CHEMISTRY
A EUROPEAN JOURNAL

Efficient Calculation of Many Stacking and Pairing Free Energies in DNA from a Few Molecular Dynamics Simulations

Chris Oostenbrink and Wilfred F. van Gunsteren*[a]

Abstract: Through the use of the one-step perturbation approach, 130 free energies of base stacking and 1024 free energies of base pairing in DNA have been calculated from only five simulations of a nonphysical reference state. From analysis of a diverse set of 23 natural and unnatural bases, it appears that stacking free energies and stacking conformations play an important role in pairing of DNA nucleotides. On the one hand, favourable pairing free energies were found for bases that do not have the possibility to form canonical hydrogen bonds, while on the other hand, good hydrogen-bonding possibilities do not guarantee a favourable pairing free energy if the stacking of the bases dictates an unfavourable conformation. In this application, the one-step perturbation approach yields a wealth of both energetic and structural information at minimal computational cost.

Keywords: DNA replication · free energy · GROMOS · hydrogen bonds · unnatural base pairs

Introduction

Free-energy calculations based on molecular dynamics (MD) simulation have been being carried out for about 20 years now.[1,2] In this period many applications of the perturbation[3] and thermodynamic integration[4] methods have been reported. Even though increases in computational power have led to impressive increases in accuracy,[5] thereby overcoming the problem of limited sampling, increasing the efficiency of free-energy and entropy calculations is still of major concern.[6–8] Free-energy differences between similar compounds can be calculated very efficiently by the one-step perturbation approach from a nonphysical reference state.[9] This method has been applied successfully to calculate relative free energies of solvation for small solutes[10,11] and relative free energies of binding to a common receptor.[12–14] Here we apply the method to obtain a massive number (about 10^3) of free energies from just a handful (five) of relatively short simulations of a double-helical DNA dodecamer, of a single DNA dodecamer strand in a random coil and in a stacked conformation and of two individual nucleotides in aqueous solution.

[a] Dr. C. Oostenbrink, Prof. Dr. W. F. van Gunsteren
Laboratory of Physical Chemistry
Swiss Federal Institute of Technology
ETH-Hönggerberg, 8093 Zurich (Switzerland)
Fax: (+41) 1-632-1039
E-mail: wfvgn@igc.phys.chem.ethz.ch

The one-step perturbation method is based on the perturbation formula by Zwanzig,[3] shown in Equation (1):

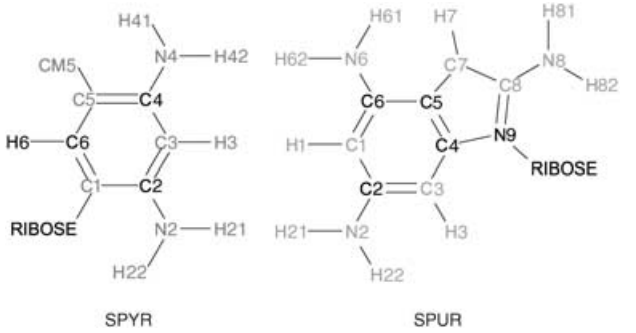
$$\Delta G_{AR} = G_A - G_R = -k_B T \ln \langle e^{-(H_A - H_R)/k_B T} \rangle_R \quad (1)$$

which states that the Gibbs free-energy difference (ΔG_{AR}) between two different states (or molecules), A and R, can be calculated from the configurational or conformational ensemble average of the Boltzmann factor ($e^{-(H_A - H_R)/k_B T}$) as calculated from the ensemble of state R. k_B is the Boltzmann constant, T is the temperature, and H_A and H_R are Hamiltonians for states A and R, respectively. Application of this formula in a single step on finite ensembles of the reference state R is only valid if the ensemble of state R shows sufficient overlap with the ensemble for state A. Because this is generally not the case, traditional free-energy perturbation (FEP) or thermodynamic integration (TI) methods split up the change from R to A into a number of small steps, using the coupling parameter approach.[15] This would require 10 to 20 simulations per free-energy difference, and thus about 10^4 simulations to obtain the large number of free energies we are interested in. Therefore, our approach has rather been to design a nonphysical reference state R that samples a configurational ensemble broad enough to show overlap not only with that of state A, but also with the ensembles of many other physically relevant states—B, C and so forth. An elegant way to do this is to make some atoms in state R "soft": that is, to remove the singularity in the nonbonded

DOI: 10.1002/chem.200401120

interaction,[16] to allow for some overlap with surrounding atoms. In this way it is possible to generate an ensemble containing both configurations similar to those of systems in which the soft atom is present and configurations similar to those of systems in which it is not.

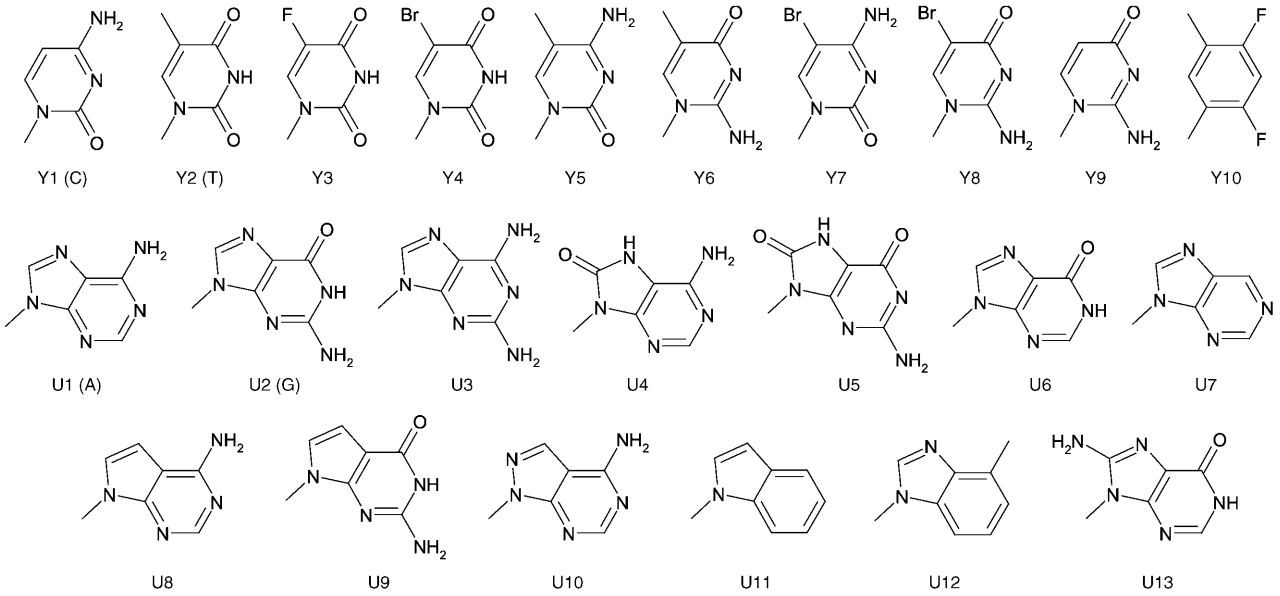
In the current application we have designed the soft pyrimidine–purine base pair depicted in Scheme 1. The position-



