Contributions of the Thymine Methyl Group to the Specific Recognition of Poly- and Mononucleotides: An Analysis of the Relative Free Energies of Solvation of Thymine and Uracil[†]

Kevin W. Plaxco and William A. Goddard, III*

Materials and Molecular Simulation Center, Beckman Institute (139-74), Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125

Received October 13, 1993; Revised Manuscript Received January 4, 1994[®]

ABSTRACT: Experimental results indicate that interactions with the 5-methyl group of thymine often account for around 1 kcal/mol of the total selectivity at A·T base pairs in protein–DNA complexes. The limited ability of methyl groups to form noncovalent interactions of this magnitude has led to the hypothesis that the energy of solvation of this hydrophobic element is responsible for the observed contribution to selectivity; however, it has not been possible to test this experimentally. We report a molecular dynamics perturbation thermodynamics (MD/PT) analysis of the relative free energy of solvation of thymine and uracil, both as the free bases and in the context of double-stranded DNA. The use of MD/PT indicates that the effect of shielding the 5-methyl group from solvent accounts for 0.90 ± 0.11 kcal/mol of the observed contribution to specificity in protein–DNA complexes. We suggest some implications of these results for the mechanism of sequence-specific DNA recognition, DNA structure, and the evolution of the deoxynucleotide synthesis pathways.

For more than a decade there has been experimental evidence that the thymine methyl group plays a major role in determining the sequence specificity of many DNA binding proteins that recognize major-groove elements (Goeddel *et al.*, 1977). This functional group has been demonstrated to contribute to the selectivity of a wide variety of protein–DNA complexes, including the *lac* repressor, in which it is estimated to account for 1.5 kcal/mol; the Cro repressor, 0.6–1.6 kcal/mol; and RNA polymerase, 0.7 kcal/mol of the total sequence selectivity at individual base pairs in the recognition sequence (Goeddel *et al.*, 1977; Takeda *et al.*, 1989; Dubendorff *et al.*, 1987).

As shown in Figure 1, the thymine methyl group projects into the major groove of DNA, where it is accessible both to proteins in protein–DNA complexes and to the solvent in free DNA. Traditionally it is considered that the sequence selectivity imparted by this functional group arises from favorable van der Waals interactions between the protein and this element. However, the ability of methyl groups to form strong intermolecular interactions is extremely limited, and estimates from gas phase studies are substantially less than the energy of selectivity observed in many protein–polynucleotide complexes [the methane–methane energy is calculated to be 0.2–0.4 kcal/mol (Novoa *et al.*, 1991), while experiments suggest values of about 0.43 kcal/mol (Matthews & Smith, 1976; Schamp *et al.*, 1958]. Because of this discrepancy it has been postulated that the bulk of the selectivity imparted by the thymine methyl group stems instead from protein-driven desolvation of this hydrophobic element. This explanation for the role that thymine methyl groups are observed to play in sequence selectivity has been accepted in the literature (Aiken & Gumport, 1981); however, due to the inability to experimentally determine relative solvation energies for highly soluble species such as mono- and polynucleotides, it has not been possible to experimentally confirm the magnitude of this solvation effect. As a result this explanation has largely failed to enter the lexicon of biology.

The improved accuracy and speed of computer simulation techniques make it practical to accurately access such important, but experimentally inaccessible, interaction energies. In order to estimate the role solvation energies play in the selectivity of sequence-specific DNA binding proteins, we used molecular dynamics pertubation thermodynamics (MD/PT) to determine the relative free energy of solvation of thymine and uracil, both as the free bases and in the context of double-stranded DNA. These results have significant implications on general issues of sequence selectivity of DNA binding, the metabolism of the deoxynucleotides, and the role of solvation in molecular recognition.

