All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data

NICOLAS FOLOPPE,* ALEXANDER D. MACKERELL, JR.

Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland 21201

Received 19 March 1999; accepted 30 August 1999

ABSTRACT: Empirical force-field calculations on biological molecules represent an effective method to obtain atomic detail information on the relationship of their structure to their function. Results from those calculations depend on the quality of the force field. In this manuscript, optimization of the CHARMM27 all-atom empirical force field for nucleic acids is presented together with the resulting parameters. The optimization procedure is based on the reproduction of small molecule target data from both experimental and quantum mechanical studies and condensed phase structural properties of DNA and RNA. Via an iterative approach, the parameters were primarily optimized to reproduce macromolecular target data while maximizing agreement with small molecule target data. This approach is expected to ensure that the different contributions from the individual moieties in the nucleic acids are properly balanced to yield condensed phase properties of DNA and RNA, which are consistent with experiment. The quality of the presented force field in reproducing both crystal and solution properties are detailed in the present and an accompanying manuscript (MacKerell and Banavali, J Comput Chem, this issue). The resultant parameters represent the latest step in the continued development of the CHARMM all-atom biomolecular force field for proteins, lipids, and nucleic acids. J Comput Chem 21: 86–104, 2000

*Present address: Center for Structural Biology, Department of Bioscience, Karolinska Institutet, S-141 57, Huddinge, Sweden

Correspondence to: A. D. MacKerell; e-mail: amackere@rx.umaryland.edu

Contract/grant sponsor: NIH; contract/grant number: 51501

This article includes a longer version available as Supplementary Material from the authors upon request or via the Internet at ftp.wiley.com/public/journals/jcc/suppmat/21/86 or http://journals.wiley.com/jcc/

CCC 0192-8651 / 00 / 020086-19

Keywords: CHARMM; force field; molecular mechanics; empirical; molecular dynamics; DNA; RNA

Introduction

Empirical force field-based computational studies are widely used methods for the investigation of a variety of properties of biological macromolecules.[1,2] In combination with growing computational resources, these methods allow for atomic detail simulations on heterogeneous systems that may contain 100,000 or more atoms. In particular, force field-based techniques offer the ability to directly analyze the relationship of structure to energetics, information that experimental approaches can only access indirectly.

Over the last several years force-field techniques have played an increasingly important role in the study of nucleic acids. Empirical force-field calculations are increasingly involved in the refinement of nucleic acid structures in conjunction with crystallographic[3,4] or NMR data.[5–7] Force field-based techniques alone can enhance the interpretation of a wide variety of biochemical and biophysical experimental data[1,2] and provide insights that may be difficult or impossible to obtain from experiment. This may be particularly true with DNA, for which the use of experimental techniques has been plagued by a number of problems. Although X-ray crystallography has yielded a wealth of information about DNA,[8–10] it is limited to the sequences that can crystallize and diffract to good resolution. Crystallization is obtained with nonphysiological solvents, and it is well documented that the observed crystal structures for a given deoxyribo-oligonucleotide may depend on the crystal packing, making it somewhat difficult to distinguish what is contributed by the intrinsic properties of the sequence and what is imposed by the crystal environment.[11–14] NMR has become increasingly powerful in deriving deoxyribooligonucleotides structures in solution; however, the accuracy of the NMR-derived structures is elusive due to the lack of long range distance restraints.[15,16] Consequently, details of the structure, dynamics, and solvation of DNA in solution remain poorly characterized, making this a particularly interesting area for the application of simulation methods. DNA is particularly amenable to computer simulations, given that duplex DNA simulations can be initiated with DNA in one of its canonical forms,[17] thereby avoiding the need for an experimentally determined structure to initiate the calculations. In addition to DNA, computational studies of small oligonucleotides[18] and of RNA[19] represent active areas of research on nucleic acids.

Not until recently have force field-based simulations of nucleic acid oligomers with an explicit representation of the aqueous solvent yielded stable structures on the nanosecond time scale.[20–23] This success has been facilitated by new force fields explicitly parametrized for simulations in the condensed phase,[24,25] along with simulations being performed with increased atom–atom nonbond truncation distances or Ewald sums-based approaches. Current tests of some of the available force fields, however, demonstrate that for nucleic acid simulations to realize their full potential further improvements of the force fields are necessary.[26,27] Limitations include improper treatment of the equilibrium between the A and B forms of DNA,[26] with CHARMM22[25] overstabilizing the A form of DNA[20,28,29] and the AMBER96[24] force field having sugar pucker and helical twist values not in agreement with canonical B values.[30] Refinement of structures based on experimental data have also highlighted the need for more accurate nucleic acid force fields.[4,7,31] Recently, a revised version of the AMBER96 nucleic acid force field (AMBER98)[30] and a nucleic acid force field from Bristol–Myers Squibb (BMS) have been presented.[32]

These observations prompted the reoptimization of the CHARMM22 all-atom nucleic acid force field, the details of which are described here. This new all-atom force field for nucleic acids will be referred to as CHARMM27, based on the version of the program CHARMM[33,34] with which it will initially be released. An important part of the development of CHARMM27 has been devoted to obtaining a force field that adequately represents the equilibrium between the A and B forms of DNA as well as the A form of RNA. This has been achieved by balancing the intrinsic energetic properties of a variety of model compounds with the overall conformational properties of DNA and RNA. This strategy is physically more relevant, although significantly more demanding, than approaches where the parameters are adjusted either purely empirically, to reproduce only experimental condensed phase properties, or to only reproduce quantum mechanical (QM) data

on model compounds. By simultaneously reproducing target data for both small model compounds and duplex DNA and RNA, a force field in which the proper combination of local contributions that yield condensed phased properties of DNA and RNA in agreement with the experiment can be achieved.

Presented is an abbreviated account of the full version of the CHARMM27 parametrization manuscript. The full version of the manuscript is contained in the Supplementary Material. In the present account, the full introduction is followed by a description of the parametrization approach used in the optimization of the CHARMM27 nucleic acid force field. Details of the calculations are included in the Methods section of the Supplementary Material. An abbreviated version of the Results and Discussion is presented, followed by a concluding section emphasizing a number of points in the present parametrization work and discussing several issues associated with force field optimization. An accompanying manuscript presents data from the application of the CHARMM27 parameters to MD simulations of DNA and RNA in solution.[35]

Parametrization Approach

POTENTIAL ENERGY FUNCTION

Empirical force fields represent an approach to computational chemistry that minimizes computational costs by using simplified models to calculate the potential energy of a system, $U(\vec{R})$, as a function of its three-dimensional structure, $\vec{R}$. The potential energy function used in the program CHARMM[33, 34] is shown in eq. (1).

$$\begin{aligned} U(\vec{R}) = & \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{UB}} K_{\text{UB}}(S-S_0)^2 \\ & + \sum_{\text{angle}} K_\theta(\theta-\theta_0)^2 \\ & + \sum_{\text{dihedrals}} K_\chi\big(1+\cos(n\chi-\delta)\big) \\ & + \sum_{\text{impropers}} K_{\text{imp}}(\varphi-\varphi_0)^2 \\ & + \sum_{\text{nonbond}} \varepsilon_{ij}\left[\left(\frac{R\min_{ij}}{r_{ij}}\right)^{12} - \left(\frac{R\min_{ij}}{r_{ij}}\right)^{6}\right] \\ & + \frac{q_i q_j}{e r_{ij}} \end{aligned} \tag{1}$$

Equation (1) includes the bond length, b, the distance between atoms separated by two covalent bonds (1,3 distance), S, the valence angle, θ, the dihedral or torsion angle, χ, the improper angle, φ, and the distance between atoms i and j, r_{ij}. Parameters, the terms being optimized in the present work, include the bond force constant and equilibrium distance, K_b and b_0, respectively, the Urey–Bradley force constant and equilibrium distance, K_{UB} and S, respectively, the valence angle force constant and equilibrium angle, K_θ and θ_0, respectively, the dihedral force constant, multiplicity and phase angle, K_χ, n, and δ, respectively, and the improper force constant and equilibrium improper angle, K_φ and φ_0, respectively. These terms are referred to as the internal parameters. Also optimized were the nonbonded or interaction parameters between atoms i and j, including the partial atomic charges, q_i, and the Lennard–Jones (LJ) well depth, ε_{ij}, and minimum interaction radius, $R_{\min,ij}$, used to treat the van der Waals (VDW) interactions. Typically, ε_i and $R_{\min,i}$ are obtained for individual atom types and then combined to yield ε_{ij} and $R_{\min,ij}$ for the interacting atoms via combining rules. In CHARMM, ε_{ij} values are obtained via the geometric mean $\varepsilon_{ij} = \text{sqrt}(\varepsilon_i * \varepsilon_j)$, and $R_{\min,ij}$ via the arithmetic mean, $R_{\min,ij} = (R_{\min,i} + R_{\min,j})/2$. The dielectric constant, e, is set to one in all calculations, corresponding to the permittivity of vacuum.