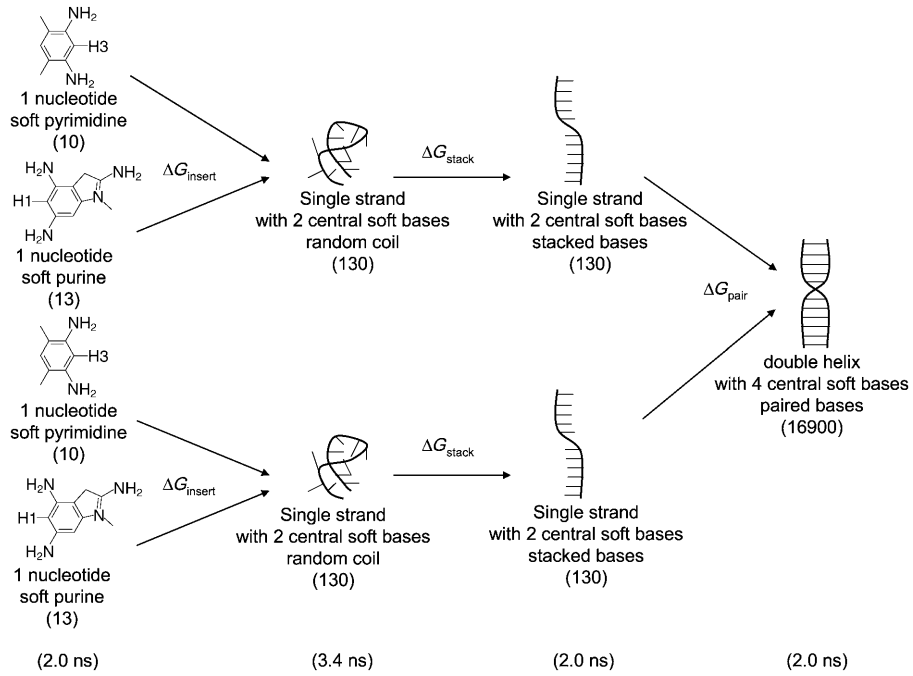
Scheme 1. Nonphysical pyrimidine (SPYR) and purine (SPUR) bases with soft atoms indicated in grey and normal atoms indicated in black. For a description of the force-field parameters used to describe the base see Table 1.

ing of the soft atoms (in grey) in these bases allowed us to generate a reference configurational ensemble that shows overlap with the configurational ensembles of the corresponding real bases (or analogues), given in Scheme 2.[17] These contain both synthetic[18–22] base analogues and bases that have been shown or postulated to be products of radiative damage on the natural bases.[23,24] The stabilities of DNA double helices with some of these compounds incorporated have been determined experimentally.[17,19,20,22] Application of Equation (1) to trajectories produced for molecules containing the soft reference bases yields the free-energy difference upon a change from the soft reference base into one of the real bases. This free-energy difference between real and soft bases can be calculated for any combination of the real bases as a function of the environment of the soft bases in a MD simulation. Four environments were considered and are illustrated in the four vertical columns of Scheme 3: 1) two simulations of a nucleotide, each containing one of the two soft bases, in water, 2) a single-strand DNA dodecamer in water, with the soft bases SPUR and SPYR in the central positions 6 and 7 along the nucleotide chain, which is in "random coil" conformation, 3) the same molecule in water, but held in a conformation in which the bases are stacked and 4) a duplex of the same molecule in water in a double-helical conformation.

The three free-energy differences involved in changing the soft bases into the real ones, corresponding to a change from one environment (columns in Scheme 3) to the next from left to right in Scheme 3, are 1) the free energy ΔG_{insert} of changing the environment of two particular bases from water into a single-strand DNA in water, 2) the free energy ΔG_{stack} of changing the environment of two particular adjacent bases in a single-strand DNA in water from a random coil into a stacked conformation and 3) the free energy ΔG_{pair} of changing the environment of two times two particular adjacent bases in two single DNA strands in water from a separated (but stacked) conformation into a double-helical, and thus paired, conformation. With the real bases of Scheme 2, this gives $13\,(\text{purines}) \times 10\,(\text{pyrimidines}) - 1 = 129$ different relative free energies, ΔG_{insert} for insertion of a purine plus pyrimidine base pair into the DNA chain and the same number of free energies ΔG_{stack} of base stacking. A theoretical number of $130 \times 130 - 1 = 16\,899$ different relative



Scheme 2. Real bases for which the stacking and pairing free energies were calculated. Compounds Y1–10 replace the soft pyrimidine(s) present in the simulations, while compounds U1–13 replace the soft purine(s) present in the simulations.



Scheme 3. Thermodynamic scheme used to calculate the insertion (ΔG_{insert}), stacking (ΔG_{stack}) and pairing (ΔG_{pair}) free energies of a combination of two adjacent base pairs in DNA. The brackets give the number of real compounds for which the free-energy difference with the indicated reference state can be calculated. The bottom line shows the lengths of the MD simulations of the different reference states.

free energies of double base pairing ΔG_{pair} can be calculated. These 1.7×10^4 relative free energies all follow from just a few (five) simulations of the reference state.

