Optimization of a Nucleic Acids united-RESidue 2-Point model (NARES-2P) with a maximum-likelihood approach

Yi He,¹ Adam Liwo,² and Harold A. Scheraga^{1,a)}

¹*Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA*

²*Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland*

(Received 16 August 2015; accepted 18 September 2015; published online 13 October 2015)

Coarse-grained models are useful tools to investigate the structural and thermodynamic properties of biomolecules. They are obtained by merging several atoms into one interaction site. Such simplified models try to capture as much as possible information of the original biomolecular system in all-atom representation but the resulting parameters of these coarse-grained force fields still need further optimization. In this paper, a force field optimization method, which is based on maximum-likelihood fitting of the simulated to the experimental conformational ensembles and least-squares fitting of the simulated to the experimental heat-capacity curves, is applied to optimize the Nucleic Acid united-RESidue 2-point (NARES-2P) model for coarse-grained simulations of nucleic acids recently developed in our laboratory. The optimized NARES-2P force field reproduces the structural and thermodynamic data of small DNA molecules much better than the original force field. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4932082>]

I. INTRODUCTION

Simulating the protein-folding process has proved to be a useful compliment to experimental studies of proteins.^{1–5} Compared to the earlier work on constructing a force field to predict protein structures,^{6,7} the computer simulation of protein-folding has been changed by the invention of hardware-customized supercomputers such as Anton.⁸ However, even with Anton,⁸ there are still limitations of the size of the biomolecules studied and the number of trajectories that can be simulated within a limited amount of time. On the other hand, nucleic acids (e.g., DNAs), as the key parts of a human biological system, are large molecules. These molecules would be too large to simulate using an all-atom force field, even with Anton.⁸ Therefore, coarse-grained force fields for nucleic acids^{9–20} are playing an important role in simulating nucleic acids in recent years.

Recently, several coarse-grained models^{9–20} were designed to simulate DNA and/or RNA folding or hybridization. Among all these models, oxDNA which was recently developed by Ouldridge and co-workers,^{11,12} can reproduce correct dependence of base-base interactions on their orientations. It has been proven to be a useful tool to simulate DNA hybridization from separated strands, and to simulate the operations of DNA tweezers in unrestricted Monte Carlo runs. However, in this model, all long-range interactions, including base-pairing and stacking, are described by designed Morse-type potentials, and other potentials are engineered to reproduce base-base interactions. Moreover, although different methods were used to simplify all-atom representation to a coarse-grained representation, in general, further optimiza-

tions/improvements of a force field are needed to produce a coarse-grained force field capable of reproducing both structural and thermodynamic properties of DNA and/or RNA molecules.

The Nucleic Acid united-RESidue 2-Point model (NARES-2P)¹⁵ was designed based on the idea that aimed at developing a minimal physics-based coarse-grained force field for nucleic acids. NARES-2P was constructed using the same bottom-up strategy as the UNited-RESidue (UNRES) force field^{21–23} for proteins. Each component of the NARES-2P energy function was fitted separately to the respective potential of mean force obtained from all-atom potential-energy surfaces, and the resulting force field can reproduce the structures of all tested double-helix DNAs as well as helix-form RNAs at medium-resolution [4–10 Å Root-Mean-Square Deviation (RMSD) from experimental structures, depending on size]. Although the NARES-2P force field has decent performance to reproduce native structures, the simulated thermodynamics are in poor agreement with the experimental data.

In this paper, different from the earlier optimization strategies,^{24–28} a completely new optimization strategy,²⁹ based on the maximum-likelihood principle, was applied to optimize the performance of the NARES-2P force field. This strategy was originally developed to calibrate the UNRES force field for proteins^{21–23} and proved to be much more stable numerically than the existing approaches such as energy-gap maximization, Z-score optimization, and hierarchical optimization.^{24–28} It has also produced a well-transferable force field despite the fact that only a small α -helical protein was used for calibration.²⁹ As opposed to energy-function optimization methods applied so far (such as energy-gap or Z-score optimization), multiple reference structures of several different systems can be considered, consistent with the fact that the

^{a)} Author to whom correspondence should be addressed. Electronic mail: has5@cornell.edu

“native structure” is an ensemble of conformations rather than a single “native” conformation. The complete target function in force-field calibration consists of the maximum-likelihood component and the sum of the squares of the differences between the computed and experimental heat capacity profile at selected temperatures (the χ^2 component). The aim of introducing the χ^2 component is to enhance the capability of the NARES-2P force field to reproduce the experimental thermodynamic data. By using this new maximum-likelihood-optimization strategy, the performance of the NARES-2P force field was significantly improved with regard to thermodynamic quantities, as shown in Sec. III.