PARAMETER OPTIMIZATION STRATEGY

The ability of eq. (1) to treat complex systems such as biomolecules in their aqueous environment is based on the quality of parameters in reproducing a variety of selected properties, referred to as target data. In addition, the exact combination of parameters is important because different sets of parameters can often reproduce selected target data in a similar way; a problem that is referred to as parameter correlation. For example, it has been shown that several sets of LJ parameters for the C and H atoms in ethane, with the C $R_{\min}$ values differing by over 0.5 Å, can all yield experimental heats of vaporization and molecular volumes of neat ethane in satisfactory agreement with experiment.[36] This is due to the large dimensionality of parameter space such that there are multiple solutions (i.e., combinations of parameters) that can reproduce a given set of target data due to correlation among the parameters. Optimization approaches applied in the present work allow for elimination of some combinations of parameters by adding more target data. For example, with the LJ parameters, an approach has been developed that includes quantum mechanical data on rare gas atoms interacting with model

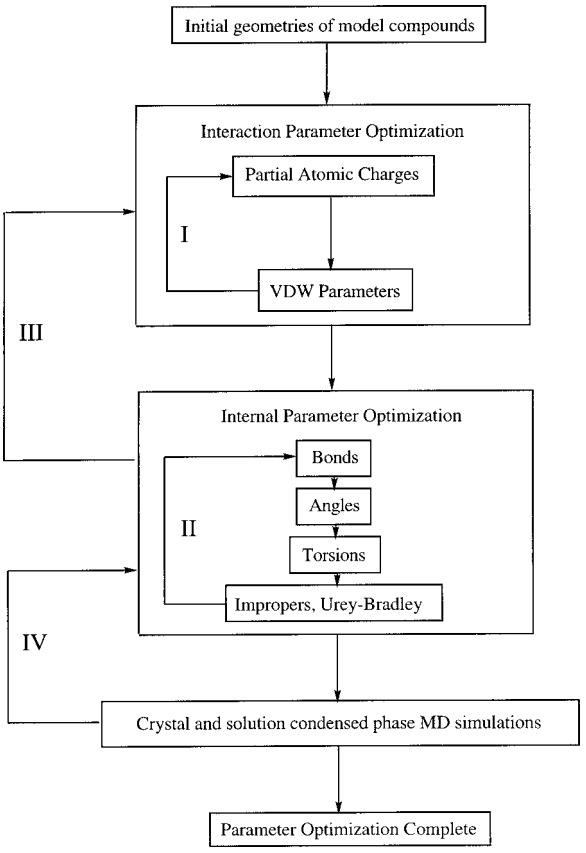

FIGURE 1. Flow diagram of the present parameter optimization. Iterative loops included in the parametrization are indicated by roman numerals.

compounds along with pure solvent properties to obtain more physically relevant parameters.[37] However, even with this additional data the presence of parameter correlation cannot be entirely eliminated. Thus, the approach used for the parameter optimization, as well as the reproduction of a selected set of target data by the parameters, can influence the quality of the force field.

The present parameter optimization study represents an extensive revision of the previously published CHARMM22 all-atom empirical force field parameters for nucleic acids.[25] Presented in Fig. 1 is a flow diagram of the parameter optimization procedure. Loops I, II, and III in Fig. 1 were included in the optimization of the CHARMM22 force field for nucleic acids, with loop IV representing an extension of that approach included in the CHARMM27 optimization.

In the CHARMM22 nucleic acid parameter optimization a variety of model compounds were selected with target data collected on those compounds. This target data included both experimental and *ab initio* data and solely acted as the basis for the parameter optimization. Empirical force field calculations were performed on the model compounds with the computed properties compared with the target data. The parameters were then manually adjusted to better reproduce the target data. Part of this process involved iterative procedures where, upon changing one class of parameters, a set of previously optimized parameters were readjusted if necessary (loops I, II, and III in Fig. 1). For example, a set of partial atomic charges would be assigned to a model compound following which dihedral parameters would be adjusted to reproduce a target potential energy surface for that model compound. The partial atomic charges would then be reinvestigated due to possible changes in geometry associated with optimization of the dihedral parameters that could effect the reproduction of the target data for the charge optimization. This approach yields a parameter set that accurately reproduces a variety of internal (e.g., geometries, vibrational spectra, conformational energetics) and interaction (e.g., interactions with water, heats, of sublimation) target data for the selected model compounds. Once the optimization procedure at the model compound level was complete, the resultant parameters were then used to perform simulations of B and Z DNA in their crystal environments, both of which yielded satisfactory agreement with experiment. At this point, the CHARMM22 parametrization was considered complete.

This approach relies on the reproduction of the small molecule target data by the force field also yielding satisfactory results on macromolecules in the condensed phase; analogous approaches have been used for the optimization of other force fields.[24, 38–40] With the CHARMM22 nucleic acid force field it was ultimately shown that simulations of duplex DNA in solution yielded A form structures, in disagreement with experiment.[26] Limitations in this approach were also observed during the optimization of the CHARMM22 all-atom force field for proteins.[41] In that work it was shown that reproduction of QM data on the energetics of the alanine dipeptide yielded conformational properties of the protein backbone in molecular dynamics (MD) simulations that disagreed with experiment. Reoptimization of the protein backbone parameters to systematically deviate from the QM energetic data led to improved properties for the protein backbone. This additional procedure is represented by loop IV in Fig. 1. The need for this additional loop may reflect limitations associated with the level of theory of the QM data as well as the simpli-

fied form of the potential energy function in eq. (1), and emphasizes the importance of including macromolecular properties as part of the target data for the parameter optimization procedure.

For the present CHARMM27 parameter optimization study, the initial parameters assigned to the model compounds were extracted directly from the CHARMM22 parameter set. The internal parameters were then optimized to reproduced geometries, vibrational spectra, and conformational energetics for the model compounds, using an iterative approach to maximize the agreement with the internal target data (loop II in Fig. 1). The partial atomic charges and LJ parameters were then iteratively adjusted using the new minimum energy geometries (loop I in Fig. 1). Partial atomic charges were adjusted using a previously applied methodology.[25, 42] In this approach the target data for optimizing the charges on specific chemical groups are minimum interaction energies and geometries between a water molecule and these chemical groups in a variety of orientations obtained from QM calculations at the HF/6-31G* level of theory. Scaling of the interaction energies and offset of the minimum interaction distances are performed to obtain charges that yield satisfactory condensed phase properties.[38, 42–44] The offsets and scaling account for a number of factors including limitations in the QM level of theory and the omission of explicit electronic polarizability in the potential energy function, as previously discussed.[41] The scaling factors and offsets mentioned above have been optimized specifically for the TIP3P water model.[43, 45] Accordingly, the CHARMM27 force field is designed to be used with the TIP3P water model. For the bases, base–base interaction energies and distances and dipole moments were also included in the charge optimization. LJ parameters of base atoms were optimized using water-model compound interactions along with crystal simulations with the crystal unitcell parameters and heats of sublimation being the target data. Using the converged interaction parameters, the internal target data for the model compounds were then rechecked, and additional optimization of the parameters performed as required until both the internal and interaction parameters had converged (loop III of Fig. 1).

Once the parameter optimization at the model compound level was complete, MD simulations of DNA crystals were performed. Results from the simulations were then compared with the macromolecular target data, including RMSD with respect to canonical A and B DNA and dihedral distributions from a survey of the Nucleic Acid Database

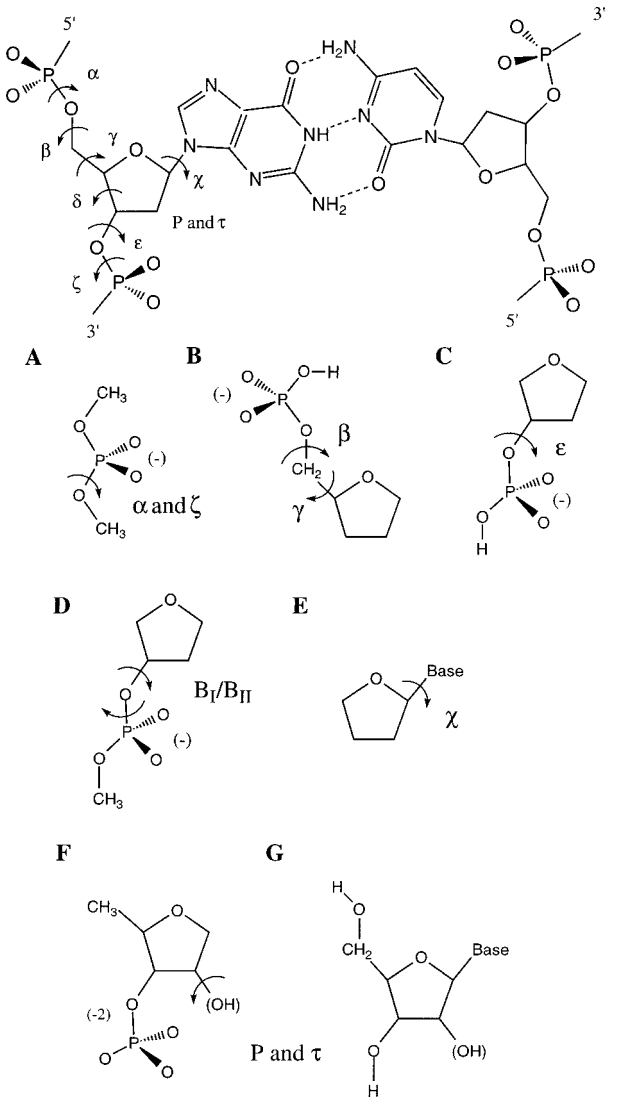

FIGURE 2. (Top) Diagram of a DNA G-C base pair showing the dihedrals considered in the parameter optimization. *P* and τ are the sugar pseudorotation angle and amplitude, respectively. Dashed lines indicate the Watson–Crick hydrogen bonds between the bases. (Bottom) Model compounds used for the optimization of the backbone dihedrals, sugar puckering, and glycosyl linkage. Included in the figures are the dihedrals that the individual models compounds were used to optimize.

