Can molecular dynamics be used to simulate biomolecular recognition? ⊘

Special Collection: Machine Learning Hits Molecular Simulations

Malin Lüking [ORCID] ; David van der Spoel [ORCID] ; Johan Elf [ORCID] ; Gareth A. Tribello ✉ [ORCID]

[Check for updates]

🌐 View Online ↗ Export Citation CrossMark

Articles You May Be Interested In

Theoretical and experimental analysis of the forced LacI-AraC oscillator with a minimal gene regulatory model

Chaos (June 2013)

Strong influence of periodic boundary conditions on lateral diffusion in lipid bilayer membranes

J. Chem. Phys. (October 2015)

Math Machines: Using Actuators in Physics Classes

Phys. Teach. (January 2018)

Can molecular dynamics be used to simulate biomolecular recognition?

View Online   Export Citation   CrossMark

Malin Lüking,[1] [iD]   David van der Spoel,[1] [iD]   Johan Elf,[1] [iD]   and Gareth A. Tribello[2,a)] [iD]

AFFILIATIONS

[1] Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, SE-75124 Uppsala, Sweden
[2] Centre for Quantum Materials and Technologies, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, United Kingdom

Note: This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.
[a)] Author to whom correspondence should be addressed: g.tribello@qub.ac.uk

ABSTRACT

There are many problems in biochemistry that are difficult to study experimentally. Simulation methods are appealing due to direct availability of atomic coordinates as a function of time. However, direct molecular simulations are challenged by the size of systems and the time scales needed to describe relevant motions. In theory, enhanced sampling algorithms can help to overcome some of the limitations of molecular simulations. Here, we discuss a problem in biochemistry that offers a significant challenge for enhanced sampling methods and that could, therefore, serve as a benchmark for comparing approaches that use machine learning to find suitable collective variables. In particular, we study the transitions LacI undergoes upon moving between being non-specifically and specifically bound to DNA. Many degrees of freedom change during this transition and that the transition does not occur reversibly in simulations if only a subset of these degrees of freedom are biased. We also explain why this problem is so important to biologists and the transformative impact that a simulation of it would have on the understanding of DNA regulation.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0146899

INTRODUCTION

The energy landscape of biological molecules often has multiple low energy states that are separated by large barriers. Interesting phenomena, such as protein folding, typically involve the molecule transitioning between these low energy states.[1] Simulating these phenomena using molecular dynamics (MD) is thus challenging as the barriers between states prevent the system from exploring all of the configuration space during the relatively short timescales that are easily accessible on computers.[2] Typically, then only those configurations in a locally ergodic region in the vicinity of the input structure are visited.

Several methods have been put forward to overcome the sampling problem described in the previous paragraph. Furthermore, several of these methods are mature and can be routinely used to extract free energies and kinetics for a wide variety of biological applications.[3–10] Many of these methods are reliant on identifying a small number of collective variables (CVs) that describe the transitions. Having suitable CVs is essential when it comes to understanding the results that have been obtained from simulations that

sample large configurational spaces. Furthermore, in many methods, the CVs play an active role in ensuring a more complete sampling of configuration space. For example, in some methods, the temperature of the collective variables[11] are raised and in others a bias that is a function of the collective variables is added.[12,13] Regardless, the net effect of both these types of simulation is to increase the amount of sampling along the CVs. The sampling along other degrees of freedom, meanwhile, is the same as it would be in an unbiased MD simulation.

It is often difficult to identify suitable CVs for running enhanced sampling calculations. Therefore, as is discussed in this special edition, there is interest in using machine learning algorithms to automate the process by which suitable coordinates for analyzing and biasing molecular dynamics simulations are identified.[14–17] However, despite all of this effort, none of the new algorithms have really taken off in the way that older methods, such as metadynamics or umbrella sampling, that use a small number of CVs have.

A possible problem with these new algorithms is that identifying CVs that can be used with the older and simpler methods is not particularly difficult for many of the problems that are actively

being studied. As an example, consider trying to simulate that quintessential rare event nucleation. It is relatively easy to think of CVs that measure the degree of crystallinity in a sample. One can thus find multiple simulation papers where the number of solid-like atoms is forced to increase using metadynamics or some other rare event method.[18–20] Nucleation can thus be studied using enhanced sampling methods that are well established and machine-learned collective variables are not required.

At the other extreme, consider trying to simulate a process such as biomineralization.[21] There is relatively little theory for describing how biological molecules interact with growing inorganic crystals. For problems like these, even if we could use a machine learning algorithm to enhance the sampling, we have no suitable theoretical framework to make sense of what was observed.

When it comes to identifying a problem that can only be solved using free energy methods that employ some form of machine learning, we thus find ourselves with something of a conundrum. If we are to use one of these algorithms, the problem we are endeavoring to solve must be simple enough that we have a body of theory that can be used to analyze any results that we obtain. At the same time, however, the problem must also be not so simple that this body of theory offers us an obvious set of collective coordinates to bias.

This paper describes an interesting problem that we believe satisfies the criteria that have been outlined earlier. As we will explain in the following paragraphs, the experimental evidence on the binding of proteins to DNA leads one to believe that this system must undergo transitions between certain states that have a well-defined physiological function. There are, however, as we will explain, many degrees of freedom that could be involved in this transition. We hope that by using machine learning researchers can untangle the effects that all these competing degrees of freedom have on the process via which the *lac* repressor binds to its specific DNA operator. We would therefore invite all researchers who have developed free energy methods that employ machine learning to attempt to solve the problem that we will lay out over the following paragraphs. If members of this research community rise to this challenge, we believe this will provide a way of benchmarking and comparing these new algorithms. Furthermore, these simulations offer a pathway to resolving longstanding problems about how the *lac* repressor identifies its specific binding sequence.

In the following sections, we first provide some background information on the problem. We then discuss results from unbiased simulations of the specifically bound complex. This part is followed by a discussion of a set of restrained MD simulations that use information from nuclear magnetic resonance (NMR) experiments to extract information on the structure of the non-specifically bound complex. We show that a number of collective variables have very different values in these two sets of simulations and the NMR structure. However, the transition between the two bound states cannot be driven by adding a metadynamics bias to these variables. We thus conclude that some more sophisticated analysis of the unbiased and NMR-restraint simulations is required. In future, it is hoped that a suitable, low-dimensional descriptor that can be used in an enhanced sampling calculation can be extracted through such analysis.

