Homologous Pairing between Long DNA Double Helices

Alexey K. Mazur
UPR9080 CNRS, Université Paris Diderot, Sorbonne Paris Cité, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France

(Received 12 October 2015; revised manuscript received 22 December 2015; published 12 April 2016)

Molecular recognition between two double stranded (ds) DNA with homologous sequences may not seem compatible with the B-DNA structure because the sequence information is hidden when it is used for joining the two strands. Nevertheless, it has to be invoked to account for various biological data. Using quantum chemistry, molecular mechanics, and hints from recent genetics experiments, I show here that direct recognition between homologous dsDNA is possible through the formation of short quadruplexes due to direct complementary hydrogen bonding of major-groove surfaces in parallel alignment. The constraints imposed by the predicted structures of the recognition units determine the mechanism of complexation between long dsDNA. This mechanism and concomitant predictions agree with the available experimental data and shed light upon the sequence effects and the possible involvement of topoisomerase II in the recognition.

DOI: 10.1103/PhysRevLett.116.158101

Mutual recognition between double stranded (ds) DNA with identical sequences is a long-standing enigma in molecular biology [1]. It is involved in processes including the premeiotic and somatic paring of homologous chromosomes [2,3], repeat-induced DNA modifications [4–6] and double strand break repair [7]. Recognition is generally assumed to occur similarly to homologous recombination, i.e., due to recruited proteins that temporarily open dsDNA and make possible the cross-stranded Watson-Crick (WC) base pairing. However, this would require proteins with very special functions, whereas so far searches including genome-wide genetic screens [8–10] have not revealed suitable candidates. Direct DNA-DNA recognition has been suggested as an alternative solution [1,11–13]. Two possible mechanisms have been considered in recent years: (1) attractive long-range electrostatic interactions between B-DNA with identical sequence-dependent conformations [14,15] and (2) a strand exchange between two dsDNA to form the PX-DNA motif used in DNA-origami nanotechnology [16,17]. These models explained available biological data and fit well with the results of *in vitro* experiments in cell-free conditions [18,19]. However, they cannot account for the phenomenon of recognition between partial homologies recently discovered by Gladyshev and Kleckner [20]. These authors studied the sequence dependence of repeat-induced point mutations (RIPs). RIPs occur in fungi cells that somehow identify and target for mutation any long repeated sequence in a genome [4,5]. Strikingly, the recognition occurs with even a 25% homology, provided that it is distributed in a series of triplets spaced by 11 or 12 base pair steps (bps) [20]. Two dsDNA with such sequences cannot form PX-DNA [21] and neither can they be structurally similar; therefore, the RIP data [20] do not fit with the mechanisms of direct recognition [14–17]. These new observations are also difficult to reconcile with any recognition via WC pairing. Indeed, the RIP data indicate that the two dsDNA remain torsionally rigid and the recognition improves with the number of active triplet frames rather than the integral homology [20]. In contrast, local melting should zero the twisting rigidity, and hybridization of continuous homologous single stranded (ss) DNA should be orders of magnitude more efficient than a pairing of the same number of base pairs in periodically spaced triplets.

In the present study I analyze the possibility of dsDNA recognition through direct binding by major grooves. It was noticed long ago that the major-groove edges of WC base pairs have complementary hydrogen bonding (H-bonding) patterns [22,23]. An infinite helical quadruplex using major-groove association between WC pairs was predicted by manual modeling [11] and discussed as an intermediate state in homologous recombination [12,24]. Experimentally, such structures were not found, but the possibility of major-groove H bonding was confirmed [25,26]. Using methods of quantum chemistry (QC), molecular mechanics (MM), and molecular dynamics (MD), I show that direct dsDNA binding by complementary major grooves should be considered as a probable pathway for direct homology recognition. The admissible recognition conformations are dictated by structural constraints and they explain experimental data better than alternative mechanisms.