(NDB)[10] of DNA and RNA crystal structures. Presented in Fig. 2 (top) is a schematic diagram of a G-C base pair that includes the dihedrals and sugar pucker terms considered in the present work. Based on deviations between the simulated and survey dihedral distributions, the dihedral parameters for, typically, one or two of the dihedrals were adjusted and the condensed phase simulations repeated. During the readjustment steps, comparisons with

the small molecule energetic target data was always performed. This iterative loop in the CHARMM27 parameter optimization constitutes loop IV in Fig. 1. During this loop, adjustment of the dihedral parameters was done to "soften" the small molecule energy surfaces (i.e., lower energy barriers) rather than moving the location of the minima in the energy surfaces and increasing energy barriers to restrict the condensed phase simulations to sample the dihedral distributions from the survey. This approach is designed to produce a force field sensitive to the environment rather than being dominated by the intrinsic conformational energetics of the nucleic acid molecule itself. Optimization of the unique parameters associated with RNA was performed following completion of the DNA parameter adjustment.

MODEL COMPOUNDS

Selection of adequate model compounds that are consistent with the ultimate application of the force field under development is essential for proper parameter optimization. The present model compounds were designed to include functional groups required to properly model the local nucleic acid environment, including the dihedrals indicated in Fig. 2 (top), while being small enough to remain computationally tractable. To select the appropriate model compounds *ab initio* calculations were performed to investigate which compounds have structural and energetic properties consistent with experimental data.[46–48] The model compounds selected from these studies are shown in Fig. 2 (bottom). For the majority of these compounds MP2 level *ab initio* data is required to properly treat experimental structural and energetic properties. Accordingly, in the present work MP2 results are used as target data whenever feasible. Note that all the compounds, excluding compound A, contain the furanose ring. This moiety was included to allow for dihedral parameter optimization to take into account contributions from changes in the furanose ring pucker, consistent with the north and south sugar puckers that both occur in DNA. In addition, the complexity of these molecules required that only a subspace of the full conformational space be sampled. This subspace was selected to be relevant to that occurring in DNA.[48]

Dimethylphosphate (DMP, compound A), was the primary model compound for optimization of the α and ζ terms. With the dihedrals β, ε, and γ, it was deemed necessary that the phosphate be included in the model compounds, yielding compound B for β and γ and compound C for ε. Preliminary studies on compounds B and C investigated their energetic properties with both a monoanionic and dianionic phosphate. The similarity of the surfaces with the different charges led to inclusion of only the monoanionic species in the present report. Parameters associated with ε and ζ were also checked using compound D, which was designed to model the B_I and B_{II} states that occur in B DNA.[9] The glycosyl linkage, χ, was modeled with compound E with the four DNA bases. This compound explicitly treats all the atoms that are included in the dihedrals describing χ. Optimization of the parameters to model the sugar puckering was performed using compounds F and G. Compound F was used to check the influence of a phosphate group on sugar pucker. With compound F, a dianionic phosphate was used to avoid problems with the proton on a monoanionic phosphate. Compound G is a nucleoside, and was studied with the standard nucleic acid bases along with imidazole as the base. QM studies have shown compound G with imidazole to have conformational properties that are consistent with a variety of experimental data.[46] Both the deoxy and ribo forms of compounds C, F, and G were included; the ribo forms were used to optimize parameters associated with the C2′ hydroxyl group and the furanose parameters in RNA.

MACROMOLECULAR TARGET DATA

As discussed previously, the present work also used macromolecular structural information as the target data. To do this, DNA and RNA duplexes were selected for condensed phased simulations and are listed in Table I. Because emphasis was placed on the DNA portion of the force field due to the sensitivity of DNA structure to environmental effects, base sequence and base composition,[17] five DNA structures were selected as target data. Two crystal structures were selected for simulations in the explicit crystal environment. The B form CGATCGATCG decamer[49] was chosen due to its high resolution and presence of several phosphodiester linkages in the B_{II} conformation and the A form GTACGTAC octamer[50] because of the relatively high content of AT base pairs in contrast to the majority of A form DNA crystal structures. During each parameter optimization cycle, the two crystals were subjected to MD simulations from which probability distributions of the backbone and glycosyl dihedrals and of the sugar pseudorotation angles and amplitudes were obtained and compared to NDB crystal survey distributions. This information was then used to adjust selected parameters

TABLE I.
DNA and RNA Duplex Structures Included as Target Data for the Parameter Optimization.

Sequence	Comment	Reference
d(CGATCGATCG)	B form crystal	49
d(GTACGTAC)	A form crystal	50
d(CGCGAATTCGCG)	Contains EcoRI recognition sequence	51, 52
d(CATTTGCATC)	NMR solution structure	53, 54
d(CTCGAG)	A to B transition	55
UAAGGAGGUGUA	RNA, 2 duplexes/asymmetric unit	56

associated with dihedrals observed to deviate significantly from the target data. In addition to the crystals, three DNA sequences were selected for additional testing in solution (Table I).[17, 51, 52] The EcoRI recognition sequence is probably the most studied DNA oligomer, making its inclusion necessary as part of the present study. The CATTTGCATC decamer was selected due to its structure being determined in solution via NMR, with emphasis on sugar puckering.[53, 54] Inclusion of the CTCGAG hexamer[55] was done to test the influence of water activity on the equilibrium between the A and B forms of DNA. During different stages of the parameter optimization solution simulations were performed on these DNA sequences to check that the results from the B DNA crystal simulations were not adversely influencing the force field, and that the parameters properly reproduced the equilibrium between the A and B forms of DNA associated with changes in water activity.[17] For condensed phase simulations of RNA the UAAGGAGGUGAU dodecamer[56] was used (Table I). Only one RNA duplex was included as target data, given the greater homogeneity of RNA duplex structures compared to DNA.[17] Details of the results from the solution simulations not included in the present manuscript are presented in the accompanying manuscript.[35]

Methods

All methods used in this study are included in Section 3 of the Supplementary Material.

Results and Discussion

Presentation of the results and discussion will be performed consistent with the flow diagram in Fig. 1 and based on the terms in eq. (1), with emphasis on the final outcome of the parametrization process. Additional details are included in Section 4 of the Supplementary Material. Except when noted, the presented results are for the converged parameters. The final parameters are included in the Appendix of the Supplementary Material, and may be obtained from the World Wide Web at www.pharmacy.umaryland.edu/˜alex.

INTERACTION TERMS

Optimization of the interaction terms involved refinement of the partial atomic charges for the sugar, the phosphate, and the bases and optimization of the LJ terms for selected atoms in the bases. This portion of the optimization procedure is shown as loop I in Fig. 1. For the furanose, tetrahydrofuran was used as the model compound for the ether oxygen, and 3-hydroxytetrahydrofuran was used for the hydroxyl charges. DMP was used as the model compound for the phosphate group and, for the bases, the individual bases themselves or methyl or ethyl substituted analogs were used. As discussed above, this charge optimization primarily involved the reproduction of HF/6-31G* minimum interaction energies and geometries between water and the various model compounds. All interaction orientations are presented in Fig. 3 of the Supplementary Material. For the bases, additional target data included base–base interaction energies, including the Watson–Crick (WC) and Hoogsteen hydrogen bond interactions,[57–61] dipole moments, stacking interactions,[57, 60, 61] and small molecule crystal calculations. For optimization of the LJ terms for selected atoms in the bases, reproduction of the heats of sublimation and unit cell parameters of uracil and 9-methyladenine were used as target data.

Presented in Table II is a comparison of the differences between the CHARMM27 or CHARMM22 empirical base to water minimum interaction energies and the *ab initio* target data. The results are presented as average differences, RMS differences

TABLE II.
Average Differences, RMS Differences, and Average Absolute Error between the Base to Water *Ab Initio* and Empirical Interaction Energies.

Base	Average Difference	RMS Difference	Average Absolute Error
CHARMM27			
Adenine	−0.04	0.12	0.08
Guanine	0.05	0.28	0.17
Cytosine	0.03	0.16	0.13
Thymine	0.05	0.28	0.21
Uracil	0.01	0.07	0.07
CHARMM22			
Adenine	0.05	0.59	0.47
Guanine	−0.29	0.83	0.56
Cytosine	0.23	0.83	0.52
Thymine	−0.33	0.62	0.60
Uracil	0.08	0.25	0.20

Average absolute error is the sum of the absolute values of the differences divided by n, the number of interactions of water with each base.

or average absolute errors over the individual bases. As may be seen, CHARMM27 is significantly improved over CHARMM22. This is due to the use of all atoms in the bases to define the electrostatic groups of unit charge, vs. groups of seven or less atoms in CHARMM22, the use of smaller radii on the base aromatic hydrogens (similar to the imidazole sidechain of histidine),[41] and additional optimization of the partial atomic charges. The resultant charges also yield good agreement with available target data for the WC and Hoogsteen base pair interactions, as shown in Table III. For the GC WC interaction, the agreement with both experiment[62] and *ab initio*[59] is good. For the AT and AU basepairs, CHARMM27 underestimates the experimental values; however, they fall in the range of the reported *ab initio* data for AT. Of note are the Hoogsteen interactions being slightly more favorable than the WC

TABLE III.
Watson–Crick and Hoogsteen Base Pairing Interaction Energies, Zero Point Energies, and Interaction Enthalpies for the Methylated Bases.

	Zero Point Vibrational				$\Delta H_{\text{Interaction}}$[a]		
Interaction Energy	Pur	Pyr	Dimer	ΔE_{vib}	C27	exp	ai
AT Watson–Crick							
−13.0	83.77	85.78	170.95	1.40	−8.99	13.0	7.8–11.9
AT Hoogsteen							
−13.3	83.77	85.78	170.84	1.29	−9.37	13.0	8.4–12.8
AU Watson–Crick							
−13.5	83.77	69.04	154.29	1.47	−9.38	14.5	
AU Hoogsteen							
−13.9	83.77	69.04	154.20	1.39	−9.89	14.5	
GC Watson–Crick							
−25.8	87.30	76.34	165.60	1.91	−20.99	21.0	19.7–25.4

Energies in kcal/mol. Zero point vibrational energies were calculated using CHARMM27 at 300 K. The 4RT correction includes the rotational (3/2RT), translational (3/2RT), and ideal gas (PV) contributions. $\Delta H_{\text{interaction}}$ calculated equals the sum of the interaction energy, the zero point energy of the dimer minus the sum of the monomer zero point energies and the 4RT correction for the rotational, translational, and ideal gas terms.