BACKGROUND

Genetic regulation and the example of the lac operon

The biomolecular system that we discuss in this paper is responsible for genetic regulation. Gene regulatory elements were first described by Jacob and Monod in a paper from 1961.[22] They describe the process via which a group of proteins that import and a galacoside called metabolize lactose, which is used as an energy source in *Escherichia coli*, are synthesized only if galactosides are present. To understand how this process works, it is important to understand the flow of information from genes to proteins. In accordance with the central dogma,[23] information is transferred from DNA and used to synthesize proteins via a messenger (today known as messengerRNA or mRNA). In bacteria, this process often involves a cluster of genes, a so-called operon. An operon is a stretch of DNA that codes for several proteins.[24] Importantly, a particular operon is under the control of an operator and a regulatory protein.[22] The way these agents control gene expression in the case of the *lac* operon is illustrated in Fig. 1. The top panel of this figure shows how the regulator, or repressor, prevents the expression of the structural genes that encode the messengers for the synthesis of a group of proteins (depicted as X, Y, and Z). As we can see the repressor, which is in cyan, binds to the operator colored in green and thus prevents the RNA polymerase, which is shown in gray, from transcribing the DNA to mRNA. If a beta-galactoside, shown as a blue pentagon, binds to the repressor, now colored in magenta, the affinity of the repressor to the operator decreases and it unbinds from the DNA. By unbinding the repressor gives way for the RNA polymerase (RNApol) to transcribe the gene.[25] As shown in the bottom panel of the figure, the mRNA produced by RNApol is translated into a protein sequence by the ribosome (for more details, see Ref. 26).

The specificity of the repressor for its operator or operator sequences is remarkable. As Gilbert and Müller-Hill have shown, the operator is just a particular stretch of DNA.[27] The repressor thus
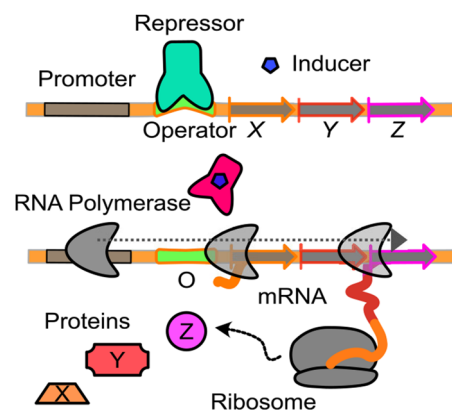

FIG. 1. A simplified representation of the *lac* repressor-operator system showing the induction of the gene expression of structural genes *X*, *Y,* and *Z* by an inducer (blue) that binds to the repressor (active-cyan, inactive-magenta). The repressor dissociates and RNA polymerase (RNA pol, gray) transcribes the gene into mRNA, which is then translated into the protein sequence by the ribosome.

works by binding more strongly to some DNA sequences than others despite these other sequences being in a 10^6 fold excess.[28]

Facilitated diffusion

The genetic regulation systems discussed in the previous paragraphs are remarkably efficient. The kinetics of both repression and induction are both fast with association rates of about 3–5 min measured *in vivo* for a strong operator.[29] When measured directly with membrane filter assays, the association rate between the repressor and operator was found to be about $7 \cdot 10^9 \mathrm{M}^{-1}\,\mathrm{s}^{-1}$. This rapid association rate is notable as it beats the theoretical limit for a simple diffusion-controlled reaction by an order of magnitude.[30] Some additional process that facilitates the association must therefore be occurring.

Figure 2 illustrates the process of facilitated diffusion that explains these fast binding rates. This figure shows how the freely diffusing repressor quickly binds the negatively charged DNA because of its net-positive charge. The fact that the repressor binds to the DNA non-specifically before it binds the operator site is critical as it reduces the dimensionality of the search and thus explains the rapid kinetics of the process.[31–33]

We can imagine facilitated diffusion in the following way: once bound, the repressor moves along the non-specific DNA to its operator site through the various mechanisms that are illustrated in Fig. 2. These mechanisms include rotation-coupled sliding, where the protein follows the DNA major groove, intersegment transfer, where a protein is transferred, without dissociation, between DNA sites that are spatially close but far apart in sequence space, and hopping, short dissociation and quick reassociation a few base-pairs downstream or upstream. All these processes are crucial for the repressor to find the target site quickly and efficiently.[34–37] In fact, it would seem that the processes the repressor performs are balanced in a way that maximizes the likelihood of encountering the operator site once the repressor is non-specifically bound to the DNA in the proximity of the operator.
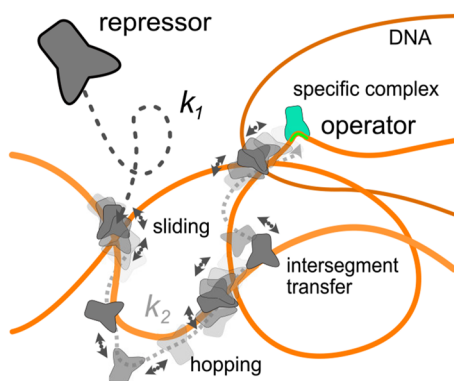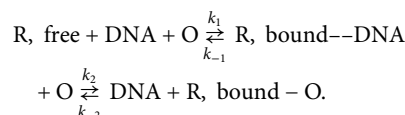
Given the mechanism for specific binding, illustrated in Fig. 2, we should be able to rationalize the kinetics of the process by considering the following sequence of two reactions:[31]

$$\text{R, free} + \text{DNA} + \text{O} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \text{R, bound--DNA}$$

$$+ \text{O} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \text{DNA} + \text{R, bound} - \text{O}.$$

The first step here describes the process by which the repressor binds non-specifically to the DNA, and has a forward rate constant, k_1, and a backward rate constant, k_{-1}. Due to the high amount of DNA though, the protein is non-specifically bound ~90% of the time.[38] The step we focus on in this paper is the second one in which the repressor transitions in a first-order reaction from a nonspecific to specific complex with DNA. As shown in Fig. 2 and discussed earlier, this search is carried out using a number of different physical mechanisms. The most important of these mechanisms for our purposes, however, is sliding because the repressor recognizes the specific site and consequently transitions into the specific complex while undergoing this process.[35,37] That recognition only takes place while the repressor is sliding is evidenced by the so-called antenna effect, which effectively increases the size of the target to the sliding length.[29,31,39]

