ray scattering data. The duplex DNA investigated is the Dickerson dodecamer [d(CGCGAATTCGCG)$_2$ (*33*)] which has served as a model system for numerous experimental and theoretical studies and for which extensive solution data are available. In particular, a large amount of high-quality NOE, RDC, CSA, and *J* coupling NMR data have been measured (*12*, *15*), and the large-angle solution X-ray scattering spectrum (*28*) is inconsistent with the structure obtained by refinement against the most NMR data (*15*). Our work not only provides a direct physical picture of the conformational space sampled by the Dickerson DNA dodecamer but also reconciles the NMR and solution X-ray scattering data.

METHODS AND THEORY

Ensemble Refinement. The ensemble refinement capabilities of Xplor-NIH (*34*, *35*) and the potentials used for ensemble refinement against RDCs, NOEs, and *J* couplings have been described previously (*30*, *31*). In this paper, we follow previous work and employ additional restraints derived from ^{31}P CSA measurements (*36*), adapted for ensemble refinement. Given a single structure, the CSA ($\Delta\delta$) is represented as

$$\Delta\delta = \sum_{\alpha,\beta} A_\alpha \sigma_\beta \cos^2(\theta_{\alpha,\beta}) \tag{1}$$

where α and β sum over the three principal moments of the alignment and CSA tensors, respectively. A_α and σ_β represent the values of the corresponding principal moments, while $\theta_{\alpha,\beta}$ is the angle between the associated principal axes. The ensemble averaged CSA is simply

$$\langle\Delta\delta\rangle_e = \sum_{i=1}^{N_e} \Gamma_i \Delta\delta_i \tag{2}$$

where Γ_i and $\Delta\delta_i$ are the weight and CSA for ensemble member i, respectively, and N_e is the ensemble size. Γ_i is normally taken to be $1/N_e$. A harmonic potential term is employed for refinement:

$$E_{CSA} = w_{CSA}(\langle\Delta\delta\rangle_e - \Delta\delta^{obs})^2 \tag{3}$$

where w_{CSA} is a scale factor and $\Delta\delta^{obs}$ is the observed CSA value.

Solution X-ray Scattering Refinement. In this section, we describe our approach for solution X-ray scattering calculations and direct ensemble refinement. Given a plane wave of X-ray radiation incident on a molecule in solution, the scattering amplitude is approximated as

$$A(\mathbf{q}) = \sum_j f_j^{eff}(q) e^{i\mathbf{q}\cdot\mathbf{r}_j} \tag{4}$$

where the sum is over all atoms, $\mathbf{q}$ is the scattering vector in reciprocal space, and $\mathbf{r}_j$ and f_j^{eff} are the position and effective scattering amplitude, respectively, of atom j. The amplitude of the scattering vector $q = |\mathbf{q}|$ is determined by the experimental scattering angle 2θ and the wavelength of the incident radiation, λ:

$$q = 4\pi \sin(\theta)/\lambda \tag{5}$$

where $\theta = 0$ is the forward scattering direction.

In the current work, boundary layer effects are neglected (*37*) so that the effective scattering amplitude can be written as

$$f_j^{eff}(q) = f_j(q) - \rho_s g_j(q) \tag{6}$$

where $f_j(q)$ is the vacuum atomic scattering amplitude, ρ_s is the bulk solvent electron density, and g_j is a scattering factor due to excluded solvent (*38*). The values of $f_j^{eff}(q)$ are precomputed using standard expressions for atomic scattering amplitudes and the solvent scattering factors (*28*, *39*).

In solution, averaging is performed over a reciprocal space solid angle such that the observed intensity is

$$I(q) = \langle |A(\mathbf{q})|^2 \rangle_\Omega \tag{7}$$

where the bracket notation denotes average over solid angle. This average can be expressed in closed form to yield the Debye formula:

$$I(q) = \sum_{i,j} f_i^{eff}(q) f_j^{eff}(q) \operatorname{sinc}(q r_{ij}) \tag{8}$$

where the sum is over all pairs of atoms, r_{ij} is the interatomic distance, and $\operatorname{sinc}(x) = \sin(x)/x$. The ensemble-averaged value of $I(q)$ is $\sum_{i=1}^{N_e} \Gamma_i I_i(q)$, where Γ_i and $I_i(q)$ are the weight and scattering intensity, respectively, of ensemble member i.

For the purposes of refinement, eq 8 is generally too expensive for use in its raw form, as it scales as the square of the number of atoms. To make the computation of $I(q)$ tractable for refinement, we employ two approximations,

including approximating eq 7 by averaging $I(q)$ computed at discrete points on the surface of a sphere, and through judicious use of atom globbing introduced previously (*40*, *41*). Recently, Gabel et al. (*42*) have presented an approach for efficiently incorporating small-angle X-ray scattering data into NMR refinement by expressing $I(q)$ as a power series in q. Such an approach can be quite effective for small angles but is unlikely to work in the current large-angle regime in which $I(q)$ contains multiple peaks and troughs.

From eq 4, one can see that the scattering amplitude due to a group of atoms is linear in the number of atoms. Hence, it makes sense to compute amplitude instead of intensity. The scattering intensity can then be obtained by numerically integrating eq 7. We found that if the points are taken uniformly on the surface of the sphere [for example, via the spiral algorithm (*43*)], relatively few points are required to obtain a good approximation to eq 8. For macromolecules, we found that $I(q)$ is well-represented by tens of points at small scattering amplitudes and up to hundreds of points at the larger values of q sampled in this study. When the number of grid points is not quite large enough, the current method seems to fail gracefully (the resulting error grows slowly with an increasing q), unlike the approach in which the scattering amplitude is expanded in terms of spherical harmonics (*27*).

Additional computational speedup is possible with this approach if we sample $I(q)$ at equally spaced values of q and if the surface grid on which $A(\mathbf{q})$ is evaluated is reused at each value of q. In this case, the contribution of atom j to $A(\mathbf{q})$ is given by

$$f_j^{\text{eff}}(q)e^{i\mathbf{q}\cdot\mathbf{r}_j} = f_j^{\text{eff}}(q)[\exp(i\Delta q\hat{\mathbf{q}}\cdot\mathbf{r}_j)]^n \qquad (9)$$

where $q = n\Delta q$, Δq is the spacing in q, and $\hat{\mathbf{q}}$ is a unit vector in the direction of $\mathbf{q}$. Thus, the exponential needs to be computed only once for each atom (for $n = 1$), and this value is then reused for all other values of q by simple multiplication.

In addition to the finite difference approximation to eq 7, we employ the globbing approximation used by others (*37*, *39*−*41*). In this approximation, the contribution of multiple atoms is approximated by a scattering center at the average atom position (weighted by the number of electrons) with the following scattering amplitude:

$$f_{\text{glob}}(q) = [\sum_{i,j} f_i^{\text{eff}}(q) f_j^{\text{eff}}(q)\ \text{sinc}(qr_{ij})]^{1/2} \qquad (10)$$

where the sum is over all atoms in the glob. We typically use globs consisting of at most three atoms. As in ref *39*, we use a multiplicative q-dependent correction factor c_{glob} to correct for the errors introduced by globbing:

$$I(q) = c_{\text{glob}}(q) I_{\text{glob}}(q) \qquad (11)$$

where $I_{\text{glob}}(q)$ is the scattering intensity obtained using the globbic scattering factors. We calculate c_{glob} using the full Debye expression so that it also corrects errors introduced by the finite grid approximation to $I(q)$.

A harmonic energy potential was used for refinement against solution-phase scattering intensity

$$E_{\text{scat}} = w_{\text{scat}} \sum_j \omega_j [\ln\langle I(q_j)\rangle_e - \ln I^{\text{obs}}(q_j)]^2 \qquad (12)$$

where w_{scat} is an overall scale factor on the energy term, ω_j is a per-q weighting, $I^{\text{obs}}(q_j)$ is the observed scattering intensity, and the sum is over all values of q_j. When experimental errors are available, it is best to set the ω_j equal to the inverse square of the error. For SAXS refinement, the energy is generally taken to be proportional to the χ^2 of intensity, i.e., with a straight linear difference and with ω_j taken to be $1/\text{err}[I^{\text{obs}}(q_j)]^2$ (*39*). In the study presented here, the experimental errors were not available and, more importantly, the region of focus is the large q/small I_q region of the scattering curve (*28*, *29*), so the natural logarithm of the difference was used with uniform error weighting ($w_j = 1$).

Atomic scattering factors and atomic volumes for DNA were obtained from D. M. Tiede (private communication) and are available for general use in the Xplor-NIH package (*34*, *35*). Unlike the case of small-angle scattering, the parametrization of I_q (through f^{eff}) for large-angle scattering is lower in quality such that quantitative agreement of peak intensity is not achieved (*28*, *29*). Peak positions do seem to be reproduced well by eq 8 using the current parametrization. We found that the quality of the $I(q)$ curve is sufficiently high that direct fitting of observed versus calculated $I(q)$ via eq 12 was successful, and we did not need to take the more involved step of fitting peak positions. Solution X-ray scattering data (*28*) were fit to a cubic spline, and two sets of data were then generated with uniform spacing in q. One, with 30 data points (N_q), was used for refinement, while a second set ($N_q = 61$) was used for plotting purposes and to monitor how well the $N_q = 30$ set represented the data. Data were unavailable for $q < 0.25$ Å^{-1}, so data points with smaller q values were given zero weight in refinement. Since the overall scale of the scattering data is unknown, experimental and calculated curves were normalized to their values at the arbitrary point where $q = 0.31$ Å^{-1}.

Potential Terms Used in Refinement. Multiple terms were included in the target function employed for ensemble refinement, including experimental NMR and X-ray scattering (see above) restraints, knowledge-based potentials of mean force, and geometrical terms.

The experimental NMR restraints were as follows. A total of 964 RDC restraints (comprising 16 ^{15}N−^{1}H, 350 ^{13}C−^{1}H, 44 ^{31}P−^{1}H, and 554 ^{1}H−^{1}H RDCs) measured in two alignment media (bicelles and phage pf1) were employed (*15*), using a weighting scheme in which terms were scaled by the inverse square of the experimental error (*30*−*32*). Twenty-two ^{31}P CSA restraints in the same two orienting media provide additional local orientational information for the phosphate group (*15*). Twenty-two measured $^{3}J_{\text{H3'-P}}$ scalar couplings restrain the ϵ torsion angle via a Karplus relationship (*44*). One hundred sixty-two NOE-derived interproton distance restraints (50 intraresidue, 108 sequential, and 4 interstrand) from ref *12* were employed. These restraints were treated as described in ref *13*.