II. METHOD

A. NARES-2P force field

In the NARES-2P model,¹⁵ depicted in Figure 1, a polynucleotide chain is represented by a sequence of virtual sugar (S) atoms that are located at the geometric centers of the sugar rings and linked by virtual bonds with attached united sugar-base (B) and united phosphate groups (P). Only the united sugar base (B) and the united phosphate (P) groups are interaction sites, while the sugar (S) centers serve only as anchor points. The energy function of the virtual-bond chain is expressed by Eq. (1),

$$U = w_{BB}^{GB} \sum_{j<i} U_{B_i B_j}^{GB} + w_{BB}^{dip} f_2(T) \sum_{j<i} U_{B_i B_j}^{dip} + w_{PP} \sum_{j<i} U_{P_i P_j} + w_{PB} \sum_{j<i} U_{P_i B_j} + w_{bond} \sum_i U_{bond}(d_i) + w_{ang} \sum_i U_{ang}(\theta_i) + w_{tor} f_2(T) \sum_i U_{tor}(\gamma_i) + w_{rot} \sum_i U_{rot}(\alpha_i, \chi_i) + U_{rest}, \quad (1)$$

in which

$$f_n(T) = \frac{\exp(1) + \exp(-1)}{\exp(\frac{nT}{T_0}) + \exp(-\frac{nT}{T_0})}. \quad (2)$$

In Equation (1), $U_{B_i B_j}^{GB}$ and $U_{B_i B_j}^{dip}$ correspond to the average van der Waals interaction energy (expressed as the anisotropic Gay-Berne potential) and electrostatic-interaction energy between the sugar-base groups, respectively. It should be noted that $U_{B_i B_j}^{dip}$ corresponds to mean-field interactions between nucleic-acid-base dipoles. $U_{P_i P_j}$ and $U_{P_i B_j}$ are the phosphate-phosphate groups interaction and phosphate group-base group interaction energies. $U_{bond}(d_i)$, $U_{ang}(\theta_i)$, $U_{tor}(\gamma_i)$, and $U_{rot}(\alpha_i, \beta_i)$ represent the energies of virtual-bond stretching, virtual-bond angle bending, rotation about backbone virtual bonds, and sugar-base rotamers, respectively. The $U_{tor}(\gamma_i)$ terms account for the correct right-handed chirality of the strands. The factors $f_n(T)$ account for the temperature dependence of the terms which have higher order than one in the generalized-cumulant expansion.²³ The term U_{rest} restricts the distance between the 5' end of one chain and the 3' end of the second chain to distances less than d_{max} (selected

to correspond to the desired monomer concentration). The w 's are energy-term weights. In our previous work,¹⁵ all the weights were set at 2.000, except w_{BB}^{GB} which was set at 0.500, to provide melting-transition temperatures comparable to experimental values. The NARES-2P model has already been built into the UNRES/MD platform,³⁰ which enables the canonical and multiplex replica-exchange simulations of nucleic acids to be carried out.

Because the parameters of the $U_{B_i B_j}^{GB}$ and $U_{B_i B_j}^{dip}$ potentials were optimized, the expressions for these energy terms are given below. The reader is referred to our earlier work for the expressions for the remaining energy terms.¹⁵ The $U_{B_i B_j}^{GB}$ term is expressed by Eq. (3),

$$U_{B_i B_j}^{GB} = 4 \left(|\varepsilon_{ij}| x_{ij}^{12} - \varepsilon_{ij} x_{ij}^6 \right), \quad (3)$$

$$\varepsilon_{ij} = \varepsilon_{ij}^{(0)} \varepsilon_{ij}^{(1)} \varepsilon_{ij}^{(2)}, \quad (4)$$

in which $\varepsilon_{ij}^{(0)}$ is a parameter corresponding to the van der Waals well depth for parallel orientation of the united bases. The functional forms of $\varepsilon_{ij}^{(1)}$ and $\varepsilon_{ij}^{(2)}$ are given by Eqs. (5) and (6),

$$\varepsilon_{ij}^{(1)} = [1 - \chi_{ij}^{(1)} \chi_{ij}^{(2)} \omega_{ij}^{(12)2}]^{-\frac{1}{2}}, \quad (5)$$

$$\varepsilon_{ij}^{(2)} = \left[1 - \frac{\chi_{ij}^{(1)'} \omega_{ij}^{(1)2} + \chi_{ij}^{(2)'} \omega_{ij}^{(2)2} - 2 \chi_{ij}^{(1)'} \chi_{ij}^{(2)'} \omega_{ij}^{(1)} \omega_{ij}^{(2)} \omega_{ij}^{(12)}}{1 - \chi_{ij}^{(1)'} \chi_{ij}^{(2)'} \omega_{ij}^{(12)2}} \right]^2. \quad (6)$$