[a] Experimental ΔH interaction energies from ref. 62 and *ab initio* ΔH interaction energies from ref. 59. The range of values are from different levels of theory used in that study.

interactions, consistent with the *ab initio* data, and the AU interaction being more favorable than the AT, consistent with the experimental trend.[62] Overall, the CHARMM27 interaction parameters are in good agreement with a variety of QM and experimental target data indicating the force field to properly represent intramolecular interactions among the different moieties comprising nucleic acids as well as between nucleic acids and their environment (see Supplementary Material). Proper treatment of both the intra and intermolecular interaction terms is particularly important for the force field properly treating the influence of environment on the structural and dynamic properties of DNA.

INTERNAL PARAMETERS

Optimization of the internal parameters involved reproduction of geometric and vibrational target data for the sugar moiety, the phosphodiester backbone, and the bases. This process is indicated in Fig. 1 as loop II. With the phosphodiester backbone, the sugar moiety, and the glycosyl linkage, a considerable part of the optimization effort involved adjustment of the dihedral parameters to simultaneously reproduce QM potential energy surfaces and probability distributions of those dihedrals in experimental crystal structures, as represented by loop IV in Fig. 1. To organize the presentation of the internal parameter optimization the results will be separated into a section describing the reproduction of the geometric and vibrational target data and a section describing the iterative optimization of selected dihedral parameters.

To allow for improved optimization of the geometries and vibrational spectra, additional atom types were added (see the topology file in the Appendix to the Supplementary Material). New atom types for the sugar and phosphodiester moieties were created for the C1′ and C5′ atoms and for the O4′ and C2′ atoms in RNA. With the bases, new atom types were created for the N3, C5, and N9 atoms in guanine, the N1, C2, and C5 atoms in thymine, and the Nl and C2 atoms in uracil. These additional atom types increase the number of parameters available for optimization, thereby allowing for improved agreement with the target data.

Reproduction of the Geometric and Vibrational Target Data

Optimization of the deoxyribose and ribose bond length and valence angle parameters was performed based on target data from a statistical analysis of high precision crystal structures of nucleosides and nucleotides.[63] In that study the deoxy and ribo structures, as well as the north and south conformations, were analyzed separately, allowing for explicit parametrization of these in the present study. To take into account the influence of base on the minimized structure a deoxy nucleoside (model compound G) was minimized with each of the four DNA bases. The same was done with the ribo nucleosides. Reported values are the average over the four DNA or RNA nucleosides (see Tables 12 and 13 of the Supplementary Material). For deoxyribose, the average absolute difference between the crystal and CHARMM27 bond lengths is 0.011 Å in the north conformation and 0.013 Å in the south conformation. For the valence angles, the average differences were 1.1° and 1.2° for the south and north conformations, respectively. Results for the ribose sugar were similar. Average absolute differences between CHARMM27 and the crystal data for the bond lengths were 0.010 Å for both the south and north conformations, and those for the angles were 1.6 and 1.0° for the south and north conformations, respectively. In the majority of cases the differences for the individual bonds and angles were less than error estimates for the experimental data, though exceptions do exist.

Bond, angle and dihedral force constants associated with the sugar moiety were initially obtained from the alkanes,[37] while those of the phosphodiester linkage were from CHARMM22. To optimize these force constants, vibrational spectra were calculated *ab initio* for the dianionic form of compound B, for compound F, and a variation of compound E with an imidazole base and a 5′ methyl group. The empirical force constants were then adjusted to reproduce the scaled *ab initio* data. Emphasis was placed on proper reproduction of the torsional and deformation modes associated with the furanose ring and of its exocyclic substituents. Tables 14 to 16 of the Supplementary Material include a detailed comparison the *ab initio* and CHARMM27 results. Final optimization of most of the dihedral parameters, however, was based on the conformational energetics, as discussed below.

A recent survey of the geometries of the nucleic acid bases[64] motivated reoptimization of the associated parameters. Presented in Table IV are RMS differences between CHARMM22 or CHARMM27 and the crystal data for the nonhydrogen atom bonds and angles. Data are also included for differences between the empirical models and *ab initio* data for the angles involving hydrogens (H-angles). Individual values for the different bases are presented in

TABLE IV.
RMS Differences of the Bonds and Valence Angles for the Methylated Bases.

	CHARMM27			CHARMM22		
Base	Bonds	Angles	H-Angles	Bonds	Angles	H-Angles
Adenine	0.005	0.2	1.3	0.012	0.9	2.7
Guanine	0.006	0.1	4.6	0.011	1.6	4.8
Cytosine	0.004	0.2	2.5	0.017	1.1	2.4
Uracil	0.005	0.3	0.3	0.019	1.1	0.5
Thymine	0.005	0.3	0.4	0.018	1.0	0.4

Rms differences with respect to crystal survey data.[64] H-angle target data based on HF/6-31G* optimized structures of the methylated bases. See Tables 18 and 19 of the Supplementary Material for original data.

Tables 18 and 19 of the Supplementary Material. The CHARMM27 values are in good agreement with the target data, in all cases being improved over CHARMM22. The largest deviations occur in the H-angle terms for the bases that contain amino groups. This is due to the assumption of planar geometries in the empirical force field.

In the present study the bases were assumed to be planar. This assumption contrasts results from *ab initio* calculations showing the base amino groups to have pyramidal character in the gas phase.[65, 66] Similar results have been obtained with the amide in the protein backbone based on calculations on N-methylacetamide; however, the amide is planar when involved in hydrogen bond interactions.[67] Based on those results it was assumed that the base amino groups would also be planar when involved in hydrogen bond interactions. This assumption was supported by *ab initio* calculations at the HF/6-31G* level on cytosine showing the presence of a single water hydrogen bonded to the N4 amino group to yield a planar structure (A. D. MacKerell, Jr., unpublished). Furthermore, in several *ab initio* studies involving hydrogen-bonded nucleic acid-base dimers planar geometries were obtained.[58–60] Thus, assuming that the base amino groups are always involved in some type of hydrogen bond, it is appropriate to treat the structures of the bases in the condensed phase as planar. Note that the force constants of the amino groups were adjusted to allow for significant deviations from planarity to occur (see Supplementary Material).

Optimization of the nucleic acid base force constants was performed via the reproduction of vibrational spectra. The amount of experimental and *ab initio* vibrational data on the bases is large (see MacKerell et al.,[25] Illich et al.,[68] Colarusso et at.,[69] and Aamouche et al.[70]), and the situation is complicated by the role of environment on the molecular vibrations. Consequently, it was decided to optimize the internal force constants based on HF/6-31G* gas phase vibrational spectra, which had been scaled by 0.9.[71] This approach may be expected to yield molecular vibrations that are representative of the experimental regimen. Detailed comparisons of the CHARMM27 and *ab initio* vibrational data, including assignments, are presented in Tables 20 to 24 of the Supplementary Material along with a detailed discussion. In summary, efforts focused on the low frequency modes, such as ring torsion and deformation modes, which make significant contributions to the molecular distortions of the bases that occur in MD simulations. The overall quality of agreement between the CHARMM27 and *ab initio* data was satisfactory, being better with the pyrimidines than the purines, due to simpler structures of the former.

Iterative Optimization of Selected Dihedral Parameters

Completion of the parameter optimization involved adjusting the dihedral parameters associated with the phosphodiester backbone, the furanose moiety, and the glycosyl linkage. This involved an iterative approach (loop IV in Fig. 1), maximizing agreement with QM potential energy surfaces for a series of model compounds, while simultaneously reproducing crystal dihedral distributions in condensed phase simulations. In previous studies we systematically investigated a variety of compounds to use as models for sugar puckering and the ε, γ, β, and χ dihedrals,[46, 48] leading to the selection of compounds G, B, C, D, and E (Fig. 2 (bottom)). For all model compounds that included the furanose moiety, dihedral energy surfaces were investigated with both the C2′ endo and C3′ endo furanose conformations. This was performed to en-

sure that conformational properties associated with both the north and south sugar puckers were adequately treated by the force field. For the condensed phase macromolecular portion of the optimization, MD simulations of the A and B crystals (Table I) were performed with the calculated dihedral distributions compared to NDB survey data. At certain stages of the optimization process, MD simulations of the EcoRI dodecamer, the CATTTGCATC decamer, and the CTCGAG hexamer were performed in solution. These acted as additional tests to ensure that the two crystal structures were not imparting undesirable characteristics on the force field due to properties particular to those structures. Results from these simulations are included in the accompanying manuscript.[35]

The dihedral parameters were initially adjusted to reproduce the *ab initio* conformational energetics of the model compounds as closely as possible for the regions populated by DNA. These parameters were then used to perform condensed phase MD simulations of A and B DNA in crystal environments, from which dihedral angle distributions were obtained and compared with the corresponding distributions from a survey of DNA crystal structures. Deviations between the MD and survey data were noted, and the dihedral parameters adjusted to enhance sampling in the MD simulations of regions poorly sampled previously. As discussed above, when it was deemed necessary to deviate from the QM model compound energy surfaces, the empirical surfaces were made "softer," such that the force field would be allowed to better sample conformational space rather than making a "harder" surface where the shape of energy wells would be narrowed and shifted to yield the correct dihedral distribution. An example of this procedure with γ is presented below. This approach ensures that the force field will not be constrained to canonical regions of conformational space, allowing for the surrounding environment, base sequence, and base composition to impact the regions of conformational space accessible to the phosphodiester backbone, the furanose moiety, and the glycosyl linkage.