Figure 1 illustrates the hypothetical association of two identical B-DNA structures by merging the major grooves. The quadruplex in the right panel has four grooves, namely, two minor grooves of the constituting double helices and two new grooves, hereafter called secondary, formed between the two juxtaposed major grooves. The interior of this structure consists of the parallel base tetrads shown in Fig. 1. They are formed by identical WC pairs linked by two new H bonds that are also called secondary. Such tetrads were experimentally confirmed for both types of WC pairs using short ssDNA

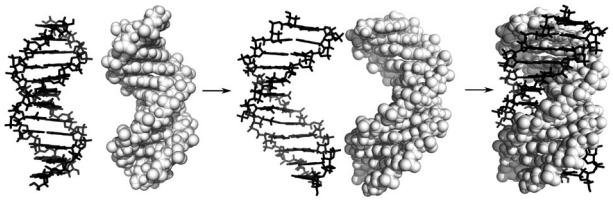

FIG. 1. Helical quadruplex formed from two WC double helices. (Left panel) Two canonical B-DNA duplexes are shown facing one another by their major grooves. (Middle panel) Their conformations in the complexed state. They are slightly stretched and untwisted to a helical pitch of about 12.9. (Right panel) A right-handed quadruplex formed by major-groove association.

hairpins [25,26]. Figure 2 and Table I reveal that the secondary H bonds are shorter and stronger than their sisters in WC pairs. These results were obtained by quantum mechanical optimizations of tetrad geometries in vacuum [27]. Earlier studies of WC base pairs indicate that such calculations reproduce experimental trends, with the energy differences scaled down due to the polar environment [47]. The stabilization energy of the GC/GC tetrad is surprisingly large; namely, the energy of two H bonds appears to be similar to that of the WC pair with three H bonds. This non pair-additive electrostatic effect is due to the large dipole moment and high polarizability of the GC pair [27]. Because of this non pair additivity, molecular mechanics significantly underestimates the strength of secondary H bonding, which is important for the interpretation of other results.

The B-DNA conformations in the left panel of Fig. 1 look predisposed for association because they resemble separated fibers of a two-strand twine. In the complex, the double helices are spun so that quadruplexes longer than one turn cannot fall apart even in the absence of the secondary H bonding. The complex is easily built by making a cylinder from stacked tetrads and then properly placing backbone strands at its surface [11], but it cannot be obtained by docking two dsDNA following the arrows in Fig. 1. To this end, the two initial structures must be untwisted to an almost flat ladder, joined, and then relaxed. Even a small untwisting of dsDNA leads to dissociation of the two strands [49]; therefore, this simple pathway is not feasible. The question is whether there exists an alternative pathway that can join the left- and right-hand states in Fig. 1. To get an idea of the transition state of such a pathway, preliminary all-atom MD simulations were run with quadruplexes of different lengths and sequences with explicit ions and water [27]. In the course of these tests, it became clear that a required transition state can be obtained by separating 5 bps at both ends of the quadruplex, which gives four B-DNA "paws" protruding from the core, and then keeping the paws wide open for the time necessary to relax the helical twist to that of B-DNA.

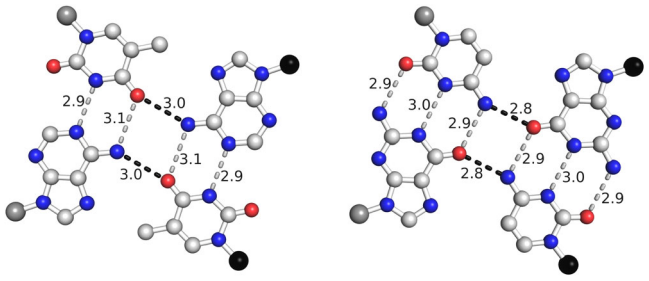

FIG. 2. Mutual recognition between identical WC base pairs via major-groove edges. Large spheres of two different colors correspond to C1′ atoms of different dsDNA. Geometry and stabilization energies (Table I) of individual base pairs and tetrads were evaluated by vacuum geometry optimizations at the MP2/6-311G(d) level of theory [27]. The computed H-bond lengths are shown in angstroms (Å). The secondary H bonds are distinguished by darker dashed lines.