The reduced distance x_{ij} is defined by Eq. (7),

$$x_{ij} = \frac{\sigma_{ij}^0}{r_{ij} - \sigma_{ij} + \sigma_{ij}^0}, \quad (7)$$

with

$$\sigma_{ij} = \sigma_{ij}^0 \left[1 - \frac{\chi_{ij}^{(1)} \omega_{ij}^{(1)2} + \chi_{ij}^{(2)} \omega_{ij}^{(2)2} - 2 \chi_{ij}^{(1)} \chi_{ij}^{(2)} \omega_{ij}^{(1)} \omega_{ij}^{(2)} \omega_{ij}^{(12)}}{1 - \chi_{ij}^{(1)} \chi_{ij}^{(2)} \omega_{ij}^{(12)2}} \right]^{-\frac{1}{2}}, \quad (8)$$

FIG. 1. The NARES-2P model of the nucleotide chain. The all-atom representation is superposed on the coarse-grained representation. The united sugar groups (S) which serve as geometric points are represented by red circles; the united phosphate groups (P) are the white circles. All P's are positioned halfway between two consecutive sugar atoms. Bases (B) are represented by ellipsoids (green circles), which are centered at the geometric centers of the B's, with dipoles located on the bases to represent their electrostatic interactions. The base-base electrostatic interactions represent the mean-field interactions of the base dipoles (red arrows) computed by integrating the Boltzmann factor over the rotation of the dipoles about the base axes (dotted lines). The backbone virtual-bond angles, θ , the virtual-bond dihedral angles, γ , are indicated. The base orientation angles α and torsional angles β , that define the location of a base with respect to the backbone are also indicated. Reprinted with permission from Y. He *et al.*, Phys. Rev. Lett. **110**, 098101 (2013). Copyright 2013 American Physical Society.

where r_{ij} is the distance between base centers, $\omega_{ij}^{(1)} = \hat{\mathbf{u}}_{ij}^{(1)} \cdot \hat{\mathbf{r}}_{ij}$, $\omega_{ij}^{(2)} = \hat{\mathbf{u}}_{ij}^{(2)} \cdot \hat{\mathbf{r}}_{ij}$, and $\omega_{ij}^{(12)} = \hat{\mathbf{u}}_{ij}^{(1)} \cdot \hat{\mathbf{u}}_{ij}^{(2)}$, where $\hat{\mathbf{u}}_{ij}^{(1)}$, $\hat{\mathbf{u}}_{ij}^{(2)}$, and $\hat{\mathbf{r}}_{ij}$ are the unit vectors of the long axis of the interacting base residues and the unit vector pointing from the center of base residue i to base residue j , respectively, σ_{ij}^0 is the cross-section of the united bases at parallel orientation, and the constants $\chi_{ij}^{(1)}$ and $\chi_{ij}^{(2)}$ are anisotropies.

The average base-base electrostatic-interaction energy is expressed by Eq. (9),

$$U_{B_i B_j}^{dip} = \frac{a_{ij}}{r_{ij}^3} W_{ij} + \frac{b_{ij}}{r_{ij}^6} [Y_{ij}^{(1)2} + Y_{ij}^{(2)2}] + \frac{d_{ij}}{r_{ij}^6} [Z_{ij}^{(1)2} + Z_{ij}^{(2)2}], \quad (9)$$

with

$$W_{ij} = \omega_{ij}^{(12)} - 3\omega_{ij}^{(1)}\omega_{ij}^{(2)}, \quad (10)$$

$$Y_{ij}^{(1)} = \sqrt{1 + 3\omega_{ij}^{(1)2} - W_{ij}^2}, \quad (11)$$

$$Y_{ij}^{(2)} = \sqrt{1 + 3\omega_{ij}^{(2)2} - W_{ij}^2}, \quad (12)$$

$$Z_{ij}^{(1)} = \sqrt{2 + 2\omega_{ij}^{(12)} + \frac{1}{2}W_{ij}^2 - \frac{3}{2}(\omega_{ij}^{(1)} + \omega_{ij}^{(2)})^2}, \quad (13)$$

$$Z_{ij}^{(2)} = \sqrt{2 - 2\omega_{ij}^{(12)} + \frac{1}{2}W_{ij}^2 - \frac{3}{2}(\omega_{ij}^{(1)} - \omega_{ij}^{(2)})^2}, \quad (14)$$

a_{ij} , b_{ij} , and d_{ij} are base-type-specific parameters for the potential energies; these parameters were optimized in the present work.