EXPERIMENTAL PROCEDURES

The magnitude of the relative free energies of solvation discussed above is 10–100 times smaller than the random thermal fluctuations and hence, long periods of dynamics must be sampled before the calculated free energy converges on the correctly averaged value. This convergence problem is particularly acute in the simulation of biologically relevant systems because the necessity of including sufficient water molecules for realistic solvation requires the study of very large molecular systems. Consequently, we have exploited molecular dynamics perturbation thermodynamics (MD/PT) (Zwanzig, 1956), which considerably speeds convergence.

The simulations reported here were conducted on both the free bases and on a full turn of B-form double-helical DNA. A helix of DNA, 1572 water molecules and 10 base pairs of

[†] Contribution No. 8854 from the Division of Chemistry and Chemical Engineering, Caltech. This work was supported by a grant from the Advanced Industrial Concepts Division (AICD) of the Department of Energy (DOE). K.W.P. has been supported in part by a National Science Foundation graduate fellowship. The KSR computer facilities were supported by NSF-GCAG (ASC-9217368). The facilities of the MSC are also supported by grants from NSF Chemistry (91-00284), DOE-CB, Allied-Signal, Asahi Chemical, Asahi Glass, BP Chemical, Chevron, BF Goodrich, Vestar, Teijin, and Xerox.

[®] Abstract published in *Advance ACS Abstracts*, February 15, 1994.

0006-2960/94/0433-3050$04.50/0

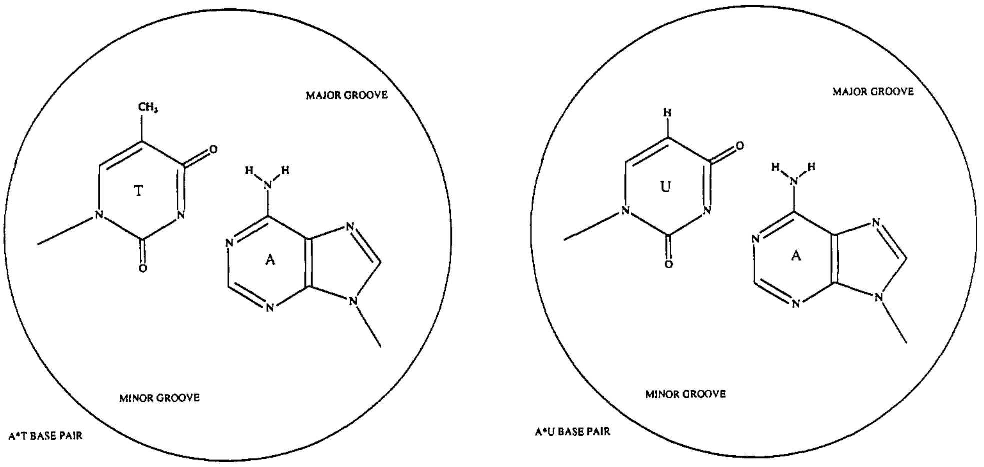

FIGURE 1: Illustration showing the difference between the A·T and A·U base pairs. Protrusion of the thymine methyl group into the major groove makes it highly solvent accessible.

dA·dT, fixed in the B conformation were placed in a box of 40.0 Å by 40.0 Å by 33.8 Å, and periodic boundary conditions (PBC) were used to generate the infinite system used in the calculations. These dimensions mimic the natural periodicity of the B conformation and simulate an infinite helix of DNA, eliminating end effects. A single T was changed to U and back to T in the MD/PT calculations. For simulations of the free base, PBC were used with 214 water molecules in an 18.9481-Å cubic box.

After equilibration, the perturbations were conducted from T to U in 200 intermediate hybrids over a period of 100 ps. The system was then reequilibrated and the process was reversed, converting U to T over a period of 100 ps. The volume of the PBC box was held fixed during these simulations so the free energies determined correspond to Helmholtz free energies. However, the differential Helmholtz and Gibbs free energies are expected to be within 0.1 kcal/mol. Issues of MD/PT implementation, force fields, and convergence are discussed in greater depth under Computational Details.