The idea of the following MD simulations is similar to some early MD studies [50], and it is explained in Fig. 3. In all subsequent modeling, only GC-alternating sequences were used. We start from a predicted barrier state and try to reach both quadruplex and unbound states in free dynamics, without any guiding restraints. If we are lucky, the trajectories in both directions will go downhill on the energy landscape. Even in this case, however, a straightforward simulation requires enormous time resources. Therefore, to obviate entropic barriers, a Maxwell demon approach is applied. A bundle of trajectories is started from the same state, with different random velocities. All trajectories are followed visually and stopped after a certain time interval (usually about 2 ns) or when an interesting local transition towards dissociation or binding occurred somewhere within the bundle. One of the final states considered to be most advanced towards dissociation or folding is selected and used as the start of a new bundle of trajectories. After several iterations in the two opposite directions, one gets two trajectories leading to unbound and quadruplex states, respectively. The trajectories are continuous in the coordinate space, with velocities periodically randomized. By inverting one of them, we obtain a pathway between the terminal states in Fig. 1 that involves only elastic deformations and does not require a base pair opening.

TABLE I. Vacuum stabilization energies, U (kcal/mol), computed by the QC and MM methods described in the Supplemental Material [27]. The energy of the WC pairing was estimated as the difference between the vacuum energy of the pair and that of constituting nucleobases. For the secondary pairing, the energy is obtained as the difference between the tetrad energy and that of the constituting base pairs. The WC pairs are denoted by the standard two-letter code. Slashes denote the secondary paring. The U_{MM} values were computed with the AMBER force field [48]. For WC pairs, these results agree with earlier data [47].

Energy	GC	GC/ GC	AT	AT/AT
U_{QC}	29.5	26.4	16.2	12.6
U_{MM}	28.3	19.5	12.9	10.6

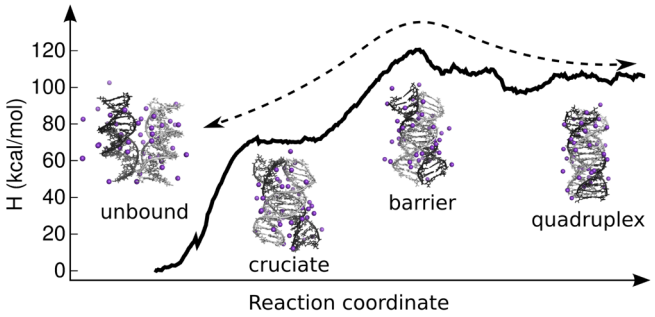

FIG. 3. The overall plan of MD simulations and the approximate energy profile along the transition pathway. Trajectories are started from a predicted barrier structure and continued in the two opposite directions. The reaction coordinate was constructed as explained in the text, with the energy profile smoothed by averaging with a sliding window of about 2 ns.

The time course of the production run is illustrated in Fig. 4. A more detailed picture is provided as animation files in the Supplemental Material [27]. Surprisingly, just 10 ns were necessary to reach both quadruplex and unbound states. During binding, only one complete additional tetrad was formed. A few ions and water molecules were sequestered between bases, which strongly complicated the formation of secondary H bonds. However, both secondary grooves, with characteristic chains of potassium ions between phosphates, were already formed. During dissociation the order of events was inverted; that is, dissociation of the tetrads preceded that of the ions and the groove opening. The starting intermediate was probably shifted towards the quadruplex state. None of the trajectories of the first bundle displayed strong trends towards dissociation, even though in two cases one boundary tetrad

