DNA Shape Dominates Sequence Affinity in Nucleosome Formation

Gordon S. Freeman,[1] Joshua P. Lequieu,[2] Daniel M. Hinckley,[1] Jonathan K. Whitmer,[1,2,3] and Juan J. de Pablo[2,3,*]
[1]*Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA*
[2]*Institute for Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA*
[3]*Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*
(Received 3 December 2013; revised manuscript received 29 July 2014; published 14 October 2014)

Nucleosomes provide the basic unit of compaction in eukaryotic genomes, and the mechanisms that dictate their position at specific locations along a DNA sequence are of central importance to genetics. In this Letter, we employ molecular models of DNA and proteins to elucidate various aspects of nucleosome positioning. In particular, we show how DNA's histone affinity is encoded in its sequence-dependent shape, including subtle deviations from the ideal straight B-DNA form and local variations of minor groove width. By relying on high-precision simulations of the free energy of nucleosome complexes, we also demonstrate that, depending on DNA's intrinsic curvature, histone binding can be dominated by bending interactions or electrostatic interactions. More generally, the results presented here explain how sequence, manifested as the shape of the DNA molecule, dominates molecular recognition in the problem of nucleosome positioning.

DOI: 10.1103/PhysRevLett.113.168101 PACS numbers: 87.15.-v, 87.10.Tf, 87.14.gk, 87.18.Wd

The human genome is comprised of DNA molecules whose total contour length is on the order of 1 m. These molecules are efficiently packed as chromatin within eukaryotic cell nuclei of only a few microns in diameter, primarily through protein-bound complexes called nucleosomes. Such complexes consist of approximately 147 base pairs (bps) of DNA encircling a disklike protein (the histone core). Negatively charged DNA experiences a strong electrostatic attraction to the positively charged histone surface. Protein-bound sites along DNA present barriers to transcription; thus, their positioning is a crucial element in the regulation of cellular function for all eukaryotic species [1–3]. In spite of being central to biology, the molecular cues that determine nucleosome binding are not fully understood [4,5]. Experimental evidence produced over the last few years suggests that nucleosome preference is directly encoded by DNA [6–9]. Different sequence motifs possess unique structural properties—intrinsic curvature, minor groove dimensions, and local flexibility—that render them more or less favorable for protein binding and nucleosome formation.

Recent studies of these structural properties have considered them individually, and several views exist of the physical origins of nucleosome positioning. One such view assumes that sequence effects on nucleosome formation can be distilled into physical variables, such as intrinsic curvature and associated deformation penalties [10–14]. This approach has been used to explain both *in vitro* affinity data and *in vivo* nucleosome positioning maps [10–13,15]. Another view attributes nucleosome affinity to the double helix's minor groove width (MGW) profile [16–18]; DNA segments that exhibit narrower minor grooves at protein contacts are thought to bind more robustly to histone residues via enhanced electrostatic attractions. A third view assumes that local mechanical flexibility dictates nucleosome formation, with DNA molecular shape or curvature being unimportant. Base stacking and "bendability" analyses from this approach lead to a simple and elegant nucleosome positioning template [19]. The seemingly disparate origins, assumptions and relative successes of these hypotheses have often been at odds, and a unified description of nucleosome positioning that reconciles different views is sorely missing. A comprehensive molecular-level analysis of nucleosome positioning that includes curvature, local mechanical flexibility, and electrostatic interactions, has never been pursued.

In this Letter, we reconcile these differing hypotheses and determine the dominant mechanism for sequence-dependent nucleosome formation. We do so by relying on a detailed model of DNA [20] and the proteins [21] [cf. Fig. 1(a)]. By incorporating the DNA properties relevant to each hypothesis into a single model [namely (1) sequence-dependent flexibility, (2) sequence-dependent intrinsic curvature, and (3) sequence-dependent minor groove widths and protein-DNA electrostatics [22]] we identify the structural properties most critical to nucleosome formation. The free energies associated with histone binding are extracted from purely molecular information, and are emergent properties of the model. To the best of our knowledge, the results presented here constitute the only available predictions of the Helmholtz free energy of a nucleosome complex, and, as such, they serve to dissect the relative contributions of different interactions into the overall assembly process. Our results indicate that DNA's shape—encoded through its sequence—is a dominant factor in determining a sequence's nucleosome affinity,

with local mechanical flexibility playing a secondary role. We show that sequence-dependent minor groove width works in concert with intrinsic curvature to dictate molecular interactions, and, depending on the intrinsic curvature of a particular sequence, binding can be dominated by bending energy or by electrostatic energy.