RESULTS

The relative free energies of gas-phase thymine and uracil were determined from the simulations to be 1.6 ± 0.2 kcal/mol. This result does not differ significantly from the relative internal free energy difference between aqueous thymine and uracil, indicating that the effect of solvation on the relative internal free energy of these species leads to a negligible contribution to the relative free energy of solvation. This observation was upheld by comparison of the solvation of internally fixed and internally mobile bases. Since no significant deviation was observed, further simulations were conducted using internally fixed bases and sugars (no internal energy contribution was calculated).

The relative free energy of solvation of uracil and thymine was determined to be

$$\Delta\Delta G^{\text{sol}}_{\text{T}\rightarrow\text{U}} = -2.34 \pm 0.03 \text{ kcal/mol} \qquad (1)$$

The values obtained in each simulation are shown in Table 1. In the context of a helix of DNA, the relative free energy of solvation of the thymine methyl group was determined to be

$$\Delta\Delta G^{\text{sol}}_{\text{T}\rightarrow\text{U}} = -0.90 \pm 0.11 \text{ kcal/mol} \qquad (2)$$

The large reduction in the relative free energy of solvation of this group is probably due to the significantly reduced solvent accessibility of this group in DNA. The solvent-accessible surface area of the methyl group is 24.4 Å^2 in the free base and 4.0 Å^2 in double-stranded DNA, which supports this general interpretation but suggests that the effect is not strictly proportional to surface area. Energy profiles for these transformations are shown in Figure 2.

Table 1: Computed Free Energy Differences from MD/PT Calculations

	transformation	relative free energy (kcal/mol)	surface area[a] (Å^2)
free bases	T → U	–2.36	24.4
	U → T	2.32	
helical DNA	T → U	–0.98	4.0
	U → T	0.83	

[a] The change in solvent-accessible area is the calculated difference in solvent-accessible surface between the two species. This was calculated using the BIOGRAF simulation package (BIOGRAF, 1992) using a probe radius of 1.4 Å.

The free energy of solvation of 1-methylthymine has been estimated from experiment (–9.1 to –12.7 kcal/mol; Clark *et al.*, 1965) and from a rather extensive simulation (–7.9 to –9.4 kcal/mol; Ferguson *et al.*, 1992). This agreement between theory and experiment indicates that the force field description is appropriate. Unfortunately, due to the very high solubility of this base, no measure has been made of the solution free energy of uracil. This limits our ability to assess the accuracy of the force field description (by comparing the results of simulation directly with experiment).

CONCLUSIONS

We have shown that the primary energetic contributor to the energy of specificity of the thymine methyl group is the cost of solvating the methyl group and not van der Waals interactions with elements in the protein.

The 0.90 ± 0.12 kcal/mol estimate of relative free energy of solvation for the thymine methyl group in the context of DNA is consistent with the observed contribution to selectivity attributed to the functional group in protein–DNA complexes. Experimental studies of protein–DNA complexes suggest a slightly larger value, 0.9–1.4 kcal/mol. This may indicate a small but real van der Waals contribution to the total energy of interaction. However, considering the uncertainties in both experiment and theory, we consider that the differential free energy of solvation is responsible for at least 80% of the effect. Historically it has been assumed that a degree of complementarity (i.e., a hydrophobic surface on each of the interacting

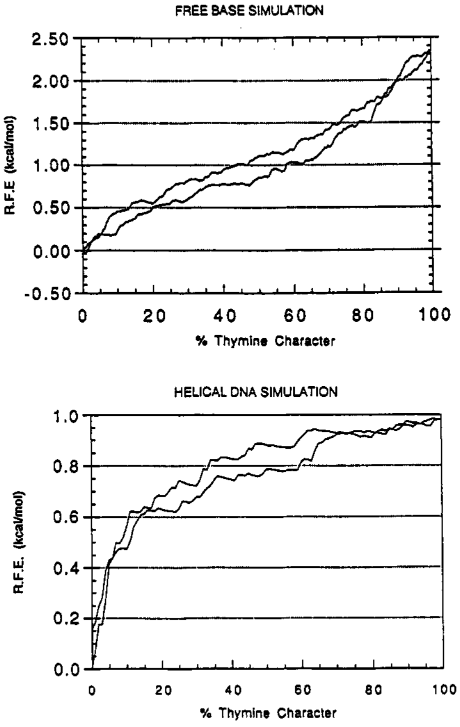

FIGURE 2: Free energy profiles for the MD/PT simulations. The plots represent the summed total free energy differences as the solute species evolves from uracil to thymine and back again. Perfect convergence would generate a single line starting at the origin. The observed relatively small degree of hysteresis indicates a high degree of convergence.