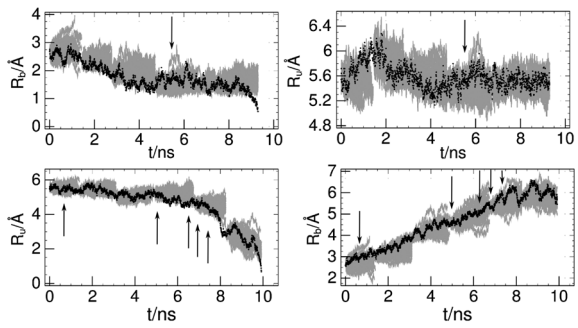

FIG. 4. Time traces of root mean square deviations (RMSDs). The RMSDs from bound (quadruplex) (R_b) and unbound (R_u) states (see Fig. 3) were computed for two double helices separately and averaged. The upper and lower panels display results for binding and dissociation, respectively. The gray bands are formed by traces of bundles of 32 trajectories computed as explained in the text. The restart points can be distinguished by narrowings of these bands. The black traces correspond to trajectories selected for continuation. The vertical arrows mark the formation and splitting of tetrads during folding and dissociation, respectively.

was split. The first of these trajectories selected for continuation towards the unbound state was not successful. The second choice worked; however, even in the fourth bundle there were trajectories that turned back to folding (see Fig. 4). The dissociation accelerated after the split of the three tetrads.

The energy profile shown in Fig. 3 was evaluated as follows. New trajectories were restarted similarly to the main run from 280 states equally spaced in time along the transition pathway, and the average total energy was evaluated for 0.5 ns after short equilibration. This profile is approximate and lacks the entropic contribution of the free energy, but it gives an estimate of shape and the order of magnitude of the values involved. Thorough calculations using umbrella sampling and the weighted histogram analysis would be prohibitively costly, and they usually give qualitatively similar profiles scaled down by 1 to 2 orders of magnitude [51]. The apparently large energies obtained are not prohibitive. First, all available data indicate that the recognition requires long incubation stages lasting from hours to weeks; therefore, the corresponding free energy barrier can well reach 10–20 kcal/mol. Second, the plateau at 70 kcal/mol mainly depends upon the type and concentration of ions. The neutralizing amount of monovalent ions used here was not meant to reproduce real conditions, which probably involve a combination of mono- and divalent ions with higher concentrations. Finally, the energy values in Fig. 3 would be much larger for alternative recognition models that include strand dissociation.

The plateau at 70 kcal/mol in Fig. 3 indicates that structures with three to four stacked tetrads can represent a metastable state with a local free energy minimum. In short quadruplexes, the tetrads are propeller twisted and slightly nonparallel, which allows the paws protruding from the core to be separated without strong bends. With the length of the tetrad stack increased, the tetrads become more parallel and a stronger bending in the paws is required. This explains the emergence of the plateau in Fig. 3 that can well transform into an energy minimum with stronger secondary H bonds corresponding to Table I. I suggest that, under appropriate ionic conditions, the free energy in this minimum is lower than that of the unbound state. In contrast, the additional bending strain responsible for the central energy barrier cannot be eliminated. With increased DNA length this barrier would broaden and eventually become a plateau. Under these assumptions, the cruciate structures with short quadruplexes of three to four stacked tetrads work as recognition units in a homologous alignment of long double helices.

Conformations of the cruciate units observed in the dynamics were used for predicting the complexes of long DNA (Fig. 5). They were built in several steps by combining all-atom and coarse grained modeling [27]. The cruciate shape of the recognition unit and the bending rigidity of the double helix impose significant limitations upon the