Results for the γ dihedral will be presented as an example of the type of compromise made at the model compound level to reproduce the condensed phase properties. Emphasis in the initial fitting of this dihedral was placed on the g^+ conformation, which is the region most populated in DNA and RNA. Shown in Fig. 3A are three empirical γ surfaces for model compound B along with the *ab initio* data. In Fig. 3B, probability distributions from MD simulations of the B form crystal using the same

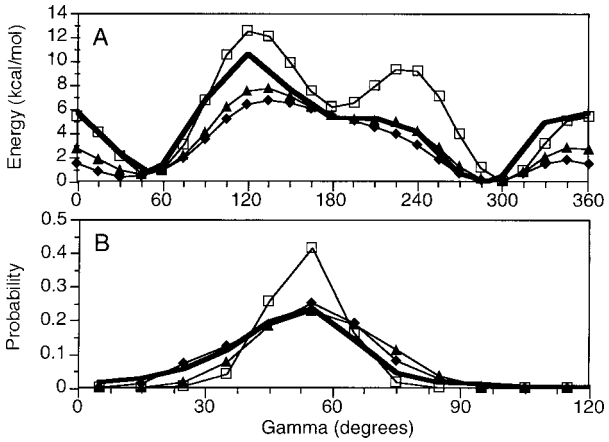

FIGURE 3. Potential energies (A) and probability distributions (B) as a function of the γ dihedral. The potential energy surfaces (A) were obtained with model compound B at the QM HF/6-31+G* (bold line) level of theory and for three empirical parameter sets designated 1 (□), 2 (▲), and 3 (♦). Backbone constraints for the surfaces in this figure were 168° for β, 298° for α and 262° for ζ. Probability distributions are from the NDB survey (bold line) for B form crystal structures and from the final 100 ps of 500 ps MD simulations of the CGATCGATCG B form crystal using the three empirical parameter sets designated 1 (□), 2 (▲), and 3 (♦). Note the change in the X-axis upon going from A (0 to 360°) to B (0 to 120°).

three parameter sets are presented along with the survey data for B form DNA structures. Note the change in the scale of the X-axis upon going from Fig. 3A to 3B. Parameter set 1 was optimized to reproduce the *ab initio* model compound data in the region of 0 to 90° (Fig. 3A). Use of that parameter set in MD simulations, however, results in a distribution of γ values much narrower than that obtained from the survey. The parameters were then adjusted to decrease the rise in energy upon departing from the minimum at 50° in the model compound (triangles in Fig. 3). This change led to better agreement between the MD and survey probability distributions, although the simulated distribution was still too narrow. This motivated additional adjustments yielding parameter set 3 (diamonds in Fig. 3), which is in the greatest disagreement with the model compound target data, but the best agreement concerning the survey data. Because the goal of the parameter development is for a force field to be used in condensed phase simulations, parameter set 3 was selected.

One point concerning the results in Fig. 3 should be emphasized. The energy surface for parameter set 3 is clearly "softer" than the *ab initio* target data,

allowing the DNA to more broadly sample conformational space in the MD simulation. The "softer" empirical surface may, in part, be a consequence of the limited sampling of the γ dihedral in the present MD simulations. It cannot be excluded that additional sampling, via longer or multiple simulations, may be required to properly sample the γ dihedral. If this were true, parameter set 1 may be the optimal choice for the final force field rather than set 3; this point is discussed in more detail in the Conclusion.

Application of the approach used for γ was applied to the remaining phosphodiester backbone dihedrals, sugar puckering, and the glycosyl linkage torsions. The *ab initio* and empirical potential energy surfaces for the remaining dihedrals along with the sugar pseudorotation angle are presented in Figs. 5 to 15 of the Supplementary Material. These efforts yielded a set of parameters that reproduced dihedral distributions from a survey of the NDB (see following paragraph). For the final parameter set, the average RMSD for the B crystal over the 600 ps of sampling (see Methods in the Supplementary Material) for all nonhydrogen atoms with respect to the experimental structure was 1.03 $\pm$ 0.08 Å, with the error being the standard deviation. For the A crystal, the corresponding values were 1.14 $\pm$ 0.08 Å.

Presented in Figs. 4 and 5 are the probability distributions from the MD simulations and the NDB survey of the A and B crystals, respectively, for the phosphodiester backbone and glycosyl linkage dihedrals and the sugar pseudorotation angle. Analysis of the figures shows the MD simulations to adequately reproduce the survey data in all cases. With the A crystal (Fig. 4) the largest discrepancies occur with ε, ζ, and χ. These differences were necessary to adequately treat the B form of DNA and to yield reasonable results for the small molecule model compound data (see Supplementary Material). In the B crystal, both the ε and ζ surfaces satisfactorily reproduce the NDB distributions, including sampling of the regions above 210° for ε and below 210° for ζ. These regions are related to the B_{II} conformer of the B form of DNA.[9, 72]

During optimization of the parameters associated with ε and ζ, the relative energies of the B_I and B_{II} conformers, based on model compound D, were included as target data (see Table 25 of the Supplementary Material). Other probability distributions of note in the B crystal are those of δ and of the sugar pseudorotation angle. Sampling of the pseudorotation angle in the MD simulation reasonably reproduces the survey distribution, extending to the region of 60 to 120°. Consistent with experiment,

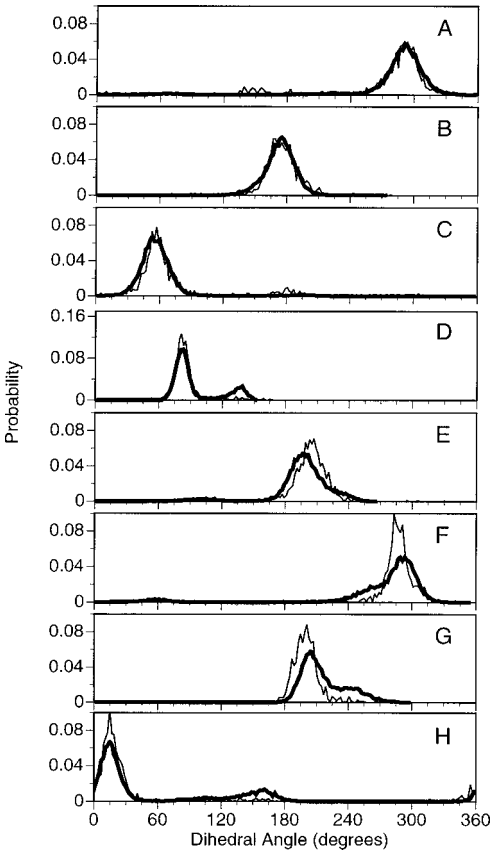

FIGURE 4. Probability distributions for dihedral angles α (A), β (B), γ (C), δ (D), ε (E), ζ (F), χ (G), and pseudorotation angle (H) from the NDB survey of A DNA crystal structures (thin lines) and from the A DNA GTACGTAC crystal simulation (bold lines).

this region is primarily sampled by pyrimidines, emphasizing the ability of CHARMM27 to properly treat base-dependent properties. Concerning δ, the MD distribution shows a bimodal distribution vs. a more continuous distribution in the crystal survey, even though the pseudorotation distribution is well modeled by the force field. The bimodal δ distribution may possibly be due to additional flexibility in the furanose ring and in the 5′ and 3′ covalent connectivity that is not properly treated in the present force field. Alternatively, sampling limitations in the MD simulations could make a contribution, and limitations in the experimental data cannot be excluded.

The correlation between sugar pucker and the glycosyl linkage, including the influence of the base, must be properly treated to account for the relation between sugar conformation and overall DNA structure. Presented in Table V are the pseudorotation angles, energies and glycosyl linkage dihe-

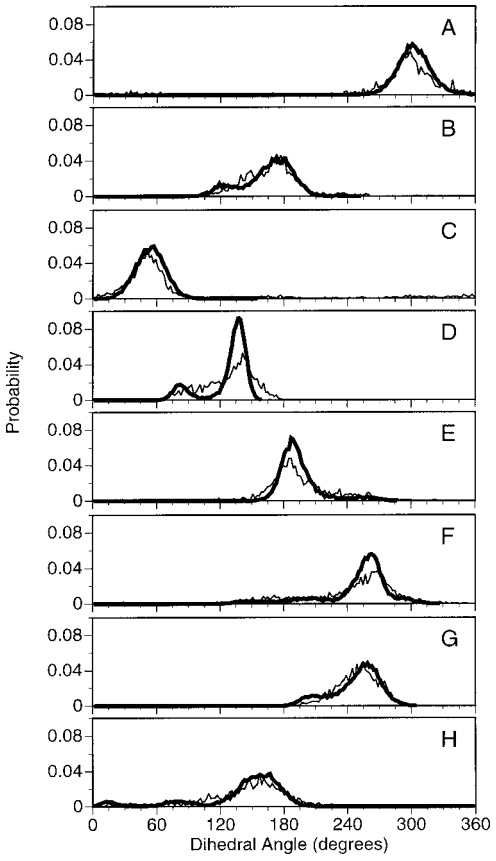

FIGURE 5. Probability distributions for dihedral angles α (A), β (B), γ (C), δ (D), ε (E), ζ (F), χ (G), and pseudorotation angle (H) from the NDB survey of B DNA crystal structures (thin lines) and from the B DNA CGATCGATCG crystal simulation (bold lines).

drals for the north and south sugar puckers from *ab initio* and CHARMM27 based on model compound G. The CHARMM27 pseudorotation angles of the north minima are all smaller than the *ab initio* values by approximately 10°, while the locations of the south minima are in good agreement with the *ab initio* values. For the south minima the pseudorotation angles are smaller in the pyrimidines vs. the purines, consistent with the *ab initio* data. The relative energies of the north and south conformers, $\Delta E_{N\text{-}S}$, are also compatible with the *ab initio* values. CHARMM27 adequately mimics the increase in $\Delta E_{N\text{-}S}$ upon going from adenine to guanine and from cytosine to thymine. With the pyrimidines, $\Delta E_{N\text{-}S}$ is less than in the *ab initio* data, while cytosine energetically favoring the north minimum is consistent with the *ab initio* result. Concerning the glycosyl linkage, the well-known correlation between sugar pucker and χ[8, 9, 63] is reproduced by CHARMM27. For both CHARMM27 and the *ab initio* data the value of χ with cytosine in the south minimum is significantly smaller than for the other bases. This property of cytosine, along with the energy of the north sugar pucker being lower than the south, has been suggested to contribute to the equilibrium between the A, B, and Z forms of DNA.[47]

TABLE V.
Deoxyribose Sugar Pseudorotation Angles and Energetics and Glycosyl Linkage Dihedrals for the North and South Sugars from *Ab Initio* and CHARMM27.