faces) is a necessary component for this mechanism of generating sequence specificity. Our results indicate that a substantial energetic contribution to specificity will occur irrespective of the nature of the recognition surface. The key ingredients are (1) methyl group desolvation and (2) no prohibitive steric interactions.

Selection against this base (i.e., selection of some other base over thymine) requires either the formation of prohibitive steric interactions or sequence-specific hydrogen bonding. Since the total number of hydrogen-bond donors and acceptors is the same in G·C and A·T base pairs, the contributions to relative solvation energies of groups other than the 5-methyl group are limited. We therefore conclude that *any protein–DNA complex that excludes solvent from a given base pair in the binding site without making sequence-specific hydrogen bonds or prohibitive steric interactions will select for AT (or TA) base pairs with an energy of selectivity of 0.9 kcal/mol*, corresponding to a 5-fold increase in K_d.

Our simulations on the free bases allow us to estimate the maximum contribution of the thymine methyl group to recognition of the free thymine nucleotide. This quantity has important implications in the evolution of deoxynucleotide metabolism. To prevent the inappropriate inclusion of dU into DNA, cellular levels of dUTP must be kept very low (typically a ratio of less than 1/300 to dTTP concentrations). This is achieved through the activity of the enzyme dUTPase, which selectively degrades any dUTP formed, rather than by the energetically more conservative process of selectively phosphorylating dTMP over dUMP. One might wonder why nature has evolved such an energetically wasteful mechanism. Our results indicate that interactions selecting thymine over uracil (which are limited to desolvation and van der Waals contacts with the thymine methyl group) provide at best a 2.8 kcal/mol selection energy. This corresponds to a selectivity of no more than 1/100, and hence dUTPase is required to obtain the 1/300 ratio. The complementary recognition, selection of uracil over thymine, can result in van der Waals interactions (steric clashes) that are significantly stronger than attractive van der Waals interactions and can easily account for the observed nucleotide ratio.

This work demonstrates to be incorrect the common view that hydrophobic interactions (such as the recognition of a thymine methyl group) are dominated primarily by van der Waals interactions. Rather, these studies indicate that the primary contributor to the energetics of such interactions is the destabilization of solvated, uncomplexed DNA containing such hydrophobic groups binding to a ligand having hydrophobic groups in the appropriate position to interact with the methyl of T. Elimination of this unfavorably solvated site leads then to enhanced bonding. Thus the driving force for the hydrophobic effort on binding of a ligand is not due to special DNA–ligand interactions in the product but rather results from DNA–water interactions in the uncomplexed system.

COMPUTATIONAL DETAILS

Thermodynamic perturbation theory (Zwanzig, 1956)

$$G(\lambda) - G(\lambda_i) = -RT \ln < e^{-(V_\lambda - V_{\lambda_i})/RT} > \lambda_i \qquad (3)$$

provides a means for calculating the free energy difference between any two states $A(\lambda)$ and $A(\lambda_i)$. However, unless the states share a significant fraction of conformation space, the average is very slow to converge. We have reduced this convergence time by calculating the relative free energy between closely related states, using a thermodynamic cycle to determine the free energy difference from some nonphysical transformations:

$$\begin{tikzcd}
(\mathrm{dT{\cdot}dA})(\mathrm{g}) \arrow[r, "\Delta G_{\mathrm{g}}(\mathrm{T} \to \mathrm{U})"] \arrow[d, "\Delta G_{\mathrm{solv}}(\mathrm{T})"] & (\mathrm{dU{\cdot}dA})(\mathrm{g}) \arrow[d, "\Delta G_{\mathrm{solv}}(\mathrm{U})"] \\
(\mathrm{dT{\cdot}dA})(\mathrm{aq}) \arrow[r, "\Delta G_{\mathrm{aq}}(\mathrm{T} \to \mathrm{U})"'] & (\mathrm{dU{\cdot}dA})(\mathrm{aq})
\end{tikzcd} \qquad (4)$$