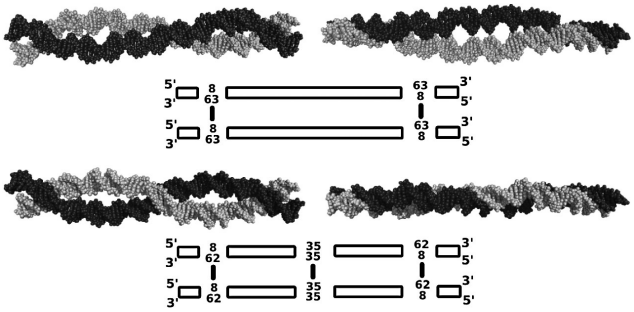

FIG. 5. Modeled complexes of long dsDNA obtained as explained in the text. The upper and lower panels demonstrate paranemic and plectonemic modes of binding, respectively. The two columns show two perpendicular views of each structure. In the schematics below the structures, dsDNA stretches are shown as rectangles interrupted by recognition units. For each unit the four bases of the central tetrad are indicated, with the secondary H bonding marked by thick vertical lines.

minimal axial separation between the pairing contacts. In the upper panels of Fig. 5, the two recognition units are separated by five helical turns. This number is identical in the two chains; therefore, the units share the same plane. For a smooth connection the intermediate helices take a particular sinuous shape. The high bending rigidity of DNA straightens these helices and pushes them close to each other against repulsive electrostatic forces. The helix-helix orientations are not optimal, according to earlier predictions and simulations [52,53], and the corresponding energy contributes to the high energy barrier of the binding. Even though the bend angles in the helical stretches are admissible, they are strained because all local bends must have concerted values and directions. With the growing distance between the recognition units, this strain is relieved and its energy can be compensated for by that of the binding. For complete homologies, the experimental recognition threshold of 200–300 bp probably corresponds to a relatively small number of specific contacts, which explains the drastic disappearance of recognition at these DNA lengths.

The upper and lower structures in Fig. 5, respectively, demonstrate paranemic and plectonemic contacts between two dsDNA. In the latter case the linear density of recognition units is two times higher. Plectonemes like that in Fig. 5 are possible when the two terminal recognition units are separated by an odd number of helical turns. In this case the middle unit is rotated by 180° and shifted by a Half-integer number of turns. In a similar plectoneme with an even number of helical turns, the central segments of the two double helices would face one another by their minor grooves. In this case, one can consider the possibility of a cruciate recognition unit formed by the hypothetical strand exchange mechanism [12]. The paranemic contacts can always be formed by loops protruding from chromosomes. In contrast, plectonemic contacts require the topoisomerase II (Top2) activity because otherwise every right-hand turn with three recognition units must be compensated for by a left-hand turn, with no recognition units and a very high entropic penalty. Interestingly, the loss or inhibition of Top2 was shown to partially compromise the pairing of homologous chromosomes in cell cultures [3]. It would be interesting to check whether this effect also plays a role in RIP.

Recognition via discrete units spaced by several helical turns represents an extension of one of the models considered by Gladyshev and Kleckner [20] and it sheds new light upon their data. Notably, the helical pitch of the quadruplex structure in Fig. 1 (12.9) is larger than that of B-DNA (10.5). Therefore, the average helical pitch in complexes shown in Fig. 5 grows with the density of the recognition units, which explains the higher RIP efficiency for sequences with periods of 11 and 12 bp as compared to 10 bp. The large distance between the recognition units makes the constraint upon the concerted twisting in the two double helices more stringent because the amplitude of thermal torsional fluctuations grows only as a square root of the chain length. Under normal temperature, the difference between 10- and 11-bp periodicities can be compensated for by thermal fluctuations for one helical turn, but not for five helical turns. In the latter case it corresponds to a rotation of 180°. This explains the strong differences in RIP activities for some periodicities that differ by only one bp. Finally, the sequence dependence of the secondary H bonding predicted by Table I might account for the examples of a strongly different RIP for homologies that differ only by the sequence [20].

