Supplementary Material to

Nicolas Foloppe[1] and Alexander D. MacKerell, Jr.* "All-atom empirical force field for nucleic acids: 1) Parameter optimization based on small molecule and condensed phase macromolecular target data." Journal of Computational Chemistry, 21:86-104, 2000.

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

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

***** To whom correspondence should be addressed.

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

Abbreviations: ABNR: adopted-basis Newton-Raphson; BSSE: basis-set superposition error; DNA: deoxyribonucleic acid; HF: Hartree-Fock; LJ: Lennard-Jones; MD: molecular dynamics; MP2: second order Møller-Plesset; NMR: nuclear magnetic resonance; NRAP: Newton-Raphson; PME: particle mesh Ewald; QM: quantum mechanics; RMS: root mean square; RMSD: RMS difference; RNA: ribonucleic acid; SD: steepest descent.

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 insure 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, A.D., Jr. and Banavali, N., J. Comp. 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.

1. 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 which 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 non-physiological 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 deoxyribo-oligonucleotides 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 which 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 experiment can be achieved.

Following the introduction, the parametrization approach used in the optimization of the CHARMM27 nucleic acid force field is described. Details of the calculations are included in the Methods Section which is followed by a Results and Discussion Section. Inclusion of the discussion with the results allows for emphasis on the actual implementation of the parametrization approach to be discussed alongside the appropriate data. A concluding section reiterates a number of points of emphasis in the present parametrization work and discusses several issues associated with force field optimization. An accompanying manuscript applies the CHARMM27 parameters to MD simulations of DNA and RNA in solution.[35]

2. Parametrization Approach

2.1 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 equation 1.

$$U(\vec{R}) = \sum_{\text{bonds}} K_b(b - b_o)^2 + \sum_{\text{UB}} K_{UB}(S - S_o)^2 + \sum_{\text{angle}} K_\theta(\theta - \theta_o)^2 +$$

$$\sum_{\text{dihedrals}} K_\chi(1 + \cos(n\chi - \delta)) + \sum_{\text{impropers}} K_{imp}(\varphi - \varphi_o)^2 + \qquad \textbf{(1)}$$

$$\sum_{\text{nonbond}} \varepsilon_{ij}\left[\left(\frac{Rmin_{ij}}{r_{ij}}\right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{e r_{ij}}$$

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 (ε_{ij} = sqrt(ε_i * ε_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.

2.2 Parameter Optimization Strategy

The ability of equation 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 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 Figure 1 is a flow diagram of the parameter optimization procedure. Loops I, II and III in Figure 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 Figure 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 Figure 1. The need for this additional loop may reflect limitations associated with the level of theory of the QM data as well as the simplified form of the potential energy function in equation 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 Figure 1). The partial atomic charges and LJ parameters were then iteratively adjusted using the new minimum energy geometries (loop I in Figure 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 Figure 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 (NDB)[10] of DNA and RNA crystal structures. Presented in Figure 2A is a schematic diagram of a G-C basepair 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 Figure 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.

2.3 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 Figure 2A, 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 Figure 2B. 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 occur in DNA, respectively. 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 (e.g. the O4'-C1'-N9-C4, O4'-C1'-N9-C8, C2'-C1'-N9-C4 and, C2'-C1'-N9-C8 dihedrals for the purines).

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.

2.4 Macromolecular Target Data

As discussed in section 2.2 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 1. Since 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 was chosen due to its high resolution and presence of several phosphodiester linkages in the B_{II} conformation and the A form GTACGTAC octamer because of the relatively high content of AT basepairs 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 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 1). The EcoRI recognition sequence is probably the most studied DNA oligomer, making its inclusion necessary as part of the present study.[17,49,50] The CATTTGCATC decamer was selected due to its structure being determined in solution via NMR, with emphasis on sugar puckering.[51,52] Inclusion of the CTCGAG hexamer[53] 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 was used (Table 1). Only one RNA duplex was included as target data given the greater homogeneity of RNA duplex structures as 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]

3. Methods.

Nucleic acid atom names and torsional angles are defined as in Saenger[17] and the same nomenclature is applied to the model compounds. The canonical A and B forms of the DNA are defined according to Arnott and Hukins[54] and the sugar pseudorotation angle and amplitude have been determined following Altona and Sundaralingam, using the same reference state for $P = 0.0°$.[55]

All empirical calculations were carried out with the CHARMM program[33,34] using a dielectric constant of 1.0. The water model in all calculations was the CHARMM-modified TIP3P.[43,45] Parameters for sodium are from Beglov and Roux[56] and the magnesium parameters are based on reproduction of the experimental free energy of solvation (B. Roux, personal communication).

3.1 Vacuum model compound calculations

QM calculations were carried out with the GAUSSIAN 94 program,[57] using the 6-31G* and 6-31+G* basis sets for neutral and negatively charges species, respectively. Torsional energy surfaces were performed at the Hartree-Fock (HF) level or with treatment of electron correlation via Møller-Plesset perturbation theory to the second order (MP2), as noted. QM energy minimizations were performed to the default tolerances in the GAUSSIAN program. Minimum interaction energies and geometries between model compounds and water were determined at the HF/6-31G* level by optimizing the interaction distance, and in some cases an interaction angle (Figure 3), with the intramolecular geometries constrained to the gas phase HF/6-31G* optimized structure for the model compound and the TIP3P geometry for water.[45] The interaction energy was determined as the total energy of the supermolecular complex minus the sum of the monomer energies; no correction for basis set superposition error (BSSE) was included.

Empirical calculations on the model compounds were carried out with no truncation of nonbond interactions, unless noted. Energy minimizations involved 50 to 200 steps of steepest descent (SD) followed by 50 to 200 steps of adopted basis Newton-Raphson (ABNR) and 50 Newton-Raphson (NR) steps to a final energy gradient of 10^{-6} kcal/mol/Å. Energy surfaces were performed by harmonically constraining the selected dihedral with a force constant of 10,000 kcal/mol/degree^2. Minimum interaction energies and geometries between model compounds and water were determined by varying the interaction distance, and in some cases an interaction angle, with the intramolecular geometries constrained to the empirical gas phase optimized structure for the model compound or the TIP3P geometry for water.[45] Interaction orientations were identical to those used in the QM calculations.

Sugar puckering surfaces were analyzed by dividing the pseudorotation space into four equally sized quadrants centered around $P = 0.0°$, $P = 90.0°$, $P = 180.0°$ and $P = 270.0°$, which are referred to as the north, east, south and west quadrants, respectively. Pseudorotation potential energy surfaces

were obtained by individually constraining one of the five furanose endocyclic dihedrals to values ranging from -40 to 40 in increments of 10 and allowing the remainder of the system to optimize. Pseudorotation angles were then calculated from the final optimized structures. The use of a single torsional constraint to enforce the pseudorotation angle was performed to allow for the amplitude of the ring to vary during the optimization. To obtain the north and south minimum structures the sugars were initially constrained to the C3'endo and C2'endo puckers, respectively, and optimized following which the constraints were removed and full optimization performed. The east barrier (P = 90.0°, O4'endo) was obtained by constraining the C1'-C2'-C3'-C4' dihedral to 0.0° with the furanose initially in the O4' conformation and optimizing the remainder of the structure. Constraints on the remaining degrees of freedom in the model compounds were applied as described in the following paragraph. In all cases an initial minimization in the presence of the constraints was followed by a minimization in the absence of the dihedral constraints, with only the appropriate sugar constraint maintained.

Torsional energy surfaces were sampled every 30° in the QM and every 15° in the empirical calculations. In all cases the furanose moiety was constrained to either the C3'endo (C4'-O4'-C1'-C2' = 0.0°) or C2'endo (C3'-C4'-O4'-C1' = 0.0°) pseudorotation angle, as noted, to avoid problems associated with switching between different furanose conformations in the energy surfaces. Additional degrees of freedom not being sampling explicitly in an energy surface were constrained to values corresponding to the A or B forms of DNA for the C3'endo and C2'endo pseudorotation angle constraints, respectively. The value of these constraints were obtained from fitting of survey data from the NDB, as described elsewhere,[48] and are as follows. For the A form α, β, γ, ε and ζ were 291, 175, 57, 205 and 287 , respectively, and for the B form they were 298, 168, 51, 187, and 262 , respectively. With the ribo analogs of compounds C, F and G alternate constraints were imposed to deal with the 2'hydroxyl group as follows. Optimizations were performed by initially minimizing with the dihedrals β, γ, ε, χ and C3'-C2'-O2'-H constrained to values of t, g^+, t, anti and 0.0 , respectively, followed by removal of those constraints and completion of the minimization. In all cases identical constraints were used for both the QM and empirical calculations.

3.2 Condensed phase simulations

Production MD simulations in the condensed phases were performed in the NPT ensemble[58] at 300 K with a time step of 0.002 ps and the Leap-Frog Integrator. All calculations were performed using SHAKE[59] to constrain covalent bonds involving hydrogens and images were generated using the CRYSTAL module[60] in CHARMM. Electrostatic interactions were treated via either atom truncation or the Ewald method.[61] Atom truncation was performed by using the force shift and force switch methods to smooth the electrostatic and LJ terms, respectively.[62] Nonbond pair lists were maintained to 14 Å, nonbond interactions were truncated at 12 Å and the LJ switching function was initiated at 10 Å. Nonbond lists were updated heuristically. Particle Mesh Ewald (PME)[63] calculations were

performed using the specified real space cutoffs with the LJ interactions truncated at the same distance. The fast Fourier transform grid densities were set to approximately 1 $Å^{-1}$. The screening parameter was determined for each system by using a 6th order smoothing spline and varying the screening parameter (kappa) from 0.20 to 0.50 and selecting the value at which the change in the energy as a function of kappa went to zero. In all cases the value of kappa was in the range of 0.28 to 0.35. RMS differences (RMSD) values relative to the experimental starting structures were determined following least-squares fitting of the specified non-hydrogen atoms.

Small molecule crystal calculations were performed with the full unitcells as the primary atoms with the symmetry of the unitcells maintained (e.g. for monoclinic systems the α and β angles were constrained to 90 while all other unitcell parameters were allowed to relax). These calculations were initiated by minimizing the entire system for 100 ABNR steps, followed by a 5 ps MD equilibration period and a MD production simulation of 50 ps. Averages and RMS fluctuations were obtained over the final 50 ps. For determination of the heats of sublimation, ΔH_{sub}, gas phase simulations of uracil and 9-methylthymine were required. These calculations were performed in an identical fashion to the crystal simulations, except that the temperature control was performed using the Nosé algorithm.[64] Vacuum simulations to obtain the gas phase energy required for determination of ΔH_{sub} from the PME simulations were performed by including all possible nonbond atom pairs.

Crystal calculations of the A and B DNA crystal structures listed in Table 1 were initiated by retrieving the experimental structures from the NDB.[10] Calculations were performed on the asymmetric unit or the minimum number of asymmetric units required to include all atoms of one nucleic acid duplex as the primary atoms. Hydrogen atoms were added to the crystal structures using the HBUILD module[65] of CHARMM and then subjected to a 50 SD step energy minimization with the non-hydrogen atoms fixed. Next, an appropriate amount of solvent was added to fill vacuum spaces in the crystal by generating the primary and image atoms (including DNA, water and counterions identified in the X-ray structure) and overlaying them with a waterbox of the same dimensions as the asymmetric unit(s). Added water molecules whose oxygen atoms were within 1.8 Å of any of the X-ray determined primary or image non-hydrogen atoms were then removed. The water deletion distance of 1.8 Å was determined by applying different values for the removal of waters, generating the starting configurations, minimizing and running 100 ps NVT simulations with the DNA constrained. The distance at which the pressure was close to 0 ATM was selected for the final system preparation. When the counterions identified in the experimental crystal structure were not adequate to neutralize the system, additional ions were added at random positions in the asymmetric unit(s). All water molecules and counterions were then energy minimized for 100 SD steps keeping the nucleic acid atoms fixed. This was followed by a 20 ps MD simulation in the NVT ensemble, with the nucleic acid atoms fixed. Following equilibration of the solvent, all atoms were subjected to a MD simulation in the NPT ensemble. The majority of crystal simulations were performed for 500 ps with analysis performed over

the final 100 ps. With the final parameters set, the final structure from the 500 ps simulation was used to initiate three 250 ps simulations that only differed by the random number seed used to assign the initial velocities to the atoms. The final 200 ps of each of these simulations were pooled and used for analysis.

The simulation of the Z DNA CGCGCG crystal was initiated from a previously equilibrated system that contains 106 water molecules, two sodium and four magnesium ions, as described elsewhere.[66] Calculations were performed using PME with a real space truncation distance of 10 Å with the MD simulations run at 288K, corresponding to the experimental regimen. Prior to the 1 ns production simulation the system was subjected to a 200 step ABNR minimization with harmonic constraints of 2.0 kcal/mol/Å^2 on all non-hydrogen atoms, a 100 ps NVT simulation with harmonic constraints 5.0 kcal/mol/Å^2 on all DNA non-hydrogen atoms and a 200 step ABNR minimization of the entire system.

Distributions of the dihedral angles and sugar pseudorotation angles in oligonucleotide crystal structures were obtained from the NDB[10] as of March 98. DNA structures containing non-standard DNA components, bound drugs or proteins were excluded. The distributions are presented as probability distributions by sorting the data into 2° bins and were obtained separately for the A, B and Z DNA families. RNA dihedral distributions include all RNA duplexes and transfer RNA structures.

4. Results and Discussion

Presentation of the results and discussion will be performed consistent with the flow diagram in Figure 1 and based on the terms in equation 1. The subsections will be separated into different portions associated with different model compounds. Except when noted, the presented results are for the converged parameters. The final parameters are included in the Appendix of the Supporting Information and may be obtained from the World Wide Web at www.pharmacy.ab.umd.edu/~alex/research.html.

4.1 Interaction terms

Optimization of the interaction terms involved refinement of the partial atomic charges for the ether oxygen in the furanose, the phosphate and the bases. Charges for the hydroxyl groups and the alkanes were consistent with the CHARMM22 set. For the LJ terms, new alkane parameters where used for the aliphatic groups[67] and LJ parameters for some of the atom types in the bases were optimized as part of the present study. The remaining terms are from CHARMM22. In the present discussion the terms interaction and nonbonded are used synonymously, referring to the partial atomic charges and LJ parameters. As emphasized by loop II in Figure 1 the interaction terms are influenced by the final values of the internal parameters.

4.1a Sugar and phosphodiester linkage

Partial atomic charges pertaining to the furanose moiety have been optimized using tetrahydrofuran (THF, complex A, Figure 3). 2'-hydroxy-tetrahydrofuran (THFOH, complex B, Figure 3) was used to test the charges of a hydroxyl group substituting the furanose at positions corresponding to the 2' or 3' carbons in nucleic acids. The hydroxyl charges were originally based on methanol and previously used directly without being assessed in the chemical context of the sugar. With THFOH, interactions with three orientations of the hydroxyl group were investigated. In all three cases the water plane was in the same plane as the C-O-H atoms. Comparison of the *ab initio* and empirical data in Table 2 shows the agreement to be satisfactory for both THF and THFOH. The minimum interaction energies were in good agreement with the scaled *ab initio* energies, while the minimum distances are all approximately 0.2 Å less than the *ab initio* values. The shorter interaction distances are required to reproduce condensed phase properties.[44] Of note was the quality of the fit for THFOH. The charges directly transferred from methanol adequately reproduce the interaction energies for a variety of orientations of the hydroxyl with respect to the tetrahydrofuran ring. The newly determined charge distribution on the furanose has been transferred to all other compounds containing this moiety.

Interactions between DMP and water (Figure 3, Complex C) in the CHARMM22 force field had those with the anionic oxygens too favorable (interactions 2, 3 and 4) and the interaction with the ester oxygen too unfavorable (interaction 1). To compensate for this imbalance the charges were

reoptimized yielding the interaction energies and geometries shown in Table 2. While the interaction with the ester oxygen (C.1 in Table 2) is still slightly too unfavorable and one of the interactions with the two anionic oxygens is slightly too favorable (C.3 in Table 2) the overall balance represents an improvement over CHARMM22. Obtaining this improvement required moving -0.02 e.u. from the anionic oxygens to the ester oxygens.

4.1b Bases

Recent high level *ab initio* data on Watson-Crick and Hoogsteen basepairing as well as other hydrogen-bonded basepairs and base stacking interactions have greatly increased the amount of target data available for charge optimization of the bases.[68-72] Results from some of these studies indicated limitations in the CHARMM22 nonbonded parameters.[69,71] These works stimulated the reoptimization of both the partial atomic charges and LJ parameters for the bases.

In addition to base-water interactions (see Figure 3 and Table 3), target data for optimization of the base charges included interaction energies for a variety of basepairs (see Table 4 and reference[71]) and dipole moments. Optimization of the LJ parameters was based on the reproduction of the heats of sublimation of the uracil and 9-methyladenine crystals along with interactions between water and aromatic base hydrogens. For example, R_{min} on the aromatic hydrogens was decreased from the value of 1.3582 Å used for benzene[73] to 1.10 Å. In the present work the flexibility in the base charges was increased by defining the unit charge group to encompass the entire base. This contrasts the CHARMM22 charges where unit charge groups of 7 atoms or less were used.

Included in Table 2 are the minimum interaction energies and geometries for base-water interactions from both the *ab initio* HF/6-31G* and CHARMM27 calculations. The empirical distances are approximately 0.2 Å shorter than the HF/6-31G* results while the interaction energies are equivalent to the scaled *ab initio* values, consistent with previous studies (see Section 4.1 and 4.1a). The largest discrepancies from these trends occur with the minimum distances between water and the aromatic base hydrogens. For example, for the Ade H8-OW and H2-OW interactions the *ab initio* distances are 2.39 and 2.49 Å, respectively, as compared to CHARMM27 values of 2.37 and 2.45 Å, respectively. Use of the standard aromatic hydrogen radius (see preceding paragraph) yielded empirical distances significantly longer than the *ab initio* values (not shown). Similar LJ parameters have previously been used for the hydrogen on the carbon between the two nitrogens in the imidazole sidechain of histidine.[41] The need for a smaller radius on the aromatic base hydrogens, as well as on imidazole, appears to be due to the carbon covalently bound to the hydrogen always being directly adjacent to one or two nitrogen atom in the rings, yielding a more polar character as compared to standard aromatic C-H groups.

An overall comparison of the present CHARMM27 and the CHARMM22 base to water interactions is presented in Table 3, where the average differences, RMSD, and average absolute errors

for the water-base interactions are reported. In all cases the CHARMM27 values are better than for CHARMM22, with the improvement being significant in all cases except adenine. This improvement is due to the use of larger unit charge groups and the new R_{min} value for the aromatic hydrogens.

Optimization of the base parameters also included the Watson-Crick and Hoogsteen basepair interactions along with the remaining base dimers defined by Hobza et al.[71] Presented in Tables 4 are the MP2/6-31G* BSSE corrected data along with their CHARMM27 counterparts. In addition to the total empirical interaction energies are the electrostatic, LJ and internal strain contributions to the basepairing interaction. In several instances the LJ and internal strain terms make significant contributions to the overall interaction energy. These contributions emphasize the need for proper treatment of the LJ and internal parameters to obtain a properly balanced force field. Comparison of the *ab initio* and CHARMM27 values in Table 4 show the agreement to generally be good, although with some differences. The largest differences occur with the GA3, GC1 and GG3 pairs. The GG3 interaction in CHARMM22 was reported to differ by -7.2 kcal/mol from the *ab initio* data; calculations in our laboratory yielded a difference of -6.5 kcal/mol. This difference is -2.2 kcal/mol in CHARMM27, a significant improvement. On the other hand, the GC1 interaction for CHARMM22 was reported to have a difference of -1.2 kcal/mol (-1.7 kcal/mol in our laboratory) which is worse with CHARMM27 (-2.5 kcal/mol).

To judge the overall improvement in CHARMM27 the data in Table 4 were subjected to linear regression analysis along with determination of the standard deviation and average absolute error, as previously performed.[71] These results are shown at the bottom of Table 4 along with the CHARMM22 values. Two sets of CHARMM22 values are reported based on calculations in the present study and values reported in Hobza et al.[71] Comparison of the CHARMM27 and CHARMM22 data show the new force field to be equivalent to or in better agreement with the *ab initio* data. This improvement indicates the CHARMM27 base charges and LJ parameters to yield a more balanced representation of the different types of hydrogen bonded base-pair dimers.

In addition to the water-base and base-base interactions, the dipole moments of the bases were also used as target data for the partial charge optimization. Presented in Table 5 are the CHARMM27 dipole moments along with those from experiment and QM calculations. The QM results include both gas phase dipole moments and those calculated in the presence of water using the AM1-SM2[74] and SCIPCM[75] reaction field models. The ordering of the CHARMM27 dipole moments is in good agreement with the QM data, with the only exception being the order of thymine and uracil. The CHARMM27 dipole moments are systematically larger than the gas phase QM values. This is required due to the omission of explicit electronic polarizability from the potential energy function (equation 1). Such an overestimation represents a mean field polarization of the bases by the surrounding condensed phase environment, an approach that is common in empirical force fields designed for the condensed phase, as discussed previously.[41] To better gauge the extent of the overestimation, dipole moments

were determined for the bases in water using two QM reaction field models. The reaction field dipole moments are all larger that the CHARMM27 values, suggesting that the extent of the overestimation of the base dipoles in the force field is reasonable. Better quantitation of the relevance of the dipole moments from the reaction field models with respect to force field models will require additional studies.