Since $\Delta G_g(\mathrm{T} \rightarrow \mathrm{U}) + \Delta G_{\mathrm{solv}}(\mathrm{U}) = \Delta G_{\mathrm{solv}}(\mathrm{T}) + \Delta G_{\mathrm{ag}}(\mathrm{T} \rightarrow \mathrm{U})$, the quantity of interest, $\Delta\Delta G = \Delta G_{\mathrm{solv}}(T) - \Delta G_{\mathrm{solv}}(U)$, can be calculated from simulations of the reactions (dT·dA)(g) → (dU·dA)(g) and (dT·dA)(aq) → (dU·dA)(aq), leading to [$\Delta\Delta G = \Delta G_g(\mathrm{T} \rightarrow \mathrm{U}) - \Delta G_{\mathrm{aq}}(\mathrm{T} \rightarrow \mathrm{U})$] much more rapidly than the direct simulation of (dU·dA)(g) → (dU·dA)(aq) and (dT·dA)(g) → (dT·dA)(aq).

The convergence time of the pertinent average can be further improved by calculating the free energy in a stepwise fashion around intermediate potentials V_λ:

$$V_\lambda = \lambda V_{\mathrm{T{\cdot}A}} + (1 - \lambda) V_{\mathrm{U{\cdot}A}} \qquad (5)$$

where V_λ represents the potential for a "hybrid molecule", λ is the coupling parameter ($0 \leq \lambda \leq 1$), and $V_{\mathrm{T{\cdot}A}}$ and $V_{\mathrm{U{\cdot}A}}$ represent the potentials of T·A and U·A base pairs respectively [reviewed in Beveridge and DiCapua (1989)]. The complete thermodynamic cycle is indicated in (4) and the perturbation results are in Figure 3.

Despite these efforts to reduce convergence times, simulations of this type remain computationally daunting. In order to undertake sufficiently long simulations to obtain the high degree of convergence noted here, the simulations were conducted using a massively parallel constant-temperature molecular dynamics simulation program we have written for the KSR 1/64 supercomputer (Plaxco, 1993).

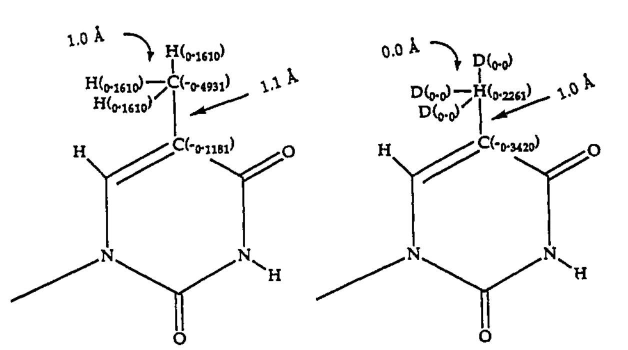

FIGURE 3: Perturbation employed in the simulations reported here, including some of the PS-GVB-derived charges. The symbol "D" represents a dummy atom with no charge or van der Waals character. As λ → 0 for the perturbation T → U, the character of the methyl hydrogens atoms disappears and the C–H bond length shrinks to zero.