B. Target function

Maximum-likelihood estimation (MLE) is a basis for fitting models to experimental data.³¹ Recently, it has been applied widely to treat evolutionary data.³² However, application of the maximum-likelihood principle to fit the experimental conformational ensembles poses a severe problem because the calculated probability-distribution function can be obtained only in the form of simulated conformations with a probability assigned to each of them. This problem was solved in our recent work by expressing the approximate target function at each experimental point as a sum of contributions from all simulated points, each of them Gaussian-weighted depending on the distance from the simulated point.²⁹ The probability of each structure, generated using multiplexed replica-exchange molecular dynamics (MREMD), can be calculated using the Weighted Histogram Analysis Method (WHAM).³³ The maximum-likelihood part of the target function ($-\log L$) is thus expressed by Eq. (15),

$$-\log L = \sum_S w_S \sum_i \sum_j \rho(j)^{-1} \times \exp \left[-\frac{RMSD(i, j)^2}{2\sigma^2} \right] (\omega_j - U_j + \ln Q), \quad (15)$$

with

$$Q = \sum_j \exp(\omega_j - \beta U_j). \quad (16)$$

In Equation (15), S denotes the sequence of a training nucleic acid, w_S is the weight of the structural data of sequence S (usually 1), indices i and j number the residues in an experimental and calculated structure, respectively, $RMSD(i, j)$ is the root-mean-square deviation of the i th structure of the experimental ensemble corresponding to a given sequence from the j th structure of the calculated ensemble, $\rho(j)$ is the density of simulated structures in the neighborhood of the simulated structure j , U_j is the energy of the simulated structure j , ω_j is the probability factor of structure j resulting by solving the WHAM equations,³³ Q is the partition function calculated from simulation data of a given sequence, and $\beta = 1/RT$, where R is the universal gas constant and T is the absolute temperature. Minimization of $-\log L$ makes the calculated ensemble as similar to the experimental ensemble as the coarse-grained model under consideration criteria.

As mentioned, in order to include thermodynamic data, the sum of the squares of the differences of the experimental and calculated heat-capacity temperature profiles has been added to the target function; the temperature profiles of the fractions of paired bases, which can be obtained from UV measurements were also included. The complete target function is expressed by Eq. (17),

$$F = -\log L + \sum_S w_S \left\{ w_C \sum_T [C^{\text{exp}}(T) - C^{\text{calc}}(T)]^2 + w_f \sum_T [f^{\text{exp}}(T) - f^{\text{calc}}(T)]^2 \right\}, \quad (17)$$

where $\log L$ is defined by Eq. (15), w_C and w_f are the weights of the sum of square terms corresponding to the deviations of the experimental and calculated heat-capacity and fraction of paired bases profiles, respectively, $C(T)$ and $f(T)$ are the values of the heat capacity and the fraction of paired bases, respectively, at temperature T and superscripts *exp* and *calc* denote the experimental and calculated values, respectively. The calculation of these quantities from the results of the MREMD simulations with NARES-2P is described in our earlier work.¹⁵

C. Optimization procedure

The implementation of the target function can be outlined as the following iterative procedure.

1. Perform MREMD simulations for the selected DNA/RNA training sequence(s) with the initial energy-function parameters. The temperature range used in this work was from 250 K to 500 K.
2. Process the resulting ensemble with WHAM in order to obtain the probabilities of the calculated conformations at the temperature at which the structures were determined (usually by NMR) and to calculate the temperature profiles of the heat-capacity and fraction of paired bases profiles.
3. With the set of conformations obtained from the NARES-2P simulation, minimize F of Eq. (17) by varying the NARES-2P energy term weights and other parameters.
4. Iterate steps 1–3 until convergence.

III. RESULTS

Two DNA duplexes, namely, T1 and UD1,^{34,35} were selected to optimize the parameters of the NARES-2P force field. In order to perform an optimization, it is necessary to have both thermodynamic data and structural data. The thermodynamic data of both duplexes were obtained previously from experiments.^{34,35} However, the structures of both duplexes are not available. It is known that both of them are double stranded DNAs. Thus, a standard B-DNA structural form was used as a template to create “native” structures for these DNA sequences to use as reference structures for both duplexes. Given the medium resolution of NARES-2P and the fact that each short DNA molecule assumes a simple right-handed double-helix structure whose geometry is virtually independent of its sequence, use of these created structures instead of actual structures does not pose a problem.