The following potential terms were used to preserve the general global features of the double helix. The α, β, and γ torsion angles were restrained to the range of values observed for right-handed DNA [−70 ± 50°, 180 ± 50°, and 60 ± 35°, respectively (*7*, *13*)] and serve only to prevent local

Table 1: Scaling of Force Constants during Refinement

	initial value	final value
experimental		
RDC (kcal mol^{-1} Hz^{-2})	0.01	1
NOE-derived distances (kcal mol^{-1} Å^{-2})	2	30
J coupling (kcal mol^{-1} Hz^{-2})	10	10
CSA (kcal mol^{-1} ppb^{-2})	0.01	0.2
X-ray scattering (kcal mol^{-1})	4000	4000
knowledge-based		
base pair hydrogen bond (kcal mol^{-1} Å^{-4})	400	1200
torsion-angle database (kcal mol^{-1})	0.2	0.2
base−base positional database (kcal mol^{-1})	0.2	0.2
α, β, and γ torsion angles (kcal mol^{-1} rad^{-2})	200	200
relative atom position (kcal mol^{-1} Å^{-2})	100	100
shape (kcal mol^{-1} Å^{-4})	10	10
orient (kcal mol^{-1} deg^{-2})	2500	25000
quartic vdw nonbonded (kcal mol^{-1} Å^{-4})	0.004	4
vdw radius scale factor	0.9	0.78
bond (kcal mol^{-1} Å^{-2})	400	1000
angle (kcal mol^{-1} rad^{-2})	200	500
improper (kcal mol^{-1} rad^{-2})	50	500

mirror images (*9*, *11*, *13*, *14*). The remaining torsion angles were not explicitly restrained. The latter torsion angles are as follows: δ, correlated to sugar pucker; glycosidic bond torsion angle χ which impacts propeller twist; and ϵ and ζ which are associated with interconversion between BI and BII DNA forms (*45*). Base pair hydrogen bond distance restraints were employed as in ref *13*. Base pair planarity restraints were used to prevent undue buckling while allowing propeller twisting to occur (*13*, *46*).

Two multidimensional knowledge-based potentials of mean force were employed. These potentials were derived from high-resolution (<2 Å) nucleic acid crystal structures and provide a gentle bias toward previously measured structures in cases where direct experimental data are absent (*13*, *14*). The base−base positional database potential relates to the relative positions of neighboring bases (intra- and interstrand) (*13*) and the torsion-angle database potential to the α, β, γ, δ, ϵ, ζ, and χ torsion angles (*14*).

Two types of potential terms were used to explicitly prevent relative motion between ensemble members, as the time scale of the NMR observables used as restraints precludes significant intraensemble rotation. It is important to note that relative rotation of ensemble members is undesirable as these degrees of freedom allow the arbitrary satisfaction of RDC data but with no physical meaning. Further restriction of relative motion is used to improve the convergence of the calculations. In this regard, we note that this study addresses the question of the minimal amount of relative motion among ensemble members required to account for the experimental data. The relative atom position potential (*30*) was applied to restrain phosphorus atoms to lie within 0.5 Å of each other. The Shape potential (*30*) was employed to more directly restrict overall rotation and intraensemble shape changes. The eigenvalues of the shape tensor of each ensemble member were restrained to lie within 1 Å^2 of each other, while the tensor orientations of the ensemble members were restrained to lie within 1° of each other. In addition, the orientations of the axes of the six shape tensors, defined by the heavy atoms of each ensemble member's central 6 bp, were restricted to lie within 10° of each other. This is a relatively soft restraint, and we found no significant changes if we allowed up to 20° of rotation for this energy term.

No explicit potential term was used to maintain the symmetry of the palindromic DNA dodecamer; thus, individual ensemble members were allowed to be asymmetric. However, all the potential terms are symmetric, and the resulting ensembles were closely checked to be symmetric to a good approximation. The usual geometrical covalent bond, angle, and improper terms were employed along with a quartic nonbonded van der Waals repulsion term to prevent atomic overlap (*9*).

Refinement Protocol. One hundred ensemble structures were calculated using a simulated annealing protocol in which dynamics and minimization were performed in torsion-angle space via the internal variable module (IVM) (*47*) of Xplor-NIH (*34*, *35*), employing its variable time step size feature. Atomic masses were set to 100 amu, except for those used to represent alignment tensors which were set to 300 amu. The structure determination protocol consisted of dynamics for 10 ps at 3000 K, followed by annealing from 3000 to 25 K in 25 K increments, with dynamics for 0.2 ps at each temperature. Final gradient minimization was performed in torsion-angle space, followed by all degrees of freedom minimization in Cartesian coordinates. The degrees of freedom for the alignment tensors were strictly restricted to physical values using the IVM throughout. Because a small percentage of the ensembles were found to not converge, only the 50 lowest-energy ensembles were used for analysis.

Force constants for the various potential terms were either scaled geometrically during refinement or held constant, while the atomic radius used in the nonbonded interaction term was scaled down such that the initial energy surface is smoothed (*9*, *48*−*50*). Additionally, the globic scattering correction c_{glob} was computed at each temperature, before dynamics. The values of the force constants for the various potential terms were chosen relative to the final value of 1000 kcal mol^{-1} Å^{-2} for the bonding potential such that the associated restraints were maximally satisfied without adversely affecting other potential terms (*9*, *11*, *13*, *48*, *49*).

Table 2: Structural Statistics for Experimental Restraints

	rms deviations between experimental and calculated values			
	$N_e = 1$	$N_e = 2$	$N_e = 4$	$N_e = 8$
RDC (Hz)[a]	0.76 ± 0.01	0.47 ± 0.01	0.44 ± 0.01	0.44 ± 0.01
CSA (ppb)	2.40 ± 0.13	1.77 ± 0.34	1.62 ± 0.23	1.78 ± 0.36
$^3J_{H3'-P}$ (Hz)	0.60 ± 0.04	0.19 ± 0.02	0.15 ± 0.02	0.13 ± 0.02
NOE (Å)	0.07 ± 0.004	0.04 ± 0.01	0.01 ± 0.005	0.01 ± 0.003
X-ray scattering[b]	0.27 ± 0.01	0.18 ± 0.02	0.22 ± 0.02	0.21 ± 0.02

[a] The weighted RDC rms deviation is defined in eq 13. [b] The values for the X-ray scattering rms deviations are normalized values and therefore unitless.

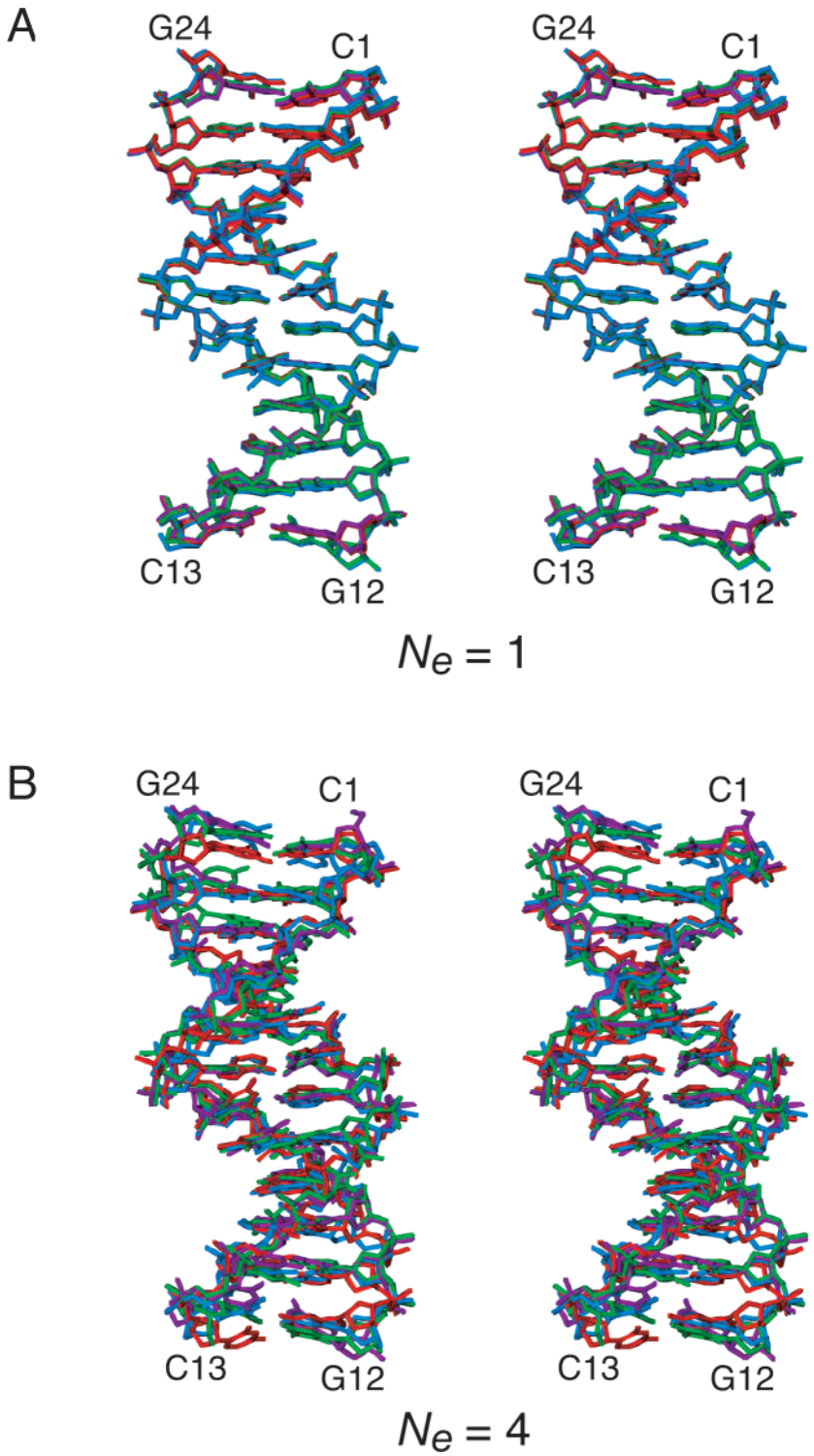

FIGURE 1: Stereoviews illustrating a comparison of (A) a best-fit superposition of four representative $N_e = 1$ structures with (B) a single representative $N_e = 4$ ensemble.

The values used in refinement are listed in Table 1.