The sequences whose nucleosome affinities are calculated here are taken from published experimental data [7,9,24], and are given partially in Fig. 1(c) and, in greater detail, in the Supplemental Material [22]. Reference sequences c1, d1, and TG are chosen to facilitate comparisons to experimental data—sequences c2 and c3 are compared to c1 [7], sequences d2 and d3 are compared to d1 [7], and sequences TGRC, TG-T, TGGA1, EXAT, and IAT are compared to TG [9,24] (sequence names are consistent with those in the original publications). Thermodynamic integration [25] is employed to determine the difference in nucleosome formation free energy, $\Delta\Delta A$, between a DNA sequence and its reference sequence. Larger positive values of $\Delta\Delta A$ are indicative of weaker affinity of a DNA sequence for the histone. Model predictions for relative nucleosome affinity, shown in Fig. 2(a) and tabulated in Table S2 [22], are in agreement with experiment ($P < 0.002$, $N = 9$), serving to establish the validity of the nucleosome model. Note that free energies reported in the figure are on the order of several k_BT, and represent a delicate balance between bending, torsion, van der Waals, and electrostatic contributions to the free energy.

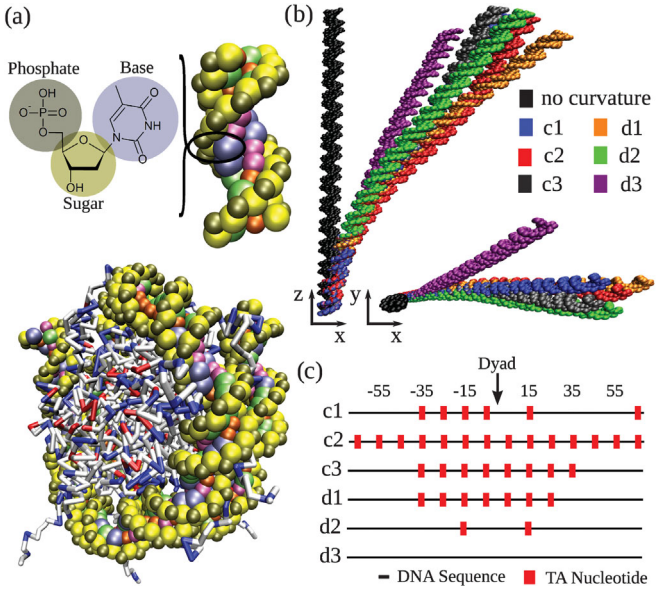

FIG. 1 (color). DNA sequence plays a crucial role in determining nucleosome affinity. (a) Coarse-graining scheme and representative nucleosome configuration. DNA is modeled using three sites per nucleotide [20]; the histone is represented with one site per amino acid, located at the center-of-mass of the side chain [21]. Nucleosomal configurations are obtained by mapping our coarse model onto the 1KX5 crystal structure [23] and then performing molecular dynamics. Note that 1KX5 is only used to provide an initial condition; no information from 1KX5 is encoded into our model. (b) Minimum-energy structure of c1, c2, c3, d1, d2, and d3 DNA sequences used in study. Differences in DNA sequence result in large variations of intrinsic curvature. A sequence with no curvature is shown for reference. (c) Sequences used in study (as represented in Ref. [7]). Red blocks denote TA nucleotides which are known to enhance DNA orientational preference.

Experiments have established that specific motifs, spaced apart by the pitch of double-stranded DNA (~10 bps), direct binding through orientational preferences in nucleosome-bound DNA [26]. For example, the strongest positioning sequences in Refs. [7] and [9] possess TA-rich motifs flanking GC-rich sequences at ~10 bp intervals [Fig. 1(c)]. Alternating TA-rich/GC-rich patterns observed by earlier studies result in alternating regions of narrow and wide minor groove widths, respectively [18]. This is consistent with recent analyses that positively charged protein residues (lysine, histidine, arginine) interact favorably with strongly negatively charged pockets created by the phosphates in a narrow minor groove [16–18]. An additional effect of periodic narrow minor grooves is a net curvature on the shape of the DNA [Fig. 1(b)].