Optimization of the LJ parameters for the base ring carbon atoms was performed using crystal calculations of the systems listed in Table 6. For uracil and 9-methyladenine the ΔH_{sub} values have been experimentally determined.[76] 1-methylthymine, for which both the crystal structure and heats of sublimation are known, was not included due to problems with a pseudosymmetry axis in the molecule, as previously discussed.[25] Presented in Table 7 are the experimental and calculated unitcell parameters for the four crystals. Note that the calculations were performed with the full unitcell as the primary atoms, versus the use of the asymmetric unit in the CHARMM22 parametrization. Stereodiagrams of all the small molecule crystals may be seen in Figure 4 of MacKerell et al.[25] In the present calculations both atom-based truncation and PME methods were used for the treatment of the electrostatic interactions. The PME method was applied with three real space cutoff distances. This was performed due to the necessity of truncating the LJ terms at the same distance as the real space cutoff. That truncation eliminates possible favorable long-range dispersion effects that could impact the calculated crystal structure. With all four crystals the unitcell parameters and volumes are reasonably reproduced by the force field (Table 7). With uracil there is a slight contraction of the unitcell, which is primarily associated with the A and B lattice parameters while the β angle decreases. In contrast, a slight increase in the volume of the unitcell of 9-methyladenine occurs. The β angle of this system is well maintained while an increase in the B face and a decrease in the C term occur. For the 9-methyladenine/1-methylthymine and 9-ethylguanine/1-methylcytosine crystals only PME calculations at the two longest real-space cutoff distances were performed. With 9-methyladenine/1-methylthymine the volume and the β angle are well maintained while an increase in the B term and a decrease in the C term compensate each other. Results for the triclinic 9-ethylguanine/1-methylcytosine crystal tend to be in poorer agreement with experiment. Significant differences occur in the α and β angles as well as the three unitcell lengths. The lack of experimental crystals of the individual molecules in the 9-ethylguanine/1-methylcytosine crystal makes systematic analysis of the cause of the differences difficult. Table 7 also includes results from the CHARMM22 force field showing CHARMM27 to generally yield a better representation of the crystals than CHARMM22, although exceptions exist.

Heats of sublimation for the uracil and 9-methyladenine crystals offers additional target data for optimization of the LJ parameters. Table 8 includes CHARMM27 ΔH_{sub} values for the two crystals for different treatments of the nonbond interactions along with CHARMM22 and experimental data. For uracil CHARMM27 overestimates ΔH_{sub} by two or more kcal/mol. Of note is the increase in ΔH_{sub} in the PME versus atom-truncation calculations and as the truncation distance in the PME calculations is increased. The increase in ΔH_{sub} with the increased truncation distance is attributable to

the dispersion portion of the LJ term and is consistent with the decrease in the unitcell volume in Table 7. With 9-methyladenine the agreement between experiment and calculations is improved as compared to uracil, with the CHARMM27 values bracketing experiment. For both uracil and 9-methyladenine the CHARMM22 values are significantly larger than both the experimental and CHARMM27 values. The improved agreement between experiment and CHARMM27 concerning both the unitcell parameters and ΔH_{sub} as compared to CHARMM22 emphasizes the improvements in the new force field.

Differences in the CHARMM27 unitcell parameters and ΔH_{sub} with respect to experiment indicate that further improvements in the force field, possibly including extension of the form of the potential energy function, are needed. For example, ΔH_{sub} for 9-methyladenine is in good agreement with experiment, however, the calculated unitcell parameters are in poorer agreement, suggesting that the ideal balance of structure and energetics has still not been reached.

Additional tests of the base parameters involved calculation of the interaction geometries and enthalpies of the Watson-Crick and Hoogsteen basepairs and base stacking interaction energies. Table 9 contains CHARMM27 basepair interaction energies, enthalpies and vibrational contributions, along with the corresponding data from experiment and *ab initio* calculations. Experimental interaction energies were determined via mass spectrometry measurements.[76] *Ab initio* calculations of the interaction enthalpies have been performed at a variety of levels of theory up to MP2/DZP//HF/6-31G*(BSSE)[70] and LMP2/cc-pVTZ//HF/cc-pVTZ.[69] In those calculations the interaction strength of the Hoogsteen AT pair is consistently more favorable than in the Watson-Crick AT pair, although the difference is always less than 1 kcal/mol. Comparison of the CHARMM27 results with the target data shows the agreement to be satisfactory. For the GC pair the empirical interaction enthalpy is in good agreement with experiment, both of which fall in the range of the *ab initio* results. For the AT and AU pairs the empirical values are significantly less favorable than the experimental values, while they fall into the range of the *ab initio* results. Recently, a corrected experimental value of 12.1 kcal/mol for the AT basepair was reported that takes into account the different conformations accessible to the AT basepair.[69] Comparison of the AT Watson-Crick and Hoogsteen pairs shows CHARMM27 to reproduce the *ab initio* trend with the AT Hoogsteen pair having an interaction enthalpy -0.38 kcal/mol more favorable than the Watson-Crick basepair. CHARMM27 also calculates the AU basepairs to have an interaction enthalpy more favorable than the AT pairs. This result is consistent with the experimental data, though the magnitude of the calculated differences is smaller. No high level *ab initio* calculations on the AU basepair are available. The necessity for such calculations is obvious from the different experimental interaction enthalpies of the AT and AU basepairs.

Hydrogen bond distances for the Watson-Crick and Hoogsteen basepairs are presented in Table 10, along with experimental data from crystal structures. The agreement overall is good with the largest difference of 0.07 Å occurring with the N4-O6 distance in GC. While comparison of the

distances with data from single crystals is limited, the quality of the agreement further indicate that the present parameters are adequately treating the hydrogen bonding interactions.

Base-base stacking interactions make a significant contribution to the stability of DNA and RNA oligonucleotides.[17,77] Such interactions must be properly treated for successful simulations of DNA and RNA duplexes. Proper treatment of the stacking interactions must include a balance with hydrogen bonded interactions between the bases. Several *ab initio* studies of base stacking interactions have been performed.[68,71,72] Of these, the work of Hobza et al.[71] is the most readily reproducable due to the use of the crystal structure of the CCAACGTTGG decamer[78] to define the relative orientations of the bases. Their *ab initio* data, at the MP2/6-31G* (BSSE corrected) level are presented in Table 11 for selected Watson-Crick pairs (HBONDED), STACKED pairs and INTERSTRAND (non WC) interacting pairs along with results from CHARMM27, including the electrostatic and LJ contributions. At the bottom of Table 11 is a summation of the different types of interacting pairs required to compare the relative energetics of the different interaction orientations. Those data include results from CHARMM22 performed as part of the present study and as reported by Hobza et al.[71] Comparison of the *ab initio* and CHARMM27 data show the trends for the different interacting pairs to be mimicked by the force field. Detailed analysis of the summations in Table 11 shows the CHARMM27 values for the HBONDED and STACKED pairs to reproduce the *ab initio* results, while the CHARMM27 INTERSTRAND interactions are too favorable. In CHARMM27 the electrostatic interactions dominate the HBONDED interactions, the LJ term dominates the STACKED interactions and there are varying contributions to the INTERSTRAND interactions. The quality of the CHARMM27 agreement with HBONDED and STACKED *ab initio* interactions suggests that the balance between the electrostatic and LJ contributions to base-base interactions in the force field is satisfactory.

The poorer agreement of the INTERSTRAND interactions is difficult to understand. For pairs where the CHARMM27 interaction energies are significantly larger then the *ab initio* (i.e. C2G20, G6G16 and C1G19) the electrostatic term dominates in two cases (C2G20 and C1G19) and the LJ dominates in the other. In other cases (i.e. A4G16 and A3T17) the electrostatics and LJ contributions are approximately equal. Problems with the *ab initio* data must also be considered. Limitations in the MP2/6-31G* level of theory with BSSE correction for the treatment of dispersion interactions have been noted.[68] Furthermore, the use of the counterpoise method to correct for BSSE may overestimate the correction, especially in the presence of a relatively limited basis set.[79] Such effects could lead to different "correction errors" for the different types of orientations. For example, the extent of orbital overlap for the HBONDED pairs is expected to be minimal as compared to the STACKED pairs. Accordingly, the influence of BSSE correction is expected to be least in the HBONDED pairs and the largest in the STACKED pairs. While these problems are beyond the scope of the present work they do indicate that additional studies are required to better quantitate stacking interactions between bases.

4.2 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. With the phosphodiester backbone, the sugar moiety and the glycosyl linkage a considerable part of the effort involved adjustment of the dihedral parameters to simultaneously reproduce QM potential energy surfaces and probability distributions of those dihedrals in experimental crystal structures. This is represented as loop IV in Figure 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 Supplemental Information). 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 N1 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.

4.2a Reproduction of the geometric and vibrational target data

Sugar and phosphodiester backbone. Optimization of the deoxyribose and ribose bond lengths and valence angle parameters was performed based on target data from a statistical analysis of high precision crystal structures of nucleosides and nucleotides.[80] Such data are ideal for the development of a force field for condensed phase simulations in that they contain condensed phase contributions averaged over a large number of compounds, thereby avoiding limitations in any single crystal structure associated with packing effects. In the study by Gelbin et al.[80] 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.

Table 12 compares the individual CHARMM27 deoxyribose bond lengths to their experimental crystal counterparts. 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, indicating that the empirical bond lengths are in reasonable agreement with their crystal references in both conformational ranges. The largest deviation between CHARMM27 and experiment is for the C5'-C4' bond, due to the equilibrium bond length being directly transferred from the aliphatic groups. Of note is the quality of

the agreement for the bonds involving the O4' atom. Because the C1'-O4' and C4'-O4' bond lengths differ significantly in the survey data, the atom type CN7B was assigned to atom C1' to distinguish between these bonds in the force field. This allowed for optimization of the two bonds individually, yielding the quality of agreement in Table 12.

Table 12 also shows the deoxyribose valence angles, comparing the individual CHARMM27 deoxyribose values to their crystal counterparts. The average absolute difference between the crystal and CHARMM27 valence angles is 1.1° in the south conformation and 1.2° in the north conformation. Therefore, the CHARMM27 angles are in reasonable agreement with their crystal references in both conformational ranges. For a majority (18 out of 28) of the valence angles the difference between the crystal average and their CHARMM equivalent falls within the experimental standard deviation. To aid in the fitting of the angles, atom C5' was assigned atom type CN8B to distinguish between angles C5'-C4'-C3' and C4'-C3'-C2'. The largest discrepancy occurs with the C4'-C3'-O3' angle in the south conformation, which has a large standard deviation in the crystal survey. This may be related to the different types of substituents at this position in nucleosides and nucleotides. This variability and the reasonable agreement for the C4'-C3'-O3' north angle and the C2'-C3'-O3' north and south angles precluded further optimization of the parameters associated with this angle.

Results on the geometry of the ribose sugar are presented in Table 13. Average absolute differences between CHARMM27 and the crystal data for the bond lengths were 0.010 Å for both the south and north conformations. With the angles the average absolute differences were 1.6 and 1.0 for the south and north conformations, respectively. The largest differences in the ribose angles occurred with terms related to the glycosyl linkage. These were in good agreement for deoxyribose (see Table 12) due to the parameter optimization being first performed on the deoxyribose sugar. To correct this problem without sacrificing the quality of the deoxyribose agreement would require the inclusion of new atoms types for the ribose species. Since the overall quality of the ribose internal geometries was deemed satisfactory, such an addition was not made.

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 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. While the size of these compounds disallows detailed analysis of the entire spectra, specific modes in the spectra can be identified and used for adjustment of the associated force constants. In particular, the torsional and deformation modes associated with the furanose ring and its exocyclic substituents were accessible. Based on these modes, angle force constants and some dihedral terms were optimized, although the final optimization of most of the dihedral parameters was based on the conformational energetics, as discussed below.

Presented in Tables 14, 15 and 16 are the CHARMM27 and QM determined vibrational spectra, including the potential energy distributions. The QM calculations were performed at the HF/6-31G* or HF/6-31+G* levels and the resultant frequencies scaled by 0.9.[81] For compound F (Table 14) there is good agreement for the torsional modes associated with the ε and ζ dihedrals (modes 1 and 2). Torsional modes associated with the sugar moiety (tRING) make contributions to vibrations at 92, 104, 228 and 238 cm^{-1} in the CHARMM27 spectra which agree well with the values of 87, 104 and 171 cm^{-1} from the QM calculation. With the sugar ring deformations (dRING), the CHARMM27 values of 526 and 637 cm^{-1} are in satisfactory agreement with the QM values of 565 and 632 cm^{-1}. Other modes of note are dC2C3O3, dC4C3O3 and dC3O3P, which are relevant to the backbone in nucleic acids. In CHARMM27, modes with contributions from dC2C3O3 or dC4C3O3 occur at 143, 304, 372 and 587 cm^{-1} which overlap the QM values of 190, 318, 422 cm^{-1}. Concerning the dC3O3P deformation, with CHARMM27 this mode occurs at 143 cm^{-1}, between the QM values of 104 and 171 cm^{-1} for modes 4 and 5.

Compound B includes the furanose moiety as well as the α, β and γ dihedrals. Table 15 shows the three lowest CHARMM27 modes (modes 1, 2 and 3), which are dominated by these terms, to be in good agreement with the QM data. Sugar furanose torsions (tRING) are again adequately represented by CHARMM27, where values of 97, 311 and 388 cm^{-1} (modes 4, 7 and 8) bracket the QM modes at 120, 148 and 206 cm^{-1}. In Compound B the furanose ring deformation frequencies are somewhat too low in the empirical model, with contributions at 584 and 594 cm^{-1} as compared to QM values of 677 and 813 cm^{-1}, but, the satisfactory agreement of these modes in Compounds F and the compound E analog (see next paragraph) precluded additional optimization of this term. Exocyclic terms of note include the dC5O5P mode at 218 cm^{-1} in CHARMM27 versus 189 cm^{-1} in the QM calculation, and the dC3C4C5 and dO4C4C5 deformation modes that make contributions at 311, 388, 446 and 497 cm^{-1} in CHARMM27 as compared to QM values of 148, 321, and 382 cm^{-1}. The final exocyclic mode of note in compound B is the scC4C5O5 scissor mode at 248 cm^{-1} in CHARMM27, in reasonable agreement with the QM value of 206 cm^{-1}.

Compound E with an imidazole base and a 5'methyl group was designed to test the influence of a base on the vibrational properties of the empirical model. Analysis of Table 16 shows CHARMM27 to be in good agreement with the QM data for the four lowest modes. This quality of agreement indicates CHARMM27 to satisfactorily treat torsional degrees of freedom associated with the glycosyl linkage and the sugar (tRING) and wagging of the C1' atom out of the plane of the imidazole moiety. Other modes associated with the glycosyl linkage are dO4C1ND1, dC2C1ND1 and the rocking of the imidazole ring (rC1ND1CG). Analysis of dO4C1ND1 and dC2C1ND1 in CHARMM27 shows only one contribution at 325 cm^{-1}, which falls in the range of the QM values at 214, 349 and 425 cm^{-1}. The CHARMM27 imidazole rocking mode occurs at 325 cm^{-1}, higher than the value of 214 cm^{-1} from the QM calculations. Finally, the sugar ring deformation modes from CHARMM27 (modes 11 and 13,

575 and 654 cm^{-1}, respectively), are in good agreement with the QM values at 566 and 651 cm^{-1}. Overall, the present vibrational analysis shows CHARMM27 to satisfactorily represent distortions associated with the furanose moiety and its exocyclic substituents including a 3'phosphate, a 5'phosphate and a base.

Bases. A recent survey of the geometries of the nucleic acid bases[82] motivated reoptimization of the associated parameters. Geometries of the bases are primarily dictated by the bond and angle equilibrium terms. 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.[83,84] 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.[85] 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 (MacKerell, Jr., A.D. unpublished). Furthermore, in several *ab initio* studies involving hydrogen bonded nucleic acid base dimers planar geometries were obtained.[69-71] 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 below).

Table 17 includes bond and valence angle RMSD data between the empirical optimized structures and target data for the methylated bases. Data are included for the bond lengths and valence angles of non-hydrogen atoms from the survey[82] and for angles involving hydrogens based on QM calculations. Individual values of the bond lengths and angles are presented in Tables 18 and 19. Data from CHARMM22 are included in Tables 17, 18 and 19 for comparison. For the non-hydrogen atoms the CHARMM27 geometries are in better agreement with the crystal survey data than is CHARMM22. This improvement is, in part, due to the CHARMM22 parameters being optimized to reproduce previous survey data.[86] Empirical angles involving hydrogen atoms (H-angles) are in good agreement with respect to the selected HF/6-31G* data. The largest discrepancy with the H-angles occurs with guanine. This difference is associated with the nonplanar *ab initio* versus planar empirical structures of the amino group and also leads to the larger differences in adenine and cytosine as compared to uracil and thymine. Overall, the CHARMM27 internal geometries of the bases are in satisfactory agreement with the target data.

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] Ilich et al.,[87] Colarusso et al.[88] and Aamouche et al.[89]) and the situation is complicated by the role of environment on the molecular vibrations. The majority of experimental data

for the bases is obtained in condensed phase environments while *ab initio* spectra are generally in the gas phase. While a detailed analysis of all available vibrational data is required to gain a clear understanding of the molecular vibrations of the nucleic acid bases, such an analysis is not within the scope of the present study. Accordingly, 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.[90] This approach may be expected to yield molecular vibrations that are representative of the experimental regime.

Vibrational data for adenine are presented in Table 20. The lowest frequency modes are in good agreement concerning both the frequencies and assignments. These modes are dominated by out-of-plane motions of the rings and their substituents. Such modes will make a significant contribution to distortion of the base occurring in MD simulations, making their correct representation critical for accurate results from MD studies. Discrepancies exist with the amino wagging and torsional modes. The amino torsional frequencies tend to be too high with empirical and *ab initio* frequencies occurring at 288 and 242 cm^{-1}, respectively, while the wagging modes are too low (i.e. empirical mode 5 occurs at 345 cm^{-1} while *ab initio* mode 6 occurs 492 cm^{-1}). Efforts to remedy these discrepancies via different combinations of dihedral and improper parameters were unsuccessful. In addition to the ring torsions at the lowest frequencies, empirical torsional modes for the 5-membered rings occur at 652 and 719 cm^{-1}, which compare well with the *ab initio* values of 652 and 694 cm^{-1}. Ring deformation modes are adequately reproduced with empirical values occurring at 468 and 531 cm^{-1} and *ab initio* modes at 512 and 518 cm^{-1}. To allow for close examination of the ring stretching modes the individual bonds were treated explicitly, except for empirical modes 22, 28, 33, and 34 where the sum of the 5- and 6-membered stretches are presented due to no individual ring stretches contributing 15% or more to the potential energy distribution. In general, the empirical and *ab initio* ring stretching frequencies are in similar ranges and the agreement of certain modes is good. For example, modes that include the C5-C6 stretch occur at 559 and 602 cm^{-1} for the empirical and *ab initio* data, respectively, C8-N9 stretching modes occur at 993 (empirical) and 1055 cm^{-1} (*ab initio*), and the C6-N1 stretching modes occur at 1469 (empirical) and 1489 cm^{-1} (*ab initio*).

Guanine is the largest of the bases and, accordingly, the most difficult to assign and fit to the *ab initio* target data. The empirical and *ab initio* data are presented in Table 21. As with adenine, the low frequencies are in satisfactory agreement; modes 2 and 3 are somewhat higher than the *ab initio* estimates and the out-of-plane wag of the amino group (gC2N) makes a significant contribution to mode 3 that is not seen in the *ab initio* data. This motion also contributes to empirical mode 15 at 673 cm^{-1}, which is in reasonable agreement with *ab initio* mode 17 at 736 cm^{-1}. A similar phenomenon is observed with cytosine (see below). Adjustment of the dihedral and improper parameters associated with these terms was not able to remove the gC2N contribution without significantly altering the higher frequency mode in the vicinity of 700 cm^{-1}. As with adenine, the guanine ring stretches, deformations and torsions all occur in the same frequency regions for the empirical and *ab initio* data. For example,

the C5-N7 stretch occurs at 1164 and 1154 cm^{-1} for the empirical and *ab initio* data, respectively. Good agreement is also observed for the amino group in-plane rocking and scissor modes. The empirical NH2 rock occurs at 980 cm^{-1} (mode 23) which is somewhat lower than the *ab initio* values of 1073 and 1127 cm^{-1} (modes 24 and 25), however, the empirical NH2 scissor modes (35 and 36) of 1638 and 1671 cm^{-1} are slightly higher than the *ab initio* values at 1612 and 1658 cm^{-1}. Overall, the empirical data is in satisfactory agreement with the *ab initio* data for the majority of the frequencies and assignments.

The simplified spectra of the pyrimidines, as compared to the purines, allowed for greater ease in interpretation of the spectra and optimization of the force constants. Results for cytosine, shown in Table 22, are in general quite good, though some of the problems present with the purines still occur. The out-of-plane wag of the amino group (gC4N) contributes to mode 2 in the empirical model but not in the *ab initio* data and the empirical amino group torsion modes are again overestimated (see mode 3), while the amino wags are underestimated (compare empirical mode 6 at 488 cm^{-1} with *ab initio* modes 6 and 8 at 519 and 532 cm^{-1}, respectively). In the central region of the spectra the agreement of the ring stretches and deformations are generally good (e.g. the C4-C5 stretch at 739 and 749 cm^{-1} for the empirical and *ab initio* data, respectively), although some significant differences are present (e.g. the C2-N3 stretch at 1572 cm^{-1} in the empirical model versus the value of 1249 cm^{-1} in the *ab initio* data).