RESULTS AND DISCUSSION

Impact of Ensemble Size on Refinement. Refinement against experimental NMR and X-ray scattering restraints was carried out for ensemble sizes N_e ranging from 1 to 8. A summary of the structural statistics is provided in Table 2. The agreement with the experimental data for $N_e = 1$ lies outside the experimental error. Large improvements for the experimental terms are seen as the ensemble size is increased to 2, and thereafter, smaller improvements in most of the experimental restraints are observed. Four representative structures calculated with $N_e = 1$ are compared to a single representative $N_e = 4$ ensemble in Figure 1. It is readily apparent that the $N_e = 1$ structures (Figure 1A) are highly precise, whereas the conformational space sampled by the $N_e = 4$ ensemble (Figure 1B) is significantly larger (see also Table 3). However, if one considers the atomic rms difference between the individual ensemble means and the overall mean (i.e., more than 50 independent calculations), these values

Table 3: Conformational Space Sampled and Coordinate Precision as a Function of Ensemble Size

N_e	intraensemble conformational space sampled (Å)[a]	interensemble coordinate precision[b] (Å)
1	–	0.35 ± 0.17
2	0.69 ± 0.05	0.52 ± 0.12
4	0.82 ± 0.05	0.47 ± 0.08
8	1.11 ± 0.13	0.67 ± 0.22

[a] The conformational space sampled by members within a given ensemble is defined as the average rms of the ensemble members to the unregularized ensemble average. [b] The interensemble coordinate precision measures the conformational space sampled by the ensemble means from 50 independent calculations. When $N_e = 1$, this is simply the average atomic rms difference between the individual structures and the restrained regularized mean. When $N_e \geq 2$, this is given by the average atomic rms difference between the regularized mean for each ensemble and the overall regularized mean.

Table 4: Heavy-Atom Structural Comparison of Ensemble-Averaged Structures with Those of Previous Studies[a]

	rms difference (Å)						
	$N_e = 2$	$N_e = 4$	$N_e = 8$	1GIP	1NAJ	1DUF	171D
$N_e = 1$	0.85	1.14	1.23	1.46	1.31	1.66	3.37
$N_e = 2$		0.54	0.90	1.04	1.21	1.34	2.84
$N_e = 4$			0.65	1.10	1.47	1.44	2.95
$N_e = 8$				1.36	1.72	1.74	3.36
1GIP					0.98	0.84	2.65
1NAJ						0.76	2.66
1DUF							2.30

[a] PDB entries 1GIP (*13*), 1NAJ (*15*), 1DUF (*12*), and 171D (*52*). 1GIP and 1DUF are calculated with the same set of NOE, RDC ($^1D_{CH}$, $^1D_{NH}$, and D_{HH} in one alignment medium), and $^3J_{H3'-P}$ restraints but differ in the representation of the nonbonded interactions. 1GIP employs a simple repulsive van der Waals term coupled with the base−base positional database potential of mean force and the multidimensional torsion-angle database potential of mean force (*13*), while 1DUF uses Lennard-Jones van der Waals and electrostatic potentials (*12*). 1NAJ and $N_e = 1$ structures are calculated using the same protocols that were used for 1DUF and 1GIP, respectively, except that the $N_e = 1$ structure includes the X-ray scattering term, and both 1NAJ and $N_e = 1$ structures make use of many more RDC (including $D_{PH3'}$ RDCs) and ^{31}P CSA restraints in two alignment media (*15*). 171D is an older structure based solely on NOE data.

are actually quite comparable to the coordinate precision for $N_e = 1$ ensembles (see Table 3).

Average structures were calculated by restrained regularization of the ensemble average structures for each value of N_e in the presence of experimental restraints (*13*, *48*, *49*, *51*). A comparison of the restrained regularized average structures for cases where $N_e = 1$ and $N_e = 4$ is shown in Figure 2. While these two structures are quite similar with an atomic rms difference of 1.1 Å, the $N_e = 1$ average structure is clearly somewhat compressed relative to the $N_e = 4$ average structure. The $N_e = 2$ and 4 overall average structures are most similar with a heavy atom rms difference of 0.54 Å, followed by the $N_e = 4$ and 8 overall average structures which differ by 0.65 Å (Table 4).

A complete comparison of the heavy atom rms differences between the overall average $N_e = 1$, 2, 4, and 8 structures and four previously published NMR structures is provided in Table 4. The latter are as follows. 1GIP (*13*) and 1DUF (*12*) are based on the same NOE, RDC ($^1D_{NH}$, $^1D_{CH}$, and D_{HH} in one alignment medium), and $^3J_{H3'-P}$ coupling restraints but differ in terms of the representation of the nonbonded contacts (see below). 1NAJ (*15*) is based on the same experimental restraints used in these calculations that include not only the experimental data used for 1GIP and 1DUF but also many more RDCs, including $D_{PH3'}$ RDCs, and ^{31}P CSA restraints from two alignment media, and makes use of the same representation of the nonbonded interactions as 1DUF; and 171D (*52*), based on NOE data only. The nonbonded interactions for 1GIP and the calculations presented here are represented by a repulsive van der Waals term together with multidimensional torsion-angle and base−base positional database potentials of mean force (*13*); 1DUF (*12*) and 1NAJ (*15*), on the other hand, make use of Lennard-Jones, van der Waals, and electrostatic terms from the CHARMM empirical energy potential (*53*−*55*). The $N_e = 1$ average structure is most similar to 1NAJ, which is not surprising since the same RDC and CSA data were employed in the calculation of the 1NAJ structure. However, of the

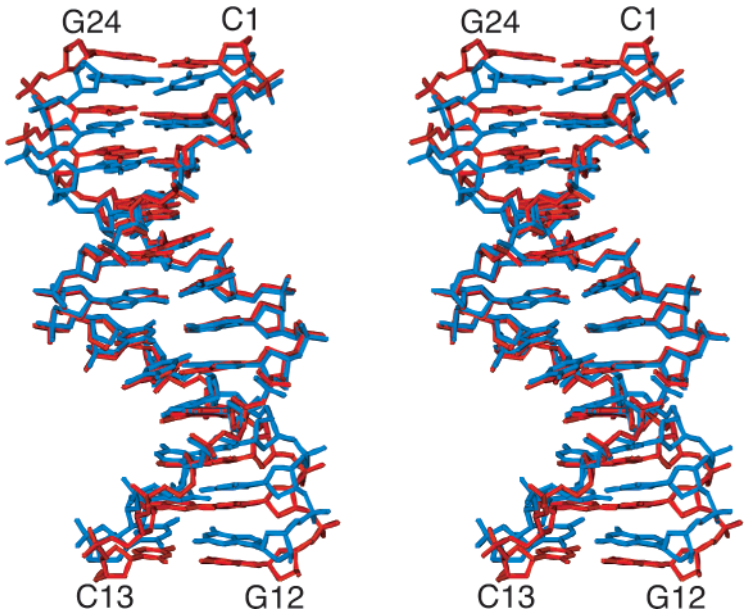

FIGURE 2: Stereoview showing a best-fit superposition of the regularized mean $N_e = 1$ (blue) and $N_e = 4$ (red) structures. For $N_e = 4$, this structure is derived from the average ensemble structures for 50 ensembles.

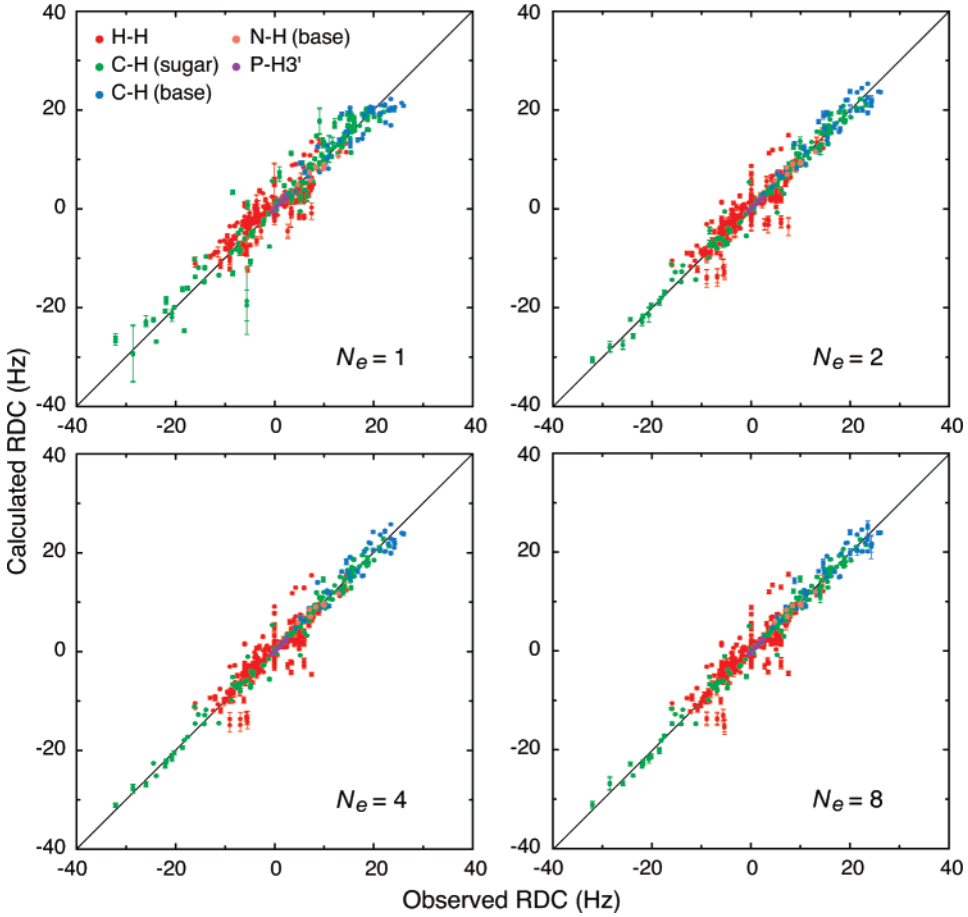

FIGURE 3: Correlation between observed and calculated RDCs for the top 50 ensembles for ensemble sizes of 1, 2, 4, and 8. The different dipolar couplings are color-coded. Error bars represent the deviation among the 50 ensembles.

previously determined structures, the $N_e \geq 2$ ensemble average structures correspond most closely to 1GIP (*13*), which, of the previously determined structures, also most closely fits the solution X-ray scattering data (*28*). This suggests that the additional $^{31}P-H3'$ RDC and ^{31}P CSA data used for both 1NAJ (*15*) and the current study bias the single $N_e = 1$ structure away from the solution X-ray scattering data, and it is only by introducing an ensemble representation ($N_e \geq 2$) that these data can be reconciled. This observation provides direct evidence of the presence of significant degrees of anisotropic motion that impacts the various experimental restraints differentially because of their different averaging properties.