Metadynamics simulations [27] using the minor groove orientation at these TA-rich motifs as an order parameter (S_{ROT}) permit direct examination of the thermodynamic forces that drive orientational preference. The order parameter (S_{ROT}) is defined as

$$S_{\mathrm{ROT}} = \left\langle \pm \arccos\left(\frac{\boldsymbol{P}\cdot\boldsymbol{B}}{\|\boldsymbol{P}\|\|\boldsymbol{B}\|}\right)\right\rangle,$$

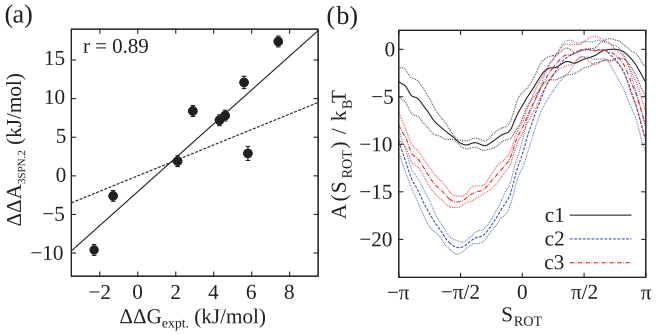

FIG. 2 (color). (a) Predicted and experimental binding free energy. The trends predicted for nucleosome binding affinity in simulations are consistent with experiments from Refs. [7,9], and [24]. The dashed line is a guide to the eye corresponding to exact agreement. (b) Binding free energy as a function of DNA orientation around the histone core. Increasing the number of favorable dinucleotide binding motifs enhances DNA orientational preference. Sequences ranked by number of favorable motifs are c2 > c3 > c1. The location of the free energy minima is consistent with the narrow minor groove at TA-rich motifs pointing inward toward the protein core.

where **B** is a vector from the center of a given base step on the sense strand to its complementary base step on the antisense strand, **P** is a vector from the center of these two base steps to the center of the protein, and the angle brackets denote an average over base steps at the −15, −5, +5, and +15 positions relative to the dyad [cf. Fig. 1(c)]. The positive sign is chosen if $(\boldsymbol{P}\times\boldsymbol{B})\cdot\boldsymbol{D}\leq 0$ (negative if >0), where $\boldsymbol{D}$ is a vector in the 5' to 3' direction along the sense strand. Notably, when $S_{\mathrm{ROT}}=-\pi/2$, the minor groove is oriented toward the protein core, and when $S_{\mathrm{ROT}}=\pi/2$, it is oriented away from it.

Figure 2(b) shows the effect of modifying the number of TA-rich motifs on DNA orientational preference for a subset of the sequences in Fig. 2(a). Sequences c1, c2, and c3, which possess TA-rich motifs separated by ∼10 bps, orient the minor groove toward the protein core at these motifs, as indicated by the free energy minima at $S_{\mathrm{ROT}}\sim-\pi/2$. Furthermore, the strength of this preference is determined by the number of favorable positioning motifs; the number of TA-rich motifs and the corresponding depth of the rotational orientation free energy minima are arranged as c2 > c3 > c1, consistent with their relative affinities for nucleosome formation. This trend is also apparent in d1, d2, and d3, which progressively purge TA motifs (Fig. S6 [22]).

Why, then, do specific sequence motifs enhance the orientational preference? We address this question by analyzing the three proposed mechanisms for nucleosome formation: intrinsic curvature, minor groove dimension, and local flexibility. As alluded to earlier, an important consequence of sequence motifs that favor nucleosome formations is the enhanced intrinsic curvature of the DNA molecule. Figure 3(a) shows the intrinsic curvature, $\langle A_f^0\rangle$, of each sequence we study here, and demonstrates that a correlation exists between nucleosome binding affinity and increasing intrinsic curvature. This result is understood through the deformation penalty incurred by nucleosomal DNA, related to its deviation from the unbound equilibrium configuration [15]; sequences with greater intrinsic curvature are believed to require less deformation to bind to the histone core than more intrinsically straight sequences.