The additional molecular simplification of uracil and thymine allowed for good agreement of both the frequencies and assignments, as shown in Tables 23 and 24. In both systems the lowest 10 empirical modes are in good agreement with the *ab initio* data. The only significant discrepancy is with mode 5 in uracil, assigned to a ring torsion in the empirical model and a ring deformation in the *ab initio* calculation. Also, modes 6 and 7 for uracil are reversed. In both uracil and thymine the empirical and *ab initio* in-plane (dC2O and dC4O) and out-of-plane wags (gC2O and gC4O) of the carbonyl groups are in good agreement. For example, the empirical C=O in-plane deformations for uracil occur at 375 cm^{-1} while the *ab initio* value is at 383 cm^{-1} and the empirical C=O wags at 712 and 783 cm^{-1} (modes 11 and 12) are in good agreement with *ab initio* values of 723 and 776 cm^{-1} (modes 10 and 12). In thymine the empirical C5 methyl in-plane (dC5-Me, mode 4, 284 cm^{-1}) and out-of-plane (gC5-Me, mode 5, 301 cm^{-1}) are in good agreement with the *ab initio* values of 267 and 289 cm^{-1}, respectively. Overall, the present force field satisfactorily reproduces HF/6-31G* scaled frequencies and assignments, indicating that deformations of the bases that occur during MD simulations will be accurately represented by CHARMM27.

4.2b 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 Figure 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 a previous study we systematically investigated four possible compounds to use as models for sugar puckering, leading to the selection of a nucleoside.[46] Similarly, the relevance of model compounds B, C, D, and E (Figure 2B) to nucleic acids has been justified by comparing the derived ε, γ, β and χ *ab initio* torsional energy profiles to the corresponding crystal distributions in nucleic acids and their components.[48]

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 in Section 2.2, 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.

In the remainder of this section results and discussion will be presented for the individual dihedral angles followed by the sugar puckering. Results include comparison of the empirical and *ab initio* torsional potential energy surfaces for both the south and north furanose puckers along with comparisons of dihedral distribution from the A and B DNA crystal simulations (Table 1) and their crystal survey counterparts. Inclusion of both the C3'endo and C2'endo furanose puckers at the model compound level was performed to represent the north and south conformational ranges populated by the sugars in nucleic acids. At certain stages during the optimization, solution MD simulations were performed on the EcoRI and CATTTGCATC sequences to insure that the B form properties were not biased by the use of one particular crystal structure. Results from these simulations are included in the accompanying manuscript.[35] For the final parameter set, the average RMSD for the B crystal over the 600 ps of sampling (see Methods) for all non-hydrogen atoms with respect to the experimental structure was 1.03±0.08 Å, with the error being the standard deviation. For the A crystal the corresponding values were 1.14±0.08 Å.

g torsion Results for the γ dihedral will be presented first as they represent a good example of the type of compromise made at the model compound level in order 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 Figure 4A are three empirical γ surfaces for model compound B along with the *ab initio* data. In Figure 4B probability distributions from MD simulations of the B form crystal using the same 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 Figure 4A to 4B. Parameter set 1 was optimized to reproduce the *ab initio* model compound data in the region of 0 to 90 (Figure 4A). 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 Figure 4). 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 Figure 4) which is in the greatest disagreement with the model compound target data, but the best agreement concerning the survey data. Since the goal of the parameter development if for a force field to be used in condensed phase simulations parameter set 3 was selected.

One point concerning the results in Figure 4 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.

Results for γ for the final parameter set are presented in Figure 5; this set differs from set 3 in Figure 4 due to changes in a number of other parameters in the force field. The *ab initio* data in Figure 5 are at the MP2/6-31+G* level. The CHARMM27 results are in satisfactory agreement with the QM data concerning the location of the minima and barriers. The empirical energy barrier at approximately 120 is lower than the QM value, due to the need to "soften" the surface in the vicinity of the g^+ minimum. Comparison of the MD and survey probability distributions shows the agreement to be good for both the A and B forms of DNA (Figures 5C and 5D, respectively).

a and z torsions Dihedral parameters associated with the phosphodiester linkage were optimized using DMP (compound A in Figure 2B). Potential energy surfaces for the O-P-O-C torsion for the final parameter set and MP2/6-31G* calculations in the presence and absence of a water molecule are shown in Figure 6A. As previously reported, the presence of a single water molecule alters the conformational energetics of DMP, as shown in Figure 6A, leading to a lowering of the energy

of the g,t conformer that may impact the equilibrium between the A and B forms of DNA.[91] This result has been reproduced at a variety of QM levels of theory, including the use of the IPCM[92] reaction field model (A.D. MacKerell, Jr., unpublished results) and other studies have shown the energetics of DMP to be altered by an aqueous environment.[93-97] While solvent contributions do affect the potential energy surface of DMP it is not clear whether the *ab initio* gas phase or solvated surface should be used as target data. Ideally, the empirical model would reproduce both the gas phase and solvated results, however, this could not be achieved with the present force field. Accordingly, a compromise was made yielding an empirical energy surface that produced energies in the g,t region of the surface (180 to 240) intermediate to the gas phase and monohydrate values. This compromise leads to a lowering of the energy barrier between the g,g and g,t conformations below the height seen in either the gas phase or monohydrate surfaces. Also of note when comparing the *ab initio* and empirical surfaces is the increase in energy as the O-P-O-C dihedral approaches 360 . In the present parameter set this increase in energy was lowered significantly as compared to the *ab initio* data to allow the α dihedral to sample regions above 300 that are significantly populated in B DNA. The resulting probability distributions for both A and B DNA are shown in Figures 6B and 6C, respectively. In both cases the crystal survey data are satisfactorily reproduced by the force field, including the shift in the maximum from approximately 285 to 300 upon going from A to B DNA. With B DNA the force field overpopulates the region of 300 to 330 , however, this is required to obtain the sampling of the region between 330 and 360 . As discussed above for γ, the need to lower the energy in the 300 to 360 at the model compound level to properly sample that region in the simulation may be due to limited sampling in the simulations.

DMP was also used for optimization of the ζ dihedral parameters. In this case the lowering of the energy in the region of 180 to 240 is even more relevant as it is significantly sampled in B form DNA. This sampling is related to the B_{II} DNA conformation.[9,98] As seen in Figure 7C the present parameters allow for sampling of ζ from approximately 120 to 240 , consistent with the NDB survey data. The locations of the maxima for both the A (Figure 7B) and B (Figure 7C) forms are reasonably well reproduced by the force field. In the case of A DNA the MD distribution is wider than that from the NDB survey. This reflects the lower empirical energy of DMP in the region of 215 to 255 relative to the QM data (see above). Such an approximation is consistent with the suggestion that environmental effects (e.g. changes in solvation or interactions with ions) may alter the intrinsic conformational energetics of the phosphodiester moiety in the backbone of DNA[91] that cannot be represented using the present form of the potential energy function.

b torsion Optimization of the parameters associated with β was performed using compound B (Figure 2B). Presented in Figure 8 are the potential energy profiles for β for the C3'endo (Figure 8A) and C2'endo (Figure 8B) puckers along with probability distributions for the A (Figure 8C) and B (Figure 8D) forms of DNA. The present force field reproduces the *ab initio* data well, although with

lower energies for $\beta < 180$. This departure from the *ab initio* energy surface was necessary to allow for β in the MD simulations to properly sample the regions occupied in the crystal surveys, as shown in Figures 8C and 8D for the A and B forms of DNA. For the A form, the MD simulation results nicely reproduces the NDB survey data. For the B form the force field in the MD simulation properly samples the region of β going down to 105 . There is a slight overpopulation by the MD simulation at 120 , however, this is not present in solution MD simulations,[35] indicating that crystal packing may cause this peak. Supporting this assertion is the presence of a β dihedral with a value of 112 in the experimental B crystal structure.

The β surfaces illustrates how the most populated regions of a dihedral can deviate significantly from the minimum in the potential energy surface at the model compound level. This is most evident with the B form DNA probability distribution. In the model compound the minimum occurs at 240 in the energy surface (Figure 8B) while the maximum in the B DNA probability distribution occurs at 177 (Figure 8D). Such differences illustrate the influence of the other contributions from the force field on the regions of conformational space being sampled, emphasizing the need for a proper balance between these different contributions.

e torsion Dihedrals associated with ε were primarily parametrized based on model compound C (Figure 2B), with additional optimization based on compound D to model the equilibrium between the B_I and B_{II} forms of DNA (see below). ε potential energy surfaces for model compound C are shown in Figure 9A and 9B for the C3'endo and C2'endo furanose puckers, respectively. Significant differences between the empirical and *ab initio* data are evident. Since the optimization procedure was initially performed on DNA, with emphasis on the B form, the C2'endo variant was the primary focus and, accordingly, is in the best agreement with the QM data (Figure 9B). The relative energies of the two minima in the force field are switched as compared to the *ab initio* data. This was done to better treat the B_I/B_{II} equilibrium, as discussed below. The empirical C3'endo surface is in significant disagreement with the *ab initio* data. While the minimum at approximately 180 is reasonably reproduced by the force field, the energy in the region from 195 to 300 in the QM surface is overestimated by the force field. Efforts to improve the quality of this surface while maintaining the C2'endo surface and satisfactorily reproducing the NDB survey data via the MD simulations were not successful. Analysis of the probability distributions shows both the A and B form MD simulation results in Figures 9C and 9D, respectively, to be in reasonable agreement with the survey data. The A form MD peak is shifted to slightly lower values than the NDB survey data (Figure 9C), possibly due to the shifted minimum at 180 in the C3'endo potential energy surface (Figure 9A). The overall shape of the MD data for the B form (Figure 9D) is in good agreement with the NDB data, including the sampling of the region above 240 associated with the B_{II} form of DNA.

B_I/B_{II} conformations. Final selection of the parameters associated with the ε and ζ dihedrals included consideration of the equilibrium between the B_I and B_{II} conformations.[9,98] Compound D,

which includes both ε and ζ without having a terminal hydrogen on either of those dihedrals, was designed for this purpose. Table 25 presents *ab initio* and empirical results for the location and relative energies of the minima associated with the B_I and B_{II} conformations in compound D. At both the HF and MP2/6-31+G* levels of theory the B_I conformer is favored over the B_{II} by values of 0.7 and 1.6 kcal/mole, respectively. An initial set of parameters yielded an energy difference, $\Delta E_{BII - BI}$, of approximately 1.0 kcal/mole. Application of these parameters in a MD simulation of the B crystal, whose experimental structure contains 4 out of 18 linkages in the B_{II} conformation, led to an underpopulation of the B_{II} region. This led to gradually decreasing the relative energy of the B_{II} conformation, ΔE_{BII-BI}, to the final value of 0.5 kcal/mole. For the final parameter set the extent of sampling of the B_{II} conformer is seen in the ζ (Figure 7C) and ε (Figure 9D) probability distributions. With ζ and ε both the extent and range of sampling of the B_{II} conformer are in satisfactory agreement with the NDB survey data. In the NDB survey the percentage of linkages in the B_{II} conformation is 13%, as compared to 7% in the B crystal simulation, indicating the force field to possibly underestimate the population of the B_{II} conformation. It should be noted that experimental studies in solution indicate the population of the B_{II} state to possibly be lower[9,99] than observed in crystal structures, although other studies suggest that high levels may be present in hydrated DNA films.[100] Further studies are required to better quantitate the population of the B_{II} conformer in solution.

c torsion Some of the largest differences between the various forms of DNA occur with χ. Therefore, proper optimization of the associated parameters is necessary to treat the different forms of DNA. The glycosyl linkage parameters were adjusted using model compound E (Figure 2B) with the four DNA bases, in combination with the condensed phase simulations. This compound was selected because it contains all the atoms involved in the χ dihedral, includes the influence of furanose puckering on χ and omits any additional functional groups on the sugar that may interact with the bases. Figures 10, 11, 12 and 13 present the QM results for the cytosine, thymine, adenine, and guanine analogs of model compound E for both the C3'endo (A) and C2'endo (B) furanose puckers, along with the empirical data. Also included in Figure 10 are the MD simulation and NDB survey probability distributions for χ for the A (Figure 10C) and B forms (Figure 10D) of DNA.

Optimization of the parameters associated with the glycosyl linkage emphasized reproduction of the potential energy surfaces in the vicinity of the global minima. Efforts were also made to reproduce the energy barrier at approximately 120 and the minimum well at 60 . Comparison of the empirical and QM χ energy surfaces for all four bases (Figures 10 through 13) shows CHARMM27 to satisfactorily reproduce both the global energy well and the barrier at 120 . For all the C2'endo empirical surfaces a well defined local minimum occurs at approximately 60 , although it is not as deep as in the *ab initio* calculations. Application of the parameters to the crystal MD simulations yielded χ distributions in good agreement with the NDB survey data for both the A (Figures 10C) and B forms (Figure 10D). In both cases there is some sampling of the alternate conformation of χ (i.e. sampling in

the region of 240 in the A form simulation), with the effect being more significant with the A form structure. This trend is not unexpected considering that in the A form experimental crystal structure two of the sixteen χ values are greater than 240 while in the experimental B structure one of the twenty χ values is less than 210 . In the EcoRI and CATTTGCATC decamer solution simulations the χ distributions were similar to those with the B crystal (Figure 10D).[35] Additional results on the quality of the parameters in representing the base dependent correlation between χ and sugar puckering are discussed below.

Sugar puckering and d . Differences in sugar puckering in the various forms of DNA and in RNA, in combination with χ, indicate these terms to be major determinants of oligonucleotide structure. The dynamics of sugar puckering are also of interest, with evidence indicating that they rapidly interconvert on NMR time scales, although direct access to details of the processes is somewhat limited.[6] Optimization of the sugar parameters is complicated by the substituents on the furanose ring. To overcome these complications two model compounds were ultimately used as the basis for the *ab initio* target data. Compound F (Figure 2B) was initially selected as it contains the 3'phosphate. Inclusion of the 3' phosphate prevents formation of an intramolecular hydrogen bond between O4' and a 3'-hydroxyl group, better mimicking the situation in nucleic acids. During the final stages of the parameter optimization model compound G with an imidazole base (referred to as G^I where the superscript indicates the identity of the base) was included as target data to account for contributions from the base. Following completion of the optimization, application of the developed parameters to model compound G with the standard DNA bases was performed as an additional test of parameters. Note that in the present study the dihedral δ was not explicitly parametrized due to its high correlation with sugar puckering, although the sampling of δ in MD simulations was investigated (see below).

Figures 14 A and B present the empirical and *ab initio* pseudorotation potential energy surfaces for Compounds F and G^I. Note that the surfaces were obtained by constraining a single furanose endocyclic dihedral, performing the optimization and extracting the pseudorotation angle from the optimized structure. This leads to an irregular distribution of points along the X-axis, however, this approach allows the amplitude to relax. Previous studies have determined the pseudorotation surface at fixed values of the amplitude,[101,102] an assumption that recent *ab initio* studies have shown to be invalid[46,47]. This approach leads to additional points in the region of 150-360 in Figures 14A and 14B due to local minima associated with different amplitudes. For example, the point at 270 corresponds to an amplitude of 4.2 . During the MD simulations these low amplitude structures were not sampled, and therefore, were assumed not to interfere with the parameter optimization process.

Compound F was instrumental in fine-tuning the locations and shapes of the north and south deoxyribose energy minima wells, as well as their relative populations, via modification of both the furanose endocyclic and exocyclic torsional terms. The 5'-carbon and 3'-oxygen in Compound F include components of δ that were exploited to adjust the relative energies of the north and south energy

minima. For compound F (Figure 14A) there are large differences between the *ab initio* and empirical pseudorotation energy surfaces. With respect to the HF/6-31+G* *ab initio* data, the location of the two empirical minima are similar and the barrier at 90 (O4'endo) is lower than the barrier at 270 (O4'exo). The energetic ordering of the north and south minima is reversed between the HF/6-31+G* and empirical data. There is better agreement between the empirical and MP2/6-31+G* *ab initio* data; the relative ordering of the north and south minima agree, although the empirical O4'endo barrier is still significantly too high. The departure between the CHARMM27 and QM data for compound F is due to the use of the survey data on the pseudorotation surface as the primary target data. Initial parameter sets in better agreement with the *ab initio* data for compound F (not shown) lead to poor pseudorotation angle probability distributions in the MD simulations. Comparison of pseudorotation distributions from MD simulations of A and B form DNA with the survey data were used to systematically alter the compound F pseudorotation energy surface, ultimately yielding the surface presented in Figure 14A.

To investigate the role of the base on the sugar pseudorotation properties the potential energy surface of model compound G^I was calculated using both the *ab initio* and empirical models. Figure 14B shows that the agreement between the *ab initio* and empirical surfaces with compound G^I is improved over compound F, though the shape of the surface between 0 and 165 still differs significantly from the *ab initio* data. Probability distributions for both the A and B forms of DNA from MD simulations (Figures 14C and 14D, respectively) are in good agreement with the NDB survey data. With both the A and B forms there is some sampling of the alternate sugar pucker (e.g. sampling in the vicinity of 165 in the A form simulation), consistent with the alternate sampling of χ conformations (see above) and the experimental crystal structures. In the A crystal structure there are two sugars with south puckers while one of the sugars in the B crystal structure is in the north conformation. Also, in the B crystal structure there are three sugars with pseudorotation values less than 120 , consistent with the sampling of sugar puckering between 60 and 120 in the MD simulation (Figure 14D). In the EcoRI and CATTTGCATC solution simulations the 60 to 120 region was also well sampled, consistent with the NDB survey data, and the minimum in the probability distribution seen at approximately 105 in the B crystal simulation is not present.[35]

Of note are similarities between the empirical potential energy surfaces for compounds F and G^I (Figures 14A and 14B, respectively) and MD probability distributions for both the A and B crystals (Figures 14C and 14D, respectively). These similarities suggest that the sugar pseudorotation potential energy surface may dominate the distributions obtained from the MD simulations. The narrow A distribution (Figure 14C) reflects the shape of the north potential energy well for both compound F and G^I. Similarly, the broader distribution of sugar pseudorotation angles in the B crystal (Figure 14D) correlates well with the overall shape of the empirical potential energy surfaces in the south region (ca. 165). The gradual increase in both empirical energy surfaces from the south minimum to the maximum

at approximately 45 correlates well with the sampling of the 60 to 120 region in the B crystal simulation. Results with Z DNA (see below) indicate that the sugar pseudorotation energy surfaces may require additional refinement. Further optimization of the sugar parameters must be performed in conjunction with additional *ab initio* data on furanose containing compounds that include the base as well as the phosphate moieties at both the 5' and 3' positions. However, as stated above, parameters yielding better agreement with the model compound target data yielded poor agreement with experiment in the macromolecular simulations. This suggests that limitations in the form of the potential energy function may contribute significantly to the problems with the sugar parametrization.

To obtain the quality of agreement between the empirical and crystal puckering distributions the flexibility of the dihedral fourier series included in equation 1 was exploited. This included the use of 4-6 fold terms for the furanose dihedrals. These high frequency terms were incorporated due to the relatively small change in any individual ring dihedral over the pseudorotation surface (e.g. all endocyclic dihedrals sample with the range of ±50). It should also be emphasized that changes in atom types associated with different substitution of the furanose were exploited to better reproduce the energetics of the different model compounds. This included the use of different C1' atom types for the pyrimidines versus the purines (see Appendix of the Supplemental Material). This approach is consistent with the CHARMM22 and CHARMM27 force fields being optimized to maximize the reproduction of selected target data at the expense of tranferability.

As stated above, δ was not parametrized explicitly in the present study. Comparison of the MD and NDB survey probability distributions therefore offers an additional means to monitor the behavior of the present force field. Shown in Figure 15A and 15B are the MD and NDB probability distributions for both the A and B crystals, respectively. For the A crystal, the distribution from the MD simulation is in excellent agreement with the NDB survey result between 60 and 105 , with a small peak in the region of the south sugars, consistent with the sugar pseudorotation distribution (Figure 14C). With the B crystal, the probability distribution from the MD simulation is much narrower than that from the survey. When considering the quality of the agreement for the sugar puckering (Figure 14D) the narrower δ distribution is difficult to understand. Possibly, additional flexibility in the furanose ring and in the 5' and 3' covalent connectivity may be present 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 can not be excluded. Further studies are required to understand this difference.