Encounters between identical DNA sequences are rare in nature, but they should be very frequent *in vitro*; therefore, one may ask why the complexes shown in Fig. 5 remain almost unnoticed in chemical laboratories. In fact, they are perhaps long known, but discarded. DNA is never stored for hours and weeks in ionic solutions because it is known to slowly aggregate and deteriorate. Rare attempts to systematically study the slow evolution of DNA samples in laboratory conditions have given very perplexing results [54]. The sequence-specific association is driven by ions, with both mono- and divalent cations probably involved. The binding shown in Fig. 5 occurs due to reversible interactions that are likely to be destroyed during dilution or penetration through gels. At the same time, a small ion excess may lead to almost irreversible nonspecific complexation. These issues are not easy to sort out and they will require further experimental investigation.

In summary, the mutual recognition between two homologous B-DNA might occur due to direct complementary H bonding of major-groove surfaces in parallel alignment. The pairing of two dsDNA results in formation of a planar cross-shaped recognition unit, with a central quadruplex of three to four bps and four B-DNA paws protruding in opposite directions. In a complex of two dsDNA, the recognition units have to be spaced by at least several helical turns; therefore, the binding requires long double helices, but only partial homology. The recognition

units are separated from the unbound state by a high energy barrier and they are stabilized by specific H bonding as well as ion-DNA interactions. Therefore, the binding takes a very long time and is very sensitive to ionic conditions. The proposed mechanism and concomitant predictions agree with earlier data and shed light upon the recent intriguing experimental results [20].

The computational resources used in this study were supported by the Initiative d'Excellence program through French State Grant No. ANR-11-LABX-0011-01 (DYNAMO).