Computational Methods

All simulations were performed by use of the GROMOS biomolecular simulation package.[25,26] The parameters were taken from the recently developed 45 A4 parameter set of the GROMOS force field.[27] The interaction parameters for the soft bases are reported in Table 1; the interactions for atoms marked "soft" were calculated by the soft-core approach,[16] with softness parameter values[10] $\alpha_{\text{LJ}}=1.51$ and $\alpha_{\text{C}}=0.5\,\text{nm}^2$. Only five simulations were required to obtain the free energies of stacking and pairing for the ten pyrimidine analogues and 13 purine analogues shown in Scheme 2. Two simulations involved a single soft base (SPYR or SPUR) nucleotide, two simulations involved a single d(CGCGAATTCGCG) strand of DNA with the two soft bases replacing the central AT bases and a fifth simulation involved the corresponding DNA duplex with four soft bases for the two central AT bases, all in explicit water. Initial coordinates were taken from the crystal structure of the Dickerson–Drew dodecamer[28–30] (d(CGCGAATTCGCG)$_2$, Protein Data Base (PDB) entry code 355D).[31] The central purine and pyrimidine in this structure were replaced by the soft bases SPUR and SPYR, respectively. The single nucleotides with SPUR or SPYR bases were solvated in truncated octahedral boxes containing 2035 and 1849 simple point charge (SPC)[32] water molecules, respectively. The single- and dual-strand simulations were performed in rectangular boxes containing 12889 and 13415 SPC water molecules. The single-base nucleotides were simulated in neutral form, the phosphate groups being replaced with OH. In the single- and double-strand DNA simulations, a neutralising amount of Na^+ ions was added, together with additional Na^+Cl^- ion pairs corresponding to a salt content of 0.1 M.

In the single-strand simulation with stacked bases, the stacking was maintained by positional restraints on the twelve C5′ backbone atoms of the chain to their crystallographic positions, through the use of a harmonic potential energy term with force constant $2.5\times10^4\,\text{kJ mol}^{-1}\text{nm}^{-2}$. To keep the soft bases paired in the double-helical DNA simulation, two attractive distance restraints were added for every base pair, as indicated in Table 1. This still allowed the bases freedom to move relative to each other, but prevented them from moving away from their partners completely.

For all simulations, initial velocities were randomly chosen from a Maxwell–Boltzmann distribution at 50 K. Periodic boundary conditions were applied. The temperature was then gradually increased during six 20 ps equilibration simulations. During this time atom-positional restraints on all solute atoms were gradually reduced. At least 2 ns of production simulation were then performed at a constant temperature of 298 K and at a constant pressure of 1 atm, and coordinates from these were stored every 0.1 ps for the free-energy analysis. Temperature and pressure were kept constant by the weak-coupling approach,[33] with relaxation times $\tau_T=0.1$ and $\tau_P=0.5$ ps and an estimated isothermal compressibility of $4.575\times10^{-4}\,(\text{kJ mol}^{-1}\text{nm}^{-3})^{-1}$. Nonbonded interactions were calculated by use of a triple-range cutoff scheme. All interactions within a cutoff distance of 0.8 nm were calculated at every time step from a pair list that was updated every fifth time step. At this point interactions between atoms (of charge groups) within 1.4 nm were also calculated and were kept constant between updates. A reaction field contribution[34] was added to the forces and energies, to account for the influence of a homogeneous medium outside the cutoff sphere of 1.4 nm with a relative dielectric constant of 66.[35]

The free energy of changing the soft bases into any of the corresponding real bases in Scheme 2 was calculated by application of the one-step perturbation formula [Eq. (1)] over the stored configurations of the trajectory. Only the interaction energies involving the atoms in the soft bases need to be reevaluated, making the postprocessing calculation efficient.

The presence of hydrogen bonds was determined by geometrical criteria. A hydrogen bond is considered to be present if the hydrogen–acceptor distance is less than 0.25 nm and the donor–hydrogen–acceptor angle is at least 135°. To analyse the hydrogen bonding between the bases in Scheme 2, a full hydrogen-bond analysis was performed on the soft bases and the presence of hydrogen bonds for every configuration of the trajectory was weighted with the exponential Boltzmann factor in Equation (1).

Results

The overall structure of the double-helical DNA dodecamer remains stable over the course of the simulation. The all-atom positional root-mean-square deviation (RMSD) from the crystal structure is around 0.4–0.5 nm, which is very similar to the results of simulations of the same dodecamer

Table 1. Nonstandard force-field parameters used to describe the artificial soft bases. Soft atoms used a softness parameter for the van der Waals interaction of $\alpha_{LJ}=1.51$ and for the Coulomb interaction of $\alpha_C=0.50\ nm^2$. If two values for C12 are specified, the first is used in apolar interactions, the second in polar interactions.[25,27] The 1–4 nonbonded van der Waals interaction parameters are equal to the normal ones. q is the partial charge on the atom.

Atom	Softness	$C6^{1/2}$ $[(kJ\,mol^{-1}\,nm^6)^{1/2}]$	$C12^{1/2}$ $[10^{-3}\ (kJ\,mol^{-1}\,nm^{12})^{1/2}]$	q $[e]$
pyrimidine				
C1	soft	0.04838	1.837	−0.2
C2	normal	0.04838	1.837	0.2
N2	soft	0.04936	1.301/2.250	−0.4
H21	soft	0.0	0.0	0.1
H22	soft	0.0	0.0	0.1
C3	soft	0.04838	1.837	−0.2
H3	soft	0.0092	0.123	0.2
C4	normal	0.04838	1.837	0.2
N4	soft	0.04936	1.301/2.250	−0.4
H41	soft	0.0	0.0	0.1
H42	soft	0.0	0.0	0.1
C5	normal	0.04838	1.837	0.0
CM5	soft	0.09805	5.162	0.0
C6	normal	0.04838	1.837	0.1
H6	normal	0.0092	0.123	0.1
purine				
C1	soft	0.04838	1.837	−0.2
H1	soft	0.0092	0.123	0.2
C2	normal	0.04838	1.837	0.2
N2	soft	0.04936	1.301/2.250	−0.4
H21	soft	0.0	0.0	0.1
H22	soft	0.0	0.0	0.1
C3	soft	0.04838	1.837	−0.2
H3	soft	0.0092	0.123	0.2
C4	normal	0.04838	1.837	0.2
C5	normal	0.04838	1.837	0.0
C6	normal	0.04838	1.837	0.2
N6	soft	0.04936	1.301/2.250	−0.4
H61	soft	0.0	0.0	0.1
H62	soft	0.0	0.0	0.1
C7	soft	0.04838	1.837	−0.2
H7	soft	0.0092	0.123	0.2
C8	soft	0.04838	1.837	0.2
N8	soft	0.04936	1.301/2.250	−0.4
H81	soft	0.0	0.0	0.1
H82	soft	0.0	0.0	0.1
N9	normal	0.04936	1.301/1.841	−0.2

Bond	Force constant $[10^6\,kJ\,mol^{-1}\,nm^{-4}]$	Ideal bond length [nm]
N–H	18.7	0.100
C–H	12.3	0.109
C–N (amino group)	10.6	0.133
5-ring inside	11.8	0.133
6-ring inside	10.8	0.139
C–CM	7.15	0.153