A second important consequence of motif spacing is the narrowing of the minor grooves at locations facing the histone protein. The width of a minor groove is believed to be inversely proportional to the strength of its interaction with positively charged residues on the histone surface [16]. Figure 3(b) shows, for sequences TG, TG-T, and TRGC, the minor groove width profiles (and the corresponding Fourier transformed intensity) from direct molecular simulations of the bulk sequences using 3SPN.2C. In particular, the minor grooves in TG ($\Delta\Delta A=0$ kJ/mol) are significantly narrower than those in TG-T ($\Delta\Delta A=17.4$ kJ/mol). Note that sequence TG also exhibits the strongest periodicity at 10 bps, which matches the pitch of DNA and optimizes favorable

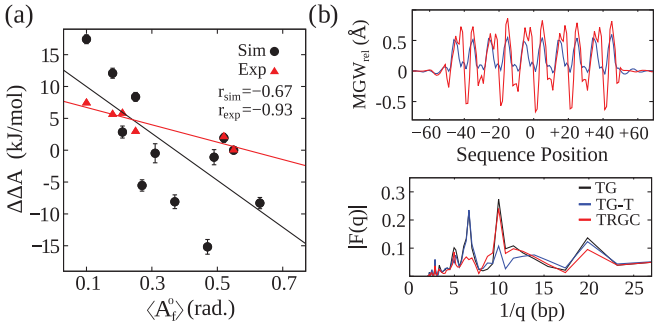

FIG. 3 (color). The role of intrinsic curvature and minor groove width. (a) Free energies of nucleosome formation, relative to sequence TG, decrease with increasing intrinsic curvature, $\langle A_f^0\rangle$, in both simulation (black circles) and experiment (red triangles) [9]. To compare all sequences, $\Delta\Delta A$ values were calculated for both c1 and d1 relative to TG; sequences c1–c3 and d1–d3 are omitted from the experimental points due a lack of analogous experimental data. (b) Affinity may be understood through the minor groove width profile, shown for sequences TG (black lines, $\Delta\Delta A=0$ kJ/mol), TG-T (blue lines, $\Delta\Delta A=17.4$ kJ/mol) and TRGC (red lines, $\Delta\Delta A=12.1$ kJ/mol). The top panel shows relative minor groove widths ($MGW_{rel}=MGW\text{-}MGW_{TG}$), and the bottom panel shows the corresponding Fourier transformed intensity, $|F(q)|$.

electrostatic interactions with the protein core, as described above. TRGC also exhibits 10 bp periodicity, and its minor grooves are narrower than those of TG ($MGW_{rel}<0$). Its binding affinity, however, is notably weaker than that of TG ($\Delta\Delta A=12.1$ kJ/mol). As we will soon demonstrate, analysis of minor groove dimensions alone is insufficient to explain differences in affinity between certain sequences (e.g., TG and TRGC).

A third hypothesis posits that nucleosome formation is guided by the local flexibility of DNA. To test this hypothesis, it is instructive to introduce two variations of our model [Fig. 4(a)]. The first, labeled "*S*" for "straight", assigns sequence-dependent flexibility to a sequence-independent shape specified by ideal B-form DNA, with no sequence-dependent curvature. That is, model *S* is intrinsically straight, regardless of the underlying sequence [e.g., the black molecule in Fig. 1(b)]. The second, labeled "*H*" for "homogeneous," maps sequence-agnostic energy parameters onto a sequence-dependent equilibrium shape. That is, model *H* possesses an identical shape to the full 3SPN.2C model, but all bonded and nonbonded interactions are the same, regardless of the underlying sequence [22]. Nucleosomal configurations for the *H* and *S* models were obtained as described in Fig. 1(a), and were found to be stable for all sequences [22].

Repeating the analysis of Fig. 2(a), we calculate the binding affinity of DNA sequences using the *S* and *H* models [Fig. 4(b)]. The results are compelling: model *H* is consistent with experimental nucleosome formation free energies [as is the full model, labeled "*A*" for "all,"