Figure 3 illustrates the free energy landscape for the reaction mechanism that was described in the last paragraph. The basin on the left of this figure represents the unbound state, which is entrop-
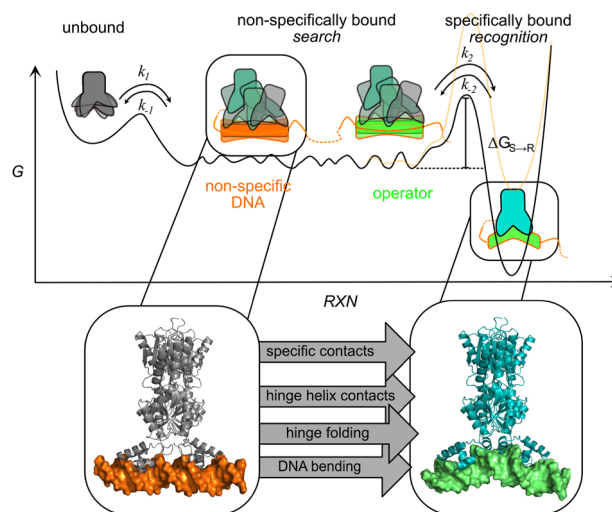
FIG. 3. Free energy along the reaction coordinate of the free to the specifically bound operator. The energies are indicative of the two-step binding mechanism that takes the free protein (gray) to the specific complex with the protein in the *recognition* state (cyan) via the non specifically bound or search state (gray/cyan). The protein binds first to non-specific DNA (orange) and eventually meets the operator (green). At the operator, the transcription factor eventually transitions to the recognition state and the specific complex between the protein and the operator forms. The scheme depicts the barriers and associated transition rates between the three states of the protein. The transparent orange barrier shows a potential barrier for a more weakly bound operator (adapted from Ref. 40). The lower part of the figure depicts which processes could, among others, play a role in determining the height of the transition barriers and the associated rate constants that separate the non-specific and the specific complexes.

FIG. 2. Free and facilitated diffusion with respective binding rates k_1 and k_2 that lead to the formation of the specific complex between the repressor (gray → cyan) and the operator (green) after interactions with the non-specific DNA (orange). During facilitated diffusion, the repressor can move along the DNA using different processes: sliding, in close contact with the DNA groves, leading to rotation around DNA, hopping, short dissociation and reassociation, and intersegment transfer between different parts of the DNA.

ically stabilized because the protein can diffuse through a larger volume of configurational space. The middle, rough part of the energy landscape illustrates the non-specifically bound state between the repressor and DNA, which the system enters while the repressor searches for the operator site. This process can be described as a thermally driven random walk in a rough potential.[41] This rough potential is illustrated by the bumpy free energy landscape in the middle part of Fig. 3. Slutsky and Mirny and others have argued that these bumps must have a height that is less than $1-2 \, k_BT$ to achieve fast enough search times when the system is in the non-specifically bound state.[42,43]

Slutsky and Mirny have also argued that the energy landscape must be rougher when the repressor is in its specifically bound conformation.[42] This rougher landscape would explain why the repressor ceases its search once the operator is found. The authors argue that the roughness increases upon specific binding because the conformation of the DNA-binding protein changes. There are thus two different conformations for the DNA-binding protein: the search state that interacts with non-specific DNA through the electrostatic interaction and the recognition state that binds at the target sequence. When the protein is in the search state it only occasionally forms hydrogen bonds to base edges and thus has a relatively smooth landscape for sliding. When the protein is in the recognition state, it tightly binds at the target sequence via hydrogen bonds and van der Waals interactions resulting in a rougher energy landscape for sliding.[42,44,45] As shown on the right of Fig. 3, this specific protein–DNA complex is characterized by a steep and deep well in the energy landscape.[46] Notice, furthermore, that Fig. 3 shows a barrier to the specific binding. This barrier exists partly because of an entropic cost associated with moving from a non-specifically bound state, where the repressor is free to slide over the full length of the DNA strand, to a specifically bound, stable state. However, the theory also suggests that there is a transition barrier between the search and the recognition states of the protein. In Fig. 3, we have thus drawn a barrier between the search and specifically bound states labeled $\Delta G_{S \rightarrow R}$[44] to illustrate the energetic cost associated with this conformational change in the protein, the DNA and the surrounding solvent. Also note the additional, higher barrier (orange line) in the plot in Fig. 3. The two different barriers represent the energy landscapes for two different operators. A stronger operator supposedly has a lower barrier for specific binding. This relates to experimental studies that have shown that the repressor misses the target site with a certain probability[29] that varies with the operator sequence.[40] It could be argued that a sequence-dependent barrier for conformational switching determines the binding probability for a given DNA sequence.

This short review brings us to the core problem that we believe should tantalize those researchers developing machine learning algorithms for sampling, namely, the processes that are responsible for the transition between non-specific and specific binding. Understanding these factors would allow one to determine why conformational change is triggered by specific sequences. To make these determinations, it is crucial to know the structure of the non-specifically bound state. It is worth noting that multiple experiments have shown that the energy landscape for sliding is sequence dependent[39,41,47] and that interactions with non-specific sites can affect the search process.[48] This suggests that much can be learned by studying the interaction between one type of repressor and a range of different DNA sequences during non-specific interaction and their transition into specific complexes.

Structural information

The lac repressor that we study in this work is one of the earliest gene regulatory elements known. LacI is natively active as a tetramer but similar sequence recognition has been demonstrated for the dimer.[29] We will thus focus on the dimer that is shown in Fig. 4(a). The dimeric LacI structure shown in this figure consists of a DNA-binding domain (DBD) and a core domain with an inducer binding site that regulates affinity to DNA.[51] Figure 4(b) zooms in on the DBD, which interacts electrostatically with the phosphate and through hydrogen bonds and van der Waals interactions with the sugars and bases of the DNA.[49] The DBD has a classical motif for a DBP involving 4 helices and 3 turns as well as an N-terminal loop. The first of these helices, helix 1 (H1) interacts with the DNA through leucine 6 and tyrosine 7[52] [Fig. 4(c)]. Helix 2 (H2) is also called the recognition helix because it forms most of the specific contacts to the DNA. This helix interacts with DNA through tyrosine 17, glutamine 18, serine 21, and arginine 22[52] [Fig. 4(c)]. Figure 4(c) also shows the histidine 29 residue that appears in the loop after H2 and can also form specific interactions with the DNA.[52,53] The third helix (H3) is the longest helix in the DBD and is not in direct contact with the DNA. However, this helix carries a number of charged residues that are likely important for the interaction.[54,55] The final and arguably the most interesting helix is the so-called hinge helix, which we investigate further in our simulations.

There are three natural, 21-bp-long operator sites that bind to the LacI dimer: O1, O2, and O3.[50] A range of synthetic operator sites can also be designed, some of which bind more strongly to LacI than the native operator sequences.[56] There are thus multiple crystal and NMR structures for LacI bound to various operator/DNA sequences that can be used as starting points for molecular modeling. When it comes to crystal structures, the strength of binding influences the resolution.[57] The only structure with resolved protein side chains (PDB-ID 1EFA[49]) consequently contains the 20-bp-long sequence OSymL that is known to bind to the protein the strongest. This structure is the starting point for our simulations.

METHODS

General methods: Starting structure, simulation details

We have created a repository on GitHub (https://doi.org/10.5281/zenodo.7627873) to accompany this paper, which contains GROMACS[58,59] input files that can be used to model the systems described in the previous section. The equilibrated structures were prepared based on the crystal structure with the PDB-ID 1EFA,[49] which is shown in Fig. 4.