Bond angle	Force constant $[kJ\,mol^{-1}]$	Ideal bond angle [degree]
5-ring inside	465	108.0
C-N-H	390	120.0
H-N-H	445	120.0
6-ring to hydrogen	505	120.0
6-ring to non-hydrogen	560	120.0
5-ring to hydrogen	575	126.0
5-ring to non-hydrogen	640	126.0
5,6 ring connection	760	132.0

Improper dihedral angle	Force constant $[kJ\,mol^{-1}\,degree^{-2}]$	Ideal improper dihedral angle [degree]
on all planar atoms	0.0510	0.0
ring torsions	0.0510	0.0

Dihedral angle	Force constant $[kJ\,mol^{-1}]$	Phase shift	Multiplicity
C-C-N-H (amino group)	33.5	−1.0	2

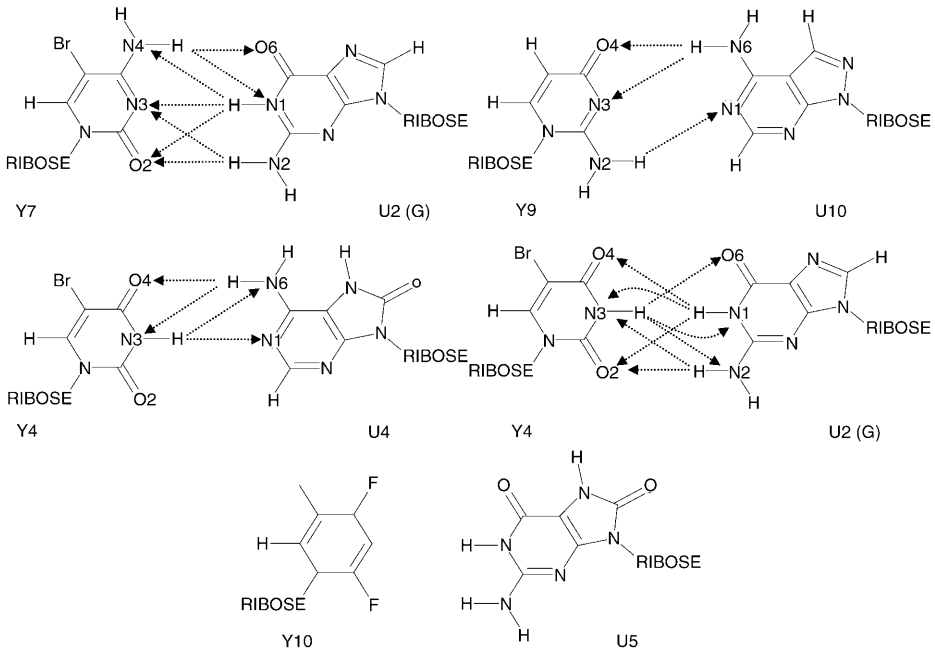
Attractive distance restraint atom pair	Force constant $[kJ\,mol^{-1}\,nm^{-2}]$	Restraint length [nm]
1 6SPUR N2–2 7SPYR N2	500	0.300
1 6SPUR N6–2 7SPYR N4	500	0.350
1 7SPYR N2–2 6SPUR N2	500	0.300
1 7SPYR N4–2 6SPUR N6	500	0.350

double helix without soft bases.[27] The occurrence of canonical (Watson–Crick) hydrogen bonds in the ten non-soft-base pairs is presented in Table 2. Because of the symmetry of the DNA duplex, two values for every hydrogen bond are given. Except between the bases of the first and last base pairs, hydrogen bonds are present for 77–98% of the time. The hydrogen-bonding patterns in the two halves of the molecule are very similar.

The distance restraints that were added to keep the bases of each of the two pairs of soft bases together do ensure stable simulations, but still allow the soft bases sufficient flexibility to sample different relative positions. As indicated by the examples in Scheme 4, noncanonical hydrogen bonds can be formed between donors and acceptors. The atom-positional fluctuations in the soft bases are slightly larger than in the surrounding bases (up to 0.2 nm versus 0.1 nm for the central "non-soft" bases).

Table 2. Occurrence of canonical (Watson–Crick) hydrogen bonds in the non-soft bases in the 2 ns MD simulation of the DNA double helix. Because of the symmetry of the DNA duplex, the same hydrogen bonds are to be expected in the first and the second halves of the double-helical DNA duplex structure.

Hydrogen-bond donor/acceptor pair	Occurrence [%] first half	second half
1 Cyt O2←12 Gua N2	56	55
1 Cyt N3←12 Gua N1	72	72
1 Cyt N4→12 Gua O6	57	61
2 Gua N2→11 Cyt O2	88	87
2 Gua N1→11 Cyt N3	97	98
2 Gua O6←11 Cyt N4	89	90
3 Cyt O2←10 Gua N2	87	81
3 Cyt N3←10 Gua N1	97	98
3 Cyt N4→10 Gua O6	89	94
4 Gua N2→9 Cyt O2	78	77
4 Gua N1→9 Cyt N3	95	98
4 Gua O6←9 Cyt N4	87	92
5 Ade N6→8 Thy O4	84	92
5 Ade N1←8 Thy N3	88	90



Scheme 4. Base pairs selected on the basis of their (decomposed) pairing free energies given in Table 4. Lowest-energy pair with canonical hydrogen bonds: Y7:U2. Lowest-energy pair without canonical hydrogen bonds: Y9:U10. Highest-energy pair with canonical hydrogen bonds: Y4:U4. Highest-energy pairs: Y4:U2 and Y10:U5. The arrows indicate possible (and observed) hydrogen bonds.

The random coil state in Scheme 3 was approximated by a 3.4 ns simulation of a single DNA strand without any restraints. Even though the sampled conformations no longer resembled the starting configuration, the soft bases were interacting with other bases (both parallel and perpendicular stacking were observed). On this timescale a complete sampling of the conformational space belonging to a true random coil cannot be expected. However, the sampling of the other reference states (columns 1, 3 and 4 in Scheme 3) is likely to be much more exhaustive. This is why the sum of ΔG_{insert} and ΔG_{stack} in Scheme 3, termed $\Delta G_{\text{insert,stacked}}$, which can be obtained by direct comparison of the free energies in the stacked single strand and in the individual bases, is probably more precise than its components ΔG_{insert} and ΔG_{stack}. Of the 130 stacking free energies (relative to those of the soft bases) calculated, those for the naturally occurring bases, together with those for the bases with the highest and lowest $\Delta G_{\text{insert,stacked}}$ values, are given in Table 3.