RDCs, Ensemble Size, and Cross Validation. Figure 3 shows the RDC correlation between observed and calculated values for ensemble sizes of 1, 2, 4, and 8. It is seen that there is visible improvement in the fit when going from $N_e = 1$ to $N_e = 4$. In particular, when $N_e = 1$, the dipolar couplings associated with the sugar atoms are least well fit and display the largest amount of dispersion.

To quantitatively assess the agreement between all observed and calculated RDCs, we make use of the following weighted RDC rms deviation given by

$$\text{rmsd}^2 = \sum_k N_k \sigma_k^{-2} \text{rmsd}_k^2 \left(\sum_k N_k \sigma_k^{-2}\right)^{-1} \qquad (13)$$

where N_k is the number of restraints in a particular RDC class, σ_k is an estimate of experimental error in that class, rmsd_k is the unweighted deviation for the class, and the sum is over all RDC classes. A plot of the weighted RDC rmsd as a function of N_e (Figure 4) clearly shows that the agreement improves dramatically with an increase in N_e from 1 to 2, with very minor improvements with further increases in N_e. The question of the optimal size of N_e therefore arises.

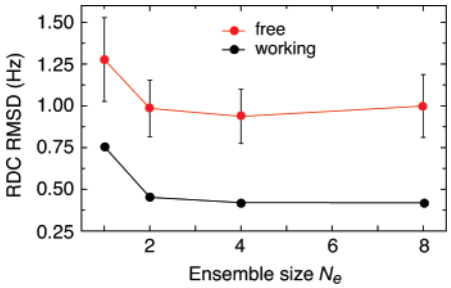

FIGURE 4: Dependence of the weighted RDC rmsd on ensemble size for working (included in the refinement, black) and free (cross-validated, red) RDCs. Cross validation was carried out by omitting 10 sets of 50 randomly chosen RDCs (of a total of 964). Error bars denote deviations across the 10 data sets.

To assess for overfitting and determine the optimal size of N_e, complete cross validation was performed (*56*) using 10 sets of refinement calculations for each ensemble size ($N_e = 1$, 2, 4, and 8), in which 50 (approximately 5%) randomly chosen RDC restraints were removed from the set used for refinement. The resulting weighted rmsd for the free RDCs (i.e., those not included in the calculations) is shown in Figure 4 which reveals a shallow minimum at $N_e = 4$. These results agree well with the results of full refinement and suggest that the $N_e = 4$ representation is the

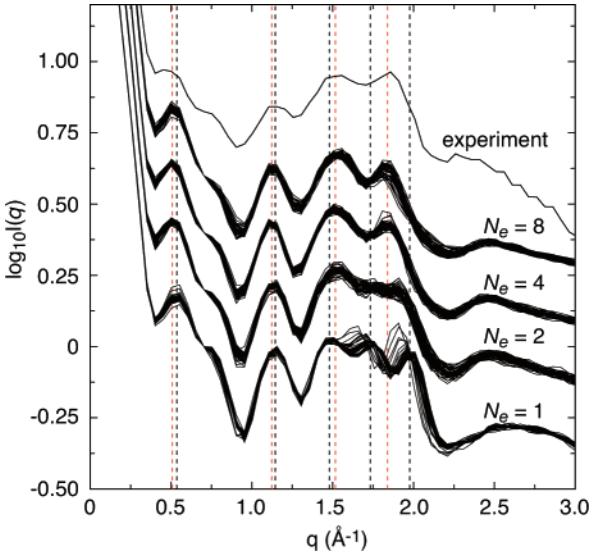

FIGURE 5: Comparison of experimental and calculated solution X-ray scattering curves. Curves from the best 50 ensembles for N_e values of 1, 2, 4, and 8 are displayed with an offset from the experimental scattering curve. Black and red vertical dashed lines represent the average peak positions for $N_e = 1$ and 4, respectively. A single, unconverged $N_e = 8$ curve was omitted from the plot.

most appropriate, given the current experimental data set. It is also interesting to note that the $N_e = 4$ calculations exhibit higher interensemble coordinate precision than the $N_e = 2$ and 8 calculations and that the standard deviation in the value of the precision for the $N_e = 4$ calculations is the smallest of any of the calculations (cf. Table 3).

Solution X-ray Scattering Results. A comparison of experimental solution X-ray scattering curves with those calculated from the various calculated ensembles is shown in Figure 5, and a comparison of peak positions is provided in Table 5. It should be noted that peak intensities are not quantitatively reproduced as discussed in Methods and Theory and seen in previous work (*28*, *29*). The experimental data display four major peaks at 0.48, 1.14, 1.54, and 1.87 Å^{-1}. Previous work (*28*, *29*) has demonstrated that the

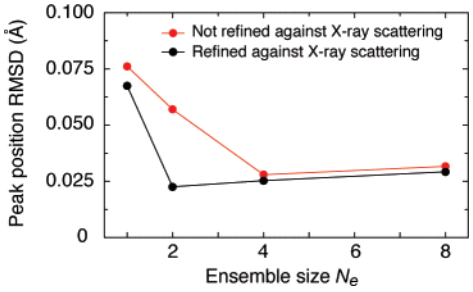

FIGURE 6: Comparison of the solution X-ray scattering peak position rms difference between observed and calculated values for $N_e = 1$, 2, 4, and 8 ensembles calculated with (black) and without (red) the X-ray scattering potential term in the refinement target function. For the $N_e = 1$ structures obtained without the X-ray scattering term, the first peak is absent and is therefore excluded from the rms deviation calculation for that point.

position of the final peak is due to interference between neighboring base pairs, and consequently, the solution X-ray scattering potential term provides an important restraint on the compression and/or extension of the DNA dodecamer. The curves for the $N_e = 1$ structures clearly contain an additional incorrect peak at 1.73 Å^{-1}, and the positions of the other four peaks exhibit significant deviations from the experimentally observed positions, particularly for the 1.87 Å^{-1} peak (Figure 5). When $N_e = 2$, an incorrect peak at 1.74 Å^{-1} is still present, albeit less prominent than that when $N_e = 1$, and the positions of the other peaks improve somewhat. When $N_e = 4$ and 8, the experimentally observed and calculated peak positions match up extremely well.

We also performed full refinement calculations without the solution X-ray scattering term for comparison and cross-validation purposes, and the resulting peak positions are included in Table 5. When $N_e = 4$, 20% of the resulting ensembles contain the spurious fourth peak, but the positions of the other peaks compare favorably with experiment. A plot of peak position rms deviation to experiment for the four major peaks, obtained from calculations both including and excluding the solution X-ray scattering term in refinement, is shown in Figure 6. The minimum for the cross-

Table 5: Peak Positions of X-ray Scattering Curves[a]

structure	peak position (Å^{-1})				
experimental[b]	0.48	1.14	1.54	–	1.87
ensemble refinement with the X-ray scattering term					
$N_e = 1$	0.54 ± 0.00	1.14 ± 0.00	1.48 ± 0.00 (46)	1.73 ± 0.034	1.98 ± 0.023
$N_e = 2$	0.51 ± 0.01	1.14 ± 0.01	1.54 ± 0.01	1.74 ± 0.02 (25)	1.90 ± 0.03 (36)
$N_e = 4$	0.51 ± 0.00	1.12 ± 0.00	1.52 ± 0.02	–	1.84 ± 0.01
$N_e = 8$	0.52 ± 0.00	1.13 ± 0.01	1.54 ± 0.02	1.68 ± 0.00 (1)	1.83 ± 0.01 (49)
ensemble refinement without the scattering term[c]					
$N_e = 1x$	–	1.11 ± 0.00	1.46 ± 0.00	1.73 ± 0.00	1.97 ± 0.00
$N_e = 2x$	0.49 ± 0.01 (27)	1.11 ± 0.01	1.47 ± 0.02	1.74 ± 0.02 (41)	1.95 ± 0.02
$N_e = 4x$	0.49 ± 0.01 (48)	1.11 ± 0.01	1.49 ± 0.01	1.73 ± 0.03 (10)	1.88 ± 0.02 (48)
$N_e = 8x$	0.52 ± 0.01	1.12 ± 0.01	1.52 ± 0.02	1.77 ± 0.00 (1)	1.83 ± 0.01 (35)
published NMR structures[d]					
1GIP	0.47 ± 0.03	1.11 ± 0.01	1.50 ± 0.00	–	1.85 ± 0.00
1NAJ	–	1.10 ± 0.00	1.43 ± 0.00	1.74 ± 0.01	1.97 ± 0.01
1DUF	–	1.08 ± 0.02	1.41 ± 0.01	1.70 ± 0.02	1.92 ± 0.02
171D	–	1.04	1.37	1.70	1.90

[a] The deviations represent the standard deviations between different ensembles and in the case of the PDB entries the standard deviations between different models deposited in that entry. The numbers in parentheses denote the number of ensembles (out of 50) which contain this peak. If there is no number, then all 50 ensembles have this peak. [b] From ref *28*. [c] The *x* denotes ensemble calculations performed without the solution X-ray scattering term. [d] PDB entries 1GIP (*13*), 1NAJ (*15*), 1DUF (*12*), and 171D (*52*).