For crystallization, the structure was stabilized by the anti-inducer ortho nitrophenyl fucoside (ONPF) that binds to the allosteric site in the core domains. Both ONPF molecules were deleted when preparing the structure for simulations. An additional preparation step was performed on the DNA. The crystal structure contains base pairs 4–20 of OSymL.[49] For the simulations, base pairs 1–3 had to be added manually to form the operator DNA. Two terminal CG base pairs were then added at the ends of the sequence
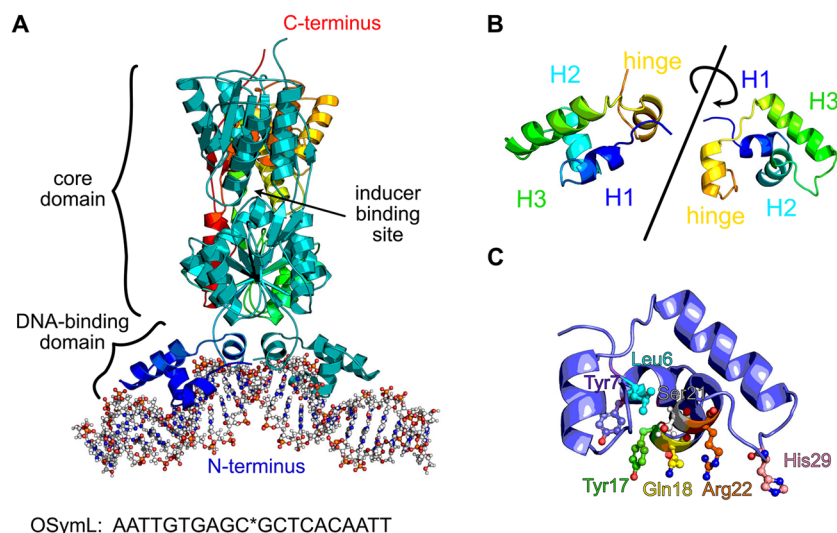
OSymL: AATTGTGAGC*GCTCACAATT

FIG. 4. (a) The starting structure for the simulations of the *lac* repressor (LacI) based on the crystal structure with PDB-ID 1EFA.[49] The dimeric repressor is shown in cartoon representation. The first monomer is shown in rainbow colors starting with blue at the N-terminus and going toward red at the C-terminus. The second monomer is colored cyan. The DNA-binding and core domains as well as the inducer binding site in the center of the core domain are labeled. The DNA is shown in ball and stick representation and the sequence of the OSymL operator is shown below it. The star marks the position that is omitted in this synthetic 20-bp operator when compared to the natural 21-bp operators O1, O2, and O3.[50] (b) The monomeric DNA binding domain (DBD) is shown with the labels H1, for helix 1, H2, for helix 2 (the recognition helix), H3, for helix 3 as well as a hinge, for the hinge helix/region that will unfold during unbinding. The left panel shows the view onto the DBD that the viewer has on the left DBD in the complex in panel (a), and the right panel shows a bird's view with the angle corresponding to the one in panel (c). (c) The specifically interacting protein side chains, shown as sticks and spheres, on the DBD that is shown in cartoon representation.

to prevent end effects on the DNA model. The DNA sequence in our simulations is thus 24 base pairs long. For simulations of a complex with a different DNA sequence (NOD), the script basemutator[60] is used to mutate the DNA structure using PyMOL.[61]

Histidine 29 [Fig. 4(c)] is a critical amino acid residue in the binding between the protein and the DNA.[52,53] There are a number of different protonation states for this residue.[62] We chose to protonate the histidine as PropKa[63] tells us it has a pKa of 7.3, which is very close to the pH of the *E. coli* cell.[64] Despite this, coarse-grained simulations have shown that the protonation of His29 is essential for the sliding motion during the search process.[37]

The systems were simulated in the NPT ensemble at 310 K, the optimal growth temperature for *E. coli*.[65] The protein and DNA were solvated in a solution with a 150 mM concentration of KCl and a 5 mM concentration of magnesium ions. Magnesium is particularly important because it is known to interact with and stabilize the DNA structure in the cell.[66] The solvation conditions in our simulations are similar to conditions used in earlier simulations of proteins and are representative of the conditions in the *E. coli* cytoplasm.[67] They are also close to the experimental conditions in the first equilibrium binding studies of LacI and the Lac operator carried out by Lin and Riggs.[68] In these experiments, the potassium and magnesium concentrations were 180 and 3 mM, respectively.

All simulations were performed with GROMACS[58,59] version 2019.6 or, in the case of metadynamics simulations, with GROMACS version 2021.3. The AMBER99SB-ws force field[69,70] was used to describe the protein interactions and the PARMBSC1[71] force field was used for DNA. Water molecule interactions were simulated using TIP4P/2005[72] and ions were simulated using the parameters

from Dang and Kollman.[73] As discussed in the supplementary material, the choice of parameters for the ions is important as standard AMBER force fields overestimate the strength of non-bonded interactions for potassium.[74]

The system was prepared in a dodecahedral box. As the LINCS algorithm[58,75] was used to restrain all intramolecular bonds involving hydrogen, we could integrate the equations of motion using a 2 fs timestep. Short-range interactions were calculated up to a cut-off of 1 nm. Electrostatic interactions were calculated explicitly below a distance of 1 nm. Beyond that distance, they were estimated using the particle-mesh-Ewald method.[76] We used the stochastic velocity rescaling thermostat[77] and Berendsen barostat[78] with frictions of 0.1 and 2.0 ps for temperature or pressure coupling, respectively. To equilibrate the system, the protein, DNA, and water was initially minimized using the steepest descent integrator. A 500 ps, NVT run at 310 K with positional restraints on the protein and DNA was then performed to equilibrate the solvation spheres. This was followed by a further 500 ps of restrained NPT equilibration at 1 bar. Most restraints on the DNA and protein were removed during production simulations although distance restraints were kept on the terminal base pair hydrogen bonds to prevent fraying of the DNA.[79]