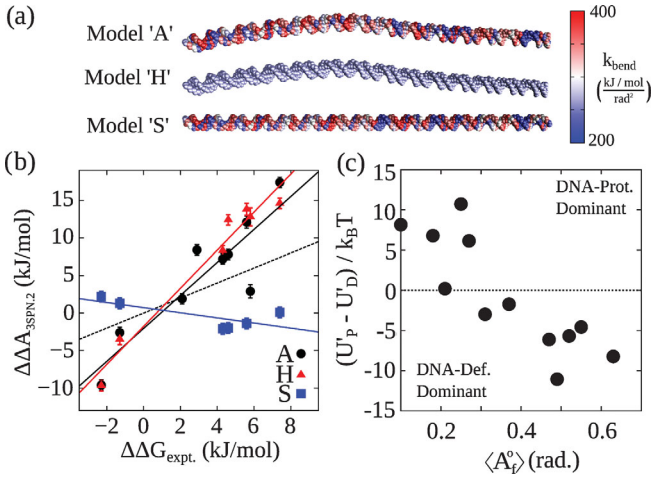

FIG. 4 (color). Identification of the dominant mechanism in nucleosome formation. (a) Schematic Description of Models "A," "H," and "S" for sequence c1. The shape corresponds to the minimum energy configuration. Colors represent local base step flexibility (blue: flexible, red: stiff), as measured by a local, base-step dependent bending constant k_{bend} [22]. (b) Predicted and experimental binding free energy for each model. Correlation coefficients for best fit lines are $r_A = 0.89$, $r_H = 0.98$, $r_S = -0.74$. The dashed line is a guide to the eye corresponding to exact agreement. (c) Relative importance of DNA-deformation and DNA-protein energy contributions. The mechanism of nucleosomal assembly is a strong function of sequence.

reproduced from Fig. 2(a)]. Model S, on the other hand, exhibits an opposite trend. These findings reveal the equilibrium shape of the unbound DNA molecule as the dominant factor determining nucleosomal sequence affinity, with the underlying mechanical properties that drive local flexibility, base stacking, base pairing, and cross stacking playing a secondary role.

Having shown that DNA shape is of central importance, we now seek to decouple the two remaining hypotheses: Is intrinsic curvature or are favorable minor groove interactions most important in nucleosome formation? To answer this, we use the A and S models to examine the balance between bending energy and electrostatic energy for different sequences. Specifically, energetic contributions to the free energy as a function of S_{ROT} are separated into DNA deformation (i.e., the penalty required to bend the DNA into the nucleosomal superhelix), $\langle U_D \rangle$, and DNA-protein interaction (which is primarily electrostatic), $\langle U_P \rangle$, for each sequence in our Letter. The corresponding averages are defined as $\langle U_i \rangle = \int U_i(S_{\mathrm{ROT}}) P(S_{\mathrm{ROT}}) \mathrm{d}S_{\mathrm{ROT}}$, where $i \in \{D, P\}$, $P(S_{\mathrm{ROT}}) = e^{-\beta A(S_{\mathrm{ROT}})}/Q$, Q is a normalizing factor, and $A(S_{\mathrm{ROT}})$ is the free energy of the full model [cf. Fig. 2(b)] for the given sequence. The primary quantity of interest is $U'_i = \langle U_i^S \rangle - \langle U_i^A \rangle$ (superscript denotes model S or A). The determination of U'_i permits the comparison of deformation energy and electrostatic energy in nucleosome formation for a given sequence relative to a reference state that lacks sequence dependent curvature and minor groove profile (i.e., the S model). Thus, U'_i indicates the importance of the ith energy contribution to nucleosome formation.

Figure 4(c) plots $\Delta U' = U'_P - U'_D$ as a function of intrinsic curvature, $\langle A_f^0 \rangle$, for each sequence in the study. The dominant mechanism depends dramatically on the sequence examined. In one regime, DNA-protein interactions dominate ($\Delta U' > 0$), and thus, DNA-protein contacts (as mediated by minor groove width) are most critical. In the other, DNA-deformation interactions are most important ($\Delta U' < 0$). Further, we demonstrate that the regime of relevance is strongly tied to the curvature of the sequence. This result reconciles the successes of both the minor groove width and intrinsic curvature hypotheses. For sequences with low intrinsic curvature, subtle deviations in DNA shape are sufficient to present minor grooves in the optimal manner to the histone protein, thereby dictating sequence affinity. For sequences with higher curvature, it is, instead, the intrinsic curvature of the DNA that imparts preferential positioning. Thus, both intrinsic curvature and minor groove dimensions play a critical role in nucleosome formation.

Our results reconcile prevailing viewpoints in the literature, often appearing to be conflictive, which state, alternately, that intrinsic curvature drives nucleosome affinity [12,13,15], or that it is driven by variations in minor groove width and associated electrostatic interactions that depend on the underlying sequence [16–18,26]. By relying on a coarse-grained but realistic representation of the nucleosome complex, we have been able to generate high-precision estimates of the free energy of binding. Analysis of those free energies show that, depending on a sequence's curvature, binding is dominated by bending penalties or by electrostatic interactions. We have also demonstrated that these characteristics of DNA work hand in hand to dictate nucleosome affinity and that the local, sequence-dependent flexibility of the DNA molecule plays a minor role. These connections emerge naturally from an accurate description of the underlying molecular interactions at the relevant length scales. Further, we have provided a mechanistic explanation for the role of sequence in dictating histone binding preference. Our results demonstrate that not only does DNA shape produce these phenomena, but accurate shape alone is a necessary and sufficient component to describe histone binding affinities. Our results are important beyond the problem of nucleosome positioning; indeed, there are many instances of DNA-protein interactions in which the shape of the DNA molecule is a critical component. Better understanding the role of DNA shape in molecular recognition will enable rational design of effective DNA-binding elements for use in therapeutic devices and genetic engineering.