Influence of base type on sugar puckering and c 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. Therefore, as an additional test of the sugar and glycosyl linkage parameters their correlation and their sensitivity to base substitution were determined in model compound G. Presented in Table 26 are the locations of the north and south

minima, the energy difference between those minima as well as the east barrier height from the *ab initio* MP2/6-31G*, CHARMM27 and CHARMM22 calculations. The CHARMM27 pseudorotation angles of the north minima are all smaller than the *ab initio* values by approximately 10 , however, they are in better agreement than CHARMM22. The locations of the south minima are in good agreement with the *ab initio* values, with smaller values in the pyrimidines versus the purines. In CHARMM27, the relative energies of the north and south conformers, ΔE_{N-S}, are compatible with the *ab initio* values. CHARMM27 nicely mimics the increase in ΔE_{N-S} upon going from adenine to guanine and from cytosine to thymine. With the pyrimidines, ΔE_{N-S} is less than in the *ab initio* data, while cytosine energetically favoring the north minimum is consistent with the *ab initio* result. In CHARMM22 the north energy minima for adenine, guanine and thymine are of lower energy, is disagreement with the *ab initio* data; the north conformation being of lower energy than the south may be assumed to contribute to CHARMM22 favoring the A form of DNA in MD simulations (see Introduction). Note that the parametrization of the sugar in CHARMM22 was based, in part, on a 3'-hydroxy-5-methyl-furanose where the base was replaced by an amino group (MacKerell, A.D., Jr. unpublished). Recent work has shown this compound to be a poor model for the sugar in nucleic acids.[46] Barrier heights, which are similar for all four bases, are consistently smaller in CHARMM27, with CHARMM22 being in better agreement with the *ab initio* data. The lower CHARMM27 barrier heights are due to the location of the maxima in the empirical model being shifted from 90 to approximately 45 , as in Figures 14B.

The four base analogs of compound G were also analyzed with respect to their glycosyl torsions and amplitudes, with the results presented in Table 27. For the glycosyl torsion, the overall agreement between CHARMM27 and the *ab initio* data for the four nucleosides is good. The empirical values of both the north and south minima are in satisfactory agreement, with larger differences occurring at the east barrier. The change in χ upon going from the north to south minima is consistent with the well known correlation between sugar pucker and χ.[8,9,80] This correlation is not reproduced by CHARMM22. 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 has been suggested to contribute to the equilibrium between the A, B and Z forms of DNA.[47] Concerning the amplitudes, the decrease between the north and south minima in the *ab initio* results is present in the empirical model although it is overestimated. The *ab initio* amplitudes are significantly lower at the east barrier, a property that is not reproduced by CHARMM27. Interestingly, CHARMM22 shows a greater decrease in the amplitude at the east barrier, in better agreement with *ab initio* data, which may be related to the better agreement of CHARMM22 with respect to the energy of the east energy barrier (Table 26). Overall, for the energetics and structure of the north and south minima CHARMM27 represents a significant improvement over CHARMM22, although the latter yields better agreement with *ab initio* for the east energy barrier.

In CHARMM27 the initial optimization of the glycosyl linkage dihedral parameters was based on model compound E (Figures 10 to 13). Facilitating the reproduction of the properties of compound E was the use of a different atom type for the C2 atom in the cytosine versus uracil and thymine. Those parameters were directly transferred to compound G, yielding the results in Tables 26 and 27. Thus, while the use of specific parameters for the different pyrimidines aided in the quality of the present force field, the ability of compound G to reproduce subtle differences in conformation and energetics is related to the proper balance between the internal and nonbond parameters in CHARMM27. It should be noted that other work has indicated that the difference in energetics of deoxycytidine and deoxythymidine can also be reproduced via the inclusion of electronic polarizability.[30]

4.3 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 Figure 1. Results on the optimization of the parameters associated with the geometry of the ribose sugar are presented in Section 4.2a (Table 13). For the final parameter set, MD simulations on the UAAGGAGGUGUA dodecamer (Table 1) yielded an RMSD for all non-hydrogen atoms in basepairs 2 through 11 of 1.9±0.6 and 5.9±0.4 Å, with respect to canonical A and B structures, showing the RNA structure to remain close to the canonical A form.

Optimization of the dihedrals for the ribose moiety first focused on the C3'-C2'-O2'-H dihedral. Figure 16 presents the *ab initio* and empirical potential energy surfaces for this dihedral in model compound C^{2OH} (Figure 2B, compound C with a hydroxyl at the 2' position). Comparison of the two surfaces shows them to be in agreement concerning both the location of the minima and the overall shape of the surfaces. Careful parametrization of the C3'-C2'-O2'-H torsion may help clarify the orientation of the 2'hydroxyl group in solution. The orientation of the 2' hydroxyl group may influence RNA properties, but the favored orientation is still a matter of debate.[46,103,104]

Optimization of the remaining ribose parameters concentrated on balancing the agreement of the empirical and *ab initio* data for model compounds F and G^{I} with a 2'hydroxyl, denoted F^{2OH} and $G^{I,2OH}$, respectively. Shown in Figure 17 are the empirical and QM potential energy surfaces as a function of pseudorotation angle for model compounds F^{2OH} and $G^{I, 2OH}$ and comparison of the MD and survey probability distributions for the ribose pseudorotation angle. The empirical and *ab initio* potential energy surfaces differ significantly, although in both cases the south energy is higher than the

north, consistent with the high population of north sugar puckering in RNA (Figure 17C). With model compound F^{2OH} (Figure 17A) the empirical east and south energies are higher than the *ab initio* values. For model compound $G^{I,\ 2OH}$ (Figure 17B) the empirical south energy is in good agreement with the *ab initio* value, although the east energy is lower. Thus, the force field overestimates the barrier height in compound F^{2OH} and underestimates it in $G^{I,\ 2OH}$. This contrasts results for deoxyribose where the force field significantly overestimates the barrier for model compound F while the barrier height is similar for model compound G (see Figures 14A and 14B, respectively). The larger differences between the two models compounds for the ribose sugar are associated with the presence of the 2'hydroxyl, which interacts differently with the phosphate in compound F^{2OH} and the hydroxyl in compound $G^{I,\ 2OH}$. The higher energy points in the empirical data in Figure 17B are due to local minima associated with low amplitude sugar puckering, as discussed above for deoxyribose.

Presented in Figure 17C is the pseudorotation angle probability distribution from the MD simulation of the RNA dodecamer and survey results from all RNA duplexes and transfer RNA crystal structures in the NDB. As may be seen the overlap of the MD and crystal probability distributions for the pseudorotation angles is good in the north region, but the MD distribution is not as narrow as observed in the experimental crystal structures. There is a small amount of sampling of the south conformation by the present force field, consistent with some occupation of that conformation in the RNA crystal structures. Thus, CHARMM27 reasonably treats the sugar puckering of RNA based on both the model compound and crystal survey target data. Differences in the shapes of the *ab initio* data in Figures 17A and 17B, however, make it clear that additional *ab initio* data on alternate model compounds and possibly the use of alternate forms of the potential energy function (see above) are required to better link the model compound and macromolecular sugar puckering properties in RNA.

Additional analysis of the quality of the new force field for simulations on RNA was done by comparing the various dihedral probability distributions from the 2 ns MD solution simulation of the RNA dodecamer with survey data. Results, presented in Figure 18, show that in all cases the agreement between the MD and NDB probability distributions is good. The largest differences occur with γ and δ. With γ the overall range of sampling is similar, but, the maximum of the MD distribution is shifted towards larger values as compared to the survey. The differences with δ are consistent with the differences between the MD and NDB results for the ribose pseudorotation angle (Figure 17C). The only other significant difference is the NDB ε distribution extending to larger values than seen in the MD simulation (Figure 18E). This difference may be due to limited sampling in the simulation as well as to contributions from various non-helical regions in transfer RNA that were included in the NDB survey. Note that the latter contribution may effect the agreement for the other dihedrals presented in Figures 17 and 18. Thus, based on both RMSD and dihedral and sugar pseudorotation angle distributions the present force field adequately models duplex RNA.

4.4 Z DNA crystal simulation

In the present work the A and B forms of DNA and 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[105] in its crystal environment was performed. The crystal contains a single duplex with 106 water molecules, two sodium and four magnesium ions, as previously described.[66] Results, presented in Figure 19, were obtained over the final 800 ps of the 1 ns simulation; the average RMS difference for all non-hydrogen atoms with respect to the crystal structure was 0.83±0.09 Å.

Analysis of the simulated probability distributions for the backbone dihedrals (Figures 19A to 19F), χ, (Figure 19G) and the pseudorotation angle (Figure 19H) are generally in satisfactory agreement with survey results, but, deviations do exist. The largest discrepancies occur with χ and the sugar pseudorotation angle. With χ (Figure 19G) a shoulder ranging from 90 to 120 is present in the MD results that is not observed in the survey. The MD pseudorotation angle distribution (Figure 19H) 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 . Discrepancies in these distributions as a function of base show the differences in χ and the portion of the pseudorotation surface below 105 to be associated with the guanines. Differences in χ are due to the terminal guanines while internal guanines cause the peak in the pseudorotation profile in the vicinity of 80 . In Z DNA internal guanines typically assume the syn conformation about the glycosyl linkage (i.e. χ approximately 60) along with north sugar conformations. The shift of the peak centered at 150 in the pseudorotation angle crystal distribution to lower values in the MD simulation is due to cytosines; cytosines also lead to the sampling of lower χ values in the MD simulation for the peak centered around 210 (Figure 19G).

Differences between the MD and survey pseudorotation distributions may be related to the pseudorotation energy surfaces associated with model compounds F and G[I]. As discussed above, the pseudorotation energy surfaces of both compounds have a maximum at 45 (see Figure 14A and 14B, respectively). It is suggested that the location of these maxima contribute to the shift in the pseudorotation angle distribution peak centered at 30 in the NDB survey (Figure 19H) to lower values in the MD simulation and also to the small peak at 80 . With χ, the shoulder in the MD distribution from 90 to 120 in Figure 19G is dominated by the terminal guanines. This may be due to the compound E χ energy surface with a guanine base (Figure 13); the force field poorly reproduces the minima in the region of 60 for both furanose puckers that may lead to increased sampling of the 90 to 120 range in the simulation. One possible contribution to limitations in the treatment of Z DNA are the presence of interactions between the C3'-H and the N3 atom is syn purines that have been observed in *ab initio* calculations (Foloppe, N. and MacKerell, Jr., A.D. Manuscript in preparation),

which may be poorly modeled by the present force field. Studies investigating this possibility are in progress. Overall, the present force field yields a reasonable representation of ZDNA, although of a lesser quality than with the A and B forms of DNA.

5. 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 amongst 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 experiment.[106] Thus, the CHARMM27 interaction parameters appear to work well in a variety of environments, including changes in water activtiy 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.[80] 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 (Tables 26 and 27). 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 allows 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. While 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 equation 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.[39,40,107] 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 equation 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 equation 1 has also involved the interaction portion of the force field, with the most common being the inclusion of electronic polarizability.[108] While improvements associated with electronic polarizability have been made,[109] cases also exist where enhanced parameter optimization has overcome limitations previously ascribed to the omission of explicit electronic polarizability.[110,111] 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 Figure 2B[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.

While 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,[112,113] suggesting that higher level QM data may be required. Alternatively, as discussed with DMP,[91] 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 Figure 14A and B 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" as 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. While 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. Since current technology disallows rigorously testing 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,[114] and use of multiple simulations[115] 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. While 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[116], 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. While 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.

This work has been financially supported by NIH grant GM51501. We also 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.

Appendix.

Included in the Appendix to the Supporting Information is 1) a Table of the model compounds used in the present study and the corresponding residue and patch name 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.ab.umd.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 Crystallog. Sect. D* 1996, *52*, 57-64.
4. Hahn, M.; Heinemann, U. *Acta Cryst*. 1993, *D49*, 468-477.
5. Ulyanov, N. B.; Schmitz, U.; Kumar, A.; James, T. L. *Biophys J* 1995, *68*, 13-24.
6. Schmitz, U.; James, T. L. *Methods in Enzymology* 1995, *261*, 3-44.
7. Rife, J. P.; Stallings, S. C.; Correll, C. C.; Dallas, A.; Steitz, T. A.; Moore, P. B. *Biophys J* 1999, *76*, 65-75.
8. Dickerson, R. E.; Drew, H. R.; Conner, B. N.; Wing, R. M.; Fratini, A. V.; Kopka, M. L. *Science* 1982, *216*, 475-485.
9. Hartmann, B.; Lavery, R. *Quarterly Reviews of Biophysics* 1996, *29*, 309-368.
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*, 751-759.
11. Jain, S.; Sundaralingam, M. *J. Biol. Chem.* 1989, *264*, 12780-12784.
12. Shakked, Z.; Guerstein-Guzikevitch, G.; Eisenstein, M.; Frolow, F.; Rabinovitch, D. *Nature* 1989, *342*, 456-459.
13. Lipanov, A.; Kopka, M. L.; Kaczor-Grzeskowiak, M.; Quintana, J.; Dickerson, R. E. *Biochemistry* 1993, *32*, 1373-1389.
14. Dickerson, R. E.; Goodsell, D. S.; Neidle, S. *Proc. Natl. Acad. Sci. USA* 1994, *91*, 3579-3583.
15. Metlzer, W. J.; Wang, C.; Kitchen, D. B.; Levy, R. M.; Pardi, A. *J. Mol. Biol.* 1990, *214*, 711-736.
16. Allain, F.; Varini, G. *J. Mol. Biol.* 1997, *267*, 338-351.
17. Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984.
18. Norberg, J.; Nilsson, L. *J. Phys. Chem.* 1996, *100*, 2550-2554.
19. Auffinger, P.; Westhof, E. *Biophys. J.* 1996, *71*, 940-954.
20. Yang, L.; Pettitt, B. M. *J. Phys. Chem.* 1996, *100*, 2550-2566.
21. Cheatham, T. E., III; Kollman, P. A. *Structure* 1997, *5*, 1297-1311.
22. Young, M. A.; Ravishanker, G.; Beveridge, D. L. *Biophys. J.* 1997, *73*, 2313-2336.
23. Flatters, D.; Young, M.; Beveridge, D. L.; Lavery, R. *J. Biomol. Struct. Dyn.* 1997, *14*, 757-765.
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. Amer. Chem. Soc.* 1995, *117*, 5179-5197.
25. MacKerell, A. D., Jr.; Wiórkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* 1995, *117*, 11946-11975.
26. Feig, M.; Pettitt, B. M. *Biophys. J.* 1998, *75*, 134-149.
27. MacKerell, A. D., Jr. Observations on the A versus B Equilibrium in Molecular Dynamics Simulations of Duplex DNA and RNA. In *Molecular Modeling of Nucleic Acids*; Leontis, N.

B., SantaLucia, J., Jr., Eds.; American Chemical Society: Washington, DC, 1998; Vol. 682; pp 304-311.
28. MacKerell, A. D., Jr. *J. Phys. Chem. B* 1997, *101*, 646-650.
29. Pastor, N.; Pardo, L.; Weinstein, H. *Biophys. J.* 1997, *73*, 640-652.
30. Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* 1999, *16*, 845-861.
31. Shui, X.; McFail-Isom, L.; Hu, G. G.; Williams, L. D. *Biochemistry* 1998, *37*, 8341-55.
32. Langley, D. R. *J. Biomol. Struct. Dyn.* 1998, *16*, 487-509.
33. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* 1983, *4*, 187-217.
34. MacKerell, A. D., Jr.; Brooks, B.; Brooks, C. L., III; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Paramerization with an Overview of the Program. In *Encyclopedia of Computational Chemistry*; P.v.R. Schleyer, N. L. A., T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer III, P.R. Schreiner, Ed.; John Wiley & Sons: Chichester, 1998; Vol. 1; pp 271-277.
35. MacKerell, A. D., Jr.; Banavali, N. *Submitted for publication* 1999.
36. MacKerell, A. D., Jr. Atomistic Models and Force Fields. In *Computational Biochemistry and Biophysics*; Watanabe, M., A.D. MacKerell, J., Roux, B., Becker, O. M., Eds.; Marcel Dekker, Inc.: New York, 1999; Vol. In Press.
37. Yin, D.; MacKerell, A. D., Jr. *J. Comp. Chem.* 1998, *19*, 334-348.
38. Jorgensen, W. L.; Tirado-Rives, J. *J. Amer.Chem.Soc.* 1988, *110*, 1657-1666.
39. Halgren, T. A. *J. Comp. Chem.* 1996, *17*, 490-519.
40. Lii, J.-L.; Allinger, N. L. *J. Comp. Chem.* 1991, *12*, 186-199.
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*, 3586-3616.
42. MacKerell, A. D., Jr.; Karplus, M. *J. Phys. Chem.* 1991, *95*, 10559-10560.
43. Reiher, W. E., III. Theoretical Studies of Hydrogen Bonding. Ph.D., Harvard University, 1985.
44. Jorgensen, W. L. *J. Phys. Chem.* 1986, *90*, 1276-1284.
45. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* 1983, *79*, 926-935.
46. Foloppe, N.; MacKerell, A. D., Jr. *J. Phys. Chem. B* 1998, *102*, 6669-6678.
47. Foloppe, N.; MacKerell, A. D., Jr. *Biophys. J.* 1999, *76*, 3206-3218.
48. Foloppe, N.; MacKerell, A. D., Jr. *Manuscript in preparation* 1999.
49. Drew, H. R.; Dickerson, R. E. *J. Mol. Biol.* 1981, *151*, 535-556.
50. Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. S. *Proc. Natl. Acad. Sci. USA* 1981, *78*, 2179-2183.
51. Weisz, K.; Shafer, R. H.; Egan, W.; James, T. L. *Biochemistry* 1992, *31*, 7477-7487.
52. Weisz, K.; Shafer, R.; Egan, W.; James, T. L. *Biochemistry* 1994, *33*, 354-366.
53. Wahl, M. C.; Rao, S. T.; Sundaralingam, M. *Biophys. J.* 1996, *70*, 2857-2866.
54. Arnott, S.; Hukins, D. W. L. *J. Mol. Biol.* 1973, *81*, 93-105.
55. Altona, C.; Sundaralingam, M. *J. Am. Chem. Soc.* 1972, *94*, 8205-8212.
56. Beglov, D.; Roux, B. *J. Chem. Phys.* 1994, *100*, 9050-9063.

57. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. Gaussian 94; C.3 ed.; Gaussian, Inc.: Pittsburgh, PA, 1996.
58. Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, R. W. *J. Chem. Phys.* 1995, *103*, 4613-4621.
59. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* 1977, *23*, 327-341.
60. Field, M. J.; Karplus, M. CRYSTAL: Program for Crystal Calculations in CHARMM; Harvard University: Cambridge, MA, 1992.
61. Ewald, P. P. *Ann. Phys.* 1921, *64*, 253-287.
62. Steinbach, P. J.; Brooks, B. R. *J. Comp. Chem.* 1994, *15*, 667-683.
63. Darden, T. A.; York, D.; Pedersen, L. G. *J. Chem. Phys.* 1993, *98*, 10089-10092.
64. Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1989.
65. Brünger, A. T.; Karplus, M. *Proteins* 1988, *4*, 148-156.
66. Jung, S.-H. Simulation of DNA and its Interactions with Ligands. Ph.D., Harvard University, 1989.
67. Yin, D.; MacKerell, A. D., Jr. *J. Phys. Chem.* 1996, *100*, 2588-2596.
68. Alhambra, C.; Luque, F. J.; Gago, F.; Orozco, M. *J. Phys. Chem. B* 1997, *101*, 3846-3853.
69. Brameld, K.; Dasgupta, S.; Goddard, W. A., III. *J. Phys. Chem. B* 1997, *101*, 4851-4859.
70. Gould, I. R.; Kollman, P. A. *J. Am. Chem. Soc.* 1994, *116*, 2493-2499.
71. Hobza, P.; Kabelac, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. *J. Comp. Chem.* 1997, *18*, 1136-1150.
72. Sponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem.* 1996, *100*, 5590-5596.
73. Jorgensen, W. L.; Severance, D. L. *J. Amer. Chem. Soc.* 1990, *112*, 4768-4774.
74. Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aid. Mol. Des.* 1992, *6*, 629-666.
75. Cossi, M.; Barone, V.; Camni, R.; Tomasi, J. *Chem. Phys. Let.* 1996, *255*, 327-335.
76. Yanson, I. K.; Teplitsky, A. B.; Sukhodub, L. F. *Biopolymers* 1979, *18*, 1149-1170.
77. Hunter, C. A. *J. Mol. Biol.* 1993, *230*, 1025-1054.
78. Privé, G. G.; Yanagi, K.; Dickerson, R. E. *J. Mol. Biol.* 1991, *217*, 177-199.
79. van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. *Chem. Rev.* 1994, *94*, 1873-1885.
80. Gelbin, A.; Schneider, B.; Clowney, L.; Hsieh, S.-H.; Olsen, W. K.; Berman, H. M. *J. Amer. Chem. Soc.* 1996, *118*, 519-529.
81. Florián, J.; Johnson, B. G. *J. Phys. Chem.* 1994, *98*, 3681-3687.
82. Clowney, L.; Jain, S. C.; Srinivasan, A. R.; Westbrook, J.; Olson, W. K.; Berman, H. M. *J. Am. Chem. Soc.* 1996, *118*, 509-518.
83. Leszczynski, J. *J. Phys. Chem. A* 1998, *102*, 2357-2362.
84. Sponer, J.; Hobza, P. *J. Phys. Chem.* 1994, *98*, 3161-3164.
85. Guo, H.; Karplus, M. *J. Phys. Chem.* 1994, *98*, 7104-7105.
86. Taylor, R.; Kennard, O. *Journal of Molecule Structure* 1982, *78*, 1-28.
87. Ilich, P.; Hemann, C. F.; Hille, R. *J. Phys. Chem. B* 1997, *101*, 10923-10938.