Results and discussion

Figure 5 shows the results from 5, 50-ns, unbiased simulations. Figures 5(a) and 5(b) show the root mean square distance (RMSD) for the core domains and the DBDs, respectively. These panels illustrate that, in accordance with earlier studies,[80] the specific complex is rather stable and the RMSD values are low. This conclusion is fur-
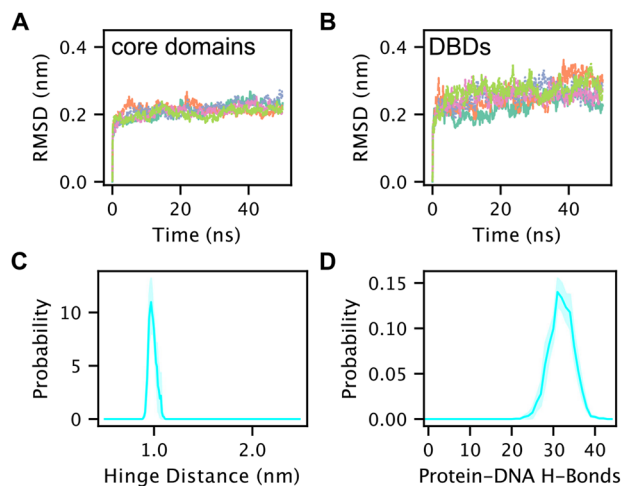
FIG. 5. Results from unbiased MD simulations of the specific complex based on the structure with PDB-ID 1EFA.[49] (a) and (b) show the RMSD for the core domains and the DBDs. (c) and (d) show estimates for the histograms together with a shaded area that indicates the 68% confidence interval on our estimate. In (c) the histogram is shown as a function of the distance between the hinge center and the center of the DNA. In (d), the histogram is shown as a function of the number of protein–DNA hydrogen bonds.

ther evidenced by Figs. 5(c) and 5(d), which show the distribution of values that were observed for the distance between the hinge helices of the protein and the center of the DNA and the distribution of values for the number of hydrogen bonds between the protein and the DNA. Both of these distributions are mono-modal, which suggests that only one state in the energy landscape—the specifically bound state—is being explored in our simulation. Much longer simulation times or some form of bias are required to get the complex to transition to the non-specifically bound state.

NMR RESTRAINTS

Some structural information on the non-specifically bound complex has been obtained from NMR experiments (PDB-ID 1OSL[81]). These experiments only considered the binding of the two DBDs to the DNA without the core domain. Therefore, valine 52 was mutated to cysteine to stabilize the DBD dimer by a disulfide bridge rather than the interactions between the two core domains.

The NMR experiments provide information on the average distances between particular pairs of atoms or groups of atoms for the non-specifically bound complex rather than a single structure. Multiple structures that have the same average distances between the atoms can then be found. PDB-ID 1OSL thus contains 20 different structures for the non-specifically bound complex, 5 of which are shown in Fig. 6(a). We can see there is a substantial overlap between these various structures for H1, H2, and H3. However, there are considerable differences in the hinge region in the center of the complex. The large flexibility of the hinge regions is likely a consequence of the removal of the core regions. Furthermore, the structure in this region is likely to be affected by the mutation of valine to cysteine. Figure 6(b) shows the structure close to the disulfide bridge that holds the two DBDs together in the experiment. Figure 6(c) shows what this part of the protein looks like when the wild-type protein is specifically bound to the DNA. We can see that the two DBDs can come closer together when the formation of the dimer is driven by the interaction between the core domains and the hydrophobic interactions between the two hinge helices.

We performed experiment-directed simulations to try to determine the structure of the non-specifically bound state in a way that resolved some of the issues with the NMR experiments described earlier. Flat-bottom harmonic oscillator potentials on the distances and angles that were identified in the NMR experiments were included as constraints in these simulations (PDB ID 1OSL[81]). To set up these restraints, we obtained NMR-STAR files[82] from the protein data bank and then used the method and scripts developed by Sinelnikova and van der Spoel[83] to construct a restraint file for GROMACS. These restraints force the system away from the initial crystal structure and toward configurations that are supposed to resemble
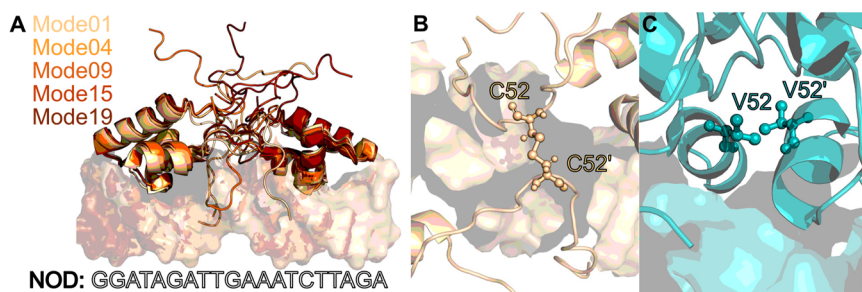
FIG. 6. The NMR structure: (a) 5 of the 20 different modes of the DNA-binding domain in the NMR structure with a non-specific DNA sequence. The monomers are truncated from residue 62 and a cysteine-bridge ensures the dimerization of the transcription factor. (b) Bird's view and zoom on the central part of panel A. The cysteine bridge between the hinge regions of the DNA-binding domains in the NMR structure is shown for mode 1 using a sticks and spheres representation. C52 and C52′ mark C52 in monomers A and B, respectively. (c) The interactions between the valine 52 residues in the two monomers in the wild-type protein. V52 and V52′ mark V52 in monomers A and B, respectively. In all figures, the protein is shown in the cartoon, the DNA in surface representation, and individual residues are shown as spheres and sticks. The figures have been produced in PyMOL based on the structures with PDB IDs 1OSL[81] and 1EFA.[49]

those seen in the experiments. Furthermore, because we remove all the restraints on residue 52, we address the most obvious issues with the mutation that were described in the previous paragraph.

Two separate restrained MD simulations were performed for the specific DNA sequence OSymL, which is known to bind strongly to the DBD of LacI, and NOD, which binds weakly to the DBD.[84] For each sequence 3, 300 ns replicas of restraint simulation with the set-up described in the previous paragraph were performed. A force constant of 1500 kJ mol^{-1} nm^{-2} was used for all NMR restraints and the time constant for averaging the restraint forces was 100 ps. Average properties were then computed from the last 50 ns of all simulations.

RMSD values of 4.7 ± 0.3 and 4.8 ± 0.2 Å were observed for the atoms in the core domains of the proteins bound to NOD and OSymL, suggesting that the restraints have some effect on the structure in these parts of the protein. Larger RMSD values of 7 ± 2.6 and 7 ± 0.8 Å are observed for the DBDs, which suggests that the structure in these parts of the protein has changed more. These changes in structure are further quantified in Fig. 7, which compares (i) the distribution for various structural properties in our restrained simulations, (ii) the distributions of those properties for the specifically bound complex, and (iii) the average value of these quantities for the 20 different modes in the PDB file that contains the NMR structure of the non-specific complex.[81] The values these structural quantities

take in the crystal structure of the specifically bound complex and the values they take for the NMR structure are also listed in Table I. Table I clearly shows that the values of all these structural quantities change when the protein changes from being specifically bound to being non-specifically bound. By contrast, Fig. 7 shows that some, but not all of these structural quantities, change when the NMR restraints are active and that some variables undergo drastic changes in values that are not seen in the experiments. The picture of unbinding that emerges from our restrained simulations is therefore incomplete.