Using these T1 and UD1 training duplexes, we optimized the energy-term weights in Eq. (1), and three parameters a_{ij} , b_{ij} , d_{ij} of $U_{B_i B_j}^{\text{dip}}$ in Eq. (9) except w_{PB} and w_{BB}^{dip} . The initial and optimized energy-term weights and the parameters of $U_{B_i B_j}^{\text{dip}}$ are summarized in Tables I and II, respectively.

It can be seen that most of the energy-term weights did not change much except for w_{PP} , which controls the phosphate–phosphate interaction. The significant decrease of the value of this energy-term weight suggests that this term was overemphasized in the original version of NARES-2P.¹⁵

The performance of the initial and optimized force field is compared in Figures 2–5. For T1, the representative structures of the largest cluster obtained from MREMD simulations at

TABLE I. Initial input and final output weight parameters of each energy term.

	w_{BB}^{GB}	w_{BB}^{dip}	w_{PP}	w_{PB}	w_{bond}	w_{ang}	w_{tor}	w_{rot}
Initial	0.500 00	2.000 00	2.000 00	2.000 000	2.000 00	2.000 00	2.000 00	2.000 00
Final	0.475 60	2.000 00	0.944 44	2.000 000	2.117 43	1.818 81	1.943 36	1.633 45

TABLE II. Initial input and final output of a_{ij} , b_{ij} , d_{ij} of $U_{B_i B_j}^{\text{dip}}$ in Equation (9).

Base pair	$a_{ij_initial}$	$b_{ij_initial}$	$d_{ij_initial}$	a_{ij_final}	b_{ij_final}	d_{ij_final}
A–A	4.75×10^{-01}	-1.00×10^{-03}	-1.06×10^{-01}	$3.41 \times 10^{+00}$	-7.00×10^{-05}	-7.00×10^{-05}
A–G	8.60×10^{-02}	-1.00×10^{-03}	-2.03×10^{-01}	3.17×10^{-01}	-8.98×10^{-01}	-7.00×10^{-05}
A–C	8.40×10^{-01}	-1.00×10^{-03}	-6.02×10^{-02}	$1.73 \times 10^{+00}$	-7.00×10^{-05}	-2.11×10^{-01}
A–T	$0.00 \times 10^{+00}$	-8.20×10^{-01}	$0.00 \times 10^{+00}$	2.50×10^{-01}	$-1.32 \times 10^{+00}$	-1.76×10^{-01}
G–G	$0.00 \times 10^{+00}$	-2.28×10^{-01}	-7.99×10^{-02}	$2.49 \times 10^{+00}$	-7.00×10^{-05}	-2.67×10^{-01}
G–C	$0.00 \times 10^{+00}$	$-2.20 \times 10^{+00}$	$0.00 \times 10^{+00}$	8.18×10^{-01}	$-1.73 \times 10^{+00}$	-7.00×10^{-05}
G–T	$0.00 \times 10^{+00}$	-7.00×10^{-03}	-2.26×10^{-01}	-1.02×10^{-01}	-8.62×10^{-01}	-4.75×10^{-01}
C–C	9.68×10^{-01}	-1.00×10^{-03}	$0.00 \times 10^{+00}$	2.24×10^{-01}	-7.00×10^{-05}	-4.95×10^{-01}
C–T	7.14×10^{-01}	-1.00×10^{-03}	$0.00 \times 10^{+00}$	$1.17 \times 10^{+00}$	-5.17×10^{-01}	-7.00×10^{-05}
T–T	5.25×10^{-01}	-1.00×10^{-03}	$0.00 \times 10^{+00}$	7.08×10^{-01}	-7.00×10^{-05}	-1.68×10^{-01}

FIG. 2. Side view of the superposition of the experimental (red) and calculated structure of the largest cluster corresponding to the initial ((a) yellow; RMSD = 4.15 Å) and final ((b) blue; RMSD = 4.04 Å) structure of the T1 duplex at $T = 300$ K.

300 K using the initial parameters are shown in Figure 2(a) and those corresponding to the final parameters are shown in Figure 2(b). The RMSD values of the structure shown in Figures 2(a) and 2(b) to the native-duplex model (generated at the all-atom level) are 4.15 Å and 4.04 Å, respectively. There is, therefore, only a slight improvement of the quality of the calculated “native-like” ensemble for this system. However, the calculated heat-capacity profile shows significant improvement of the agreement with the experimental curve^{34,35} with

FIG. 3. Comparison of the heat capacity curves calculated from MREMD simulations of the T1 duplex with the initial (blue) and optimized (red) NARES-2P parameters with the experimental curve (black).