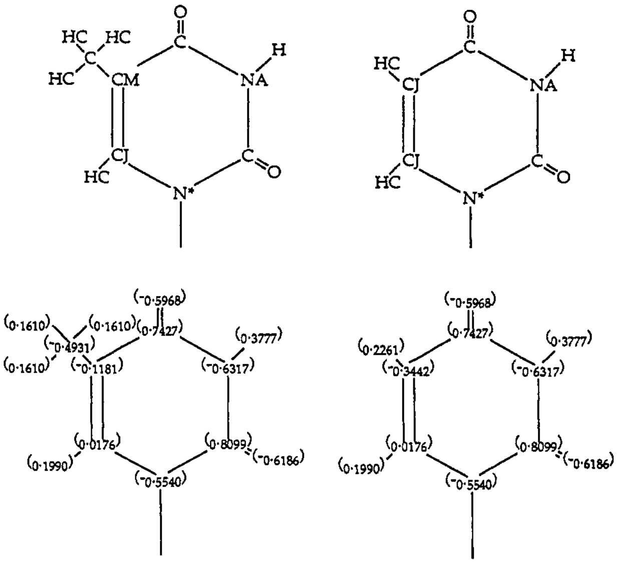

FIGURE 4: AMBER atom types and PS-GVB derived point charges for uracil and thymine.

The simulations used the TIP3P potential for water molecules (Jorgensen *et al.*, 1983) and the AMBER force field (Weiner *et al.*, 1984) to describe DNA (the AMBER atom-types are shown in Figure 4). All pyrimidine hydrogens and hydrogens bonded to heteroatoms were explicitly included in the simulations. Thymine and uracil charges derived from pseudospectral-generalized valence bond calculations (PS-GVB) (Ringnalda *et al.*, 1993) are shown in Figure 4. The charges of acidic phosphate oxygens were reduced by $0.5q_e$ to reproduce counterion effects.

Ten base pairs of poly(dA·dT) were fixed in the B conformation [using BIOGRAF (1992)] and placed in a periodic box of dimension 40 Å by 40 Å by 33.8 Å. The species under study were solvated by removing all water molecules within 1.4 Å of any solute atom from a 1.0 g/mL bath of TIP3P water preequilibrated with 100 ps of dynamics. These waters were reequilibrated with the solute molecule for at least 50 ps. All van der Waals and electrostatic interactions of all base pairs and all waters were included.

All simulations were conducted using (*i*) a 1-fs time step, (*ii*) the shake algorithm (Ryckaert & Berendsen, 1977) to maintain rigid water bonds and angles, and (*iii*) a spline cutoff to scale nonbond interactions over the range 0–12.0 Å. Preequilibration was judged by monitoring heat capacities and was generally complete within 25–50 ps.

MD/PT analyses (conducted with charges that differed slightly from the remainder of the work recorded here) were found to be highly convergent when conducted using 200 intermediate hybrids over 100 ps of dynamics, as illustrated in Figure 5. For each hybrid 0.25 ps of equilibration was followed by 0.25 ps of data collection. For the free base perturbation, a second set of simulations was conducted for 200 ps, including 0.5 ps of equilibration followed by 0.5 ps of data collection per hybrid, and without scaling bond lengths (*i.e.*, the C–D bond length remained 1 Å; see Figure 3). No significant deviation was observed and all remaining simulations were conducted for 100 ps with bond scaling. For all cases, the perturbation sequence was conducted independently from both directions (λ = 0 → 1 and λ = 1 → 0). The various values reported are the average of these two simulations and the standard deviation of this average is reported as the uncertainty. This reflects errors in convergence, not in the force field description.