88. Colarusso, P.; Zhang, K.; Guo, B.; Bernath, P. F. *Chem. Phys. Let.* 1997, *269*, 39-48.
89. 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*, 10063-10074.
90. Scott, A. P.; Radom, L. *J. Phys. Chem.* 1996, *100*, 16502-16513.
91. MacKerell, A. D., Jr. *J. Chim. Phys.* 1997, *94*, 1436-1447.
92. Barone, V.; Cossi, M.; Tomasi, J. *J. Comp. Chem.* 1998, *19*, 404-417.
93. Jayaram, B.; Mezei, M.; Beveridge, D. L. *J. Comp. Chem.* 1987, *8*, 917-942.
94. Jayaram, B.; Mezei, M.; Beveridge, D. *J. Am. Chem. Soc.* 1988, *110*, 1691-1694.
95. Jayaram, B.; Ravishanker, G.; Beveridge, D. L. *J. Phys. Chem.* 1988, *92*, 1032-1034.
96. Alagona, G.; Ghio, C.; Kollman, P. A. *J. Am. Chem. Soc.* 1985, *107*, 2229-2239.
97. Florián, J.; Strajbl, M.; Warshel, A. *J. Amer. Chem. Soc.* 1998, *120*, 7959-7966.
98. Dickerson, R. E. *Methods in Enzymology* 1992, *211*, 67-111.
99. Szyperski, T.; Ono, A.; Fernández, C.; Iwai, H.; Tate, S.-i.; Wüthrich, K.; Kainosho, M. *J. Amer. Chem. Soc.* 1997, *119*, 9901-9902.
100. Pichler, A.; Rudisser, S.; Mitterbock, M.; Huber, C. G.; Winger, R. H.; Liedl, K. R.; Hallbrucker, A.; Mayer, E. *Biophys J* 1999, *77*, 398-409.
101. Nilsson, L.; Karplus, M. *J. Comp. Chem.* 1986, *7*, 591-616.
102. Brameld, K. A.; Goddard, W. A., III. *J. Amer. Chem. Soc.* 1999, *121*, 985-993.
103. Auffinger, P.; Westhof, E. *J Mol Biol* 1997, *274*, 54-63.
104. Gyi, J. I.; Lane, A. N.; Conn, G. L.; Brown, T. *Nucleic Acids Res* 1998, *26*, 3104-10.
105. Gessner, R. V.; Quigley, G. J.; Wang, A. W.-J.; van der Marel, G. A.; van Boom, J. H.; Rich, A. *Biochemistry* 1985, *24*, 237-240.
106. Luo, R. D. L.; Head, M. S.; Given, J. A.; Gilson, M. K. *Biophys. Chem.* 1999, *78*, 183-193.
107. Hagler, A. T.; Maple, J. R.; Thacher, T. S.; Fitzgerald, G. B.; Dinur, U. Potential energy functions for organic and biomolecular systems. In *Computer Simulation of Biomolecular Systems*; van Gunsteren, W. F., Weiner, P. K., Eds.; ESCOM: Leiden, 1989; pp 149-167.
108. Gresh, N. *J. Chim. Phys.* 1997, *94*, 1365-1416.
109. Meng, E. C.; Cieplak, P.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* 1994, *116*, 12061-12062.
110. MacKerell, A. D., Jr. *J. Phys. Chem.* 1995, *99*, 1846-1855.
111. Rizzo, R. C.; Jorgensen, W. L. *J. Amer. Chem. Soc.* 1999, *121*, 4827-4836.
112. Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Amer. Chem. Soc.* 1997, *119*, 5908-5920.
113. Halgren, T. A. *J. Comp. Chem.* 1999, *20*, 730-748.
114. Wu, X.; Wang, S. *J. Phys. Chem. B* 1998, *102*, 7238-7250.
115. Caves, L. S. D.; Evanseck, J. D.; Karplus, M. *Protein Science* 1998, *7*, 649-666.
116. Schlenkrich, M.; Brickmann, J.; MacKerell, A. D., Jr.; Karplus, M. Empirical Potential Energy Function for Phospholipids: Criteria for Parameter Optimization and Applications. In *Biological Membranes: A Molecular Perspective from Computation and Experiment*; Merz, K. M., Roux, B., Eds.; Birkhaüser: Boston, 1996; pp 31-81.
117. Grzeskowiak, K.; Yanagi, K.; Prive, G. G.; Dickerson, R. E. *J. Biol. Chem.* 1991, *266*, 8861-8883.
118. Langlois D'Estaintot, B.; Dautant, A.; Courseille, C.; Precigoux, G. *Eur. J. Biochem.* 1993, *213*, 673-682.

119. Schindelin, H.; Zhang, M.; Bald, R.; Fuerste, J.-P.; Erdmann, V. A.; Heinemann, U. *J. Mol. Biol.* 1995, *249*, 595-603.
120. Pranata, J.; Wierschke, S. G.; Jorgensen, W. L. *J. Am. Chem. Soc.* 1991, *113*, 2810-2819.
121. DeVoe, H.; Tinoco, I., Jr. *J. Mol. Biol.* 1962, *4*, 500.
122. Brown, R. B.; Godfrey, P. D.; McNaughton, D.; Pierlot, A. P. *J. Amer. Chem. Soc.* 1988, *110*, 2329-2330.
123. Stewart, R. F.; Jensen, L. H. *Acta Cryst*. 1967, *23*, 1102-1105.
124. Stewart, R. F.; Jensen, L. H. *J. Chem. Phys.* 1964, *40*, 2071-2075.
125. Frey, M. N.; Koetzle, T. F.; Lehmann, M. S.; Hamilton, W. C. *J. Chem. Phys.* 1973, *59*, 915-924.
126. O'Brien, E. J. *Acta Cryst*. 1967, *23*, 92-106.
127. Seeman, N. C.; Rosenberg, J. M.; Suddath, F. L.; Park, J. J.; Rich, A. *J. Mol. Biol.* 1976, *104*, 109-144.
128. Rosenberg, J. M.; Seeman, N. C.; Day, R. O.; Rich, A. *J. Mol. Biol.* 1976, *104*, 145-167.

Table 1) DNA and RNA duplex structures included as target data for the parameter optimization

Sequence	Comment	Reference
d(CGATCGATCG)	B form crystal	117
d(GTACGTAC)	A form crystal	118
d(CGCGAATTCGCG)	Contains EcoRI recognition sequence	49,50
d(CATTTGCATC)	NMR solution structure	51
d(CTCGAG)	A to B transition	53
UAAGGAGGUGUA	RNA, 2 duplexes/asymmetric unit	119

Table 2) Comparison of *ab initio* and empirical minimum interaction energies and geometries between selected model compounds and water

Interaction	*Ab Initio*			CHARMM27		
	R_{min}	A_{min}	E_{min}	R_{min}	A_{min}	E_{min}
A.1) THF-HW	2.04		-5.97	1.86		-5.91
B.1) THFOH-OW[a]						
60	2.04		-6.35	1.85		-6.27
180	2.01		-6.49	1.84		-6.39
300	2.02		-6.44	1.84		-6.47
B.2) THFOH-HW[a]						
60	2.08		-4.68	1.87		-4.80
120	2.07		-4.89	1.87		-4.99
300	2.04		-5.80	1.84		-5.59
C.1) DMP-HW[b]	2.02		-7.37	1.89		-7.19
C.2) DMP-HW[b]	1.85		-13.32	1.66		-13.24
C.3) DMP-OW[b]	3.73		-10.48	3.60		-11.08
C.4) DMP-OW[b]	3.38		-16.64	3.29		-16.68
D.1) Ade N1-HW	2.10		-6.99	1.89		-6.97 (0.92)
D.2) Ade H2-OW	2.49		-1.51	2.45		-1.55 (-0.27)
D.3) Ade N3-HW	2.12		-7.09	1.90		-7.07 (0.88)
D.4) Ade H62-OW	2.00		-5.30	1.85		-5.35 (1.06)
D.5) Ade H61-OW	2.08		-4.66	1.89		-4.54 (0.76)
D.6) Ade N7-HW	2.08		-7.09	1.89		-7.16 (0.90)
D.7) Ade H8-OW	2.39		-2.95	2.37		-3.27 (-0.05)
D.8) Ade H9-OW	2.01		-7.25	1.84		-7.25 (1.09)
E.1) Gua H1-OW	2.04		-7.17	1.89		-7.21 (0.47)
E.2) Gua H21-OW	2.08		-8.12	1.89		-8.25 (0.76)
E.3) Gua H22-OW	2.03		-6.11	1.87		-6.24 (0.91)
E.4) Gua N3-HW	2.15		-4.54	1.92		-4.72 (0.50)
E.5) Gua O6-HW	2.07	137	-6.26	1.80	155	-5.47 (0.61)
E.6) Gua O6-HW	1.92	113	-9.98	1.74	106	-9.94 (1.56)
E.7) Gua N7-HW	2.22		-5.14	1.95		-5.05 (0.45)
E.8) Gua H8-OW	2.41		-2.88	2.36		-2.81 (-0.02)
E.9) Gua H9-OW	2.01		-6.83	1.84		-6.85 (1.08)
F.1) Ura H1-OW[c]	1.98		-8.18	1.80		-8.25 (1.52)
F.2) Ura O2-HW[c]	2.08		-5.27	1.80		-5.20 (0.71)
F.3) Ura H3-OW[c]	1.96		-6.81	1.83		-6.83 (1.14)
F.4) Ura O4-HW[c]	2.07		-5.30	1.79		-5.22 (0.79)
F.5) Ura H5-OW	2.43		-2.23	2.43		-2.17 (-0.23)
F.6) Ura H6-OW	2.33		-3.95	2.37		-4.04 (-0.04)
G.1) Thy H1-OW	1.99		-7.69	1.82		-7.51 (1.25)
G.2) Thy O2-HW	1.96	112	-7.86	1.77	103	-8.33 (1.73)

G.3) Thy O2-HW	2.00	119	-6.48	1.78	104	-6.35 (1.03)
G.4) Thy H3-OW	1.97		-6.56	1.83		-6.54 (1.12)
G.5) Thy O4-HW	1.99	120	-6.63	1.76	107	-6.73 (1.24)
G.6) Thy O4-HW	2.04	135	-6.82	1.79	140	-6.29 (0.60)
G.7) Thy H6-OW	2.37		-4.33	2.34		-4.30 (-0.10)

Table 2, continued						
H.1) Cyt H1-OW[c]	2.01		-6.52	1.84		-6.43 (1.05)
H.2) Cyt O2-HW[c]	2.03	113	-8.69	1.79	108	-8.45 (0.67)
H.3) Cyt O2-HW[c]	1.91	114	-10.06	1.72	107	-10.37 (2.04)
H.4) Cyt N3-HW[c]	2.06		-9.71	1.88		-9.71 (0.92)
H.5) Cyt H41-OW[c]	2.03		-5.72	1.86		-5.81 (1.02)
H.6) Cyt H42-OW[c]	2.13		-6.09	1.92		-6.00 (0.56)
H.7) Cyt H5-OW	2.57		-2.96	2.47		-2.84 (-0.33)
H.8) Cyt H6-OW	2.36		-4.23	2.35		-4.11 (0.00)

See Figure 3 for interaction orientations. Minimum energies, E_{min}, in kcal/mole, minimum distances, R_{min}, in Å and minimum angles, A_{min}, in degrees. *Ab initio* interaction energies scaled by 1.16.

a) Dihedral angle (degrees) defining the conformation of the hydroxyl group relative to the furanose in 2'-hydroxy-tetrahydrofuran.

b) *Ab initio* results from MacKerell et al.[25]

c) *Ab initio* results from Pranata et al.[120]

Table 3) 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 (see Table 2).

Table 4) Interaction energies for 26 hydrogen bonded DNA basepairs [a]

Basepair	*Ab initio*	CHARMM27				Difference
		Total	Elec	LJ	Internal	
ATRH	-13.2	-12.98	-12.99	-0.02	0.03	0.22
ATRWC	-12.4	-12.39	-11.66	-0.07	-0.66	0.01
ATH	-13.3	-13.25	-13.45	0.06	0.15	0.05
ATWC	-12.4	-12.66	-12.10	0.03	-0.60	-0.26
AC1	-14.3	-13.70	-12.36	0.06	-1.40	0.60
AC2	-14.1	-13.48	-13.88	-0.35	0.75	0.63
GA1	-15.7	-14.38	-12.02	-0.59	-1.77	1.32
GA2	-10.4	-11.71	-12.39	-0.34	1.02	-1.31
GA3	-15.2	-12.94	-14.88	-0.91	2.85	2.26
GA4	-11.1	-12.69	-11.23	0.14	-1.61	-1.59
GT1	-14.7	-13.50	-13.22	-0.01	-0.27	1.20
GT2	-14.3	-13.02	-12.34	-0.10	-0.59	1.28
GCWC	-25.4	-25.69	-24.58	0.52	-1.63	-0.29
GC1	-13.9	-16.35	-15.01	-0.23	-1.11	-2.45
TC1	-11.6	-10.73	-9.22	-1.19	-0.32	0.88
TC2	-11.8	-11.21	-10.02	-1.10	-0.10	0.59
GG1	-24.0	-22.42	-21.30	0.42	-1.54	1.59
GG3	-17.1	-19.28	-17.37	-0.21	-1.69	-2.18
GG4	-10.3	-11.53	-9.86	-0.21	-1.46	-1.23
AA1	-11.5	-11.31	-10.00	0.36	-1.67	0.19
AA2	-11.0	-10.64	-10.84	-0.03	0.23	0.37
AA3	-10.0	-9.87	-11.51	-0.42	2.06	0.13
TT1	-10.6	-9.61	-9.22	-0.17	-0.23	0.99
TT2	-10.6	-10.02	-9.84	-0.03	-0.15	0.58
TT3	-10.5	-9.21	-8.63	-0.27	-0.31	1.29
CC	-18.8	-18.19	-16.69	-0.31	-1.19	0.61

	R	SD	A	B	AAE
C27	0.96	1.15	-0.34	0.96	0.93
C22	0.89	1.97	-0.33	0.96	1.34
C22, Hobza	0.92	1.78	0.14	1.00	1.0

Energies in kcal/mole. Total interaction energies determined as the difference between the total energy of the minimized dimer and the sum of the minimized monomer energies. Electrostatic, Lennard-Jones (LJ) and Internal energy contributions were obtained by taking the respective energies for the minimized dimer and subtracting the sum of the respective energies for the two monomers. Dimer optimizations involved building the dimers followed by a 200 step ABNR minimization with harmonic force constants of 1.0 kcal/mole on all nonhydrogen atoms, followed by 200 ABNR steps without constraints. Forces at the end of the minimizations were generally less than 0.1 kcal/mole/Å with the largest value being 0.34 kcal/mole/Å for the TC1 dimer.

a) 26 basepairs as described in Figure 1 and Table 1 of Hobza et al.[71]

b) Correlation coefficent, R, standard deviation, SD, y-intercept, A, and slope of linear regression and average absolute error, AAE (kcal/mol) for the present force field (CHARMM27) and the CHARMM22 force field[25] as calculated in our laboratory (C22) and as reported by Hobza et al. (C22, Hobza). The C22 (Hobza) results correspond to C23 data reported by Hobza et al.

Table 5) Dipole moments of the nucleic acid bases from CHARMM27, experiment, *ab initio* gas phase and *ab initio* reaction field calculations

Base	C27	Exp[a]	AI[b]	AI[c]	AI[d]	AI[e]	AI[f]
Adenine	2.91	3	2.46	2.55	2.56	3.10	3.26
Guanine	7.59	-	6.79	6.27	6.49	8.55	8.64
Cytosine	7.85	-	7.06	6.45	6.65	8.98	8.84
Thymine	4.50	-	4.59	4.01	4.31	6.21	5.74
Uracil	4.29	3.9	4.72	-	-	6.46	5.91

Units in debye

a) See references[121,122]

b) HF/6-31G*

c) MP2/6-31G*, see reference[72]

d) MP2/AUG_CC_PVDW, see reference[72]

e) AM1/SM2

f) HF/6-31G*/SCIPCM

Table 6) Small molecule crystals simulated to test the base parameters.

Crystal	Space Group	Asymmetric units/unitcell[a]	Ref.
Uracil	P21/a	4	123
9-methyladenine	P21/c	4	124
9-methyladenine,1-methylthymine[a]	P21/m	2	125
9-ethylguanine, 1-methylcytosine[a]	P-1	2	126

a) The asymmetric units of uracil and 9-methyladenine are comprised of a single molecule while those of 9-methyladenine,1-methylthymine and 9-ethylguanine, 1-methylcytosine are comprised of the hydrogen bonded purine-pyrmidine pair of molecules.

Table 7) Experimental and calculated unitcell parameters for uracil, 9-methyladenine, 9-methyladenine/1-methylthymine and 9-ethylguanine/1-methylcytosine.

System	A	B	C	α	β	γ	Volume
Uracil							
exper	11.938	12.376	3.655	-	120.54	-	465.12
Atom	11.39±0.43	11.82±0.21	3.57±0.08	-	104.0±6.7	-	462.4±10.7
P8	11.81±0.94	11.83±0.21	3.60±0.08	-	102.3±17.9	-	464.6±11.1
P12	10.80±0.28	11.86±0.21	3.58±0.08	-	93.3±5.3	-	455.5±10.3
P22	10.97±0.39	11.79±0.18	3.58±0.07	-	96.3±9.2	-	453.6±9.2
C22	13.25±0.57	11.92±0.20	3.49±0.07	-	124.1±3.7	-	454.4±7.6
9-methyladenine							
exper	7.67	12.24	8.47	-	123.26	-	664.9
Atom	8.10±0.28	12.25±0.49	7.92±0.15	-	120.94±1.60	-	673.1±18.7
P8	7.75±0.25	12.73±0.30	8.03±0.18	-	120.96±2.78	-	677.6±16.3
P12	7.69±0.20	12.67±0.24	7.99±0.19	-	120.65±2.75	-	667.9±16.3
P22	7.67±0.25	12.67±0.27	7.99±0.21	-	120.77±2.91	-	665.9±14.3
C22	7.61±0.16	12.61±0.23	8.08±0.15	-	121.60±2.09	-	660.1±13.6
9-methyladenine/1-methylthymine							
exper	8.304	6.552	12.837	-	106.83	-	668.5
P12	8.24±0.11	7.14±0.12	11.80±0.18	-	105.9±1.6	-	666.8±12.8
P22	8.23±0.16	7.19±0.16	11.84±0.28	-	105.8±2.3	-	672.9±15.7
C22	8.27±0.15	6.90±0.15	12.14±0.33	-	104.5±2.3	-	669.5±15.0
9-ethylguanine/1-methylcytosine							
exper	8.838	11.106	7.391	107.49	87.30	91.27	691.1
P12	8.21±0.43	10.01±0.60	9.55±1.21	99.4±5.0	82.5±4.2	109.0±7.4	711.5±22.2
P22	8.25±0.48	10.18±0.74	9.08±1.57	100.3±6.3	84.9±5.1	103.9±9.1	697.8±22.0
C22	12.07±0.78	11.47±1.05	7.92±0.99	113.2±12.8	131.4±7.3	76.9±4.0	699.6±19.3

Distances in Å, angles in degrees and volumes in $Å^3$. Errors represent the rms fluctuations. Atom indicates atom based truncation using the 14 Å list generation, 12 Å for the nonbond interaction truncation and 10 Å for initiation of the switchng function. P indicates the Particle Mesh Ewald method with real space cutoffs of 8, 12 or 22 Å represented by P8, P12 and P22, respectively. C22 calculations were performed using PME with the 22 Å real space truncation distance.

Table 8) Experimental and calculated heats of sublimation for uracil and 9-methyladenine.

Truncation	Experimental	Calculated
Uracil		
Atom	28.8,29.1	30.9
P8		31.3
P12		32.6
P22		33.2
C22, P22		34.7
9-methyladenine		
Atom	32,33	33.2
P8		31.0
P12		33.1
P22		33.7
C22, P22		35.1

Energies in kcal/mole. Heats of sublimation were determined as in MacKerell et al.[25] Experimental data from reference.[76] Atom indicates atom based truncation using the 14 Å list generation, 12 Å for the nonbond interaction truncation and 10 Å for initiation of the switching function. P indicates the Particle Mesh Ewald method with real space cutoffs of 8, 12 or 22 Å represented by P8, P12 and P22, respectively.

Table 9) Watson-Crick and Hoogsteen basepairing interaction energies, zero point energies and interaction enthalpies for the methylated bases.

Interaction Energy	Zero Point Vibrational			ΔE_{vib}	ΔH interaction[a]		
	Pur	Pyr	Dimer		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	87.3	76.34	165.60	1.91	-20.99	21.0	19.7-25.4

Energies in kcal/mole. 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. $\Delta H_{interaction}$ 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 reference[76] and *ab initio* ΔH interaction energies from reference.[70] The range of values are from different levels of theory used in that study.