We can better understand the incomplete picture that emerges from the simulations with the NMR restraints by working through each panel of Fig. 7 in turn. Figure 7(a) shows the distribution for the number of specific contacts between the DNA bases and protein side chains. In Fig. 7(a), the top plot is for NOD and the bottom is for OSymL. NMR studies showed that when the specific complex is formed, leucine 6, tyrosine 7, tyrosine 17, glutamine 18, serine 21, arginine 22, and histidine 29 interact with the DNA.[52,81] The contact map CV that was used to construct Fig. 7(a) was calculated using the distances between these residues and the atoms in the DNA bases that are closest to them in the specific complex (see the supplementary material for a more detailed description and Table S1 for the contact parameters for different systems). Figure 7(a) clearly shows that the number of specific contacts decreases
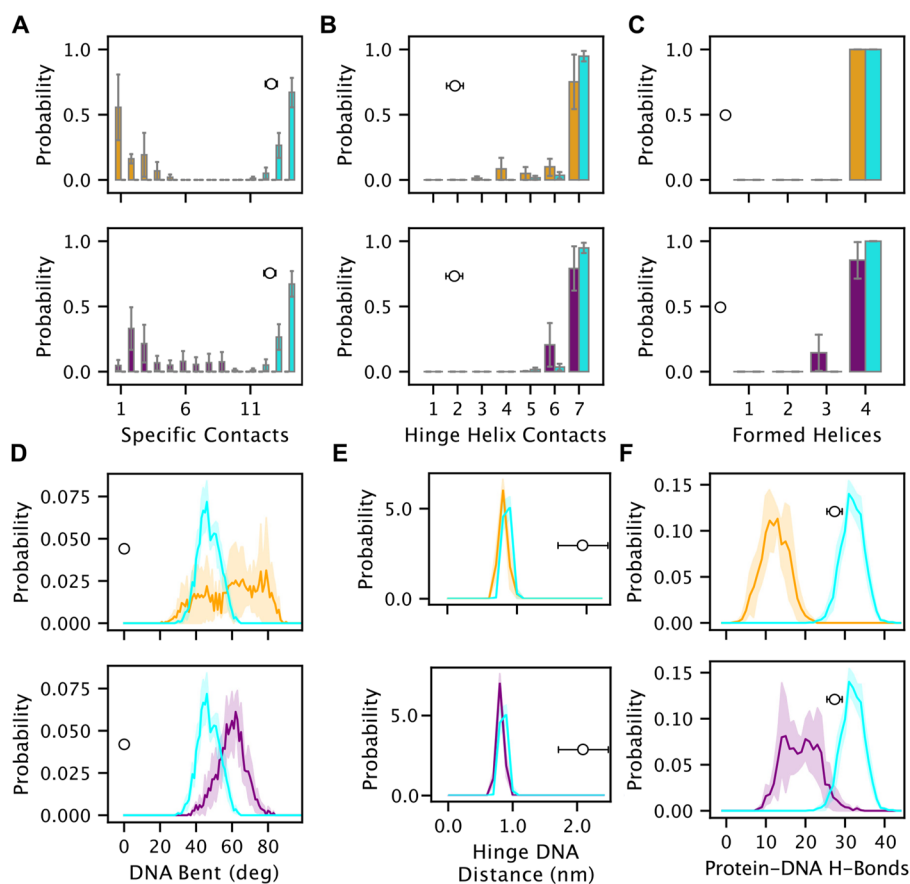

FIG. 7. A comparison between results from MD simulations starting from the specific complex containing the non-specific sequence NOD (orange) and the operator sequence OSymL (purple) with distance restraints based on NMR data and unbiased runs (cyan) of the specific complex with OSymL. Panels (a)–(f) show the average distributions of different collective variables that could be used to visualize the unbinding process. The error bars/shaded areas show the standard error of the mean between the five replicas. The data used to construct the histograms is taken from unbiased simulations of 50 ns length (see Fig. 5) and the last 50 ns of the NMR simulations. In all plots, the mean and standard deviation for the distribution of CV values in the 20 modes in the NMR structure of the non-specific complex is shown as a circle with whiskers.

TABLE I. The values the collective variables that we have investigated in this work take for the experimental NMR structure of the non-specifically bound state and the experimental crystal structure of the specifically bound state. There are multiple configurations in the PDB file that describe the NMR structure so we calculated the CVs for all the reported structures and reported the mean and standard deviation of these values.

Collective variable	Non-specific complex (NMR)	Specific complex (crystal)
No. of specific contacts	12.6 ± 0.58	14
No. of hinge helix contacts	1.9 ± 0.75	7
No. of alpha helices in hinge region	0.1 ± 0.1	4
DNA bent (deg)	1.1 ± 0.4	41
Hinge helix to DNA distance (nm)	2.1 ± 0.5	1
No. of Protein–DNA hydrogen bonds	27 ± 5	33

substantially when the NMR restraints are active. However, this behavior is at odds with the behavior that is seen in the experiments. The experimental structures suggest that many of these contacts are still present when the protein is non-specifically bound and it is unclear whether this is an artifact in the experiment or a real feature of the non-specifically bound state.

Figures 6(a) and 4(a) show the structures of the non-specifically bound and the specifically bound complexes, respectively. The largest differences in these structures appear in the central part of the protein. As has been discussed in previous experimental and simulation work,[81,85,86] when the protein is non-specifically bound this hinge region is unstructured and flexible. By contrast, when the protein binds, two alpha-helices form that are bound together tightly.[49] Figures 7(b) and 7(c) show histograms for collective variables that measure these differences. Figure 7(b) shows histograms as a function of the number of contacts between the left and right hinge region for the specifically bound state (top and bottom) and the non-specifically bound state with NOD (top) and OSymL (bottom). The CV here measures the number of native contacts involving arginine 51, valine 52, alanine 53, glutamine 54, glutamine 55, leucine 56, and alanine 57 that are formed. We can see that all these contacts are present when the protein is specifically bound. This result is in agreement with the experimental information in Table I, which shows that all these contacts are formed when the protein is specifically bound. However, Table I also shows that many of these contacts are broken when the protein is non-specifically bound. This observation is at odds with what we see in our simulations—the majority of these contacts remain intact in the simulations when the NMR restraints are active. The introduction of the NMR restraints is thus doing little to disrupt the structure of the protein in this region.