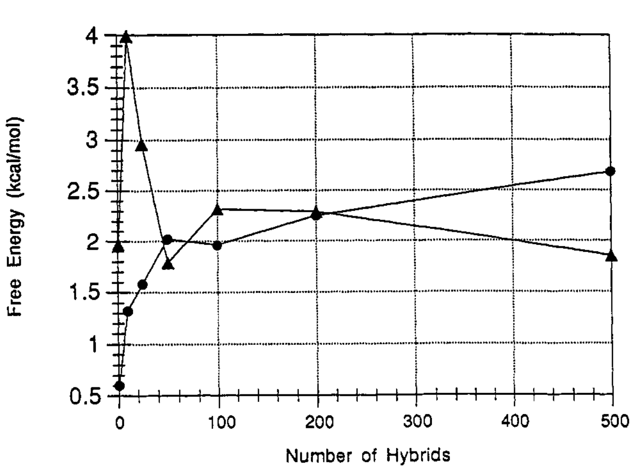

FIGURE 5: Free energy results for the U → T perturbation (circles) and the T → U perturbation (triangles), as a function of the number of intermediate hybrids sampled. With no hysteresis, the two results would be equal (leading to about 2.3 kcal/mol); the optimum seems to be 200 hybrids or 0.5 ps/hybrid. The lack of convergence for 500 hybrids is attributed to the extremely short equilibration time (0.125 ps) for each hybrid.

The determination of solvent-accessible surfaces was made with the BIOGRAF simulation package (BIOGRAF, 1992) using a 1.4-Å radius probe. The coordinates of B-form DNA were taken as indicated in BIOGRAF.

ACKNOWLEDGMENT

We thank Dr. Murco N. Ringnalda (Schrödinger) for performing the PS-GVB calculations used in this study and Dr. Steve Benner (ETH, Zurich) for suggesting the role of solvation in stabilizing the double helix. We wish to thank Dr. Steve Breit of Kendall Square Research for help in optimizing the MD/PT code. The simulations were carried out on the KSR in the MSC and on the KSR at the NSF-Cornell Supercomputer Center.

REFERENCES

Aiken, C. R., & Gumport, R. I. (1991) *Methods Enzymol. 208*, 433–457.

Beveridge, D. L., & DiCapua, F. M. (1989) *Annu. Rev. Biophys. Biophys. Chem. 18*, 431–493.

BIOGRAF/POLYGRAF (1992) Molecular Simulations, Inc., Burlington, MA.

Clark, L. B., Pesche, G. G., & Tinoco, I. (1965) *J. Phys. Chem. 69*, 3615.

Dubendorff, J. W., Dehaseth, P. L., Rosendahl, M. S., & Caruthers, M. H. (1987) *J. Biol. Chem. 262*, 892–898.

Ferguson, D. M., Pearlman, D. A., Swope, W. C., & Kollman,

P. A. (1992) *J. Comput. Chem. 13*, 362–370.

Goeddel, D. V., Yansula, D. B., & Caruthers, M. H. (1977) *Nucleic Acids Res. 4*, 3038–3055.

Jorgenson, W. L., Chandrasekhar, J., Madura, J. L., Impey, R.W., & Klein, M. (1983) *J. Chem. Phys. 79*, 926–935.

Matthews, G. P., & Smith, E. B. (1976) *Mol. Phys. 32*, 1719–1725.

Novoa, J. J., Whangbo, M.-H., & Williams, J. M. (1991) *J. Chem. Phys. 94*, 4835–4841.

Plaxco, K. W. (1993) Protein–DNA Interactions: Molecular Modelling of Structure and Energetics, Ph.D. Thesis Biology, California Institute of Technology, Pasadena, CA.

Ringnalda, M. N., Langlois, J.-M., Greeley, B. H., Russo, T. V., Muller, R. P., Marte, B., Won, Y., Donnelly, R. E., Jr., Pollard, W. T., Miller, G. H., Goddard, W. A., III, & Freisner, R. A.(1993) PS-GVB, v0.08, Schrödinger, Inc.

Ryckaert, G. C., & Berendsen, H. J. C. (1977) *J. Comput. Phys. 23*, 327.

Schamp, H. W., Jr., Mason, E. A., Richardson, A. C. B., & Altman, A. (1958) *Phys. Fluids 1*, 329–335.

Takeda, Y., Sarai, A., & Rivera, V. M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 80*, 439–443.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Jr., & Weiner, P. (1984) *J. Am. Chem. Soc. 106*, 765–783.

Zwanzig, R. W. (1956) *J. Chem. Phys. 22*, 1420.