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences,

Materials Sciences and Engineering Division. G. S. F., D. M. H., and J. K. W. were partially supported by a NHCRI training grant to the Genomic Sciences Training Program, T32HC002760. D. M. H. was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256259. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. The authors gratefully acknowledge Professor Anita Scipioni for assistance in calculating intrinsic DNA curvature, the University of Chicago Research Computing Center, the UW-Madison Center for High Throughput Computing, and the Open Science Grid for computing resources.

*depablo@uchicago.edu

[1] C. Jiang and B. F. Pugh, Nat. Rev. Genet. **10**, 161 (2009).
[2] J. J. Wyricki, F. C. Holstege, E. G. Jennings, H. C. Causton, D. Shore, M. Grunstein, E. S. Lander, and R. A. Young, Nature (London) **402**, 418 (1999).
[3] C. Viallant, L. Palmeira, G. Chevereau, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo, Genome Res. **20**, 59 (2010).
[4] N. Kaplan, I. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, Nat. Struct. Mol. Biol. **17**, 918 (2010).
[5] Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl, Nat. Struct. Mol. Biol. **17**, 920 (2010).
[6] K. Struhl and E. Segal, Nat. Struct. Mol. Biol. **20**, 267 (2013).
[7] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. C. Thåström, Y. Field, I. K. Moore, J. P. Z. Wang, and J. Widom, Nature (London) **442**, 772 (2006).
[8] K. Brogaard, L. Xi, J. P. Wang, and J. Widom, Nature (London) **486**, 496 (2012).
[9] T. E. Shrader and D. M. Crothers, Proc. Natl. Acad. Sci. U.S.A. **86**, 7418 (1989); J. Mol. Biol. **216**, 69 (1990).
[10] V. Miele, C. Vaillant, Y. d'Aubenton Carafa, C. Thermes, and T. Grange, Nucleic Acids Res. **36**, 3746 (2008).
[11] A. V. Morovoz, K. Fortney, D. A. Gaykalova, V. M. Studitsky, J. Widom, and E. D. Siggia, Nucleic Acids Res. **37**, 4707 (2009).
[12] C. Vaillant, B. Audit, and A. Arneodo, Phys. Rev. Lett. **99**, 218103 (2007).
[13] G. Chevereau, L. Palmeira, C. Thermes, A. Arneodo, and C. Vaillant, Phys. Rev. Lett. **103**, 188103 (2009).
[14] N. B. Becker and R. Everaers, Structure **17**, 579 (2009).
[15] A. Scipioni, S. Pisano, C. Anselmi, M. Savino, and P. De Santis, Biophys. Chem. **107**, 7 (2004).
[16] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, Nature (London) **461**, 1248 (2009).
[17] E. P. Bishop, R. Rohs, S. C. Parker, S. M. West, P. Liu, R. S. Mann, B. Honig, and T. D. Tullius, ACS Chem. Biol. **6**, 1314 (2011).
[18] S. West, R. Rohs, R. S. Mann, and B. Honig, J. Biomol. Struct. Dyn. **27**, 861 (2010).
[19] E. N. Trifonov, Phys. Life Rev. **8**, 39 (2011).
[20] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo, J. Phys. Chem. **139**, 144903 (2013); G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer, and J. J. de Pablo, J. Phys. Chem. (in press).
[21] W. Li, P. G. Wolynes, and S. Takada, Proc. Natl. Acad. Sci. U.S.A. **108**, 3504 (2011).
[22] See Supplemental material at http://link.aps.org/supplemental/10.1103/PhysRevLett.113.168101 for additional simulation details and results.
[23] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, J. Mol. Biol. **319**, 1097 (2002).
[24] H. Cao, H. R. Widlund, T. Simonsson, and M. Kubista, J. Mol. Biol. **281**, 253 (1998).
[25] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, San Diego, 2002), pp. 168–172.
[26] H. R. Drew and A. A. Travers, J. Mol. Biol. **186**, 773 (1985).
[27] A. Laio and F. L. Gervasio, Rep. Prog. Phys. **71**, 126601 (2008).