The conclusion that the NMR restraints do little to disrupt the structure of the protein in the hinge region is further evidenced by Fig. 7(c), which offers an analysis of the secondary structure content for this part of the protein. We measured the helix content using the ALPHARMSD CV that was developed by Pietrucci and Laio.[87] Figure 7(c) confirms that 4 helices are formed by the hinge region when the protein is specifically bound and that they remain intact when the NMR restraints are active.[84,88] By contrast, in the NMR structure, these helices are not present.

Figure 4(a) shows that the DNA is bent in the specifically bound complex. However, the DNA is straight in the NMR structure of the nonspecific complex.[89] Experimental studies observe bending upon specific binding to LacI for all operators,[49,84] which suggests that DNA bending is a feature of specific binding. In Fig. 7(d), the histogram of DNA bending angles that were visited in our simulations of the specific complex and our simulations that included the NMR restraints are shown. We can see (top panel) that in the simulations with the NMR restraints, the bending angle of the DNA is similar to or even larger than the angle that is seen for the specific complex. In other words, the DNA does not straighten.

Figure 7(e) shows histograms for what is perhaps the most obvious CV for measuring the degree of binding between the protein and the DNA: the distance between the hinge helix and the central base of the DNA. Naively, one would expect this distance to be small when the protein is specifically bound and large when the protein is non-specifically bound as it is for the crystal and NMR structures, respectively (see Table I). When the NMR restraints are active, this CV does not change. The protein remains trapped in a conformation with bent DNA that has the helices inserted into the minor groove.

The final panel of Fig. 7 [panel (f)], shows the effect the NMR restraints have on the number of hydrogen bonds between the protein and the DNA. We can see that there are far fewer hydrogen bonds when the NMR restraints are active. We would expect the number of hydrogen bonds between the protein and the NMR to decrease when the system moves from being specifically bound to non-specifically bound, but Table I shows this is not what is seen in the experiments. Consequently, introducing the restraints from the NMR simulations breaks many hydrogen bonds that the experiment tells us are not broken in the non-specifically bound configuration.

To summarize, in the structure that emerges from the simulation with the NMR restraints many of the specific contacts and hydrogen bonds between the protein and the DNA are broken. It would appear that these restraints sever contacts and hydrogen bonds that are present in the structures of the specifically and non-specifically bound configurations that have been determined from the experiment, which is not ideal. To make matters worse, disruptions to the structure of the hinge helices and changes in the degree of DNA bending that are seen in the NMR structure of the non-specifically bound state are not seen in our simulations. When the NMR restraints are active, the repressor's hinge helices are left trapped in a tight interaction with the minor groove. If anything, the breaking of the specific contacts and hydrogen bonds between the protein's DNA binding domains and the DNA makes the interactions in the center of the complex even more pronounced. This suggests that the order in which hinge helix unfolding, DNA straightening, and contact breaking occur is important. If that is the case, then it is not surprising that the structures for the non-specifically bound complex that emerge from a simulation that was started from the specifically bound state contain features that would not be expected to occur in the non-specifically bound state. The NMR restraints did not drive the system over the barrier between the specifically bound and non-specifically bound state as they serve only to make the non-specifically bound state more thermodynamically favorable. What these restraints do when the system is in the specifically bound state—far from the non-specifically bound state—is unclear. To achieve transitions between the two states, some other method is required.

METADYNAMICS

We used well-tempered metadynamics[90,91] simulations to try to force the protein to transition from the specifically bound configuration to a non-specifically bound configuration. GROMACS 2021.3[58] patched with PLUMED 2.7.2[92] was used to perform these simulations. In all calculations, the bias factor was set equal to 10 and hills with an initial height of 1.3 kJ mol^{-1} were added every 100 MD steps. The width of these Gaussian hills was set using the geometric adaptive framework from Branduardi *et al.*[93] with a sigma parameter of 0.25 nm. The minimum allowed width for the hills was always equal to 0.05 CV units.

The difficulty with metadynamics simulations is to identify a suitable CV for driving the system between the states of interest. As detailed in the following sections, we tried running metadynamics with a number of different CVs but were unable to obtain converged results. The comparison between the distributions of configurations that were visited in the metadynamics simulations with the distributions that were obtained from the unbiased simulations, the simulations that were performed with the NMR restraints and the NMR structure, that is presented in the next few sections, is, nevertheless, useful as it offers further insights into the qualitative behaviors of this system.

CMAP

Figure 8 shows the results from three simulations in which a metadynamics bias was applied to the number of specific contacts between the protein and the DNA.[94] Histograms as a function of the number of specific contacts are shown in Fig. 8(a). This figure shows that the metadynamics bias does cause the number of contacts to decrease slightly. Figure 8(b) also shows that the metadynamics bias causes protein–DNA hydrogen bonds to break. Fewer protein–DNA hydrogen bonds break during the metadynamics simulation than in the simulation with the NMR restraints. In this regard, the structures visited during the metadynamics simulations are more consistent with the structure of the non-specifically bound state that emerges from the NMR experiments. However, Fig. 8(c) illustrates that there are still enormous differences between the ensembles of structures that are emerging from simulations and experiments. This figure shows the distribution of the distance between the hinge helix and the DNA. We can see that the distances that are explored in the metadynamics simulation are very close to the distances that are observed in unbiased simulations and that they are much shorter than the distances that are seen in the experiment. Consequently, even though the number of specific contacts decreases, it seems unlikely that this simulation is exploring states that correspond to the non-specifically bound configuration.

BIASING THE EXTENDED CONTACT MAP

As discussed in the section on the simulations with the NMR restraints and as shown in Figs. 6(b) and 6(c), the NMR experiment shows that the hinge region of the DBD undergoes significant structural changes upon unbinding. We incorporated these structural changes in our metadynamics simulations by biasing a new contact map that measured the number of contacts between the protein and the DNA as well as the number of contacts between the hinge helices of the two DBDs. Results from these simulations are shown in Fig. 9.

Figures 9(a) and 9(b) show that the metadynamics bias causes the number of hinge–helix contacts and the number of specific contacts to decrease. Furthermore, there appear to be correlations between these two groups of contacts. In other words, the hinge–helix contacts appear to break at the same time as the DNA–DBD contacts.

Figure 9(c) shows how this new metadynamics bias is not driving unbinding but that new configurations of the system can be observed. This panel shows the histogram as a function of the distance between the center of the hinge helices and the center of the DNA and should be compared to Fig. 8(c). We once again see here that when the bias acts only on the DNA–protein contacts this distance does not change much from the value it has in the specifically bound state but that two additional peaks for shorter and longer distances emerge.