[1] N. Kleckner and B. M. Weiner, Cold Spring Harbor Symp. Quant. Biol. **58**, 553 (1993).
[2] B. D. McKee, Biochim. Biophys. Acta **1677**, 165 (2004).
[3] B. R. Williams, J. R. Bateman, N. D. Novikov, and C. T. Wu, Genetics **177**, 31 (2007).
[4] A. T. Hagemann and E. U. Selker, Cold Spring Harbor Monogr. Arch. **32**, 586 (1996).
[5] E. U. Selker, Adv. Genet. **46**, 439 (2002).
[6] J. L. Rossignol and G. Faugeron, Experientia **50**, 307 (1994).
[7] A. Barzel and M. Kupiec, Nat. Rev. Genet. **9**, 27 (2008).
[8] J. R. Bateman and C. T. Wu, Genetics **180**, 1329 (2008).
[9] J. P. Blumenstiel, R. Fu, W. E. Theurkauf, and R. S. Hawley, Genetics **180**, 1355 (2008).
[10] E. F. Joyce, B. R. Williams, T. Xie, and C. T. Wu, PLoS Genet. **8**, e1002667 (2012).
[11] S. McGavin, J. Mol. Biol. **55**, 293 (1971).
[12] J. H. Wilson, Proc. Natl. Acad. Sci. U.S.A. **76**, 3641 (1979).
[13] D. Zickler and N. Kleckner, Cold Spring Harbor Persp. Biol. **7**, a016626 (2015).
[14] A. A. Kornyshev and S. Leikin, Phys. Rev. Lett. **86**, 3666 (2001).
[15] A. A. Kornyshev, Phys. Chem. Chem. Phys. **12**, 12352 (2010).
[16] N. C. Seeman, Nano Lett. **1**, 22 (2001).
[17] X. Wang, X. Zhang, C. Mao, and N. C. Seeman, Proc. Natl. Acad. Sci. U.S.A. **107**, 12547 (2010).
[18] G. S. Baldwin, N. J. Brooks, R. E. Robson, A. Wynveen, A. Goldar, S. Leikin, J. M. Seddon, and A. A. Kornyshev, J. Phys. Chem. B **112**, 1060 (2008).
[19] C. Danilowicz, C. H. Lee, K. Kim, K. Hatch, V. W. Coljee, N. Kleckner, and M. Prentiss, Proc. Natl. Acad. Sci. U.S.A. **106**, 19824 (2009).
[20] E. Gladyshev and N. Kleckner, Nat. Commun. **5**, 3509 (2014).
[21] Z. Shen, H. Yan, T. Wang, and N. C. Seeman, J. Am. Chem. Soc. **126**, 1666 (2004).
[22] P. O. Löwdin, in *Electronic Aspects of Biochemistry*, edited by B. Pullman (Academic Press, New York, 1964), p. 167.
[23] H. E. Kubitschek and T. R. Henderson, Proc. Natl. Acad. Sci. U.S.A. **55**, 512 (1966).
[24] A. Lebrun and R. Lavery, J. Biomol. Struct. Dyn. **13**, 459 (1995).
[25] A. Kettani, R. A. Kumar, and D. J. Patel, J. Mol. Biol. **254**, 638 (1995).
[26] N. Zhang, A. Gorin, A. Majumdar, A. Kettani, N. Chernichenko, E. Skripkin, and D. J. Patel, J. Mol. Biol. **312**, 1073 (2001).
[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.116.158101, which includes Refs. [28–46], for additional details, comments, protocols of computations, and trajectory movies.
[28] A. R. Srinivasan, R. R. Sauers, M. O. Fenley, A. H. Boschitsch, A. Matsumoto, A. V. Colasanti, and W. K. Olson, Biophys. Rev. Lett. **1**, 13 (2009).
[29] A. A. Granovsky, FIREFLY version 8.1.0, http://classic.chem.msu.su/gran/firefly.
[30] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. J. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su *et al.*, J. Comput. Chem. **14**, 1347 (1993).
[31] I. S. Joung and T. E. Cheatham, III, J. Phys. Chem. B **112**, 9020 (2008).
[32] J. Wang, P. Cieplak, and P. A. Kollman, J. Comput. Chem. **21**, 1049 (2000).
[33] A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, Biophys. J. **92**, 3817 (2007).
[34] M. Zgarbova, F. J. Luque, J. Sponer, T. E. Cheatham, M. Otyepka, and P. Jurecka, J. Chem. Theory Comput. **9**, 2339 (2013).
[35] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).
[36] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[37] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[38] A. K. Mazur, J. Comput. Chem. **18**, 1354 (1997).
[39] A. K. Mazur, J. Chem. Phys. **111**, 1407 (1999).
[40] A. K. Mazur, J. Am. Chem. Soc. **120**, 10928 (1998).
[41] A. K. Mazur, J. Phys. Chem. B **102**, 473 (1998).
[42] A. K. Mazur, Biophys. J. **91**, 4507 (2006).
[43] A. K. Mazur, Phys. Rev. Lett. **105**, 018102 (2010).
[44] A. K. Mazur, Phys. Rev. E **84**, 021903 (2011).
[45] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked *et al.*, J. Mol. Biol. **313**, 229 (2001).
[46] A. K. Mazur, J. Phys. Chem. B **113**, 2077 (2009).
[47] J. Sponer, P. Jurecka, and P. Hobza, J. Am. Chem. Soc. **126**, 10142 (2004).
[48] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).
[49] Z. Bryant, M. D. Stone, J. Gore, S. B. Smith, N. R. Cozzarelli, and C. Bustamante, Nature (London) **424**, 338 (2003).
[50] S. C. Harvey and H. A. Gabb, Biopolymers **33**, 1167 (1993).
[51] L. V. Bock, C. Blau, G. F. Schroder, I. I. Davydov, N. Fischer, H. Stark, M. V. Rodnina, A. C. Vaiana, and H. Grubmuller, Nat. Struct. Mol. Biol. **20**, 1390 (2013).
[52] P. Varnai and Y. Timsit, Nucleic Acids Res. **38**, 4163 (2010).
[53] R. Cortini, A. A. Kornyshev, D. J. Lee, and S. Leikin, Biophys. J. **101**, 875 (2011).
[54] G. P. Brewood, J. J. Deltow, and J. M. Schurr, Biochemistry **49**, 3367 (2010).