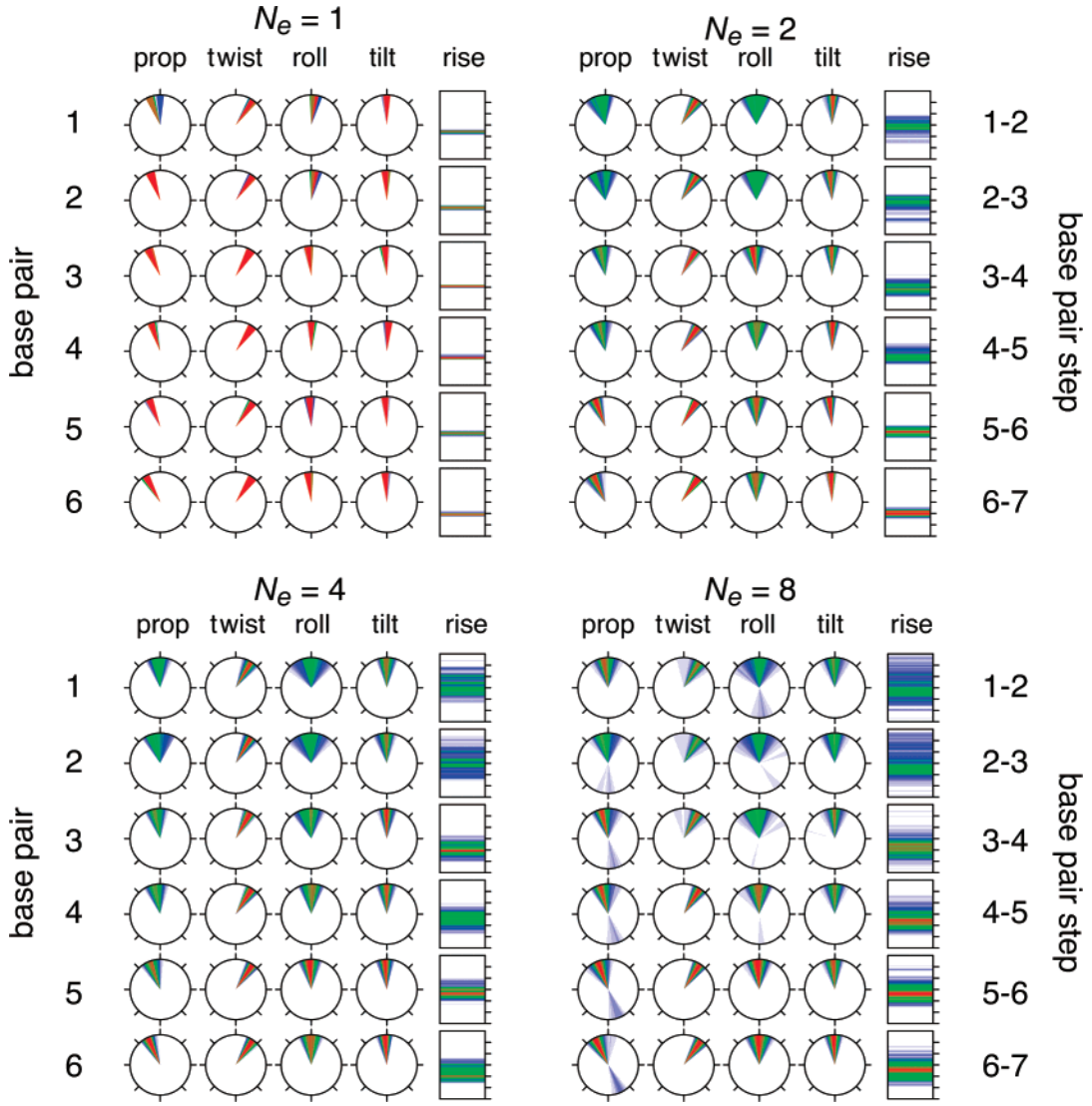

FIGURE 7: Distribution of helicoidal parameters describing the dodecamer double helix as a function of ensemble size. Angular parameters are plotted on clockfaces where the top corresponds to 0°, the right side to 90°, and the bottom to 180°. Helical rise is plotted on a bar chart with ticks at 1 Å intervals, with the bottom being at 1 Å. The figure was generated by accumulating the output from the Curves program (*57*) from each of the structures of the top 50 ensembles for N_e values of 1, 2, 4, and 8. The highest probability is denoted with red, followed by green and then blue. Abbreviations: prop, propeller twist; twist, helical twist.

validated curve is at $N_e = 4$, consistent with the results obtained from RDC cross validation (Figure 4). All the $N_e = 1$ structures calculated without the X-ray scattering term ($N_e = 1x$ in Table 5) are missing the first peak and contain the spurious fourth peak. However, when $N_e = 4$ and 8, good cross validation with the solution X-ray scattering data is obtained. Thus, if the ensembles with spurious and/or missing peaks are filtered out, $N_e \geq 4$ refinement using RDC and other NMR restraints produces structures that are fully consistent with the solution X-ray scattering data. Note that the converse, namely, cross validation of RDC and CSA restraints against the X-ray scattering data, is not expected since the former are very sensitive to local bond vector orientations and the latter reflects global overall structure.

Structural Analysis of Helical Parameters. A detailed analysis of helical properties was obtained using the Curves program (*57*), and the results from the top 50 ensembles for each ensemble size are displayed in Figure 7. The average helical twist and rise for the $N_e = 4$ ensemble calculations are provided in Table 6 (values for the other ensemble sizes are given in the Supporting Information). Relatively small changes in the ensemble mean values of the helical twist, roll, and tilt parameters, all of which are close to those of canonical B-DNA, are seen along the nucleotide sequence. The first two base pair steps are a little underwound which is not unexpected, but the mean helical twist angles for the central 6 bp steps are within ±2° of 36°, characteristic of B-DNA. The ensemble mean base pair tilt and roll angles are very close to 0°, with the exception of the roll angle for the central base pair step which is somewhat larger (mean of ca. −9°). Mean propeller twist exhibits a trend toward

Table 6: Average Values of Helical Rise and Twist for the $N_e = 4$ Ensemble Calculations Obtained by Refinement against NMR and X-ray Scattering Data

base pair step	helical rise (Å)	helical twist (deg)
1−2	4.3 ± 1.0	31.6 ± 9.6
2−3	3.1 ± 0.6	30.7 ± 3.2
3−4	3.6 ± 0.6	34.3 ± 3.6
4−5	3.8 ± 0.4	35.8 ± 3.6
5−6	3.4 ± 0.5	36.9 ± 3.3
6−7	2.9 ± 0.3	36.5 ± 3.9

Table 7: Comparison of Average Rise for the Central 10 bp for Various Structures[a]

structure[b]	average helical rise (Å)
1GIP	3.40 ± 0.09
1DUF	3.26 ± 0.05
1NAJ	3.18 ± 0.17
$N_e = 1$	3.17 ± 0.23
$N_e = 4$	
pos-db and LAXS	3.41 ± 0.22
LAXS without pos-db	3.39 ± 0.36
pos-db without LAXS	3.28 ± 0.17
no pos-db and no LAXS	3.27 ± 0.28

[a] Abbreviations: pos-db, base−base positional database potential of mean force; LAXS, experimental large-angle X-ray scattering data. [b] PDB entries 1GIP (*13*), 1DUF (*12*), and 1NAJ (*15*). The $N_e = 4$ ensemble calculations include all RDC ($^1D_{CH}$, $^1D_{NH}$, D_{HH}, and $D_{PH3'}$ in two alignment media) and ^{31}P CSA restraints used for 1NAJ and $N_e = 1$ structures with the nonbonded term represented by a repulsive van der Waals potential and multidimensional torsion-angle database of mean force, with or without the base−base positional database potential (pos-db) as indicated, in the presence or absence of the large-angle X-ray scattering restraints as noted.

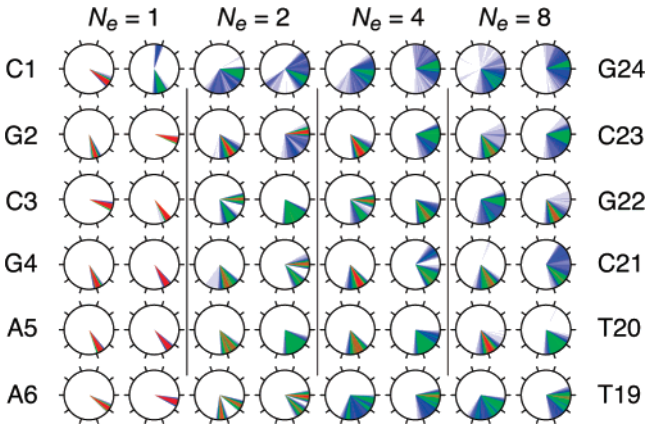

FIGURE 8: Dials representation of the distribution of sugar pseudorotation angles *P* describing sugar pucker as a function of ensemble size. Red indicates the highest probability of a given pucker value, while blue indicates the lowest non-zero probability. Each dial represents an angle as a position around its edge, with the top being 0°, the right side to 90°, and the bottom 180°. The tick marks denote the following sugar pucker conformations: C3′-endo (18°), C4′-exo (54°), O1′-exo (90°), C1′-exo (126°), C2′-endo (162°), C3′-exo (198°), C4′-endo (234°), O1′-endo (270°), C1′-endo (306°), and C2′-exo (342°). The pseudorotation angles for nucleotides 7−18 are related to those shown by symmetry.

more negative values at the center of the dodecamer: when $N_e = 4$, the average values range from −3° for the terminal base pairs to −19° for the central base pair. While the distribution of all parameters broadens when one moves to a larger ensemble size, some remain in a relatively small range, such as the helical twist for the central base pair step (base pairs 6 and 7), taking a value of 37 ± 4° when $N_e = 8$. In general, average values differ only slightly with larger ensemble size, with the exception of base pair rise; while there is a population near the $N_e = 1$ values (∼3 Å), the distribution to larger values when $N_e \geq 2$ leads to larger average rise values.

Average values of the helical rise for the central 10 bp are listed in Table 7 for the 1GIP (*13*) 1DUF (*12*), 1NAJ (*15*), and $N_e = 1$ structures, together with those for the $N_e = 4$ ensemble calculated for all combinations of the base−base positional database potential of mean force (pos-db) and X-ray scattering term on and off. It can now be understood why the 1GIP structure reproduces the X-ray scattering term better than those structures determined from a larger quantity of experimental NMR data such as the 1NAJ and $N_e = 1$ structures (cf. Table 5). The 1GIP structure (*13*) does not include the ^{31}P−H3′ RDC and ^{31}P CSA data, unlike the 1NAJ (*15*) and $N_e = 1$ structures, but does include the pos-db term. The latter term provides a gentle restraint that biases base−base distances to those of high-quality X-ray crystal structures, thereby overcoming the limitations of traditional representations of nonbonded contacts for DNA which have a tendency to either compress (Lennard-Jones and electrostatic potentials) or expand (van der Waals repulsion potential) the DNA (*13*). As a result, the average base pair rise for 1GIP is close to that of the $N_e = 4$ structures that include both the solution X-ray scattering data and pos-db potential. The 1DUF structure (*12*) was calculated using the same experimental restraints that were used for the 1GIP structure but does not include the pos-db, the result being that the 1DUF structure is compressed by the Lennard-Jones and electrostatic potential terms used in that calculation. The $N_e = 1$ structures include the X-ray and pos-db terms but could not resist the compression tendency arising from the ^{31}P−H3′ RDC and ^{31}P CSA restraints. Increasing the ensemble size to 4 allows the RDC and CSA restraints to be satisfied with a base pair rise that is consistent with the X-ray scattering data. Under these circumstances, the effect of the pos-db term is minimal; that is, the average base pair

Table 8: Fraction of Structures Taking the BI Form, Averaged over All Ensembles

	fraction BI form			
base step	$N_e = 2$	$N_e = 4$	$N_e = 8$	predicted[a]
C1pG2	0.72 ± 0.25	0.81 ± 0.17	0.65 ± 0.15	0.64
G2pC3	0.82 ± 0.24	0.75 ± 0.04	0.66 ± 0.10	0.78
C3pG4	0.65 ± 0.23	0.75 ± 0.00	0.67 ± 0.06	0.64
G4pA5	0.50 ± 0.00	0.53 ± 0.08	0.58 ± 0.08	0.63
A5pA6	1.00 ± 0.00	0.75 ± 0.00	0.78 ± 0.06	0.88
A6pT7	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.06	1.00
T7pT8	1.00 ± 0.00	0.99 ± 0.06	0.93 ± 0.07	1.00
T8pC9	1.00 ± 0.00	0.89 ± 0.13	0.84 ± 0.06	0.91
C9pG10	0.63 ± 0.22	0.75 ± 0.00	0.69 ± 0.11	0.64
G10pC11	1.00 ± 0.00	0.79 ± 0.09	0.79 ± 0.07	0.78
C11pG12	0.88 ± 0.22	0.78 ± 0.08	0.85 ± 0.10	0.64

[a] The predicted solution values are taken from ref *58* and were derived empirically on the basis of cross correlation between sequential H2′(*i*)−H6/H8(*i* + 1), H2″(*i*)−H6/H8(*i* + 1), and H6/H8(*i*)−H6/H8(*i* + 1) distances (from X-ray structures and NOE data), the $\epsilon - \zeta$ values in X-ray structures, and ^{31}P chemical shifts in solution.