Table 3. Insertion and stacking free energies for selected pairs of adjacent bases in the middle of the single-strand DNA dodecamer as obtained by one-step perturbation and MD simulation. Out of 130 combinations of 13 purines and ten pyrimidines at the sixth and seventh positions in the DNA strand, only the naturally occurring bases and the base combinations with the lowest and highest stacking free energies were selected. ΔG_{insert} is calculated as the difference in the free-energy change from soft to real bases of the random coil single-strand state and the single-base nucleotide state (see Scheme 3). ΔG_{stack} is calculated as the difference in the free-energy change from soft to real bases of the stacked single-strand state and the random coil single-strand state (Scheme 3). $\Delta G_{\text{insert,stacked}}$ is calculated as the difference in the free-energy change from soft to real bases of the stacked single-strand state and the single-base nucleotide state (Scheme 3).

Nucleotide sequence position						ΔG_{insert}	ΔG_{stack}	$\Delta G_{\text{insert,stacked}}$
1–4	5	6	7	8	9–12	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
$(CG)_2$	A	A	T	T	$(CG)_2$	27.6	−23.8	3.8
$(CG)_2$	A	A	C	T	$(CG)_2$	42.7	−28.5	14.2
$(CG)_2$	A	G	T	T	$(CG)_2$	54.2	−23.6	30.6
$(CG)_2$	A	G	C	T	$(CG)_2$	52.8	−58.1	−5.3
$(CG)_2$	A	U13	C	T	$(CG)_2$	12.2	−66.0	−53.8
$(CG)_2$	A	G	Y9	T	$(CG)_2$	39.1	+18.0	57.1

Of the 16900 free energies of double base pairing that can be obtained from the double-helical DNA simulation and the single-strand one (see Scheme 3), only 1024 were evaluated. These correspond to one base pair containing all purine–pyrimidine combinations of the naturally occurring bases (A, C, G, T) and another base pair containing alternative bases at the two central (sixth and seventh) positions in the dodecamer duplex. The SPUR soft base at position 6 in the first strand can be changed to A or G, and the SPYR soft base at the pairing position 7 in the second strand to C or T, which yields four possibilities for the natural base pair in position 6 of strand 1 and position 7 of strand 2. A similar calculation for the ten possible natural and alternative bases at position 7 in the first strand and for the 13 possible natural and alternative bases at position 6 of the second strand yields 130 possibilities for the natural or alternative base pair in position 7 of strand 1 and position 6 of strand 2. Combining these possibilities yields $4\times130=520$ possibilities. Interchanging the position of the natural and alternative or natural base pairs brings the number of possible double base pairs to $2\times520=1040$. However, in this calculation, $4\times4=16$ double base pairs consisting of natural bases are doubly counted, which leads to a total of $1040-16=1024$ double base pairs of the type described. The distribution of these 1024 free energies (relative to those of the soft bases) is given in Figure 1. Because of the symmetry of the double helix, the free energy for a set of two base pairs U*i*:Y*j* and Y*k*:U*l* should theoretically be identical to the free energy for the set U*l*:Y*k* and Y*j*:U*i*. However, the timescale (2 ns) of the duplex simulation is not sufficient to sample all conformational bending modes of the double helix and to ensure that on average the molecule is purely symmetric in a conformational manner as well. We therefore took the two combinations of the base pairs together in the ensemble

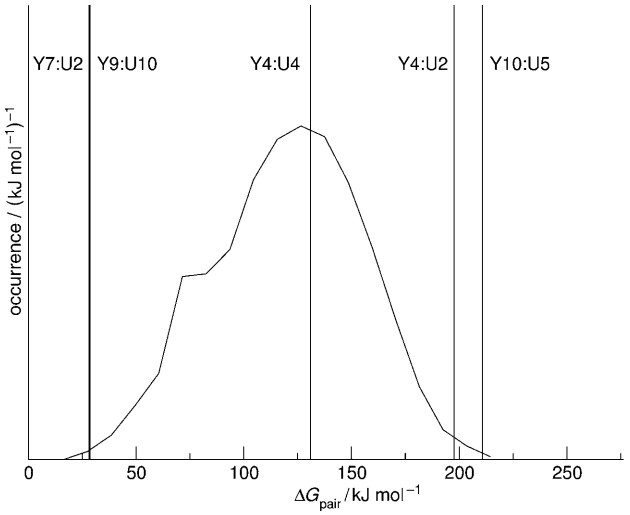


Figure 1. Distribution of ΔG_{pair}, the free energy of base pairing (relative to that of pairing of the soft bases) for 1024 combinations of the two base pairs. The values for the combinations of base pairs are twice the decomposed ΔG_{pair} values according to Table 4. Vertical lines indicate the ΔG_{pair} values of the base pairs shown in Scheme 4.

Table 4. Decomposed ΔG_{pair} values in kJ mol^{-1} as obtained from MD simulation. For every combination of bases, the double-base-pairing free energy is decomposed into single-base-pair contributions by least-squares fitting; 130 values are obtained from 520 independent free energies.

	Pyrimidine									
Purine	Y1 (C)	Y2 (T)	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
U1 (A)	64	52	41	54	58	35	55	40	29	64
U2 (G)	24	83	65	94	19	83	13	88	45	92
U3	51	47	39	46	48	32	47	35	30	61
U4	77	64	50	65	79	47	77	50	24	80
U5	35	90	78	95	42	74	37	76	51	105
U6	46	84	73	87	43	68	38	70	55	94
U7	53	52	42	54	52	37	49	39	36	61
U8	58	48	29	48	60	30	57	33	23	59
U9	30	90	68	95	30	71	24	73	40	94
U10	51	37	21	38	54	16	52	19	11	51
U11	42	58	45	61	41	36	40	38	32	57
U12	39	64	49	71	45	60	43	61	40	78
U13	39	86	75	88	39	78	34	80	58	95

average of Equation (1), doubling the statistics for any combination of Ui:Yj and Yk:Ul for which $i \neq l$ and $j \neq k$ ($1024-16=1008$ cases). The number of free-energy values is then reduced to $1024-(1008/2)=520$.

The 520 calculated values of ΔG_{pair} correspond to the simultaneous base pairing of two base pairs. This number of ΔG_{pair} values is still too large to list them all. We therefore decomposed all these double-base-pairing free energies into contributions from the individual bases, which reduces the number of values to 130, but at the expense of neglecting the influence of the neighbouring base pairs on a base-pairing free energy. The resulting free energies of single base pairing are given in the matrix in Table 4. We note that the values are relative to the base pairing of the soft bases. For comparison with experimental results only differences between the listed single-base-pairing free energies should be considered.