FIG. 4. Side view of the superposition of the experimental (red) and calculated structure of the largest cluster corresponding to the initial ((a) yellow; RMSD = 3.51 Å) and final ((b) blue; RMSD = 3.64 Å) structure of the UD1 duplex at $T = 300$ K.

respect to that obtained with the initial parameters (Figure 3). A remarkable indication of the improvement is that, as opposed to the results obtained with the original version of NARES-2P,¹⁵ the calculated heat-capacity curve now has a single band (Figure 3), consistent with the experimental observation.^{34,35} Moreover, the melting temperature, reported from the heat capacity curve generated by the initial parameters, was 339 K which deviated about 6.2° from the experimental value, 332.8 K, while the difference is only 2° with the optimized parameters which predicted 331 K as the melting temperature.

Similar results were obtained for the UD1 duplex (it should be noted that optimization was run using the experimental data of the T1 and the UD1 system together) except that the average structure slightly deteriorated after optimization

FIG. 5. Comparison of the heat capacity curves calculated from MREMD simulations of the UD1 duplex with the initial (blue) and optimized (red) NARES-2P parameters with the experimental curve (black).

FIG. 6. Comparison of the fraction of paired-bases from MREMD simulations of the (a) T1 and (b) UD1 duplexes with the initial (blue) and optimized (red) NARES-2P parameters with the experimental curve (black).

(RMSD = 3.64 Å compared to 3.51 Å with initial parameters; Figure 4). However, the melting temperature differs only by 9° from the experimental value after optimization, compared to 55° before optimization and the heat-capacity profile calculated with the optimized parameters has only one peak, compared to two peaks with the initial parameters (Figure 5).

In addition to the temperature profile of the heat-capacity, the temperature profile of the fraction of paired bases was also optimized in the maximum-likelihood optimization [Eq. (17)]. As shown in Figure 6, the fraction of paired bases generated from optimized parameters was much smoother and closer to the experimental curve. The optimized parameters produce more stable base pairs at low temperatures, and clean dissociation of all base pairs at high temperatures. It can be seen from Figure 6(b) that the fraction of the paired-bases of UD1

generated by the initial parameters was not fully broken up at high temperatures.

The optimized force field was subsequently tested with 16 DNA duplexes, which were not used in optimization. The results are summarized in Table III. As shown in Table III, the melting temperatures obtained with the optimized parameters are closer to the experimental values. For 17 out of 18 systems (including T1 and UD1), the melting temperatures obtained using the optimized parameters are within 25° deviation to experimentally reported melting temperatures; and for 12 out of 18 systems the agreement with the experimental melting temperatures is improved compared to that obtained by using the initial parameters. The improvement is especially remarkable for T4 and UD3 whose melting temperatures calculated with the original version of NARES-2P deviated from the experimental counterparts by more than 44°.

TABLE III. Calculated and experimental melting temperatures and standard enthalpies and entropies of unfolding of the small double-stranded DNA molecules studied in Refs. 34 and 35 using the optimized parameters from maximum-likelihood optimization.

Sequence ^a	C_T^b (μM)	T_m^c (K)		$\Delta H^{\circ d}$ (kcal/mol)		$\Delta S^{\circ d}$ (cal/(mol K))	
		Calc.	Expt.	Calc.	Expt.	Calc.	Expt.
C1 CTACGCATTCC	75	333.0	332.9	55.9	84.5	170.8	232.3
C2 CTAACGGATGC	75	363.0	333.1	86.9	85.3	246.0	234.4
C3 CTATTGGCGAC	75	340.0	333.7	55.0	83.8	153.1	229.5
C4 CGTATTCAGGC	75	337.0	332.2	89.5	88.6	259.3	245.0
C5 CAATACGCCTC	75	347.0	332.4	82.3	77.7	234.7	212.2
T1 TTCATAGCCGT	75	331.0	332.8	74.0	80.4	220.9	219.9
T2 TTCCGTAGCAT	75	333.0	332.6	67.9	74.9	192.2	203.6
T3 TCGGATAAGT	75	356.0	332.6	67.5	79.7	194.5	217.7
T4 TCGGCTATTGT	75	355.0	332.9	67.2	77.6	187.7	212.4
T5 TACTCCGCATT	75	333.0	334.1	73.3	74.0	215.5	199.9
T6 TAGACCGCAAT	75	357.0	334.3	79.5	74.1	223.5	199.8
T7 TATCGTTGCCT	75	346.0	334.6	82.5	74.0	250.7	199.4
UD1 GCCCAGCG	100	348.0	331.0	58.4	57.6	162.9	172.4
UD2 AGACCTAGT	100	336.0	319.2	45.4	56.0	136.5	154.0
UD3 CCATGTCCC	100	337.0	326.0	74.3	65.9	214.9	181.0
UD4 TGCACGCTA	100	348.0	331.2	55.8	55.0	162.9	145.0
UD5 ATGCTCATGC	100	342.0	330.1	66.6	65.0	191.8	176.0
UD17 CTCGGGAACGCC	75	358.0	344.0	94.4	100.1	264.0	269.0

^aLeft: the symbol of the sequence from the respective reference. Right: the sequence. Only the sequence of one strand is given; the second chain is complementary.