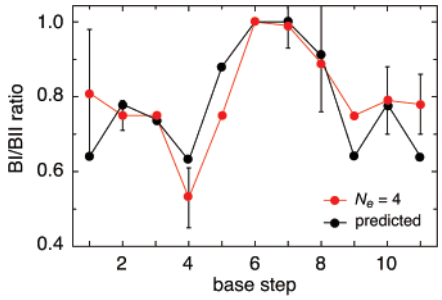

FIGURE 9: Comparison of the ratio of BI to BII phosphate backbone conformations observed in the $N_e = 4$ ensembles (red) with empirically predicted values in solution (black; ref *58*) for the different base steps. Error bars denote the deviation in the values obtained for the top 50 ensembles.

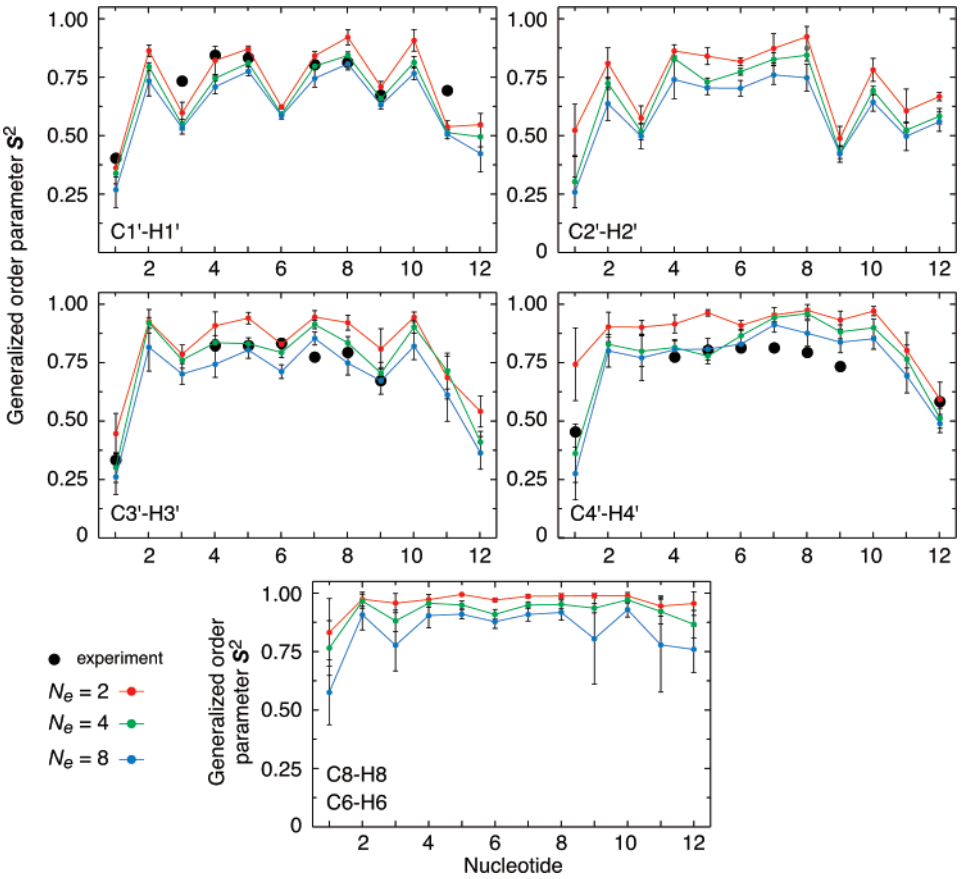

FIGURE 10: Generalized order parameters (S^2) calculated from the top 50 ensembles with ensemble sizes of 2 (red), 4 (green), and 8 (blue). Experimental order parameters derived from ^{13}C relaxation data (*60*) are shown as black circles. Order parameters for residues 13−24 are related by symmetry to those plotted. Error bars indicate variation among the 50 computed ensembles.

rise values are unaffected by the presence or absence of the pos-db term.

For each ensemble, a restrained regularized average was calculated, and the overall bend angle, calculated using the Curves program (*57*), was determined using only the center 10 bp. The resulting values are 16 ± 1°, 12 ± 4°, 11 ± 4°, and 14 ± 10° when N_e = 1, 2, 4, and 8, respectively. The $N_e \geq 2$ values are consistent with those found previously (*12*, *13*, *15*).

Deoxyribose Conformation. Sugar pseudorotation phase angles (*P*; *57*) were computed for the ensembles, with results shown in Figure 8. The $N_e = 1$ structures are quite precise, with evidence of the idealized B-form 2′-endo conformation at $P = 162°$ and the 1′-exo form at $P = 126°$, in addition to smaller pucker values. In ref *15*, an attempt to take sugar motion into account was made by fitting a linear combination of 2′-endo and 3′-endo sugar puckers to the RDC data. Here we see that large variations in the puckers do indeed occur when $N_e \geq 2$, possibly including sampling the 3′-endo form. For some of the nucleotides, distinct subpopulations are present, while for others, there is a more continuous distribution. An example of the presence of distinct subpopulations is seen in nucleotides C3 and A6. It is noteworthy that, for both nucleotides, the structural bifurcation present when $N_e = 2$ and 4 is much less pronounced when $N_e = 8$.

BI−BII Phosphate Conformations. Crystallographic data have shown that the sugar−phosphate backbone can adopt two possible conformations, the more common BI form, typical of canonical B-DNA, and a rarer BII form (*45*). These two conformations are defined by the ϵ and ζ torsion angles: negative values of $\epsilon - \zeta$ correspond to the BI form, while positive values correspond to the BII form (*45*). When $N_e = 1$, only the BI state is observed. When $N_e \geq 2$, both BI and BII forms are sampled, and the average percentage of BI form in an ensemble, averaged over the top 50 ensembles, is provided in Table 8, with a visual representation for the $N_e = 4$ ensemble shown in Figure 9. The nucleotide base step sequence dependencies for the $N_e = 4$ and 8 ensembles are in excellent agreement with predicted solution values derived empirically on the basis of cross correlation between sequential H2′(i)−H6/H8($i + 1$), H2″(i)−H6/H8($i + 1$), and H6/H8(i)−H6/H8($i + 1$) distances (from X-ray structures and NOE data), the $\epsilon - \zeta$ values in X-ray structures, and ^{31}P chemical shifts in solution (*58*). The population of the BI form is seen to vary from ∼60% for GpA base steps to 100% for ApT and TpT base steps (Table 7 and Figure 9).

Order Parameters. Generalized order parameters (S^2) can be directly calculated from the ensembles using the equation

$$S^2 = \sum_{i=1}^{N_e}\sum_{j>1}^{N_e} \Gamma_i \Gamma_j [3\cos^2(\mathbf{u}_i \cdot \mathbf{u}_j) - 1] \qquad (14)$$

where $\mathbf{u}_i$ is a unit vector along the bond in question in ensemble member i and Γ_i is the weight associated with ensemble member i (*31*, *59*). S^2 values for representative C−H sugar and base bonds are displayed in Figure 10. The average value of S^2 decreases slightly with a larger ensemble size, but the pattern as a function of nucleotide number remains consistent. On the basis of cross-validation results,

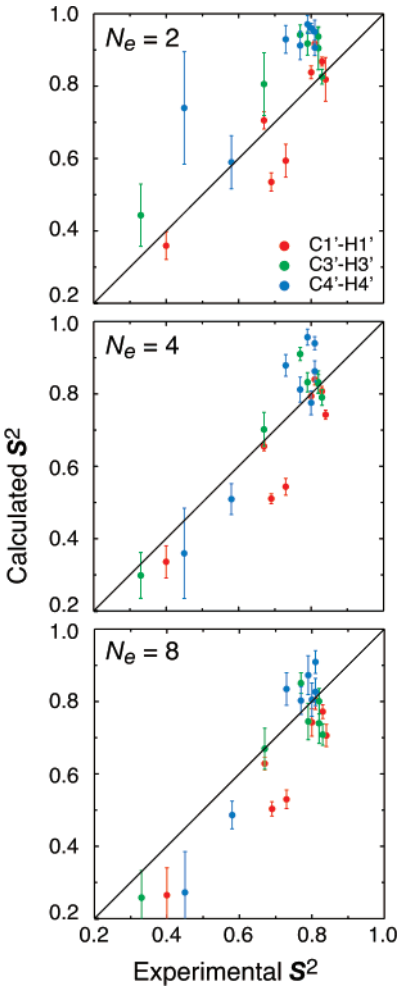

FIGURE 11: Correlation between calculated and experimental S^2 order parameters. The experimental values were taken from ref *60* and are derived from ^{13}C relaxation NMR studies on the Dickerson DNA dodecamer using an axially symmetric rotational diffusion model. The Pearson correlation coefficients between observed and calculated values are 0.80 for $N_e = 2$, 0.89 for $N_e = 4$, and 0.91 for $N_e = 8$. Error bars indicate variation among the 50 computed ensembles.

we believe that the S^2 values obtained from the $N_e = 4$ ensemble are probably the most reliable with average values (excluding the terminal bases) of 0.93 ± 0.04 for the base C−H bonds and 0.85 ± 0.06, 0.82 ± 0.07, 0.69 ± 0.14, and 0.71 ± 0.12 for the C4′−H4′, C3′−H3′, C2′−H2′, and C1′−H1′ sugar C−H bonds, respectively. Interestingly, not only do the C1′−H1′ and C2′−H2′ bonds exhibit the largest motion (smallest S^2 values) but they also display the most marked alternating pattern that is reproduced for all ensemble sizes.