Table 10) Hydrogen bond distances for the Watson-Crick and Hoogsteen basepairs

Basepair	Distances, Å		
AT Watson-Crick	N6-O4	N3-N1	
C27	2.90	2.88	
Exp[a]	2.93	2.85	
AT Hoogsteen	N6-O4	N7-N3	
C27	2.91	2.87	
Exp[b]	2.86	2.93	
AU Watson-Crick	N6-O4	N3-N1	
C27	2.89	2.87	
Exp[a]	2.93	2.85	
AU Hoogsteen	N6-O4	N7-N3	
C27	2.90	2.86	
Exp[b]	2.86	2.93	
GC Watson-Crick	O2-N2	N3-N1	N4-O6
C27	2.85	2.92	2.84
Exp[c]	2.86	2.95	2.91

a) see reference [127]

b) see reference [125]

c) see reference [128]

Table 11) *Ab initio* and CHARMM27 interaction energies for selected Watson-Crick basepairs, intrastrand stacking interactions and interstrand interactions based on the crystal structure of the CCAACGTTGG BDNA duplex.[a]

Orientation type[b]	*Ab initio*	CHARMM27 Total	CHARMM27 Elec	CHARMM27 LJ	Difference
HBONDED					
A2T18:	-13.4	-12.10	-12.95	0.85	1.30
A4T17:	-13.2	-11.34	-11.58	0.24	1.86
C1G20:	-24.5	-25.13	-25.59	0.46	-0.63
C2G19:	-26.0	-25.24	-25.41	0.18	0.76
C2G19:	-27.6	-25.28	-26.89	1.61	2.32
STACKED					
A3A4:	-6.3	-6.55	2.11	-8.66	-0.25
A4C5:	-4.8	-4.32	3.07	-7.38	0.48
C2A3:	-2.4	-1.92	3.00	-4.92	0.48
C1C2:	0.0	2.54	7.37	-4.84	2.54
C5G6:	-5.3	-5.88	-0.82	-5.06	-0.58
G9G10:	-2.9	0.55	10.10	-9.55	3.45
G6T7:	-4.9	-6.43	1.62	-8.05	-1.53
T8G9:	-6.0	-7.30	-2.80	-4.51	-1.30
T7T8:	-3.0	-3.32	2.37	-5.70	-0.32
INTERSTRAND					
A3G19:	-2.7	-3.82	-2.04	-1.78	-1.12
A4T18:	-0.9	-1.19	-0.17	-1.01	-0.29
C2G20:	-3.5	-5.42	-4.14	-1.29	-1.92
C5T17:	0.2	0.37	1.02	-0.64	0.17
G6G16:	-4.6	-7.54	-1.92	-5.62	-2.94
A4G16:	-4.5	-5.75	-3.13	-2.62	-1.25
A3T17:	-1.9	-3.32	-1.66	-1.66	-1.42
C5C15:	1.4	1.72	2.67	-0.96	0.32
C1G19:	-5.4	-7.59	-5.82	-1.77	-2.19
C2T18:	-2.5	-3.94	-0.54	-3.41	-1.44

SUMMATION	A.I.	C27	C22[b]	C22(Hobza)
HBONDED	-209.4	-198.2	-199.8	-205.9
STACKED	-71.0	-65.3	-72.8	-53.2
INTERSTRAND	-48.8	-73.0	-79.0	-79.6
H+S+I (sum)	-329.2	-336.5	-351.6	-338.7

Energies in kcal/mole. Structures for the interaction energy calculated were obtained by taking the non-hydrogen atom coordinates for the bases from the crystal structure of CCAACGTTGG,[78] adding hydrogens and minimizing the hydrogens for 100 ABNR steps with all non-hydrogen atoms fixed.
a) Interaction pairs selected from Table 4 of Hobza et al.[71]

b) HBONDED indicates Watson-Crick basepair interactions, STACKED indicates interactions between adjacent intrastrand bases and INTERSTRAND indicates interactions between base i in strand one and base i + 1 or base i - 1 in strand 2 (i.e. the base in strand one and either one of the two bases in strand two adjacent to the base that would normally be in a Watson-Crick basepairing interaction with the base in strand one).
c) CHARMM22 values calculated in the present study and C22(Hobza) are those reported by Hobza et al. (see Table 4 legend).

Table 12) Comparison of the X-ray derived and CHARMM27 deoxyribose bond lengths (Å) and valence angles (degrees).

	South				North			
Bond	X-Ray	C27	σ	Δ	X-Ray	C27	σ	Δ
C1'-C2'	1.518	1.530	0.010	0.012	1.519	1.536	0.010	0.017
C2'-C3'	1.516	1.513	0.008	0.003	1.518	1.510	0.012	0.008
C3'-C4'	1.529	1.538	0.010	0.009	1.521	1.534	0.010	0.013
C4'-O4'	1.446	1.458	0.010	0.012	1.449	1.458	0.009	0.009
O4'-C1'	1.420	1.431	0.011	0.011	1.418	1.434	0.012	0.016
C3'-O3'	1.435	1.425	0.013	0.010	1.419	1.425	0.006	0.006
C5'-C4'	1.512	1.536	0.007	0.024	1.509	1.535	0.011	0.026
C1'-N1/N9	1.468	1.466	0.014	0.002	1.488	1.477	0.013	0.011
(H)O5'-C5'	1.418	1.435	0.025	0.017	1.423	1.435	0.011	0.012
Angle								
C1'-C2'-C3'	102.5	102.8	1.2	0.3	102.4	102.5	0.8	0.1
C2'-C3'-C4'	103.1	104.7	0.9	1.6	102.2	101.6	0.7	0.6
C3'-C4'-O4'	106.0	105.6	0.6	0.4	104.5	104.6	0.4	0.1
C4'-O4'-C1'	110.1	110.5	1.0	0.4	110.3	109.9	0.7	0.4
O4'-C1'-C2'	105.9	106.3	0.8	0.4	106.8	105.8	0.5	1.0
C2'-C3'-O3'	109.4	110.5	2.5	1.1	112.6	110.5	3.3	2.1
C4'-C3'-O3'	109.7	114.4	2.5	4.7	112.3	110.3	2.0	2.0
C5'-C4'-C3'	114.1	113.6	1.8	0.5	115.7	114.0	1.2	1.7
C5'-C4'-O4'	109.3	110.1	1.9	0.8	109.8	111.2	1.1	1.4
O4'-C1'-N1/N9	108.0	108.4	0.7	0.4	108.3	109.6	0.3	1.3
C2'-C1'-N1/N9	114.3	113.8	1.4	0.5	112.6	114.1	1.9	1.5
(H)O5'-C5'-C4'	110.9	113.7	1.7	2.8	111.0	113.6	2.5	2.6
C1'-N9-C4	126.3	127.1	1.2	0.8	123.9	125.0	1.0	1.1
C1'-N1-C2	117.8	119.4	1.8	1.6	117.5	117.5	1.4	0.0

X-Ray refers to the mean values (standard deviation σ) obtained from statistical analysis of crystal structures of nucleosides and nucleotides,[80] and C27 refers to their CHARMM27 counterpart. The comparison is provided for the deoxyribose in either the north or the south conformation. Δ is the absolute difference between the mean X-ray value and the corresponding CHARMM27 value.

Table 13) Comparison of the X-ray derived and CHARMM27 ribose bond lengths (Å) and valence angles (degrees).

	South				North			
Bond	X-Ray	C27	σ	Δ	X-Ray	C27	σ	Δ
C1'-C2'	1.526	1.513	0.008	0.013	1.529	1.522	0.011	0.007
C2'-C3'	1.525	1.526	0.011	0.001	1.523	1.522	0.011	0.001
C3'-C4'	1.527	1.538	0.011	0.011	1.521	1.533	0.010	0.012
C4'-O4'	1.454	1.458	0.010	0.004	1.451	1.456	0.013	0.005
O4'-C1'	1.415	1.420	0.012	0.005	1.412	1.424	0.013	0.012
C3'-O3'	1.427	1.436	0.012	0.009	1.417	1.439	0.014	0.022
C5'-C4'	1.509	1.538	0.012	0.029	1.508	1.533	0.007	0.025
C2'-O2’	1.412	1.419	0.013	0.007	1.420	1.416	0.010	0.004
C1'-N1/N9	1.464	1.474	0.014	0.010	1.483	1.482	0.015	0.001
(H)O5'-C5'	1.424	1.434	0.016	0.010	1.420	1.435	0.009	0.015
Angle	X-Ray	C27	σ	Δ	X-Ray	C27	σ	Δ
C1'-C2'-C3'	101.5	101.7	0.8	0.2	101.3	101.3	0.7	0.0
C2'-C3'-C4'	102.6	103.4	1.0	0.8	102.6	101.0	1.0	1.6
C3'-C4'-O4'	106.1	105.3	0.8	0.8	104.0	104.2	1.0	0.2
C4'-O4'-C1'	109.7	109.8	0.7	0.1	109.9	109.1	0.8	0.8
O4'-C1'-C2'	105.8	107.6	1.0	1.8	107.6	107.3	0.9	0.3
C1'-C2'-O2’	111.8	112.1	2.6	0.3	108.4	110.6	2.4	2.2
C3'-C2'-O2’	114.6	112.5	2.2	2.1	110.7	111.2	2.1	0.5
C2'-C3'-O3'	109.5	113.4	2.2	3.9	113.7	113.8	1.6	0.1
C4'-C3'-O3'	109.4	112.8	2.1	3.4	113.0	110.8	2.0	2.2
C5'-C4'-C3'	115.2	113.9	1.4	1.3	116.0	114.9	1.6	1.1
C5'-C4'-O4'	109.1	110.8	1.2	1.7	109.8	110.5	0.9	0.7
O4'-C1'-N1/N9	108.2	111.7	0.8	3.5	108.5	111.6	0.7	3.1
C2'-C1'-N1/N9	114.0	111.4	1.3	2.6	112.0	111.8	1.1	0.2
(H)O5'-C5'-C4'	111.7	113.9	1.9	2.2	111.5	113.0	1.6	1.5
C1'-N9-C4	127.4	126.9	1.2	0.5	126.3	126.3	2.8	0.0
C1'-N1-C2	118.5	119.1	1.1	0.6	116.7	118.9	0.6	2.2

X-Ray refers to the mean values (standard deviation σ) obtained from statistical analysis of crystal structures of nucleosides and nucleotides,[80] and C27 refers to their CHARMM27 counterpart. The comparison is provided for the deoxyribose in either the north or the south conformation. Δ is the absolute difference between the mean X-ray value and the corresponding CHARMM27 value.
See legend of Table 12 for definitions.

Table 14) Vibrational data on Compound F

	Charmm27			*Ab initio*		
	Freq.	Assignment		Freq.	Assignment	
1	33	tZET	64	24	tEPS	84
		tEPS	32			
2	92	tEPS	49	49	tZET	76
		tRING'	21		EPS	16
		tZET	16			
3	104	tRING'	49	87	tRING'	79
		tZET	18			
4	143	dC3O3P	70	104	dC3O3P	36
		dC4C3O3	16		tRING	20
					tRING'	19
5	213	tCH3	46	171	tRING	52
		dC5C4C3	16		dC3O3P	25
6	228	tCH3	34	190	CH3	21
		tRING	30		dC4C3O3	15
7	238	tRING	37	232	CH3	76
		tCH3	16			
8	304	rPO3'	18	265	rPO3	32
		rPO3	16		sO3P	18
		dC2C3O3	16			
9	372	dC5C4C3	32	318	dC5C4C3	37
		rPO3'	24		dC4C3O3	19
		dC4C3O3	16			
10	421	rPO3	35	344	rPO3'	41
		dC5C4O4	27		dC5C4O4	21
					asPO3'	16
11	452	dC5C4O4	34	422	dC5C4O4	27
		rPO3	15		dC2C3O3	19
					asPO3	17
12	526	asPO3'	53	457	asPO3	33
		dRING'	21		asPO3'	22
13	535	asPO3	59	497	asPO3'	38
14	545	sPO3	57	517	asPO3	30
					rPO3	21
15	559	asPO3'	27	555	sPO3	87
16	587	dC2C3O3	29	565	dRING'	25
		rPO3	17		dC3O3P	18
17	637	dRING	57	632	dRING	59

18	788	sO3P	24	674	sO3P	49
19	817	rC1H2	24	809		
		rkC2'H2	24			
		sO3P	16			

Table 14 continued

20	855	sC1O4	18	835	wC1H2	43
21	888	sC3C4	15	860	sC4O4	29
					rCH3	17
					sC3C4	15
22	933	sC1O4	22	913	sPO3	34
		rCH3	20		sC1C2	29
23	988	rkC2'H2	29	919	sPO3	64
					sC1C2	17
24	1010	sC1C2	39	961	rCH3'	31
		rCH3	16			
25	1030	sPO3	48	1010	sC1O4	30
					sC4C5	19
26	1053			1065	rCH3	26
27	1068	rC4H'	22	1077	asPO3	65
		rC3H	20		asPO3'	19
28	1076	rCH3'	26	1096	sC1O4	24
		sC4C5	24		sC4O4	23
29	1098	rC1H2	30	1111	asPO3'	72
		sC2C3	17		asPO3	19
30	1114	sC3O3	36	1121	sC4C5	18
31	1161	twC1H2	90	1146	rC1H2	18
		rC1H2	-32		rkC2'H2	17
		twC2H2	15			
32	1182	twC2H2	54	1165	sC3O3	61
					rC1H2	26
33	1241	rC3H	26	1187	wC2H2	33
		rC4H'	24		rC1H2	29
		twC2H2	19			
34	1262	asPO3	88	1204	rC1H2	38
					wC1H2	27
35	1266	asPO3'	84	1267	wC1H2	39
					C2H2	37
					rC4H'	16
36	1304	wC2H2	62	1287	rC4H'	30
					rC3H	22
37	1325	rC4H'	40	1332	rC4H'	22
					rC3H	22
					rC3H'	20
					rC4H'	20
38	1343	wC1H2	44	1361	C1H2	50
					rC3H'	16

39	1384	dCH3s	70	1374	C1H2	43
					dCH3s	22
					rC4H'	15
40	1410	dCH3s	23	1394	dCH3s	43
		rC4H'	19		rC3H'	25
41	1431	dCH3s	75	1410	rC4H'	28
					rC3H	22
42	1437	dCH3a'	78	1460	dCH3a'	87

Table 14 continued

43	1446	scC1H2	91	1466	dCH3s	86
44	1474	scC2H2	95	1470	scC2H2	88
45	1548	rC3H'	51	1507	scC1H2	92
		rC3H	20			
46	2845	sCH3	95	2837	sCH2	84
47	2852	sCH2	99	2838	sCH3	48
					asCH3'	19
48	2864	sC4H	74	2843	asCH2	94
		sC3H	20			
49	2872	sC3H	78	2856	sC4H	61
		sC4H	20		sCH3	17
50	2893	asCH2	99	2866	sC3H	64
51	2902	asCH3	98	2885	sCH2	52
					asCH2	23
					sC3H	16
52	2905	asCH3'	98	2907	asCH3'	71
					sCH3	23
53	2910	sCH2	98	2951	asCH3	90
					sCH2	28
54	2943	asCH2	100	2968	asCH2	71

Frequencies in cm^{-1}. Symbols represent; s, stretching modes; as, asymmetric stretching modes; d, bends; w, out-of-plane deformations (wags); r, rocking modes, t, torsional modes, tw, twisting modes, and sc, scissor modes. Only internal coordinates contributing 15% or more to the potential energy distribution are reported.

Table 15) Vibrational data on Compound B, dianionic form.

	Charmm27			*Ab initio*		
	Freq.	Assignment		Freq.	Assignment	
1	29	tALPHA	72	33	tALPHA	91
		tGAMMA	31			
2	53	tGAMMA	72	55	tGAMMA	55
		tALPHA	27		tBETA	32
3	72	tBETA	118	86	tBETA	56
					tGAMMA	17
4	97	tRING	61	120	tRING	40
		tRING'	19		tBETA	28
					tRING'	28
5	218	dC5O5P	55	148	dC3C4C5	21
		rPO3'	17		tRING'	20
					tRING	17
6	248	scC4C5O5	42	189	dC5O5P	47
		rPO3	16			
7	311	tRING'	46	206	tRING'	37
		dO4C4C5	30		scC4C5O5	29
		dC3C4C5	21			
8	388	rPO3'	24	321	dC3C4C5	27
		dC3C4C5	21		rPO3'	22
		tRING'	16		rPO3	16
		tRING	15			
9	446	rPO3'	28	356	rPO3'	32
		dC3C4C5	27		rPO3	25
		rPO3	22		dPO3as	17
10	468	rPO3	37	382	dO4C4C5	27
11	497	dPO3s	50	463	sO5P	25
		dO4C4C5	20		dPO3as	20
12	544	dPO3as	50	498	dPO3as	45
		dPO3as	28		dPO3as	21
13	553	dPO3as	47	524	dPO3as	28
		dPO3as	20		rPO3	19
14	584	dRING'	20	555	dPO3s	47
15	594	dRING'	38	608	sO5P	43
		dRING	21		dPO3s	31
16	670	dRING	37	677	dRING	44
		sC3C4	16		dRING'	33
17	721	sC4O4	30	743	rC2H2	29
					sC3C4	19
18	778	sO5P	39	793	sC4C5	28
		sPO3	16			

19	802	rC3H2	43	813	sC4O4	18
		sC1O4	19		dRING'	17
					rC2H2	17
20	847	rC2H2	26	862	sC4O4	23
		rC1H2	18		sC3C4	20
21	888	sC1O4	25	888	sC1C2	46
		rC5H2	24		sC2C3	34

Table 15 continued

22	920	sC2C3	20	913	sPO3	90
		rC1H2	19			
		rC3H2	17			
23	933	sC1C2	34	932	rC3H2	24
					rC1H2	16
24	987	sC5O5	20	972	sC1C2	23
		sPO3	15		sC2C3	16
					sC4C5	16
25	1026	rC4H4'	15	1005	rC5H2	26
					sC1O4	23
26	1034	sPO3	36	1040	rC4H4	19
27	1049	wC2H2	17	1073	asPO3	89
28	1058	rC1H2	27	1090		
		rC2H2	22			
29	1105	twC1H2	58	1109	asPO3'	87
30	1128	twC2H2	35	1114	sC1O4	27
		sC5O5	17		sC4O4	19
					rC5H2	18
31	1142	twC2H2	52	1154	sC5O5	53
32	1172	twC3H2	64	1204	rC1H2	25
					twC3H2	21
					twC2H2	18
33	1247	rC4H4	32	1236	twC1H2	64
34	1262	asPO3	68	1239	twC3H2	27
		asPO3'	21		twC2H2	23
35	1265	wC2H2	38	1256	twC5H2	58
		sC2C3	15			
36	1266	asPO3'	66	1298	wC2H2	35
		asPO3	18		twC2H2	18
37	1285	rC4H4'	34	1315	wC3H2	36
		wC3H2	29		rC4H4	25
					wC1H2	23
38	1301	wC1H2	50	1332	wC3H2	29
		sC1O4	16		wC2H2	29
					rC4H4'	21
39	1381	wC3H2	20	1355	rC4H4'	30
		twC5H2	16		rC4H4	17
40	1394	scC5H2	45	1385	wC1H2	52
		twC5H2	22			
41	1402	twC5H2	40	1394	wC5H2	71
		scC5H2	25			
		wC5H2	17			
42	1465	scC3H2	65	1459	scC3H2	51
		scC2H2	25		scC2H2	40

43	1480	scC2H2	71	1473	scC5H2	94
		scC3H2	24			
44	1522	scC1H2	95	1478	scC3H2	40
					scC2H2	34
					scC1H2	22
45	1647	wC5H2	63	1502	scC1H2	73
		scC5H2	19		scC2H2	20

Table 15 continued

46	2855	sC5H2	98	2778	sC5H2	59
					asC5H2	39
47	2857	sCH2	92	2801	sCH2	58
					asCH2	34
48	2858	sCH2	68	2822	sCH2	60
		sC4H4	30		asCH2	38
49	2860	sC4H4	62	2841	sC4H4	56
		sCH2	36		sCH2	26
50	2886	asC5H2	98	2855	sCH2	46
					sC4H4	27
					asCH2	23
51				2886	asC5H2	54
	2898	asCH2	93		sC5H2	41
52	2902	asCH2	80	2935	asCH2	61
		sCH2	20		sCH2	39
53	2907	sCH2	77	2954	asCH2	65
		asCH2	22		sCH2	35
54	2944	asCH2	100	2976	asCH2	65
					sCH2	35

See Table 14 legend for definitions.

Table 16) Vibrational data on Compound E with an imidazole base and a 5' methyl group.