Discussion

About 70% of the 130 values of $\Delta G_{insert,stacked}$ lie—like the values of the naturally occurring bases (Table 3)—within the range from -6 to 31 kJ mol^{-1}. These values correspond to the stacking of two bases in between A and T, relative to the sequence $(CG)_2$, A, SPUR, SPYR, T, $(CG)_2$. Experimentally, the stacking of small aromatic compounds has been studied extensively, but only limited data exist for stacking adjacent bases in a DNA strand.[36,37] Recent "dangling residue" experiments[37] cannot be compared to our data, because these experiments study the stacking of a single base on top of a DNA double helix, and this will still have considerably more conformational freedom than a pair of bases in the middle of a DNA strand. Possible cooperative effects from repetitive stacking will not be represented by these experiments either. The variation in the values listed in Table 3 might seem rather large, but since the values involve three stacking interfaces, the variations are of the same order of magnitude as those from the dangling residue experiments. The values agree reasonably well with values from early quantum-mechanical calculations[38] on double base pairs, which were surprisingly shown to correlate with the melting temperatures of double helices, indicating that the stacking free energies play an important role in DNA stability.[39] The values of $\Delta G_{insert,stacked}$ show the same correlation: bases that are known to pair well also stack well.

For the sequence with the lowest stacking free energy (A, U13, C, T) it is interesting to note that ΔG_{stack} is fairly comparable to that for the (A, G, C, T) sequence. The difference in $\Delta G_{insert,stacked}$ between the two sequences comes mostly from ΔG_{insert}. A particular advantage of the one-step perturbation method is that structural information can be obtained from the simulation of the reference state by picking out those configurations that contribute most to the ensemble average in Equation (1). The strongest contributing conformations for three base stacking sequences are depicted in Figure 2. From the graphical representation of the stacking sequence with the lowest free energy $\Delta G_{insert,stacked}$ (Figure 2B) it becomes clear that an intramolecular hydrogen bond from the 8-amino group in U13 to the O5′ of the backbone is mainly responsible for the favourable value of ΔG_{insert} (Table 3). The relative orientation to the pyrimidine Y1 (C) is highly comparable to the relative G/C orientation (Figure 2A). The positive hydrogen atoms in the cytidine NH_2 group position themselves between the negatively charged N and O atoms of the purine, while the positively charged carbonyl carbons of the cytidine orient themselves towards the negatively charged N-ring atoms of the purine. The configuration contributing most to the free energy $\Delta G_{insert,stacked}$ of the least favoured stacking sequence A, G, Y9, T (Figure 2C) reveals that such a stacking conformation is not pos-

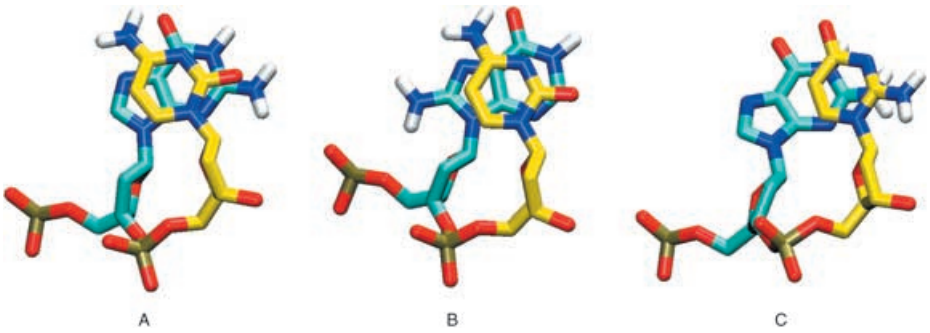


Figure 2. Structures from the simulation of the stacked single DNA chain that contribute most to the free energy $\Delta G_{\text{insert,stacked}}$ of sequences: A) A,G,C,T; B) A,U13,C,T; C) A,G,Y9,T. (See also Table 3). Atoms have been coloured according to the real bases with purine carbons in blue and pyrimidine carbon atoms in yellow.

sible for Y9. Instead, the pyrimidine ring rotates away from the purine because otherwise both the carbonyl and the amino dipoles of the two bases would be exactly aligned.

Use of the decomposed (single base pair) values of ΔG_{pair} in Table 4 to reproduce the original double-base-pair free energies (ΔG_{pair}) of Figure 1 gives a root-mean-square deviation of 14 kJ mol^{-1} over all 520 pairs, indicating that such a decomposition is indeed rather approximate. We note that the value of ΔG_{pair} no longer includes the direct stacking energy within a single strand. One can also calculate a value $\Delta G_{\text{insert,paired}}$ by comparing the free energies in the double helix with those of the single-base nucleotides. A decomposition of the values obtained in this way would mean averaging out of the different stacking energies. By decomposing the values of ΔG_{pair}, we neglected the influence of the upper and lower neighbouring bases on the pairing partner.

Notwithstanding their uncertainties and approximate character, the decomposed pairing free energies of Table 4 still contain a wealth of information. On consideration of the four naturally occurring bases, it is clear that the G:C base pair is most preferred, followed by the A:T base pair. Experimentally, the pyrimidines Y3 and Y4 are known to form more stable DNA double helices than Y2 (T) when pairing to U1 (A). Similarly, Y5 and Y7 are experimentally favoured over Y1 (C) when pairing to U2 (G).[17] The decomposed free energies in Table 4 do indeed show these trends. Surprisingly enough, the free energy for A:C pairing is—at 12 kJ mol^{-1}—not so much higher than for A:T pairing. In fact the discriminating power of adenosine seems much smaller than that of guanidine: it pairs most favourably with pyrimidines 9 and 6, which do not even show the expected hydrogen-bond partners. However, the soft bases are flexible enough to allow for favourable conformations other than the canonical Watson–Crick base pairs[40] (see also Scheme 4).

Guanidine, on the other hand, shows the largest variation of all the purines in its pairing free energy. One of the lowest free-energy base pairs (Y7:G) and one of the highest free-energy base pairs (Y4:G) each involve this purine. Notably, both Y4 and Y7 are Br-containing pyrimidines. Apparently the rather bulky Br atom finds a "comfortable" niche, while also maintaining a good interaction with guanidine in the Y7:G pair, causing Y7 even to be preferred over Y1 (C). For Y4, no such a combination of favourable interactions can be found, leading to a very high free energy of pairing. The two bases forming this pair find themselves rather in configurations in which they are shifted with respect to each other, or in which one moves slightly out of plane to interact with the adjacent base pairs. Scheme 4 and Table 5 display the hydrogen-bonding patterns for these base pairs. The hydrogen-bond percentages have been obtained by reweighting the hydrogen-bond occurrence between the soft bases with the Boltzmann factor in Equation (1) for every configuration in the reference simulation. These percentages will not correspond to the hydrogen-bond occurrences in a true simulation of the "non-soft" bases, but they do indicate the important hydrogen bonds seen in the configurations contributing most to the pairing free energy.