Also shown in Figure 10 are the experimentally determined S^2 values for the C1′−H1′, C3′−H3′, and C4′−H4′ bond vectors derived from recent ^{13}C relaxation measurements on the DNA dodecamer (*60*). The latter reflect motions on the subnanosecond time scale. The experimental and calculated S^2 values are consistent with one another and highly correlated for the $N_e = 4$ and 8 ensembles as shown in Figure 11. The Pearson correlation coefficient between observed and calculated S^2 values is 0.80 for the $N_e = 2$ ensembles but is increased to 0.89 and 0.91 for the $N_e = 4$ and 8 ensembles, respectively. Thus, one can conclude that the $N_e = 4$ and 8 ensembles reproduce the experimentally observed amplitudes of motion for the C1′−H1′, C3′−H3′, and C4′−H4′ bond vectors in solution. It is worth noting that the calculated S^2 values for the dodecamer are also consistent with experimental values for several other similarly sized DNA duplexes (*61*, *62*).

To understand the origin of the trend in S^2 values observed for base and sugar C−H bonds, it is instructive to visually look at a pictorial representation of the conformational space sampled by ensemble members within a given nucleotide. This is illustrated in Figure 12 by the stereoview of the A6 nucleotide taken from a representative $N_e = 4$ ensemble. It is readily seen that, while displaced, the base C8−H8 bond vector takes the same orientation in each ensemble member, and hence the large value of S^2. In contrast, sugar pucker motion within the deoxyribose results in significant angular reorientation of the C1′−H1′ bond, thereby accounting for its low S^2 value.

B Factors. Another metric of the amplitude of atomic motions is provided by the thermal *B* factors. The *B* factor for atom *i*, B_i, can be computed as $B_i = 8\pi^2\langle r_i^2\rangle_e$, where $\langle r_i^2\rangle$ is the mean-square displacement of atom *i*. Figure 13 shows histograms for the distribution of the average *B* factors, excluding the terminal base pairs, for the base atoms, the deoxyribose sugars, and the phosphate groups of the $N_e = 4$ ensembles. As expected, the atomic displacements correlate with the amplitudes of the various base, sugar, and phosphate motions depicted in Figures 7−9 and increase in the following order: base atoms < sugar atoms < phosphate atoms with average *B* values of 36 ± 10, 44 ± 15, and 50 ± 27 Å^2 and ranges of ∼20−70, ∼20−105, and ∼20−130 Å^2, respectively. The magnitudes of the *B* factors for the $N_e = 4$ ensembles are similar to those observed for the 1.6 Å resolution crystal structure (*33*) which displays average *B* values of 27 ± 8, 42 ± 12, and 51 ± 10 Å^2 and corresponding ranges of 9−47, 18−73, and 28−79 Å^2, respectively. Both the NMR and X-ray *B* factors are also consistent with the results of molecular dynamics simulations on the DNA dodecamer in water (*63*).

Concluding Remarks. In this paper, we have shown that a realistic picture of motional amplitudes within the Dickerson DNA dodecamer can be obtained by ensemble refinement against NMR (RDC, ^{31}P CSA, $^3J_{H3'-P}$ and NOE) and solution X-ray scattering data. A single structure representation ($N_e = 1$) fails to account for all the experimental data since it cannot reproduce the solution X-ray scattering curves (Figure 5 and Table 5) and does not describe well the extensive RDC data sets (Figure 3 and Table 2). Further, when $N_e = 1$, the introduction of ^{31}P−H3′ RDCs and ^{31}P CSA restraints increases the disagreement with the solution X-ray scattering data, providing direct evidence of the presence of significant anisotropic motion. All the experimental data can be reconciled by ensemble refinement to represent the amplitude of atomic motions. Cross validation against both the RDC (Figure 4) and X-ray scattering (Figure 6) data suggests that for the current experimental data, an ensemble size of 4 is optimal. Indeed, with ensemble sizes of 4 or 8, the solution X-ray scattering data can be reproduced on the basis of the RDC and CSA restraints without including the X-ray scattering term in the refinement if a small number of ensembles with an incorrect peak pattern are filtered out (Table 5). Although RDC, ^{31}P CSA, and X-scattering data sample motions from the picosecond to millisecond time

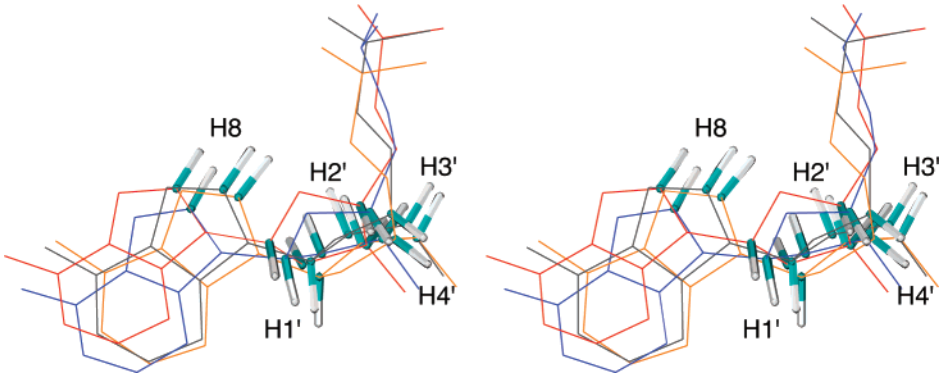

FIGURE 12: Stereoview of the A6 nucleotide for a representative $N_e = 4$ ensemble. C−H bond vectors, represented by green and white rods, are labeled for the C8−H8, C1′−H1′, C2′−H2′, C3′−H3′, and C4′−H4′ bonds.

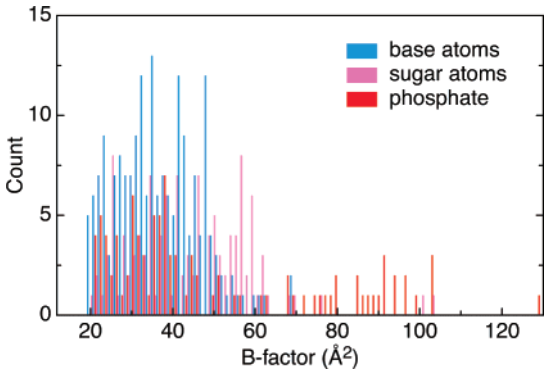

FIGURE 13: Histogram of the *B* factors for the base (blue), sugar (lilac), and phosphate (red) atoms averaged over the top 50 $N_e = 4$ ensembles.

scales, the motional amplitudes observed in the $N_e = 4$ or 8 ensemble-refined structures agree reasonably well with the available experimental order parameters for the C1′−H1′, C3′−H3′, and C4′−H4′ bond vectors derived from ^{13}C relaxation NMR studies which sample motions only on the subnanosecond time scale.

Analysis of multiple ensembles permits one to analyze the conformational space sampled by various helical parameters (helical twist and rise and base pair tilt, roll, and propeller twist), the deoxyribose sugars, and the sugar−phosphate backbone. The motional amplitudes, characterized by calculated generalized order parameters, are smallest for the bases and largest for the sugars (in particular, the C1′−H1′ and C2′−H2′/2″ vectors). While the average helical parameters are typical of B-DNA, quite large rms fluctuations within an ensemble are observed. For the $N_e = 4$ ensembles, excluding the terminal base pairs where fraying may occur, the rms fluctuations in helical twist, rise, tilt, roll, and propeller twist range from ∼3° to 4°, ∼0.3 to 0.6 Å, ∼5° to 25°, ∼9° to 18°, and ∼15° to 30°, respectively. The deoxyribose rings sample a range of sugar puckers from pure C2′-endo to C1′-exo with some evidence of rare C3′-endo forms. In addition, both BI and BII phosphate conformations are observed, with the BI conformation being predominant (ranging from 50 to 100%, depending on the base step). The fraction of the BI and BII forms computed for the $N_e = 4$ and 8 ensembles is found to be fully consistent with empirically predicted values (*58*) for the different base steps. The motional amplitudes observed in the current work are important since they demonstrate that although DNA behaves as a rigid polymer in terms of persistence length, quite extensive conformational fluctuations are observed at the base pair level, thereby permitting DNA to readily adopt a local conformation that optimally matches that of the binding surface on DNA binding proteins.

ACKNOWLEDGMENT

We thank D. Tiede for providing guidance and parameters for the solution X-ray scattering calculations and A. Bax and Z. Wu for providing the experimental NMR restraints. This study utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health (http://biowulf.nih.gov).

SUPPORTING INFORMATION AVAILABLE

Eight tables detailing helical parameters and sugar and torsion-angle parameters for $N_e = 1$, 2, 4, and 8 ensembles. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Williams, L. D., and Maher, L. J. (2000) Electrostatic mechanisms of DNA deformation, *Annu. Rev. Biophys. Biomol. Struct. 29*, 497−521.
2. Manning, G. S. (2006) The persistence length of DNA is reached from the persistence length of its null isomer through an internal electrostatic stretching force, *Biophys. J.* (in press).
3. Travers, A. A. (2004) The structural basis of DNA flexibility, *Philos. Trans. R. Soc. London, Ser. A 362*, 1423−1438.
4. Patikoglou, G., and Burley, S. K. (1997) Eukayotic transcription factor-DNA complexes, *Annu. Rev. Biophys. Biomol. Struct. 26*, 289−325.
5. Bewley, C. A., Gronenborn, A. M., and Clore, G. M. (1998) Minor groove-binding architectural proteins: Structure, function and DNA recognition, *Annu. Rev. Biophys. Biomol. Struct. 27*, 105−131.
6. Widom, J. (1998) Structure, dynamics and function of chromatin *in vitro*, *Annu. Rev. Biophys. Biomol. Struct. 27*, 285−327.
7. Saenger, W. (1984) *Principles of Nucleic Acid Structure*, Springer, New York.
8. Egli, M. (2004) Nucleic acid crystallography: Current progress, *Curr. Opin. Chem. Biol. 8*, 580−591.
9. Clore, G. M., and Gronenborn, A. M. (1989) Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy, *CRC Crit. Rev. Biochem. Mol. Biol. 24*, 479−564.
10. Feigon, J., Sklenar, V., Wang, E., Gilbert, D. F., Macaya, R. F., and Schultze, P. (1992) ^{1}H-NMR spectroscopy of DNA, *Methods Enzymol. 211*, 235−253.
11. Gronenborn, A. M., and Clore, G. M. (1989) Analysis of the relative contributions of the nuclear Overhauser interproton distance restraints and empirical energy function in the calculation of oligonucleotide structures using restrained molecular dynamics, *Biochemistry 28*, 5978−5984.