	Charmm27			*Ab initio*		
	Freq.	Assignment		Freq.	Assignment	
1	33	tCHI	80	27	tCHI	51
		tSUGA'	18		tSUGA'	37
2	57	tSUGA'	66	63	tSUGA'	49
		wGLYC	16		wGLYC	35
					tCHI	15
3	76	wGLYC	73	111	wGLYC	29
					tCHI	25
					tSUGA	22
4	206	tSUGA	51	191	tSUGA	52
		tCH3	36		wGLYC	27
5	222	tCH3	58	214	rC1ND1CG	48
		tSUGA	21		dO4C1ND1	22
6	314	dC5C4C3	67	241	tCH3	83
7	325	dO4C1ND1	32	308	dC5C4C3	40
		rC1ND1CG	28		dC5C4O4	36
8	356	sC1ND1	28	349	sC1ND1	21
					dC2C1ND1	19
9	424	dC5C4O4	24	425	dC2C1ND1	17
					dO4C1ND1	17
10	502	dC5C4O4	34	455	dC5C4O4	40
					dC5C4C3	18
11	575	dRING'	58	566	dRING'	45
					rC3H2	15
12	616	wHE1	50	608	tIMID'	88
13	654	dRING	30	651	dRING	24
					tIMID	16
14	686	wHG	38	657	tIMID	84
		tIMID	30			
15	695	wHD2	26	748	wHG	87
16	718	tIMID	21	770	dRING	25
					dO4C1ND1	15
17	768	sC4O4	21	799	sC3C4	29
		sC3C4	18		sC4C5	16
18	830	tIMID'	28	859	sC4O4	31
		rC2H2	22		rC2H2	16
19	852			882	wHD2	65
					wHE1	36
20	887	dIMID'	42	897	dIMID'	85
21	922	dIMID'	23	902	rCH3	20
		rC3H2	22		sC2C3	16

					sC3C4	16
22	943	sC2C3	22	906	sC2C3	27
		rCH3	21		sC1C2	26
23	950	wHD2	40	908	wHE1	62
		wHG	38		wHD2	27

Table 16 continued

24	972	dIMID	31	955	rCH3'	32
					rC2H2	19
25	990	rCH3'	27	979	sC2C3	21
		rC2H2	19			
26	1005	wC3H2	16	1013	dIMID	24
		rC1H1	16		sND1CG	18
					sCE1ND1	16
27	1032	sCD2NE2	32	1046	sC4C5	30
		rHE1	21		sC1O4	20
		sNE2CE1	20			
28	1041	rC4H'	27	1077	rHG	34
		rCH3'	18			
29	1051	sC4C5	20	1099	sCD2NE2	51
30	1064			1105	rHD2	21
31	1070	rC1H1'	19	1128	rCH3	24
					sC4C5	18
32	1088	rHG	40	1153	sC1O4	30
		rHD2	33		sC4O4	24
33	1104	rC1H1	15	1195	twC2H2	35
34	1117	twC3H2	40	1203	twC3H2	49
		twC2H2	22			
35	1136	twC2H2	45	1244	rHE1	47
		twC3H2	26			
36	1207	rHE1	38	1257	sND1CG	27
		rHD2	33			
37	1221	wC3H2	40	1302	wC3H2	41
		rC4H	20		rC4H'	17
38	1272	wC2H2	30	1304	rHD2	21
		rC1H1'	19		sCE1ND1	18
39	1300	wC2H2	31	1333	wC2H2	62
		sC1C2	16			
40	1318	rC4H'	24	1347	rC4H'	25
		sC4O4	17			
41	1336	rC1H1	25	1368	rC1H1'	36
					rC4H'	28
42	1396	dCH3s	35	1382	sNE2CE1	25
		rC4H	26		rC1H1'	17
43	1417	dCH3s	57	1389	dCH3s	34
					rC4H	30
44	1427	dCH3a	87	1401	rC1H1	41
45	1431	dCH3a'	77	1419	dCH3s	58
					rC4H	21
46	1461	sCD2NE2	29	1468	scC2H2	44
		rHG	18		dCH3a'	26

					scC3H2	23
47	1479	scC2H2	80	1470	dCH3a'	61
		scC3H2	16		scC2H2	16
48	1485	scC3H2	79	1478	dCH3a	85
		scC2H2	16			
49	1526	sNE2CE1	22	1486	scC3H2	59
		sCGCD2	20		scC2H2	34

Table 16 continued						
50	1639	dIMID	26	1517	sCGCD2	23
		sCGCD2	26		rHG	21
51	1648	sCE1ND1	29	1550	sNE2CE1	36
		sND1CG	22		sCGCD2	33
		rC1ND1CG	18			
52	2846	sCH3	92	2884	sC4H4	82
53	2853	sCH2	95	2891	sCH3	85
54	2863	sC4H4	86	2906	sCH2	71
					asCH2	24
55	2864	sC1H	95	2910	sCH2	93
56	2897	asCH2	94	2941	sC1H	92
57	2902	asCH3	93	2948	asCH3	90
58	2904	asCH3'	92	2956	asCH2	58
					asCH3'	23
59	2906	sCH2	94	2962	asCH3'	74
					asCH2	20
60	2943	asCH2	100	2976	asCH2	81
					sCH2	18
61	3061	sCE1HE1	99	3092	sCD2HD2	77
					sCGHG	23
62	3159	sCD2HD2	66	3117	sCGHG	76
		sCGHG	33		sCD2HD2	22
63	3166	sCGHG	66	3140	sCE1HE1	99
		sCD2HD2	33			

See Table 14 legend for definitions.

Table 17) RMS differences of the bonds and valence angles for the methylated bases.

Base	CHARMM27			CHARMM22		
	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.[82] See Table S4 of the Supplemental Information for original data.

Table 18) Comparison of the methylated base geometries from crystal survey data[a] and the CHARMM27 and CHARMM22 force fields.

	Survey	CHARMM27		CHARMM22	
Adenine		Calc.	Diff.	Calc.	Diff.
Bonds					
N1-C2	1.339	1.335	-0.004	1.330	-0.009
C2-N3	1.331	1.336	0.005	1.331	0.000
N3-C4	1.344	1.343	-0.001	1.357	0.013
C4-C5	1.383	1.384	0.001	1.382	-0.001
C5-C6	1.406	1.410	0.004	1.416	0.010
C6-N1	1.351	1.354	0.003	1.348	-0.003
C5-N7	1.388	1.390	0.002	1.387	-0.001
N7-C8	1.311	1.309	-0.002	1.315	0.004
C8-N9	1.373	1.379	0.006	1.371	-0.002
N9-C4	1.374	1.383	0.009	1.407	0.033
C6-N6	1.335	1.345	0.010	1.344	0.009
N9-C9	1.464	1.468	0.004	1.470	0.006
RMSD			0.005		0.012
Angles					
C6-N1-C2	118.6	118.4	-0.2	119.7	1.1
N1-C2-N3	129.3	129.4	0.1	128.9	-0.4
C2-N3-C4	110.6	110.7	0.1	110.8	0.2
N3-C4-C5	126.8	126.8	0.0	126.6	-0.2
C4-C5-C6	117.0	116.9	-0.1	116.9	-0.1
C5-C6-N1	117.7	117.8	0.1	117.2	-0.5
C4-C5-N7	110.7	110.9	0.2	111.8	1.1
C5-N7-C8	103.9	103.9	0.0	103.1	-0.8
N7-C8-N9	113.8	113.9	0.1	115.1	1.3
C8-N9-C4	105.8	105.5	-0.3	104.7	-1.1
C5-C4-N9	105.8	105.7	-0.1	105.2	-0.6
N3-C4-N9	127.4	127.5	0.1	128.3	0.9
C6-C5-N7	132.3	132.2	-0.1	131.2	-1.1
N1-C6-N6	118.6	118.7	0.1	119.9	1.3

Table 18 continued						
	C5-C6-N6	123.7	123.5	-0.2	122.9	-0.8
	C4-N9-C9	126.3	126.6	0.3	126.5	0.2
	C8-N9-C9	127.7	127.9	0.2	128.8	1.1
	RMSD			0.2		0.9
Guanine						
	Bonds					
	N1-C2	1.373	1.372	-0.001	1.376	0.003
	C2-N3	1.323	1.324	0.001	1.331	0.008
	N3-C4	1.350	1.350	0.000	1.356	0.006
	C4-C5	1.379	1.384	0.005	1.395	0.016
	C5-C6	1.419	1.421	0.002	1.416	-0.003
	C6-N1	1.391	1.390	-0.001	1.397	0.006
	C5-N7	1.388	1.389	0.001	1.382	-0.006
	N7-C8	1.305	1.308	0.003	1.312	0.007
	C8-N9	1.374	1.380	0.006	1.367	-0.007
	N9-C4	1.375	1.380	0.005	1.404	0.029
	C2-N2	1.341	1.323	-0.018	1.333	-0.008
	C6-O6	1.237	1.232	-0.005	1.233	-0.004
	N9-C9	1.459	1.467	0.008	1.469	0.010
	RMSD			0.006		0.011
	Angles					
	C6-N1-C2	125.1	125.3	0.2	125.5	0.4
	N1-C2-N3	123.9	123.8	-0.1	122.7	-1.2
	C2-N3-C4	111.9	112.2	0.3	113.8	1.9
	N3-C4-C5	128.6	128.4	-0.2	126.8	-1.8
	C4-C5-C6	118.8	118.8	0.0	119.4	0.6
	C5-C6-N1	111.5	111.6	0.1	111.9	0.4
	C4-C5-N7	110.8	111.0	0.2	110.6	-0.2
	C5-N7-C8	104.3	104.3	0.0	104.2	-0.1
	N7-C8-N9	113.1	113.2	0.1	114.9	1.8
	C8-N9-C4	106.4	106.2	-0.2	104.8	-1.6
	N9-C4-C5	105.4	105.3	-0.1	105.5	0.1
	N3-C4-N9	126.0	126.3	0.3	127.7	1.7
	C6-C5-N7	130.4	130.3	-0.1	130.0	-0.4
	N1-C2-N2	116.2	116.3	0.1	113.4	-2.8
	N3-C2-N2	119.9	119.9	0.0	124.0	4.1
	N1-C6-O6	119.9	119.9	0.0	120.8	0.9
	C5-C6-O6	128.6	128.5	-0.1	127.3	-1.3
	C4-N9-C9	126.5	126.7	0.2	126.2	-0.3
	C8-N9-C9	127.0	127.1	0.1	129.0	2.0
	RMSD			0.1		1.6
Cytosine						
	Bonds					

N1-C2	1.397	1.407	0.010	1.440	0.043
C2-N3	1.353	1.358	0.005	1.374	0.021
N3-C4	1.335	1.338	0.003	1.332	-0.003
C4-C5	1.425	1.429	0.004	1.435	0.010
C5-C6	1.339	1.342	0.003	1.341	0.002
C6-N1	1.367	1.366	-0.001	1.379	0.012
C2-O2	1.240	1.241	0.001	1.240	0.000

Table 18 continued						
	C4-N4	1.335	1.333	-0.002	1.341	0.006
	N1-C1	1.470	1.470	0.000	1.472	0.002
	RMSD			0.004		0.017
	Angles					
	C6-N1-C2	120.3	120.5	0.2	119.0	-1.3
	N1-C2-N3	119.2	118.8	-0.4	118.6	-0.6
	C2-N3-C4	119.9	120.3	0.4	120.6	0.7
	N3-C4-C5	121.9	121.7	-0.2	121.9	0.0
	C4-C5-C6	117.4	117.6	0.2	118.0	0.6
	C5-C6-N1	121.0	121.1	0.1	121.9	0.9
	N1-C2-O2	118.9	119.2	0.3	121.3	2.4
	N3-C2-O2	121.9	121.9	0.0	120.1	-1.8
	N3-C4-N4	118.0	118.0	0.0	117.8	-0.2
	C5-C4-N4	120.2	120.3	0.1	120.3	0.1
	C2-N1-C1	118.8	118.8	0.0	119.4	0.6
	C6-N1-C1	120.8	120.7	-0.1	121.6	0.8
	RMSD			0.2		1.1
Uracil						
	Bonds					
	N1-C2	1.381	1.389	0.008	1.429	0.048
	C2-N3	1.373	1.371	-0.002	1.389	0.016
	N3-C4	1.380	1.378	-0.002	1.391	0.011
	C4-C5	1.431	1.434	0.003	1.445	0.014
	C5-C6	1.337	1.343	0.006	1.347	0.010
	C6-N1	1.375	1.371	-0.004	1.388	0.013
	C2-O2	1.219	1.225	0.006	1.228	0.009
	C4-O4	1.232	1.226	-0.006	1.227	-0.005
	N1-C1	1.469	1.473	0.004	1.473	0.004
	RMSD			0.005		0.019
	Angles					
	C6-N1-C2	121.0	121.0	0.0	119.9	-1.1
	N1-C2-N3	114.9	114.7	-0.2	114.8	-0.1
	C2-N3-C4	127.0	127.4	0.4	126.9	-0.1
	N3-C4-C5	114.6	114.5	-0.1	115.2	0.6
	C4-C5-C6	119.7	119.5	-0.2	119.6	-0.1
	C5-C6-N1	122.7	122.8	0.1	123.5	0.8
	N1-C2-O2	122.8	123.0	0.2	125.1	2.3
	N3-C2-O2	122.2	122.2	0.0	120.1	-2.1
	N3-C4-O4	119.4	119.9	0.5	119.9	0.5
	C5-C4-O4	125.9	125.6	-0.3	124.9	-1.0
	C2-N1-C1	117.7	117.8	0.1	118.8	1.1
	C6-N1-C1	121.2	121.2	0.0	121.3	0.1
	RMSD			0.3		1.1

Thymine					
Bonds					
N1-C2	1.376	1.388	0.012	1.428	0.052
C2-N3	1.373	1.372	-0.001	1.388	0.015
N3-C4	1.382	1.380	-0.002	1.391	0.009
C4-C5	1.445	1.444	-0.001	1.456	0.011
C5-C6	1.339	1.339	0.000	1.349	0.010

Table 18 continued					
C6-N1	1.378	1.373	-0.005	1.388	0.010
C2-O2	1.220	1.226	0.006	1.228	0.008
C4-O4	1.228	1.229	0.001	1.228	0.000
C5-C5M	1.496	1.497	0.001	1.498	0.002
N1-C1	1.473	1.473	0.000	1.473	0.000
RMSD			0.005		0.018
Angles					
C6-N1-C2	121.3	121.2	-0.1	120.0	-1.3
N1-C2-N3	114.6	114.5	-0.1	114.8	0.2
C2-N3-C4	127.2	127.1	-0.1	127.0	-0.2
N3-C4-C5	115.2	115.3	0.1	115.3	0.1
C4-C5-C6	118.0	118.2	0.2	119.0	1.0
C5-C6-N1	123.7	123.6	-0.1	123.8	0.1
N1-C2-O2	123.1	123.0	-0.1	125.1	2.0
N3-C2-O2	122.3	122.5	0.2	120.1	-2.2
N3-C4-O4	119.9	119.6	-0.3	119.5	-0.4
C5-C4-O4	124.9	125.1	0.2	125.2	0.3
C4-C5-C5M	119.0	118.7	-0.3	117.9	-1.1
C6-C5-C5M	122.9	122.6	-0.3	123.0	0.1
C2-N1-C1	118.2	117.7	-0.5	118.8	0.6
C6-N1-C1	120.2	121.1	0.9	121.3	1.1
RMSD			0.3		1.0

Distances in Å and angles in degrees. C1 and C9 are the methyl carbon atoms for the pyrimidines and purines, respectively.

a) Crystal survey data are the mean values of Clowney et al., 1996.

Table 19) Angles for the hydrogens on the methylated bases from ab initio data and the CHARMM27 and CHARMM22 force fields.

	Ab initio	CHARMM27		CHARMM22	
Adenine		Calc.	Diff	Calc.	Diff
N3-C2-H2	116.1	115.3	-0.8	115.5	-0.6
N1-C2-H2	115.1	115.3	0.2	115.7	0.6
C6-N6-H61	118.1	117.5	-0.6	115.8	-2.3
H61-N6-H62	118.7	121.3	2.6	124.6	5.9
C6-N6-H62	119.5	121.2	1.7	119.6	0.1
N7-C8-H8	125.2	124.8	-0.4	126.5	1.3
N9-C8-H8	121.2	121.3	0.1	118.3	-2.9
RMSD			1.3		2.7
Guanine					
C2-N2-H21	113.8	123.6	9.8	121.5	7.7
H21-N2-H22	114.5	121.1	6.6	123.7	9.2
C2-N2-H22	118.0	115.4	-2.6	114.8	-3.2
C2-N1-H1	119.6	119.9	0.3	120.4	0.8
C6-N1-H1	113.9	114.9	1.0	114.1	0.2
N7-C8-H8	125.7	124.8	-0.9	126.6	0.9
N9-C8-H8	121.4	122.0	0.6	118.5	-2.9
RMSD			4.6		4.8
Cytosine					
C4-C5-H5	122.1	119.9	-2.2	119.6	-2.5
C6-C5-H5	122.3	122.4	0.1	122.5	0.2
C5-C6-H6	123.1	123.5	0.4	123.0	-0.1
N1-C6-H6	116.6	115.4	-1.2	115.1	-1.5
RMSD			1.3		1.5
Uracil					
C2-N3-H3	115.7	115.5	-0.2	116.3	0.6
C4-N3-H3	116.5	117.0	0.5	116.8	0.3
C4-C5-H5	118.3	118.2	-0.1	118.2	-0.1
C6-C5-H5	122.5	122.3	-0.2	122.2	-0.3
C5-C6-H6	122.7	122.6	-0.1	122.2	-0.5
N1-C6-H6	115.1	114.6	-0.5	114.2	-0.9
RMSD			0.3		0.5
Thymine					
C2-N3-H3	115.8	115.8	0.0	116.3	0.5
C4-N3-H3	116.5	117.0	0.5	116.7	0.2
C5-C6-H6	122.3	122.2	-0.1	122.1	-0.2
N1-C6-H6	114.7	114.2	-0.5	114.1	-0.6
RMSD			0.4		0.4

Angles in degrees. *Ab initio* data based on HF/6-31G(d) optimized geometries obtained as part of the present study.

Table 20) Vibrational data on adenine.

	CHARMM27			*Ab initio*		
	Frequency	Assignment	Frequency	Assignment		
1	180	tR6a'	65	167	tR6a'	62
		wC6N	21		pucR6	18
		bfly	16			
2	214	bfly	38	206	bfly	47
		tR6a	19		tR6a	30
					tNH2	15
3	279	dC6N	75	242	tNH2	86
4	288	tNH2	39	272	dC6N	51
		tR6a	20			
		wNH2	20			
5	345	wNH2	68	298	tR6a	46
					bfly	24
					tR5'	22
6	374	tNH2	43	492	wNH2	45
		tR6a	18		wH9	38
		bfly	17			
7	458	pucR6	40	498	wH9	48
		tR6a'	26		wNH2	21
		wN6	15			
8	468	dR6a'	19	512	dR6a	30
		dR6a	17		wNH2	27
9	501	wH9	65	518	dR6a'	58
		tR5	20			
10	531	dR6a'	24	557	tR6a'	31
		sN9C4	21		pucR6	21
					tR5'	20
11	559	sC5C6	24	602	dR5	28
		sC5N7	18		sC5C6	24
					dR6a	17
12	645	dR6a	31	653	tR5	91
		dR5	21			
		sN3C4	19			
13	652	tR5	66	694	wC6N	50
					tR5'	31
14	719	tR5'	35	702	sN3C4	21
		tR6a	19			
		wC6N	16			
15	801	pucR6	30	809	pucR6	49
		wH2	19		tR5'	24
		tR5'	17		wC6N	22
16	833	dR5'	24	882	dR6	46

		sC4C5	21		dR6a'	17
		sC8N9	16			
17	847	wH8	105	903	wH8	102

Table 20 continued

18	858	dR6a'	23	922	dR5'	69
		dR6	21			
		dR5'	17			
19	967	rNH2	38	1007	rNH2	43
					sC6N1	29
20	977	wC2H	88	1009	wC2H	107
		tR6	16			
21	993	sC8N9	35	1055	sC8N9	61
		dN9H	18		dN9H	26
22	1028	sR5*	30	1121	sN9C4	17
					dR5	15
23	1094	sC2N3	36	1213	sN1C2	29
					sC5N7	20
24	1140	dC2H	25	1232	rNH2	21
		sN1C2	21		dC8H	17
25	1194	dC8H	26	1272	sC2N3	42
		dC2H	18		dC8H	22
		sN7C8	17			
26	1236	sN3C4	22	1332	dC2H	27
		sC5N7	19		sN1C2	20
27	1320	dN9H	21	1347	dC8H	21
		dR6	19		sC5N7	20
28	1412	sR5*	27	1408	dN9H	38
		sR6*	19		dC2H	27
					sC8N9	18
29	1469	sC6N1	21	1418	sN9C4	29
					sC4C5	23
30	1552	dC8H	22	1489	sC6N1	23
					dC2H	17
					sC6N6	17
31	1585	scNH2	18	1549	sN7C8	57
		dC8H	17			
32	1634	scNH2	52	1612	scNH2	47
33	1657	sR6*	34	1638	sC5C6	21
		sR5*	20			
34	1697	sR6*	28	1642	scNH2	30
					sC4C5	17
35	3120	sC2H	99	3050	sC2H	100
36	3121	sC8H	99	3103	sC8H	99
37	3445	sNH2s	99	3455	sNH2s	100
38	3455	sN9H	99	3518	sN9H	100
39	3563	sNH2a	100	3571	sNH2a	100

Symbols represent; s, stretching modes; d, bends; w, out-of-plane deformations (wags); x, torsional deformations; r, rocking modes, t, torsional modes, tw, twisting modes, and sc, scissor modes. R represents ring modes for the 5-membered (R5) and 6-membered rings (R6). Only internal coordinates contributing 15% or more to the potential energy distribution are reported. *Modes 22, 28, 33 and 34 are dominated by ring stretches, however, there are individual contributions of 15 % or more. Instead the sum of the 5-membered (sR5) and 6-membered ring stretches (sR6) are presented.

Table 21) Vibrational data on guanine.