Table 5. Occurrence of hydrogen bonds (arrows in Scheme 4) in the DNA duplex for selected base pairs. Hydrogen-bond occurrences are weighted for every configuration of the reference trajectory by the Boltzmann probability [Eq. (1)] from the corresponding free-energy calculation.

Hydrogen-bond donor/acceptor pair	Occurrence [%]	Hydrogen-bond donor/acceptor pair	Occurrence [%]
Y7 O2←U2 N1	14	Y9 N2→U10 N1	13
Y7 O2←U2 N2	15	Y9 N3←U10 N6	16
Y7 N3←U2 N1	24	Y9 O4←U10 N6	26
Y7 N3←U2 N2	15	Y4 O2←U2 N2	17
Y7 N4→U2 N1	16	Y4 O2←U2 N1	14
Y7 N4←U2 N1	16	Y4 N3←U2 N2	16
Y7 N4→U2 O6	23	Y4 N3→U2 N2	16
Y4 N3→U4 N1	24	Y4 N3←U2 N1	26
Y4 N3→U4 N6	12	Y4 N3→U2 N1	25
Y4 N3←U4 N6	16	Y4 N3→U2 O6	12
Y4 O4←U4 N6	21	Y4 O4←U2 N1	17

The decomposed pairing free energies of the five base pairs shown in Scheme 4 are indicated in the distribution ΔG_{pair} in Figure 1 as well. The base pairs Y7:U2 and Y9:U10 have the lowest pairing free energy (Table 4). Base pair Y7:U2 does show a canonical hydrogen-bonding pattern, while base pair Y9:U10 does not. Base pair Y10:U5 shows the most unfavourable pairing energy, closely followed by the Y4:U2 base pair. Finally, Y4:U4 is the base pair that has the highest free energy of pairing, while still being able to form Watson–Crick hydrogen bonds. It involves, again, the Br-containing pyrimidine and a purine in which an additional carbonyl is present in the 8-position. Indeed, Y4 shows a much more favourable pairing energy

with U10, even though this pair has identical hydrogen-bonding possibilities. If we compare the pairing free energies for purines U1, U8 and U10, we see that they show the same free-energy profile in "pyrimidine space", with U10 being 14–20 kJ mol^{-1} more favourable, except when pairing to Y5 and Y7, where it is still preferred over adenosine (U1) by 3–4 kJ mol^{-1}. This agrees with the experimental finding that U10 pairs to Y2 more favourably than U1.[19] However, U8:Y2 is experimentally found to give less-stable DNA double helices than U1:Y2,[20] while the simulations rather tend towards similar or more favourable pairing energies for U8. Interestingly enough, the purines U7 and U11 also show a profile similar to that of U1, indicating that the hydrogen-bonding capabilities do not affect ΔG_{pair} for U1 (A) so much. However, "purine" U12, a completely neutral steric analogue of adenosine, shows a different pairing pattern, indicating that if there is a substituent on the 6-position, it should be able to form hydrogen bonds. A comparison to experiment based on melting free energies can be made. The U12:C pair is less stable than the A:T base pair by 19 kJ mol^{-1}, the U12:T pair by 20 kJ mol^{-1}, and the U12:Y10 pair by 15 kJ mol^{-1}. In Table 4 we find -13, $+12$ and $+26$ kJ mol^{-1}, respectively. This indicates a prediction of the wrong sign for the U12:C pair. However, the experimental sequence always has a C:G base pair adjacent to the pair under investigation. Even though the rest of the sequence is not identical to the simulated sequence, this knowledge allows us to go back to the values of ΔG_{pair} before decomposition and compare with the double-pairing energies. In this case the pairing energy for the U12:C pair is $+10$ kJ mol^{-1} relative to A:T, U12:T gives $+24$ kJ mol^{-1} and U12:Y10 $+35$ kJ mol^{-1}. From this example it once again becomes clear that the pairing free energy is greatly affected by the type of adjacent base pairs.

The pairing pattern of U12 with all pyrimidines lies somewhat in between those of U1 and U6, which is in turn very similar to that of U13. Comparison of U6 and U13 shows that their pairing free energies are very similar, as opposed to the stacking free energy of U13, as discussed before. As can be seen from Figure 2B, U13 already stacks very similarly to U2 (Figure 2A) in the single strand. For pairing, the base no longer has to turn into a different conformation. The pairing free energies of U13 are on average higher than those for U2. Experimentally this purine is expected to form more stable pairs in Hoogsteen-base-paired parallel-stranded DNA,[41] but this hydrogen-bond pattern is not included in these simulations.

The above examples support the view of Kool et al.[42,43] that the pairing free energies of base pairs are strongly affected by the stacking and steric properties of the individual bases. Substituents at positions far from the base–base interface have major influences on the pairing free energies, while in some cases removal of the hydrogen-bonding capabilities does not affect the selectivity as much. In fact, the most important atomic property in the purines seems to be the protonation state at N1, rather than the substitution pattern at positions 2 and 6. U2, U6, U9 and U13 have pairing properties different from the rest of the purines. Most probably, a combination of both steric effects and hydrogen bonding together leads to optimal pairing free energies, as is also described in the recent work by Fonseca Guerra and Bickelhaupt, based on density functional calculations.[44,45]

Conclusion

From only five MD simulations involving one to four soft bases, we calculated a large number of stacking and pairing free energies. From 1024 out of 16900 theoretical double-base-pairing free energies, we were able to construct a pyrimidine–purine matrix of single-base-pairing free energies. A large influence of the neighbouring base pairs on the base-pairing free energy is indicated by an RMSD value of 14 kJ mol^{-1} when the double base-pairing free energies are back-calculated from the decomposed (single base) values. Still, the decomposed values do agree with several experimental findings and can be used to obtain an initial indication of base-pairing free energies.

A more detailed analysis of specific base pairs reveals that unexpected combinations can give rise to favourable pairing free energies, due to shifted hydrogen-bonding patterns and stacking free energies. In agreement with experiment[42] it turns out that the hydrogen-bonding properties of the bases play only a limited role in the pairing free energies. Base orientations that are dictated by stacking properties and resulting steric effects seem to be as important to explain the base–base interactions.

Acknowledgements

Financial support by the National Center of Competence in Research (NCCR) (Structural Biology) of the Swiss National Science Foundation (SNSF) is gratefully acknowledged.