^bMonomer concentration; in NARES it was simulated by setting the appropriate distance restraint between chains [$C_T(\mu\text{M}) = 3 \times 10^{33} / (4\pi N_A d_{\text{max}}^3)$], where $N_A = 6.022 \times 10^{23}$ is the Avogadro number and d_{max} is the upper boundary of the distance between the two chains (in Å) (see the text under Eq. (2) in the main text for explanation).

^cExcess thermodynamic quantities.

^dMultiple peaks occur in the calculated heat-capacity curve; the position of the dominant band is reported.

FIG. 7. The differences between the calculated and the experimental melting temperatures, entropy, and enthalpy obtained using initial parameters (red) and optimized parameters (blue) to the corresponding experimental values.

Similar improvement of the agreement between the calculated and experimental values is observed for entropy and enthalpy. As shown in Table III and Figure 7, the optimized force field gives better agreement of the calculated entropies for 12 out of 18 systems and of the calculated enthalpies for 13 out of 18 systems. It can, therefore, be concluded that the optimized force field gives better agreement between the calculated and experimental thermodynamic quantities of DNA melting transition.

IV. CONCLUSION

Force fields, both coarse-grained and all-atom, should be able to reproduce not only structural but also thermodynamic data. Compared to all-atom force fields, the coarse-grained force fields enable us to run simulations at orders of magnitude larger time- and size-scale and thus, to obtain converged thermodynamic quantities from simulations. Therefore, optimization of a coarse-grained force field is one of the critical steps in force field development processes. To improve the performance of the NARES-2P force field with regard to the agreement of both structure and thermodynamics, the recently developed variant of the modified maximum-likelihood optimization method²⁹ was used to optimize the NARES-2P force field whose previous version performed well with regard to reproducing the structures of DNA duplexes but its capability to reproduce experimental thermodynamic quantities was more limited. The optimized force field was tested with 16 small DNA molecules not used in optimization. It was shown that better agreement between the calculated and experimental melting temperatures, entropies, and enthalpies

was obtained for the majority of systems (Table III and Figure 7).

On the other hand, the heat capacity curves obtained with the optimized force field are still broader than experimental curves, which suggests that the force field does not produce sufficiently cooperative melting transitions. Also, in Table III, it is clear that NARES-2P tends to predict higher melting temperatures compared to experimental data. The reason for this might be that, apart from the well-depths, the cross-sections (σ^o) and anisotropies (χ) of the base-base Gay-Berne potentials [Eqs. (3)–(8)] need to be optimized or multi-body terms of base-base interactions need to be included in the force field. As demonstrated in our previous work,^{27,36} multi-body terms can significantly increase the structural stabilization of helical structure in proteins with UNRES. These terms can provide strong interactions to stabilize the secondary structures, and will quickly vanish as temperature increases. Thus, introduction of multibody terms will certainly not only increase the stability of the double stranded DNA structure but also improve significantly the melting temperatures as well as other thermodynamic properties obtained from NARES-2P force field. Both points are now being addressed in our laboratory.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health Grant No. GM-14312 and National Science Foundation Grant No. MCB-10-19767, by Grant No. DEC-2012/06/A/ST4/00376 from the Polish National Science Centre (NCN), and by Grant Mistrz 7./2013 from the Foundation for Polish Science. Computational resources were provided

by (a) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdansk, (b) the National Science Foundation at the National Institute for Computational Sciences (<http://www.nics.tennessee.edu/>), (c) our 952-processor Beowulf cluster at the Baker Laboratory of Chemistry, Cornell University, and (d) our 184-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