	Pseudorotation angles			
	P_n		P_s	
	a.i.	C27	a.i.	C27
Adenine	7.0	−3.1	168.3	164.4
Guanine	9.6	−0.5	168.6	165.2
Cytosine	8.8	−0.6	162.1	162.2
Thymine	12.4	0.9	162.7	161.0
	Pseudorotation energetics			
	$\Delta E_{N\text{-}S}$		B	
	a.i.	C27	a.i.	C27
Adenine	0.4	0.6	4.2	2.6
Guanine	0.7	1.0	4.3	2.9[a]
Cytosine	−0.3	−0.2	4.0	1.9
Thymine	0.9	0.2	4.0	2.2
	Glycosyl torsion[b]			
	P_n		P_s	
	a.i.	C27	a.i.	C27
Adenine	192	192	230	225
Guanine	198	204	233	235
Cytosine	195	194	207	207
Thymine	198	200	231	225

Pseudorotation angles (deg.) P_n and P_s correspond, respectively, to the north and south energy minima. $\Delta E_{N\text{-}S}$ (kcal/mol) is the energy of the north minimum minus the energy of the south minimum. B (kcal/mol) is the energy of the O4′ endo conformation relative to the global energy minimum (north or south). *Ab initio* data (a.i.) at the MP2/6-31G* level of theory.[47]

[a] Guanine barrier computed with β constrained to 180.0.

[b] Glycosyl torsions in degrees. *Ab initio* calculations (a.i.) at the MP2/6-31G* level of theory.[47]

RNA DIHEDRAL PARAMETRIZATION

Optimization of the dihedral parameters for RNA was performed following completion of the

DNA portion of the force field. This was based on the assumption that a set of nucleic acid parameters that represent both the A and B forms of DNA would also be appropriate for RNA. This was verified by preliminary simulations of RNA using a first-order approximation of the dihedral parameters unique to the ribose moiety showing them to yield the expected A form RNA structure (not shown). Additional optimization of the dihedral parameters was performed to reproduce model compound conformational energetics while maximizing agreement with crystal survey data on the dihedral probability distributions in RNA, consistent with loop IV in Fig. 1 (see Figs. 16 and 17 of the Supplementary Material). For the final parameter set, MD simulations on the UAAGGAGGUGUA dodecamer (Table I) yielded an RMSD for all nonhydrogen atoms in base pairs 2 through 11 of 1.9 ± 0.6 and 5.9 ± 4 Å, with respect to canonical A and B structures, showing the RNA structure to remain close to the canonical A form.

Presented in Fig. 6 are the dihedral and sugar pseudorotation angle probability distributions from the RNA MD simulation and the NDB survey data on RNA structures. For the majority of terms the agreement is good, with the range of values sampled in the simulation corresponding to that in the survey. The largest discrepancies occur with δ and the pseudorotation angle. In both cases the MD distributions are somewhat offset from the NDB survey. The survey distribution for the pseudorotation angle is non-Gaussian, suggesting that limitations in the RNA experimental structure determinations may be present. Because the overall range of pseudorotation angles sampled in the MD simulation agrees with the survey data and the satisfactory behavior of CHARMM27 in treating the deoxy sugars, additional optimization of the parameters was not performed. Thus, based on both RMSD and dihedral and sugar pseudorotation angle distributions the present force field adequately models duplex RNA.

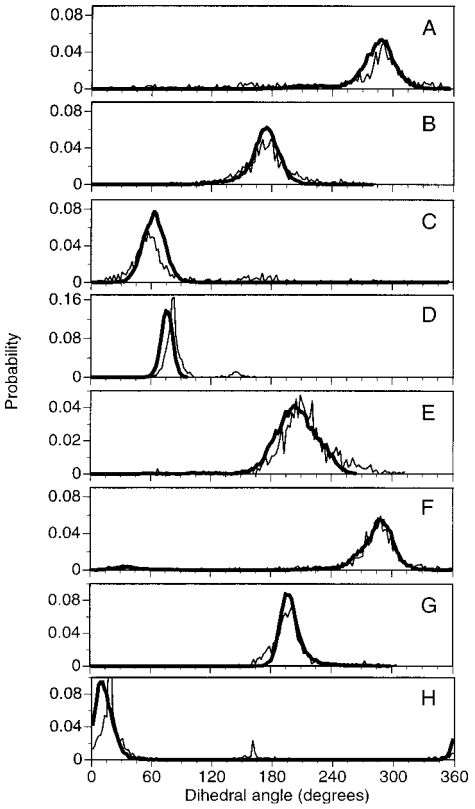

FIGURE 6. Probability distributions of dihedral angles α (A), β (B), γ (C), δ (D), ε (E), ζ (F), χ (G), and pseudorotation angle (H) from the NDB survey for RNA duplex and transfer RNA crystal structures (thin lines) and from the r(UAAGGAGGUGUA) RNA dodecamer solution simulation (bold line).

Z DNA CRYSTAL SIMULATION

In the present work the A and B forms of DNA and A form RNA were considered explicitly during the parameter optimization, while Z DNA was not. To determine the applicability of CHARMM27 for simulations of Z DNA and perform an additional test of the generality of the force field, a 1-ns MD simulation of the Z DNA CGCGCG hexamer[73] in its crystal environment was performed. Results, presented in Fig. 7, were obtained over the final 800 ps of the 1-ns simulation; the average RMS difference for all nonhydrogen atoms with respect to the crystal structure was 0.83 ± 0.09 Å.

Analysis of the simulated probability distributions for the backbone dihedrals, χ, and the pseudorotation angle are generally in satisfactory agreement with survey results (Fig. 7), but deviations do exist. The largest discrepancies occur with χ and the sugar pseudorotation angle. With χ, a shoulder ranging from 90 to 120° is present in the MD results, which is not observed in the survey. The MD pseudorotation angle distribution shows a small peak in the region of 80° that is not present in the survey, and the survey peak in the vicinity of 30° is shifted to lower values in the simulation, as is the larger peak centered around 150°. Overall, the present force field yields a reasonable representation of Z DNA, although of a lesser quality than with the A and B forms of DNA. An indepth discussion of the limitations in the treatment of Z DNA

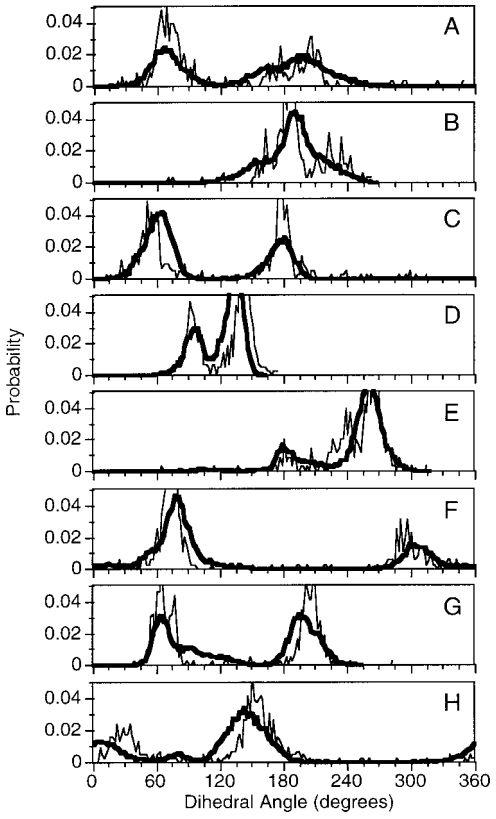

FIGURE 7. Probability distributions for dihedral angles α (A), β (B), γ (C), δ (D), ε (E), ζ (F), χ (G), and pseudorotation angle (H) from the NDB survey of Z DNA crystal structures (thin lines) and from the Z DNA CGCGCG hexamer crystal simulation (bold lines).

and possible causes are included in the Supplementary Material.

Conclusion

Presented is the CHARMM27 all-atom force field for molecular modeling and simulation studies of nucleic acids in the condensed phase. Extensive optimization of the parameters combined with the availability of additional target data allowed for significant improvements over the CHARMM22 nucleic acid force field. For MD simulations, CHARMM27 now yields B-like DNA structures in aqueous solution, while correctly changing to an A form conformation in low-water activity environments.[32, 35] Furthermore, CHARMM27 properly yields A form RNA in solution, and reasonably reproduces the structure of Z DNA in its crystal environment. The ability to treat these different nucleic acids is associated with the overall balance between and among the interaction and internal terms in the force field.