12. Tjandra, N., Tate, S., Ono, A., Kainosho, M., and Bax, A. (2000) The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase, *J. Am. Chem. Soc. 122*, 6190−6200.
13. Kuszewski, J., Schwieters, C. D., and Clore, G. M. (2001) Improving the accuracy of NMR structures of DNA by means of a database potential of mean force describing base-base positional interactions, *J. Am. Chem. Soc. 123*, 3903−3918.
14. Clore, G. M., and Kuszewski, J. (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, *J. Am. Chem. Soc. 125*, 1518−1525.
15. Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V. B., and Bax, A. (2003) Overall structure and sugar dynamics of a DNA dodecamer from homo- and heteronuclear dipolar couplings and ^{31}P chemical shift anisotropy, *J. Biomol. NMR 26*, 297−315.
16. Schmitz, U., Kumar, A., and James, T. L. (1992) Dynamic interpretation of NMR data: Molecular dynamics with weighted time-averaged restraints and ensemble R-factor, *J. Am. Chem. Soc. 114*, 10654−10656.
17. Schmitz, U., and James, T. L. (1995) How to generate accurate solution structures of double-helical nucleic acid fragments using nuclear magnetic resonance and restrained molecular dynamics, *Methods Enzymol. 261*, 3−44.
18. Tonelli, M., and James, T. L. (1998) Insights into the dynamics nature of DNA duplex structure via analysis of nuclear Overhauser effect intensities, *Biochemistry 37*, 11478−11487.
19. Görler, A., Ulyanov, N. B., and James, T. L. (2000) Determination of the populations and structures of multiple conformers in an ensemble from NMR data: Multiple-copy refinement of nucleic acid structures using floating weights, *J. Biomol. NMR 16*, 147−164.
20. Bonvin, A. M., and Brünger, A. T. (1995) Conformational variability of solution nuclear magnetic resonance structures, *J. Mol. Biol. 250*, 80−93.
21. Bonvin, A. M., and Brünger, A. T. (1996) Do NOE distances contain enough information to assess the relative populations of multi-conformer structures, *J. Biomol. NMR 7*, 72−76.
22. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.
23. Bax, A., Kontaxis, G., and Tjandra, N. (2001) Dipolar couplings in macromolecular structure determination, *Methods Enzymol. 339*, 127−174.
24. Grishaev, A., and Bax, A. (2005) Weak alignment NMR: A hawk-eyed view of biomolecular structure, *Curr. Opin. Struct. Biol. 15*, 563−570.
25. Boibouvier, J., Delaglio, F., and Bax, A. (2003) Direct observation of dipolar couplings between distant protons in weakly aligned nucleic acids, *Proc. Natl. Acad. Sci. U.S.A. 100*, 11333−11338.
26. Tiede, D. M., Zhang, R., Chen, L. X., Yu, L., and Lindsey, J. S. (2004) Structural characterization of modular supramolecular architectures in solution, *J. Am. Chem. Soc. 126*, 14054−14062.
27. Svergun, D., Barberato, C., and Koch, M. H. J. (1995) CRYSOL: A program to evaluate X-ray scattering of biological macromolecules from atomic coordinates, *J. Appl. Crystallogr. 28*, 768−773.
28. Zuo, X., and Tiede, D. M. (2005) Resolving conflicting crystallographic and NMR models for solution-state DNA with solution X-ray diffraction, *J. Am. Chem. Soc. 127*, 16−17.
29. Zuo, X., Cui, G., Merz, K. M., Zhang, L., Lewis, F. D., and Tiede, D. M. (2006) X-ray diffraction "fingerprinting" of DNA structure in solution for quantitative evaluation of molecular dynamics simulation, *Proc. Natl. Acad. Sci. U.S.A. 103*, 3534−3539.
30. Clore, G. M., and Schwieters, C. D. (2004) How much backbone motion in ubiquitin is required to be consistent with dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation, *J. Am. Chem. Soc. 126*, 2923−2938.
31. Clore, G. M., and Schwieters, C. D. (2004) Amplitudes of protein backbone dynamics and correlated motions in a small α/β protein: Correspondence of dipolar coupling and heteronuclear relaxation measurements, *Biochemistry 43*, 10678−10691.
32. Clore, G. M., and Schwieters, C. D. (2006) Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small α/β protein: A unified picture of high probability, fast atomic motions in proteins, *J. Mol. Biol. 355*, 879−886.
33. Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1981) Structure of a B-DNA dodecamer: Conformation and dynamics, *Proc. Natl. Acad. Sci. U.S.A. 78*, 2179−2183.
34. Schwieters, C. D., Kuszewski, J., Tjanda, N., and Clore, G. M. (2003) The Xplor-NIH NMR molecular structure determination package, *J. Magn. Reson. 160*, 66−74.
35. Schwieters, C. D., Kuszewski, J., and Clore, G. M. (2006) Using Xplor-NIH for NMR molecular structure determination, *Prog. Nucl. Magn. Reson. Spectrosc. 48*, 47−62.
36. Wu, Z., Tjandra, N., and Bax, A. (2001) ^{31}P chemical shift anisotropy as an aid in determining nucleic acid structures in liquid crystals, *J. Am. Chem. Soc. 123*, 3617−3618.
37. Svergun, D. I., Petoukhov, M. V., and Koch, M. H. J. (2001) Determination of domain structure of proteins from X-ray solution scattering, *Biophys. J. 80*, 2946−2953.
38. Fraser, R. D. B., MacRae, T. P., and Suzuki, W. (1978) An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules, *J. Appl. Crystallogr. 11*, 693−694.
39. Grishaev, A., Wu, J., Trewheela, J., and Bax, A. (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data, *J. Am. Chem. Soc. 127*, 16621−16628.
40. Chacon, P., Moran, F., Diaz, J. F., Panto, E., and Andreu, J. M. (1998) Low resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm, *Biophys. J. 74*, 2760−2775.
41. Guo, D. Y., Blessing, R. H., and Langs, D. A. (2000) Globbic approximation in low-resolution direct-methods phasing, *Acta Crystallogr. D56*, 1148−1155.
42. Gabel, F., Simon, R., and Sattler, M. (2006) A target function for quaternary structural refinement for small angle scattering and NMR orienational restraints, *Eur. Biophys. J. 35*, 313−327.
43. Staff, E. B., and Kuijlaars, A. B. J. (1997) Distributing many points on a sphere, *The Mathematical Intelligencer 19*, 5−11.
44. Sklenar, V., and Bax, A. (1997) Measurement of ^{1}H-^{31}P NMR coupling constants in double-stranded DNA fragments, *J. Am. Chem. Soc. 109*, 7525−7526.
45. Hartmann, B., Piazzola, D., and Lavery, R. (1993) BI-BII transitions in DNA, *Nucleic Acids Res. 21*, 561−568.
46. Murphy, E. C., Zhurkin, V. B., Louis, J. M., Cornilescu, G., and Clore, G. M. (2001) Structural basis for SRY-dependent 46-X,Y sex reversal: Modulation of DNA bending by a naturally occurring point mutation, *J. Mol. Biol. 31*, 481−499.
47. Schwieters, C. D., and Clore, G. M. (2001) Internal coordinates for molecular dynamics and minimization in structure determination and refinement, *J. Magn. Reson. 152*, 288−302.
48. Nilges, M., Gronenborn, A. M., Brünger, A. T., and Clore, G. M. (1988) Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints: Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2, *Protein Eng. 2*, 27−38.
49. Clore, G. M., and Gronenborn, A. M. (1998) New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. U.S.A. 95*, 5891−5898.
50. Kuszewski, J., Schwieters, C. D., Garrett, D. S., Byrd, R. A., Tjandra, N., and Clore, G. M. (2004) Completely automated macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments, *J. Am. Chem. Soc. 126*, 6258−6273.
51. Clore, G. M., Brünger, A. T., Karplus, M., and Gronenborn, A. M. (1986) Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination: A model study of crambin, *J. Mol. Biol. 191*, 523−551.
52. Schweitzer, B. I., Mikita, T., Kellogg, G. W., Gardner, K. H., and Beardsley, G. P. (1994) Solution structure of a B-DNA dodecamer containing the anti-neoplastic agent arabinosylcytosine: Combined use of NMR, restrained molecular dynamics and full relaxation matrix refinement, *Biochemistry 33*, 11460−11475.
53. Brooks, S. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM: A program for macromolecular energy, minimization and dynamics calculations, *J. Comput. Chem. 4*, 187−217.
54. Nilsson, L., and Karplus, M. (1986) Empirical energy functions for energy minimization and dynamics of nucleic acids, *Comput. Chem. 7*, 591−616.
55. Nilsson, L., Clore, G. M., Gronenborn, A. M., Brünger, A. T., and Karplus, M. (1986) Structure refinement of oligonucleotides

by molecular dynamics with nuclear Overhauser effect interprotn disrtance restraints: Application to 5′d(CGTACG)$_2$, *J. Mol. Biol. 188*, 455−475.
56. Clore, G. M., and Garrett, D. S. (1999) R-factor, free R and complete cross-validation for dipolar coupling refinement of NMR structures, *J. Am. Chem. Soc. 121*, 9008−9012.
57. Lavery, R., and Sklenar, H. (1989) Defining the structure of irregular nucleic acids: Conventions and principles, *J. Biomol. Struct. Dyn. 6*, 655−667.
58. Heddi, B., Foloppe, N., Bouchemal, N., Hantz, E., and Hartmann, B. (2006) Quantification of DNA BI/BII backbone states in solution: Implications for DNA overall structure and recognition, *J. Am. Chem. Soc. 128*, 9170−9177.
59. Lipari, G., and Szabo, A. (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. I. Theory and range of validity, *J. Am. Chem. Soc. 104*, 4546−4559.
60. Boisbouvier, J., Wu, Z., Ono, A., Kainosho, M., and Bax, A. (2003) Rotational diffusion tensor of nucleic acids from ^{13}C NMR relaxation, *J. Biomol. NMR 27*, 133−142.
61. Spielmann, H. P. (1998) Dynamics of a bis-intercalator DNA complex by ^{1}H-detected natural abundance ^{13}C NMR spectroscopy, *Biochemistry 37*, 16863−16876.
62. Kojima, C., Ono, A., Kainosho, M., and James, T. L. (1998) DNA duplex dynamics: NMR relaxation studies of a decamer with uniformly ^{13}C-labeled purine nucleotides, *J. Magn. Reson. 135*, 310−333.
63. Duan, Y., Wilkosz, P., Crowley, M., and Rosenberg, J. M. (1997) Molecular dynamics simulation study of DNA dodecamer d(CGCGAATTCGCG) in solution: Conformation and hydration, *J. Mol. Biol. 272*, 553−572.

BI061943X