- ¹Y. Duan and P. A. Kollman, *Science* **282**, 740 (1998).
- ²A. Liwo, M. Khalili, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2362 (2005).
- ³A. Rojas, A. Liwo, D. Browne, and H. A. Scheraga, *J. Mol. Biol.* **404**, 537 (2010).
- ⁴R. Zhou, G. G. Maisuradze, D. Suñol, T. Todorovski, M. J. Macias, Y. Xiao, H. A. Scheraga, C. Czaplowski, and A. Liwo, *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18243 (2014).
- ⁵R. O. Dror, T. J. Mildorf, D. Hilger, A. Manglik, D. W. Borhani, D. H. Arlow, A. Philippson, N. Villanueva, Z. Yang, M. T. Lerch, W. L. Hubbell, B. K. Kobilka, R. K. Sunahara, and D. E. Shaw, *Science* **348**, 1361 (2015).
- ⁶G. Nemethy and H. A. Scheraga, *Biopolymers* **3**, 155 (1965).
- ⁷H. A. Scheraga, *Adv. Phys. Org. Chem.* **6**, 103 (1968).
- ⁸D. E. Shaw, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Lerardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. M. Deneroff, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S. C. Wang, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, and K. J. Bowers, *Commun. ACM* **51**, 91 (2008).
- ⁹A. Savelyev and G. A. Papoian, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 20340 (2010).
- ¹⁰A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, *J. Chem. Phys.* **132**, 035105 (2010).
- ¹¹T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, *Phys. Rev. Lett.* **104**, 178101 (2010).
- ¹²T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, *J. Chem. Phys.* **134**, 085101 (2011).
- ¹³J. C. Araque, A. Z. Panagiotopoulos, and M. A. Robert, *J. Chem. Phys.* **134**, 165103 (2011).
- ¹⁴M. C. Linak, R. Tourdot, and K. D. Dorfman, *J. Chem. Phys.* **135**, 205102 (2011).
- ¹⁵Y. He, M. Maciejczyk, S. Oldziej, H. A. Scheraga, and A. Liwo, *Phys. Rev. Lett.* **110**, 098101 (2013).
- ¹⁶O. Gonzalez, D. Petkevičiūtė, and J. H. Maddocks, *J. Chem. Phys.* **138**, 055102 (2013).
- ¹⁷D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo, *J. Chem. Phys.* **139**, 144903 (2013).
- ¹⁸T. Cragolini, P. Derreumaux, and S. Pasquali, *J. Phys. Chem. B* **117**, 8047 (2013).
- ¹⁹N. Korolev, D. Luo, A. Lyubartsev, and L. Nordenskiöld, *Polymers* **6**, 1655 (2014).
- ²⁰C. Maffeo, T. T. M. Ngo, T. Ha, and A. Aksimentiev, *J. Chem. Theory Comput.* **10**, 2891 (2014).
- ²¹A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *J. Comput. Chem.* **18**, 849 (1997).
- ²²A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Oldziej, and H. A. Scheraga, *J. Comput. Chem.* **18**, 874 (1997).
- ²³A. Liwo, C. Czaplowski, J. Pillardy, and H. A. Scheraga, *J. Chem. Phys.* **115**, 2323 (2001).
- ²⁴A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 5482 (1999).
- ²⁵A. Liwo, P. Artukowicz, S. Oldziej, C. Czaplowski, M. Makowski, and H. A. Scheraga, *J. Phys. Chem. B* **108**, 16918 (2004).
- ²⁶S. Oldziej, A. Liwo, C. Czaplowski, J. Pillardy, and H. A. Scheraga, *J. Phys. Chem. B* **108**, 16934 (2004).
- ²⁷A. Liwo, M. Khalili, C. Czaplowski, S. Kalinowski, S. Oldziej, K. Wachucik, and H. A. Scheraga, *J. Phys. Chem. B* **111**, 260 (2007).
- ²⁸Y. He, Y. Xiao, A. Liwo, and H. A. Scheraga, *J. Comput. Chem.* **30**, 2127 (2009).
- ²⁹B. Zaborowski, D. Jagieła, C. Czaplowski, A. Hałabis, A. Lewandowska, W. Żmudzińska, S. Oldziej, A. Karczyńska, C. Omieczynski, T. Wirecki, and A. Liwo, *J. Chem. Inf. Model.* **55**, 2050 (2015).
- ³⁰C. Czaplowski, S. Kalinowski, A. Liwo, and H. A. Scheraga, *J. Chem. Theory Comput.* **5**, 627 (2009).
- ³¹G. A. F. Seber and C. J. Wild, *Nonlinear Regression* (Wiley-Interscience, 2003).
- ³²K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, *Mol. Biol. Evol.* **28**, 2731 (2011).
- ³³S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *J. Comput. Chem.* **13**, 1011 (1992).
- ³⁴C. B. Hughesman, R. F. B. Turner, and C. Haynes, *Biochemistry* **50**, 2642 (2011).
- ³⁵C. B. Hughesman, R. F. B. Turner, and C. A. Haynes, *Biochemistry* **50**, 5354 (2011).
- ³⁶A. Liwo, S. Oldziej, C. Czaplowski, U. Kozłowska, and H. A. Scheraga, *J. Phys. Chem. B* **108**, 9421 (2004).