[1] J. P. M. Postma, H. J. C. Berendsen, J. R. Haak, *Faraday Symp. Chem. Soc.* **1982**, *17*, 55–67.
[2] M. Mezei, D. L. Beveridge, *Ann. N.Y. Acad. Sci.* **1986**, *482*, 1–23.
[3] R. W. Zwanzig, *J. Chem. Phys.* **1954**, *22*, 1420–1426.
[4] J. G. Kirkwood, *J. Chem. Phys.* **1935**, *3*, 300–313.
[5] M. R. Shirts, J. W. Pitera, W. C. Swope, V. S. Pande, *J. Chem. Phys.* **2003**, *119*, 5740–5760.
[6] W. P. Reinhardt, M. A. Miller, L. M. Amon, *Acc. Chem. Res.* **2001**, *34*, 607–614.
[7] C. Chipot, D. A. Pearlman, *Mol. Simul.* **2002**, *28*, 1–12.
[8] C. Peter, C. Oostenbrink, A. van Dorp, W. F. van Gunsteren, *J. Chem. Phys.* **2004**, *120*, 2652–2661.
[9] H. Y. Liu, A. E. Mark, W. F. van Gunsteren, *J. Phys. Chem.* **1996**, *100*, 9485–9494.
[10] H. Schäfer, W. F. van Gunsteren, A. E. Mark, *J. Comput. Chem.* **1999**, *20*, 1604–1617.
[11] J. W. Pitera, W. F. van Gunsteren, *J. Phys. Chem. B* **2001**, *105*, 11264–11274.
[12] B. C. Oostenbrink, J. W. Pitera, M. M. H. Van Lipzig, J. H. N. Meerman, W. F. van Gunsteren, *J. Med. Chem.* **2000**, *43*, 4594–4605.
[13] C. Oostenbrink, W. F. van Gunsteren, *J. Comput. Chem.* **2003**, *24*, 1730–1739.
[14] C. Oostenbrink, W. F. van Gunsteren, *Proteins* **2004**, *54*, 237–246.

[15] D. L. Beveridge, F. M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.
[16] T. C. Beutler, A. E. Mark, R. C. Van Schaik, P. R. Gerber, W. F. van Gunsteren, *Chem. Phys. Lett.* **1994**, *222*, 529–539.
[17] "Synthese und Paarungseigenschaften einer neuen Klasse von Oligodeoxynucleotiden mit einem die Nucleobasen enthaltenden Phosphat-Rückgrat": W. Czechtizky, Ph.D Thesis (ETH: No. 14239), Zürich, **2001**.
[18] V. Nair, S. D. Chamberlain, *Synthesis* **1984**, 401–403.
[19] F. Seela, K. Kaiser, *Helv. Chim. Acta* **1988**, *71*, 1813–1823.
[20] F. Seela, H. Berg, H. Rosemeyer, *Biochemistry* **1989**, *28*, 6193–6198.
[21] R. X.-F. Ren, N. C. Chaudhuri, P. L. Paris, S. Rumney, IV, E. T. Kool, *J. Am. Chem. Soc.* **1996**, *118*, 7671–7678.
[22] K. M. Guckian, J. C. Morales, E. T. Kool, *J. Org. Chem.* **1998**, *63*, 9652–9656.
[23] J. J. Conlay, *Nature* **1963**, *197*, 555–557.
[24] H. Kamiya, K. Miura, H. Ishikawa, H. Inoue, S. Nishimura, E. Ohtsuka, *Cancer Res.* **1992**, *52*, 3483–3485.
[25] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, I. G. Tironi, *Biomolecular Simulation: The GROMOS96 Manual and User Guide*, Vdf Hochschulverlag AG an der ETH Zürich, Zürich, **1996**.
[26] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, W. F. van Gunsteren, *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
[27] T. A. Soares, P. H. Hünenberger, M. A. Kastenholz, V. Kräutler, T. Lenz, R. D. Lins, C. Oostenbrink, W. F. van Gunsteren, *J. Comput. Chem.* **2005**, *26*, 725–737.
[28] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, R. E. Dickerson, *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 2179–2183.
[29] R. E. Dickerson, H. R. Drew, *J. Mol. Biol.* **1981**, *149*, 761–786.
[30] H. R. Drew, R. E. Dickerson, *J. Mol. Biol.* **1981**, *151*, 535–556.
[31] X. Shui, L. McFail-Isom, G. G. Hu, L. D. Williams, *Biochemistry* **1998**, *37*, 8341–8355.
[32] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, J. Hermans in *Intermolecular Forces* (Ed.: B. Pullman), Reidel, Dordrecht, The Netherlands, **1981**, pp. 331–342.
[33] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, *J. Chem. Phys.* **1984**, *81*, 3684–3690.
[34] I. G. Tironi, R. Sperb, P. E. Smith, W. F. van Gunsteren, *J. Chem. Phys.* **1995**, *102*, 5451–5459.
[35] A. Glättli, X. Daura, W. F. van Gunsteren, *J. Chem. Phys.* **2002**, *116*, 9811–9828.
[36] N. Colocci, M. D. Distefano, P. B. Dervan, *J. Am. Chem. Soc.* **1993**, *115*, 4468–4473.
[37] K. M. Guckian, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, D. C. Tahmassebi, E. T. Kool, *J. Am. Chem. Soc.* **2000**, *122*, 2213–2222.
[38] R. L. Ornstein, R. Rein, D. L. Breen, R. D. MacElroy, *Biopolymers* **1978**, *17*, 2341–2360.
[39] O. Gotoh, Y. Tagashira, *Biopolymers* **1981**, *20*, 1033–1042.
[40] J. D. Watson, F. H. C. Crick, *Nature* **1953**, *171*, 737–738.
[41] E. Cubero, A. Avino, B. G. de la Torre, M. Frieden, R. Eritja, F. J. Luque, C. González, M. Orozco, *J. Am. Chem. Soc.* **2002**, *124*, 3133–3142.
[42] E. T. Kool, J. C. Morales, K. M. Guckian, *Angew. Chem.* **2000**, *112*, 1046–1068; *Angew. Chem. Int. Ed.* **2000**, *39*, 990–1009.
[43] E. T. Kool, *Curr. Opin. Chem. Biol.* **2000**, *4*, 602–608.
[44] C. Fonseca Guerra, F. M. Bickelhaupt, *Angew. Chem.* **2002**, *114*, 2194–2197; *Angew. Chem. Int. Ed.* **2002**, *41*, 2092—2095.
[45] C. Fonseca Guerra, F. M. Bickelhaupt, *J. Chem. Phys.* **2003**, *119*, 4262–4273.

Received: November 5, 2004
Published online: May 6, 2005