	CHARMM27			*Ab initio*		
	Frequency	Assignment	Frequency	Assignment		
1	133	tR6a'	53	137	tR6a'	77
		bfly	29			
2	191	bfly	27	162	pucR6	51
		tR5	24		bfly	22
		tR6a'	23			
		tR6a	19			
3	259	gC2N	33	196	tR6a	75
		tNH2	25		bfly	26
		tR6a'	19			
		tR6a	17			
4	303	tNH2	51	306	dC6O	17
		tR6a	22		dR6a'	17
5	304	dC6O	50	320	tNH2	46
		dC2N	22		wNH2	37
6	333	dC2N	24	330	dC2N	47
					dC6O	24
7	395	bfly	29	353	tR5	27
		tR6a	17		bfly	19
		tR5	15			
8	454	dR6a	29	470	dR6a	65
9	463	wNH2	91	516	dR6a'	38
		tNH2	-19		gN9H	20
10	517	gN9H	64	519	gN9H	79
		tR5'	23			
11	547	dR6a	25	547	wNH2	48
					tNH2	32
12	559	gN1H	88	589	gN1H	77
		tNH2	17			
13	590	dR6a'	36	611	sC5C6	15
		sC4N9	33			
14	651	dC2N	18	650	tR5'	75
		dC6O	16			
15	673	gC2N	34	657	tR5'	21
		tR5'	15		dC6O	18
					dC2N	15
16	684	tR5'	48	711	tR5	37
		gN9H	16		gC6O	20
17	736	dR5'	17	741	gC2N	46
		sC5N7	16		gC6O	28
		sC2N	15			

18	745	tR5	38	784	gC6O	38
		pucR6	38		pucR6	30
					tR5	18
19	767	gC6O	110	817	dR6’	29
					sC5N7	17
					dR5'	16

Table 21 continued

20	843	dR5	49	872	gC8H	102
		sC8N9	16			
21	876	gC8H	106	930	dR5	80
22	898	sC2N3	36	1032	sN1C2	38
		dR6t	17			
		sN1C2	15			
23	980	rNH2	60	1046	sC8N9	59
					dN9H	27
24	1004	sC8N9	29	1073	sN1C6	33
		dN9H	22		rNH2	21
25	1045	sN7C8	27	1127	rNH2	36
		sN1C6	17		sC2N3	16
26	1164	sC5N7	18	1154	sC5N7	19
		dN9H	16		dC8H	17
		sN1C6	16			
		sN3C4	16			
27	1221	dC8H	32	1283	dC8H	33
		dR5'	15		sC5N7	17
28	1272	dN9H	21	1312	dN1H	22
		sN1C6	16		sC5C6	15
		sN3C4	15	1330	dN1H	19
29	1332	dN1H	29			
		sC2N	24			
30	1420	dN9H	24	1382	dN9H	38
		sC5N7	16		sC8N9	21
31	1484	sC2N3	22	1419	sC4C5	28
					sC4N9	18
32	1559	dC8H	37	1518	dN1H	19
		sN7C8	20		sN7C8	17
33	1592	sN1C2	19	1563	sN7C8	46
		scNH2	17			
		dN1H	17			
34	1614	sC4C5	31	1603	sN3C4	21
		dN1H	16		sC4C5	18
35	1638	scNH2	39	1612	scNH2	35
					sC2N3	27
36	1671	scNH2	26	1658	scNH2	52
					sC2N3	20
37	1830	sCO	36	1799	sCO	76
		dR6a	16			
38	3120	sC8H	99	3107	sC8H	99
39	3440	sNH2	63	3433	sNH2	99

		sN1H	36			
40	3448	sN1H	62	3464	sN1H	99
41	3453	sN9H	97	3518	sN9H	100
		sNH2	35			
42	3562	sNH2a	99	3535	sNH2a	100

See legend of Table 20 for definitions.

Table 22) Vibrational data on cytosine.

	CHARMM27			*Ab initio*		
	Frequency	Assignment	Frequency	Assignment		
1	162	tRa	50	142	tRa	95
		pucR	25			
2	193	tRa	40	199	pucR	37
		gC4N	33		tRa'	36
3	309	tNH2	50	225	tNH2	88
		wNH2	39			
4	350	dC4N	59	352	dC4N	61
5	424	tRa'	52	389	tRa'	62
		tNH2	19		pucR	19
6	488	tNH2	29	519	wNH2	54
		wNH2	25		dRa	15
		tRa'	25			
		pucR	23			
7	529	dCO	22	522	dCO	32
					dRa	26
8	544	gN1H	67	532	dRa	36
					wNH2	20
					dCO	15
9	568	dRa'	29	563	dRa'	75
		sC4N4	17			
10	607	dRa	38	599	gN1H	82
		dCO	20			
11	706	gC5H	22	721	gC5H	32
		gC4N	21		pucR	23
					gC4N	19
12	739	sC4C5	28	749	sC4C5	23
		dRa'	20		sN1C2	20
		dRa	18			
		sN1C2	17			
13	772	gC2O	41	769	gC5H	49
		gN1H	27		gC4N	44
		gC6H	26			
		gC5H	18			
14	859	gC4N	27	794	gC2O	81
		gC2O	26		pucR	18
		pucR	19			
15	874	sC2N3	36	915	sN1C2	30
		sN1C2	19			
16	953	sC6N1	42	966	dR	53
		dR	24		sC4C5	25

17	983	gC6H	65	998	gC6H	98
		gC5H	53			
18	1026	rNH2	62	1096	rNH2	27
		sN3C4	15		sC6N1	18

Table 22 continued

19	1034	dR	29	1103	dC5H	26
					rNH2	25
20	1168	dC6H	38	1185	dC6H	29
		sC5C6	27		sC6N1	20
		dC5H	22		dC5H	16
21	1299	dC5H	24	1249	sC2N3	44
		sC4N4	20			
22	1417	dN1H	56	1333	sC4N4	21
					dC6H	19
					dC5H	17
23	1503	dC6H	34	1425	dN1H	41
		dC5H	24			
24	1568	sN3C4	29	1475	dC5H	19
		sN1C2	16		dC6H	18
					sN3C4	15
					sC4N4	15
25	1572	sC2N3	18			
				1560	sN3C4	26
26	1636	dNH2	72			
				1626	scNH2	83
27	1725	sC5C6	27	1672	sC5C6	37
					sN3C4	18
28	1753	sC2O	41	1785	sC2O	77
29	2996	sC5H	83	3061	sC6H	69
		sC6H	16		sC5H	31
30	2997	sC6H	83	3080	sC5H	69
		sC5H	16		sC6H	31
31	3444	sNH2	99	3455	sNH2	99
32	3458	sN1H	99	3492	sN1H	100
33	3563	sNH2a	99	3571	sNH2a	99

See legend of Table 20 for definitions.

Table 23) Vibrational data on uracil.

	CHARMM27			*Ab initio*		
	Frequency	Assignment	Frequency	Assignment		
1	151	pucR	46	150	pucR	54
		tRa	40		tRa	48
2	181	tRa'	95	161	tRa'	107
3	375	dC2O	31	383	dC4O	36
		dC4O	26		dC2O	32
		sN3C4	16			
		sC2N3	16			
4	402	tRa	64	386	tRa	66
		pucR	16		pucR	21
5	521	pucR	32	505	dRa	76
		gN1H	30			
		gC5H	20			
6	523	sN1C2	15	528	gN1H	91
		dC2O	15			
		dC4O	15			
7	580	gN1H	51	533	dC2O	29
					dRa'	25
					dC4O	19
8	593	dRa	24	548	dRa'	42
					dC4O	16
9	605	dRa	35	659	gN3H	90
		dRa'	19		pucR	17
		dC4O	17			
10	640	gN3H	110	723	gC4O	34
					gC6H	22
					pucR	19
11	712	gC5H	47	746	sC4C5	32
		gC4O	37		sN1C2	17
12	757	sC4C5	30	776	gC2O	92
		sN1C2	29			
		dRa'	18	815	gC6H	52
13	783	gC2O	70		gC4O	47
		pucR	24	950	sN1C2	23
14	843	sC2N3	26		sC4C5	20
		sN3C4	18			
15	949	dR	41	970	dR	76
		sC6N1	34			
16	1010	dR	27	1000	gC5H	90
		dC6H	17		gC6H	26
17	1010	gC6H	72	1061	sC6N1	36

		gC4O	32		dC6H	24
18	1112	dC5H	48	1182	dC5H	37
		dC6H	26		dC6H	15
		sC5C6	15			

Table 23 continued

19	1311	dC6H	29	1215	sN3C4	24
		dN1H	19		dC6H	23
		dC5H	19			
20	1383	dN3H	60	1381	dC5H	29
					dC6H	20
21	1397	dN1H	37	1402	dN1H	20
		dC2O	16			
22	1449	sN1C2	19	1407	dN3H	57
		sC2N3	19			
23	1527	dRa'	15	1485	dN1H	39
					sC6N1	22
24	1600	sC5C6	49	1663	sC5C6	62
25	1795	sC4O	46	1795	sC4O	56
					sC2O	18
26	1925	sC2O	67	1811	sC2O	55
					sC4O	23
27	2996	sC6H	89	3074	sC5H	90
28	2999	sC5H	89	3100	sC6H	89
29	3454	sN3H	55	3466	sN3H	100
		sN1H	45			
30	3462	sN1H	55	3499	sN1H	100
		sN3H	45			

See legend of Table 20 for definitions.

Table 24) Vibrational data on thymine.

	CHARMM27			*Ab initio*		
	Frequency	Assignment	Frequency	Assignment		
1	114	tRa'	85	106	tRa'	104
2	147	tCH3	61	150	pucR	50
		pucR	21		tRa	37
3	149	tCH3	35	154	tCH3	82
		tRa	30			
		pucR	27			
4	284	dC5-Me	76	267	dC5-Me	73
5	301	gC5-Me	57	289	gC5-Me	77
		pucR	21		pucR	16
6	377	dC4O	24	385	dC2O	36
		dC2O	23		dC4O	28
7	389	tRa	37	386	tRa	64
		gC6H	18		pucR	21
8	476	dR	44	445	dR	73
		sC6H	18			
9	488	gN1H	49	521	gN1H	91
10	577	dC2O	26	536	dR	58
		dC4O	18			
		sN1C2	16			
11	607	dR	41	593	dC2O	29
					dC4O	23
12	613	gC6H	56	659	gN3H	90
		gN1H	30		pucR	16
13	674	gN3H	100	705	sC4C5	39
		gC4O	22		sC6H	15
		pucR	-15			
14	725	gC2O	31	764	gC4O	47
		gC4O	25		gC2O	46
15	742	dR	28	778	gC2O	55
		sC6H	18		gC4O	33
16	769	sN1C2	28	785	dR	48
		dR	27		sC6H	20
		sC4C5	16			
17	804	gC2O	40	936	gC6H	95
		pucR	30			
		gC4O	21			
18	860	sC2N3	24	951	sN1C2	25
		dR	19			
19	1001	dCH3	51	1008	dCH3	54

		dCH3a	32		dR	16
20	1010	dCH3a	58	1062	dCH3a	84
		dCH3	33			
21	1025	sC6N1	46	1132	sC6N1	25
		dC6H	20		sN3C4	17

Table 24 continued

22	1174	dR	25	1170	sC6H	25
		sC6H	24		sC6N1	25
		dC6H	19			
		sN3C4	17			
23	1223	dC6H	38	1216	dC6H	21
					sC2N3	18
24	1384	dN3H	45	1364	dC6H	47
		dC2O	17			
25	1402	dN1H	34	1403	dN3H	58
26	1408	dCH3	75	1413	sC2N3	21
27	1410	dCH3a	58	1418	dCH3	83
		dCH3	32			
28	1433	dCH3	68	1457	dCH3a	100
		dCH3a	29			
29	1458	dN1H	21	1476	dCH3	86
		sN1C2	19			
		sC2N3	15			
30	1571	dR	27	1488	dN1H	32
		sC4C5	22			
31	1665	sC5C6	37	1693	sC5C6	63
		sC4O	25			
32	1822	sC4O	32	1784	sC4O	73
		sC5C6	22			
		dR	18			
33	1928	sC2O	63	1806	sC2O	67
34	2904	sCH3	100	2894	sCH3	100
35	2958	sCH3a	96	2953	sCH3a	100
36	2959	sCH3	96	2957	sCH3	100
37	3000	sC5-Me	98	3068	sC5-Me	99
38	3453	sN3H	51	3467	sN3H	100
		sN1H	48			
39	3461	sN1H	51	3501	sN1H	100
		sN3H	48			

See legend of Table 20 for definitions.

Table 25) Relative energies and e and z values associated with the B_I and B_{II} conformations computed using model compound D.

Method	B_I		B_{II}		$\Delta E_{BII - BI}$
	ε	ζ	ε	ζ	
HF/6-31+G*	200.9	270.0	262.7	173.1	0.68
MP2/6-31+G*	194.4	274.2	267.4	161.4	1.55
C27	188.2	261.2	259.0	176.0	0.46

Dihedrals in degrees and energies in kcal/mole where $\Delta E_{BII - BI}$ is the total energy of the B_{II} conformer minus that of the B_I conformer. All minimizations performed at the stated level of theory with the furanose allowed to optimize in the south pucker.

Table 26) Deoxyribose pseudorotation angles and energetics: Comparison between *ab initio* and the CHARMM27 and CHARMM22 force fields.

Pseudorotation angles	P_n			P_S		
	a.i.	C27	C22	a.i.	C27	C22
Adenine	7.0	-3.1	-7.1	168.3	164.4	144.0
Guanine	9.6	-0.5	-5.0	168.6	165.2	144.1
Cytosine	8.8	-0.6	-1.2	162.1	162.2	149.9
Thymine	12.4	0.9	0.6	162.7	161.0	156.8
Energetics		ΔE_{N-S}			B	
	a.i.	C27	C22	a.i.	C27	C22
Adenine	0.4	0.6	-2.9	4.2	2.6	3.9
Guanine	0.7	1.0	-3.8	4.3	2.9[a]	4.6[a]
Cytosine	-0.3	-0.2	0.2	4.0	1.9	1.9
Thymine	0.9	0.2	-0.2	4.0	2.2	2.2

Pseudorotation angles (deg.) P_n and P_S correspond, respectively, to the north and south energy minima. ΔE_{N-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.

Table 27) Sugar amplitudes and glycosyl torsions in nucleosides: comparison between *ab initio* and the CHARMM27 and CHARMM22 force fields.

Base	North			South			East		
Amplitude	a.i.	C27	C22	a.i.	C27	C22	a.i.	C27	C22
Ade	39.7	38.1	42.2	36.7	31.4	38.0	14.3	34.8	27.8
Gua	39.4	38.1	42.1	36.7	31.2	37.3	14.4	34.4	28.4
Cyt	39.4	38.8	41.3	37.4	31.8	37.2	19.8	34.8	28.1
Thy	39.3	38.7	40.6	37.7	32.3	35.6	20.5	35.1	28.4
Glycosyl torsion	a.i.	C27	C22	a.i.	C27	C22	a.i.	C27	C22
Ade	192	192	186	230	225	202	220	202	201
Gua	198	204	190	233	235	205	222	214	200
Cyt	195	194	195	207	207	222	201	194	215
Thy	198	200	200	231	225	227	224	210	218

Amplitudes and glycosyl torsions in degrees. North, south and east refer to the north energy minimum, the south energy minimum and the east energy barrier of the pseudorotation cycle, respectively. *Ab initio* calculations (a.i.) at the MP2/6-31G* level of theory.[47]

Figure Legends

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

Figure 2) A) Diagram of a DNA G-C basepair 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. B) 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.

Figure 3) Interaction orientations between model compounds and water used in the adjustment of the intermolecular portion of the force field. In each case the water-model compound complexes were studied individually (i.e. monohydrates) with the only optimized degree(s) of freedom being the represented distances and, in selected cases, the shown angle.

Figure 4) 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 (G), 2 (H) and 3 (F). 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 (G), 2 (H) and 3 (F). Note that change in the X-axis upon going from A (0 to 360) to B (0 to 120).

Figure 5) Potential energies for the C3'endo (A) and C2'endo (B) furanose puckers and probability distributions for A form (C) and B form (D) DNA as a function of γ. The potential energy surfaces (A and B) were obtained with model compound B at the QM MP2/6-31+G* (thin line) level of theory and for the CHARMM27 parameter set (bold line). Probability distributions are from the NDB survey for A form (C, thin line) and B form (D, thin line) crystal structures and from the GTACGTAC A form crystal (C, bold line) and CGATCGATCG B form crystal (D, bold line) simulations.

Figure 6) Potential energies for model compound A and probability distributions for A form (B) and B form (C) DNA as a function of α. The potential energy surfaces (A) were obtained at the QM MP2/6-31G* level of theory in the absence (thin line) and presence (1) of a single water molecule and with the CHARMM27 parameter set (bold line). In the surfaces ζ remained in the gauche orientation.

Probability distributions are from the NDB survey for A form (B, thin line) and B form (C, thin line) crystal structures and from the GTACGTAC A form crystal (B, bold line) and CGATCGATCG B form crystal (C, bold line) simulations.

Figure 7) Potential energies for model compound A and probability distributions for A form (B) and B form (C) DNA as a function of ζ. The potential energy surfaces (A) were obtained with model compound G at the QM MP2/6-31G* level of theory in the absence (thin line) and presence (1) of a single water molecule and with the final empirical parameter set (bold line). In the surfaces α remained in the gauche orientation. Note that the potential energy surfaces are identical to Figure 5. Probability distributions are from the NDB survey for A form (B, thin line) and B form (C, thin line) crystal structures and from the GTACGTAC A form crystal (B, bold line) and CGATCGATCG B form crystal (C, bold line) simulations.

Figure 8) Potential energies for the C3'endo (A) and C2'endo (B) furanose puckers and probability distributions for A form (C) and B form (D) DNA as a function of β. The potential energy surfaces (A and B) were obtained with model compound B at the QM MP2/6-31+G* (thin line) level of theory and for the CHARMM27 parameter set (bold line). Probability distributions are from the NDB survey for A form (C, thin line) and B form (D, thin line) crystal structures and from the GTACGTAC A form crystal (C, bold line) and CGATCGATCG B form crystal (D, bold line) simulations.

Figure 9) Potential energies for the C3'endo (A) and C2'endo (B) sugar puckers and probability distributions for A form (C) and B form (D) DNA as a function of ε. The potential energy surfaces (A and B) were obtained with model compound C at the MP2/6-31+G* (thin line) level of theory and for the CHARMM27 parameter set (bold line). Probability distributions are from the NDB survey for A form (C, thin line) and B form (D, thin line) crystal structures and from the GTACGTAC A form crystal (C, bold line) and CGATCGATCG B form crystal (D, bold line) simulations.

Figure 10) Potential energies as a function of χ for the C3'endo (A) and C2'endo (B) furanose puckers and probability distributions for A form (C) and B form (D) DNA. The potential energy surfaces (A and B) were obtained with model compound E with a cytosine base at the MP2/6-31G* (thin line) level of theory and for the CHARMM27 parameter set (bold line). Probability distributions are from the NDB survey for A form (C, thin line) and B form (D, thin line) crystal structures and from the GTACGTAC A form crystal (C, bold line) and CGATCGATCG B form crystal (D, bold line) simulations.

Figure 11) Potential energies as a function of χ for the C3'endo (A) and C2'endo (B) furanose puckers of model compound E with a thymine base at the MP2/6-31G* (thin line) level of theory and for the CHARMM27 parameter set (bold line).

Figure 12) Potential energies as a function of χ for the C3'endo (A) and C2'endo (B) furanose puckers of model compound E with an adenine base at the MP2/6-31G* (thin line) level of theory and for the CHARMM27 parameter set (bold line).

Figure 13) Potential energies as a function of χ for the C3'endo (A) and C2'endo (B) furanose puckers of model compound E with a guanine base at the MP2/6-31G* (thin line) level of theory and for the CHARMM27 parameter set (bold line).

Figure 14) Potential energies as a function of the pseudorotation angle for model compounds F (A) and G with an imidazole base (B) and probability distributions for A form (C) and B form (D) DNA. The potential energy surfaces were obtained at the QM HF/6-31+G* (thin line) and MP2/6-31+G* (1) levels in A and the MP2/6-31G* level in B (1) and for the final empirical parameter set in both A and B(J). For the empirical energy surface in B β was constrained to 180 . Probability distributions are from the NDB survey for A form (C, thin line) and B form (D, thin line) crystal structures and from the GTACGTAC A form crystal (C, bold line) and CGATCGATCG B form crystal (D, bold line) simulations.

Figure 15) Probability distributions as a function of δ for A form (A) and B form (B) DNA. Probability distributions are presented from the NDB survey for A form (A, thin line) and B form (B, thin line) crystal structures and from the GTACGTAC A form crystal (A, bold line) of the CGATCGATCG B form crystal (B, bold line) simulations.

Figure 16) Potential energies as a function of the H-O2'-C2'-C3' dihedral for model compound C^{2OH} at the QM HF/6-31+G* level (1) and for the CHARMM27 parameter set (-J-). The surfaces were obtained with the furanose constrained to the C3'endo pucker, α constrained to 200 and ε constrained to 180 .

Figure 17) Potential energies as a function of the pseudorotation angle for model compound F^{2OH} (A) and model compound G^{2OH} with a imidazole base (B) and probability distributions for RNA (C). The potential energy surfaces were obtained at the QM MP2/6-31G* level (1) and for CHARMM27 (J) in both A and B. Probability distributions are presented from the NDB survey for RNA duplex and tRNA

crystal structures (C, thin line) and from the r(UAAGGAGGUGUA) RNA dodecamer solution simulation (bold line).

Figure 18) Probability distributions of dihedral angles α (A), β (B), γ (C), δ (D), ε (E), ζ (F) and χ (G) from the NDB survey for RNA duplex and tRNA crystal structures (thin lines) and from the r(UAAGGAGGUGUA) RNA dodecamer solution simulation (bold line).

Figure 19) 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).