Improved interaction parameters allow for more accurate nonbond interactions between nucleic acids and their environment and between different moieties within the nucleic acids themselves. The quality of the interaction parameters is evident from the good agreement with a variety of target data, including interactions with water, base–base interactions, dipole moment, crystal geometries, and heats of sublimation. Recent calculations of the binding free energies of bases in chloroform using a continuum model of the solvent show the CHARMM27 base nonbonded parameters to yield good agreement with the experiment.[74] Thus, the CHARMM27 interaction parameters appear to work well in a variety of environments, including changes in water activity required to treat the equilibrium between the A and B forms of DNA.

Internal parameters are significantly improved over CHARMM22. Geometries of the sugars, including their exocyclic substituents, are in excellent agreement with small molecule crystal survey data. Vibrational properties of these moieties are in good agreement with *ab initio* data concerning both the frequencies and assignments. The importance of the improved representation of bond lengths and valence angles in nucleic acid force fields has been shown in recent refinements of experimental structures.[7, 31] It was found that using the bond lengths and valence angles derived from Gelbin et al.[63] led to better agreement between the calculated structures and the experimental data. The good agreement of CHARMM27 with the Gelbin et al. survey data, along with the physically relevant force constants, should be seen in this context.

Notable is the ability of the force field to account for subtle changes in the energetics at the nucleoside level as a function of base (Table V and Tables 26 and 27 in the Supplementary Material). This includes energetic stabilization of the north conformation over the south by the cytosine nucleoside, and the presence of an A-type conformation of the glycosyl linkage in the south conformation in that same nucleoside. These properties have been suggested to contribute to the equilibrium between the A, B, and Z forms of DNA.[47] The ability of the force field to reproduce base-dependent properties on the model compound level may facilitate studying these properties in oligonucleotides.

Central to the quality of the CHARMM27 nucleic acid force field was the simultaneous inclusion of target data based on small model compounds and condensed phase data from DNA and RNA. This al-

lows for calibration of the contributions of different moieties in nucleic acids, as judged by calculations on the model compounds, to the overall structure and energetics of the macromolecules. Ideally, both the small molecule and macromolecular target data would be accurately reproduced. Although this has only partially been achieved in the present work, knowledge of the quality of the agreement at the model compound level has two advantages; (1) it allows for understanding of possible contributions from the force field to results from modeling and MD studies, and (2) it indicates where improvements in the force field can be achieved. Several factors may be contributing to the inability to simultaneously reproduce both small model compound and macromolecular target data. These include the form of the potential energy function, the quality of the target data, and the ability to effectively sample conformational space in the MD simulations.

The form of the potential energy function in eq. (1) represents one of the simplest mathematical models used in molecular mechanics. To date, extensions of the form of the function have mostly involved additional internal terms.[39, 40, 75] These extended models have been successful in treating small molecules, typically in the gas phase, however, they have not led to improvements over biological force fields that use potential energy functions identical or similar to eq. (1).[24, 38, 41] The success of these biomolecular force fields is due to enhanced optimization of the parameters in the potential energy function, a goal we have attempted to extend in the present study. Extension of the energy function in eq. (1) has also involved the interaction portion of the force field, with the most common being the inclusion of electronic polarizability.[76] Although improvements associated with electronic polarizability have been made,[77] cases also exist where enhanced parameter optimization has overcome limitations previously ascribed to the omission of explicit electronic polarizability.[78, 79] Additional work is requried to determine if simultaneous agreement with both the model compound and macromolecular target data may be obtained via the addition of electronic polarizability or other terms in the potential energy function.

High-quality model compound target data is essential for accurate force field optimization. Prior to performing the present work, a large number of *ab initio* calculations had to be performed on the model compounds shown in Fig. 2 (bottom)[46–48] to generate and validate the target data. During those studies the relevance of both the QM level of theory and the composition of the model compounds was tested. For the furanose containing compounds it was shown that the MP2/6-31G* level of theory (MP2/6-31+G* level for charged species) yields satisfactory agreement with experimental data; accordingly, that level of theory was primarily used as the *ab initio* target data in the present study. Although use of MP2 is an improvement over HF treatment, some studies indicate that larger basis sets and alternative treatments of electron correlation can impact the calculated energetic properties,[80, 81] suggesting that higher level QM data may be required. Alternatively, as discussed with DMP,[82] the presence of solvent can significantly alter conformational energetics. Ideally, solvation effects should be taken into account by the force field via the explicit inclusion of solvent, but in certain cases it is necessary to include solvation contributions to the target data.[41] Although the size and composition of the model compounds used in the present study was tested, it may be necessary to use even larger compounds. This is indicated by the deviation between the empirical and *ab initio* energetic data for the sugar pseudorotation model compounds (see Fig. 14A and B of the Supplementary Material for compounds F and G^I, respectively). Thus, future efforts are required to better assess the influence of QM methods and model compound composition on potential energy data that, when applied directly to macromolecular MD simulations, yield better agreement with macromolecular target data.

With several of the model compounds the empirical energy surfaces had to be made "softer" compared to the *ab initio* surfaces to allow for reproduction of the crystal dihedral distributions by the MD simulations. The best examples were the model compounds associated with α, ζ, and γ. Although this may be related to the QM method and model compound composition, the "softening" of these surface may be due to the present assumption that MD simulations of a few sequences on a nanosecond time scale should reproduce survey data from a large number of crystal structures. Comparison with the survey data, however, may require simulations over time scales much greater than a nanosecond on a wide variety of DNA and RNA sequences to adequately sample the conformational space observed in the survey results. Because current technology disallows rigorously testing of these limitations, it is important that users of the force field are aware of the assumption, and interpret results accordingly. It is expected that increases in computational power, algorithmic advances,[83] and use of multiple simulations[84] will allow for the present assumption to be tested more rigorously.

The present parameter optimization approach may be compared to AMBER96[24] as well as with two recently published force fields for nucleic acids—the BMS force field[32] and the revised AMBER98.[30] AMBER96 was based primarily on small molecule data, with the majority of parameters directly transferred from small model compounds (e.g., alkanes or dimethylether) with additional optimization of the parameters performed to reproduce DNA-based small molecule (e.g., DMP, the bases, deoxyadenosine) target data. No condensed phase simulations of oligonucleotides were included in the optimization process. The BMS force field was optimized to reproduce crystal survey data and the influence of environment on the equilibrium between the A and B forms of DNA.[32] Parameter adjustment in that work was done primarily in an empirical fashion, with only a few direct comparison of model compound empirical and *ab initio* data performed. The second new force field is a revision of the AMBER96 nucleic acid force field (AMBER98).[24, 30] Revisions involved additional optimization of selected dihedrals associated with the sugar moiety and the glycosyl linkage to improve the sugar pseudorotation angle distribution and the overall helical twist obtained from MD simulations. Comparison are made between AMBER98 and *ab initio* data for the four DNA nucleosides with respect to sugar puckering, χ and γ; however, no other details concerning the remaining degrees of freedom or nonbond interactions are reported. The AMBER98 force field did yield improvements in the targeted properties; however, sensitivity of the force field to environmental conditions appeared to be sacrificed. Although BMS, AMBER98, and CHARMM27 all rely on condensed-phase MD simulations at the final stages of the optimization, only with CHARMM27 is careful evaluation of the contributions of individual moieties describing all torsional degrees of freedom in the nucleic acids performed. Such information is essential for an understanding of the balance between different aspect of the force field that combine to yield the obtained condensed phase properties. Additional aspects of these force fields with respect to DNA and RNA duplex solution simulations are presented in the accompanying manuscript.[35]

It is hoped that the present work will extend the applicability of empirical force field approaches to study biological systems, including refinement of nucleic acid structures based on NMR data. Extensive validation of the force field for solution simulations is presented in the accompanying manuscript.[35] The present parameters were designed to be compatible with the CHARMM all-atom force fields for proteins[41] and lipids,[85] allowing for simulations of nucleic acid–protein and nucleic–acid lipid complexes. A refined version of the lipid force field is in progress (A. D. MacKerell, Jr. and S. Feller, work in progress). The CHARMM27 nucleic acid force field represents a careful and systematic optimization of empirical force field parameters. Although the level of rigor has made evident a number of limitations, such knowledge will enhance its utility by allowing the user to better understand its strengths and weaknesses as required for its application.

Acknowledgments

We thank the NSF PACI program, DOD ASC Major Shared Resource Computing, and High Performance Computing, the Pittsburgh Supercomputing Center, and NCI's Frederick Biomedical Supercomputing Center for providing computational resources. Appreciation to Drs. T. Cheatham, M. Feig, D. Langely, L. Nilsson, B. M. Pettitt, and D. Strahs for preliminary tests of the force field during its development; to N. Banavali and Dr. N. Pastor for helpful discussions; and C. Zardecki of the Nucleic Acids Database.

Supplementary Material

Included in the Supplementary Material is a full version of the present manuscript including a detailed description of the methods and a thorough analysis and discussion of the results. Included in the Appendix to the Supplementary Material is (1) a table of the model compounds used in the present study and the corresponding residue and patch names in the CHARMM topology file; (2) the CHARMM27 topology file; and (3) the CHARMM27 parameter file. The topology and parameter tables are presented in CHARMM format, allowing for their direct use in the program CHARMM. The topology and parameter files may also be accessed via A.D.M.'s web page at www.pharmacy.umaryland.edu/˜alex. CHARMM may be obtained via the following email address: marci@tammy.harvard.edu.

References

1. Brooks, C. L., III; Karplus, M.; Pettitt, B. M. Proteins, A Theoretical Perspective Dynamics, Structure, and Thermodynamics; John Wiley and Sons: New York, 1988; vol. LXXI.

2. McCammon, J. A.; Harvey, S. C. Dynamics of Proteins and Nucleic Acids; Cambridge University Press: New York, 1987.
3. Parkinson, G.; Vojtechovsky, J.; Clowney, L.; Brünger, A.; Berman, H. M. Acta Crystallogr Sect D 1996, 52, 57.
4. Hahn, M.; Heinemann, U. Acta Crystallogr 1993, D49, 468.
5. Ulyanov, N. B.; Schmitz, U.; Kumar, A.; James, T. L. Biophys J 1995, 68, 13.
6. Schmitz, U.; James, T. L. Methods Enzymol 1995, 261, 3.
7. Rife, J. P.; Stallings, S. C.; Correll, C. C.; Dallas, A.; Steitz, T. A.; Moore, P. B. Biophys J 1999, 76, 65.
8. Dickerson, R. E.; Drew, H. R.; Conner, B. N.; Wing, R. M.; Fratini, A. V.; Kopka, M. L. Science 1982, 216, 475.
9. Hartmann, B.; Lavery, R. Q Rev Biophys 1996, 29, 309.
10. 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.
11. Jain, S.; Sundaralingam, M. J Biol Chem 1989, 264, 12780.
12. Shakked, Z.; Guerstein–Guzikevitch, G.; Eisenstein, M.; Frolow, F.; Rabinovitch, D. Nature 1989, 342, 456.
13. Lipanov, A.; Kopka, M. L.; Kaczor–Grzeskowiak, M.; Quintana, J.; Dickerson, R. E. Biochemistry 1993, 32, 1373.
14. Dickerson, R. E.; Goodsell, D. S.; Neidle, S. Proc Natl Acad Sci USA 1994, 91, 3579.
15. Metlzer, W. J.; Wang, C.; Kitchen, D. B.; Levy, R. M.; Pardi, A. J Mol Biol 1990, 214, 711.
16. Allain, F; Varini, G. J Mol Biol 1997, 267, 338.
17. Saenger, W. Principles of Nucleic Acid Structure; Springer-Verlag: New York, 1984.
18. Norberg, J.; Nilsson, L. J Phys Chem 1996, 100, 2550.
19. Auffinger, P.; Westhof, E. Biophys J 1996, 71, 940.
20. Yang, L.; Pettitt, B. M. J Phys Chem 1996, 100, 2550.
21. Cheatham, T. E., III; Kollman, P. A. Structure 1997, 5, 1297.
22. Young, M. A.; Ravishanker, G.; Beveridge, D. L. Biophys J 1997, 73, 2313.
23. Flatters, D.; Young, M.; Beveridge, D. L.; Lavery, R. J Biomol Struct Dynam 1997, 14, 757.
24. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, J. K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
25. MacKerell, A. D., Jr.; Wiórkiewicz–Kuczera, J.; Karplus, M. J Am Chem Soc 1995, 117, 11946.
26. Feig, M.; Pettitt, B. M. Biophys J 1998, 75, 134.
27. MacKerell, A. D., Jr. In Molecular Modeling of Nucleic Acids; Leontis, N. B., SantaLucia, J., Jr., Eds.; American Chemical Society: Washington, DC, 1998; p. 304, vol. 682.
28. MacKerell, A. D., Jr. J Phys Chem B 1997, 101, 646.
29. Pastor, N.; Pardo, L.; Weinstein, H. Biophys J 1997, 73, 640.
30. Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. J Biomol Struct Dynam 1999, 16, 845.
31. Shui, X.; McFail–Isom, L.; Hu, G. G.; Williams, L. D. Biochemistry 1998, 37, 8341.
32. Langley, D. R. J Biomol Struct Dynam 1998, 16, 487.
33. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; Stales, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
34. MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In Encyclopedia of Computational Chemistry; Shleyer, P. v. R.; Clark, T; Allinger, N. L.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F., III; Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; p. 271, vol. 1.
35. MacKerell, A. D., Jr.; Banavali, N. J Comput Chem 2000, 21, 105.
36. MacKerell, A. D., Jr. In Computational Biochemistry and Biophysics; Watanabe, M.; MacKerell, A. D., Jr.; Roux, B.; Becker, O. M., Eds.; Marcel Dekker, Inc.: New York, to appear.
37. Yin, D.; MacKerell, A. D., Jr. J Comput Chem 1998, 19, 334.
38. Jorgensen, W. L.; Tirado–Rives, J. J Am Chem Soc 1988, 110, 1657.
39. Halgren, T. A. J Comp Chem 1996, 77, 490.
40. Lii, J.-L.; Allinger, N. L. J Comp Chem 1991, 12, 186.
41. MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. J Phys Chem B 1998, 102, 3586.
42. MacKerell, A. D., Jr.; Karplus, M. J Phys Chem 1991, 95, 10559.
43. Reiher, W. E., III Theoretical Studies of Hydrogen Bonding. Ph.D., Harvard University (1985).
44. Jorgensen, W. L. J Phys Chem 1986, 90, 1276.
45. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. J Chem Phys 1983, 79, 926.
46. Foloppe, N.; MacKerell, A. D., Jr. J Phys Chem B 1998, 102, 6669.
47. Foloppe, N.; MacKerell, A. D., Jr. Biophys J 1999, 76, 3206.
48. Foloppe, N.; MacKerell, A. D., Jr. J Phys Chem B, in press.
49. Grzeskowiak, K.; Yanagi, K.; Prive, G. G.; Dickerson, R. E. J Biol Chem 1991, 266, 8861.
50. Langlois D'Estaintot, B.; Dautant, A.; Courseille, C.; Precigoux, G. Eur J Biochem 1993, 213, 673.
51. Drew, H. R.; Dickerson, R. E. J Mol Biol 1981, 151, 535.
52. Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. S. Proc Natl Acad Sci USA 1981, 78, 2179.
53. Weisz, K.; Shafer, R. H.; Egan, W.; James, T. L. Biochemistry 1992, 31, 7477.
54. Weisz, K.; Shafer, R.; Egan, W.; James, T. L. Biochemistry 1994, 33, 354.
55. Wahl, M. C.; Rao, S. T.; Sundaralingam, M. Biophys J 1996, 70, 2857.
56. Schindelin, H.; Zhang, M.; Bald, R.; Fuerste, J.-P.; Erdmann, V. A.; Heinemann, U. J Mol Biol 1995, 249, 595.
57. Alhambra, C.; Luque, F. J.; Gago, F.; Orozco, M. J Phys Chem B 1997, 101, 3846.
58. Brameld, K.; Dasgupta, S.; Goddard, W. A., III J Phys Chem B 1997, 101, 4851.
59. Gould, I. R.; Kollman, P. A. J Am Chem Soc 1994, 116, 2493.
60. Hobza, P.; Kabelac, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. J Comp Chem 1997, 18, 1136.
61. Sponer, J.; Leszczynski, J.; Hobza, P. J Phys Chem 1996, 100, 5590.
62. Yanson, I. K.; Teplitsky, A. B.; Sukhodub, L. F. Biopolymers 1979, 18, 1149.

63. Gelbin, A.; Schneider, B.; Clowney, L.; Hsieh, S.-H.; Olsen, W. K.; Berman, H. M. J. J Am Chem Soc 1996, 118, 519.
64. Clowney, L.; Jain, S. C.; Srinivasan, A. R.; Westbrook, J.; Olson, W. K.; Berman, H. M. J Am Chem Soc 1996, 118, 509.
65. Leszczynski, J. J Phys Chem A 1998, 102, 2357.
66. Sponer, J.; Hobza, P. J Phys Chem 1994, 98, 3161.
67. Guo, H.; Karplus, M. J Phys Chem 1994, 98, 7104.
68. Illich, P.; Hemann, C. F.; Hille, R. J Phys Chem B 1997, 101, 10923.
69. Colarusso, P.; Zhang, K.; Guo, B.; Bernath, P. F. Chem Phys Lett 1997, 269, 39.
70. Aamouche, A.; Ghomi, M.; Grajcar, L.; Baron, M. H.; Romain, F.; Baumruk, V.; Stepanek, J.; Coulombeau, C.; Jobic, H.; Berthier, G. J Phys Chem A 1997, 101, 10063.
71. Scott, A. P.; Radom, L. J Phys Chem 1996, 100, 16502.
72. Dickerson, R. E. Methods Enzymol 1992, 211, 67.
73. Gessner, R. V.; Quigley, G. J.; Wang, A. W.-J.; van der Marel, G. A.; van Boom, J. H.; Rich, A. Biochemistry 1985, 24, 237.
74. Luo, R. D. L.; Head, M. S.; Given, J. A.; Gilson, M. K. Biophys Chem 1999, 78, 183.
75. Hagler, A. T.; Maple, J. R.; Thacher, T. S.; Fitzgerald, G. B.; Dinur, U. In Computer Simulation of Biomolecular Systems; van Gunsteren, W. P., Weiner, P. K., Eds.; ESCOM: Leiden, 1989; p. 149.
76. Gresh, N. J Chim Phys 1997, 94, 1365.
77. Meng, E. C.; Cieplak, P.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1994, 116, 12061.
78. MacKerell, A. D., Jr. J Phys Chem 1995, 99, 1846.
79. Rizzo, R. C.; Jorgensen, W. L. J Am Chem Soc 1999, 121, 4827.
80. Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. J Am Chem Soc 1997, 119, 5908.
81. Halgren, T. A. J Comp Chem 1999, 20, 730.
82. MacKerell, A. D., Jr. J Chim Phys 1997, 94, 1436.
83. Wu, X.; Wang, S. J Phys Chem B 1998, 102, 7238.
84. Caves, L. S. D.; Evanseck, J. D.; Karplus, M. Protein Sci 1998, 7, 649.
85. Schlenkrich, M.; Brickmann, J.; MacKerell, A. D., Jr.; Karplus, M. In Biological Membranes: A Molecular Perspective from Computation and Experiment; Merz, K. M., Roux, B., Eds.; Birkhaüser: Boston, 1996; p. 31.