Further evidence of partial unbinding is observed in Fig. 9(d). This figure shows histograms as a function of the number of protein–DNA hydrogen bonds. The inset configurations beside this panel show illustrative configurations from each of the three modes in the distribution that was observed in the metadynamics simulation. We can see that the peak with 38 hydrogen bonds corresponds to the specifically bound complex. In structures that appear in the
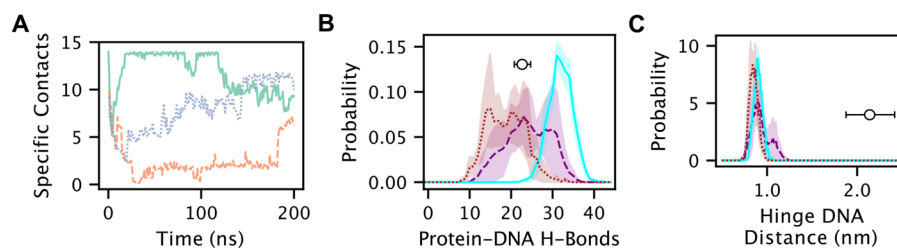

FIG. 8. Results from metadynamics simulations using the number of specific contacts between protein and DNA as a collective variable. (a) shows the time series for the number of specific contacts for the three replicas (1-green; 2-orange, dashed; 3-blue, dotted). (b) and (c) show estimates for the distribution of configurations that are sampled in the three metadynamics simulations (purple, dashed), the unbiased (cyan), and the simulations with the NMR restraints (brown, dotted) with a 68% confidence interval (shaded area) and the averages from the 20 modes in the NMR structure as a dot with whiskers. In panel (b), the histogram is shown as a function of the number of protein–DNA hydrogen bonds that are present. In panel (c), the histogram is shown as a function of the distance between the center of the hinge helices and the center of the DNA strand.
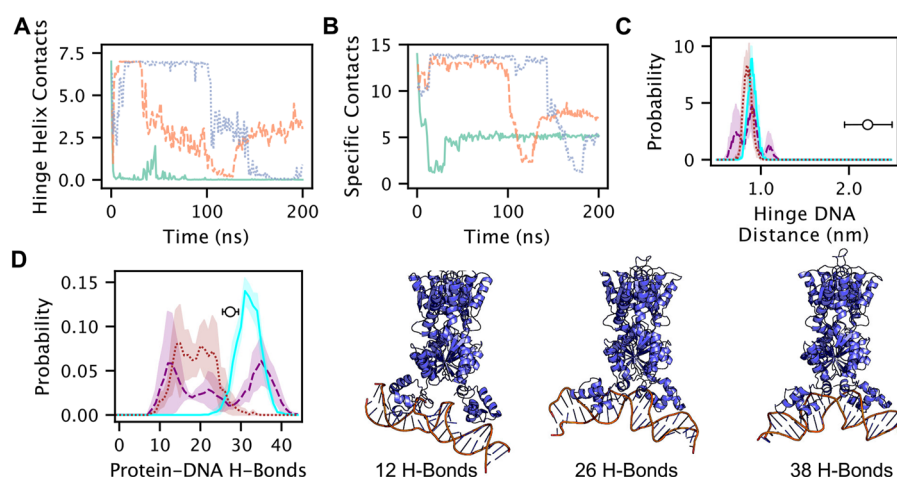
FIG. 9. Results from metadynamics simulations using contacts between the two hinge regions and specific contacts between protein and DNA as a collective variable. (a) and (b) show the time series of the contacts between the hinge helices and the specific contacts for the three replicas (1-green; 2-orange, dashed; 3-blue, dotted). (c) and (d) show an estimate for the distribution of configurations that are sampled in the three metadynamics simulations (purple, dashed), the unbiased (cyan), the simulations with the NMR restraints (brown, dotted) with a 68% confidence interval (shaded area) and the averages from the 20 modes in the NMR structure as a dot with whiskers. In panel (c), the histogram is shown as a function of the distance between the center of the hinge helices and the center of the DNA strand. In panel (d), the histogram is shown as a function of the number of protein–DNA hydrogen bonds that are present. In (d) some representative snapshots of the protein–DNA complex for different numbers of hydrogen bonds are shown.

peak around 26 hydrogen bonds, the two hinge helices have begun to drift apart and the DNA has started to straighten. Finally, in structures that have only about 12 hydrogen bonds we often find that one DBD has drifted away from the DNA.

In short, Fig. 9 demonstrates that the system is driven further when both the hinge–helix and the specific contacts are biased together. However, we are still clearly missing some critical slow degrees of freedom because we do not see the system returning to the specific complex and the hinge DNA distance does not increase by a great deal.

BIASING THE CONTACT MAP AND THE SECONDARY STRUCTURE

The NMR experiments indicate that when the protein is non-specifically bound, the hinge–helices completely unfold so there is no alpha-helical structure in these regions. We found that these helices did not unfold in the metadynamics simulations that were run in the previous section. As shown in the supplementary material (Fig. S1), the hinge helices partially unfolded but the ALPHARMSD variable[87] never fell to values of zero. We thus ran a new set of metadynamics simulations in which a metadynamics bias that was a function of both the contact map and the ALPHARMSD was added. The results from this simulation are shown in Fig. 10.

Figure 10(a) shows that the metadynamics bias can drive the hinge helices to unfold and, in some cases, to fold during the simulation. Furthermore, Fig. 10(b) shows that the number of contacts between the two hinge helices changes substantially as the hinge helices fold and unfold. However, Fig. 10(c) shows that all the activity associated with the folding and unfolding of the hinge helices is not accompanied by any significant change in the degree of binding

between the protein and the DNA. The number of specific contacts between the DNA and the protein remains resolutely high, which suggests that the non-specifically bound state is not visited.

This conclusion, that the non-specifically bound state is not visited in these simulations, is confirmed by Figs. 10(d) and 10(e). Figure 10(d) shows the histogram as a function of the distance between the center of the hinge helices and the center of the DNA together with the histograms that were obtained from the unbiased and NMR restraint simulation and the average value of this quantity from the NMR experiment. One can see that the histogram we obtain from the metadynamics simulation is similar to the result we get from the unbiased simulations and the NMR simulations. During most of the simulation, the distance between the DNA and the protein remains resolutely low.

Figure 10(e) shows that metadynamics does a better job of determining what happens to the amount of hydrogen bonding between the protein and the DNA than the simulations with the NMR restraints. We can see that most of these hydrogen bonds remain intact during the metadynamics simulations but that this behavior is also most consistent with the structures that are observed in the NMR experiments. The reason these bonds remain intact becomes clear when we look at the representative snapshot that is shown in Fig. 10. This figure clearly shows that the bias drives the hinge helices to move away from the DNA and to unfold. However, H1, H2, and H3 remain bound to the DNA strand.

We can thus conclude that unbinding was not observed with this combination of collective variables. It would be easy to argue that this is simply because the simulation was not run for long enough. However, we obtained similar results from longer simulations with similar combinations of CVs. Furthermore, the average value of the bias for the last 100 steps of the simulation is
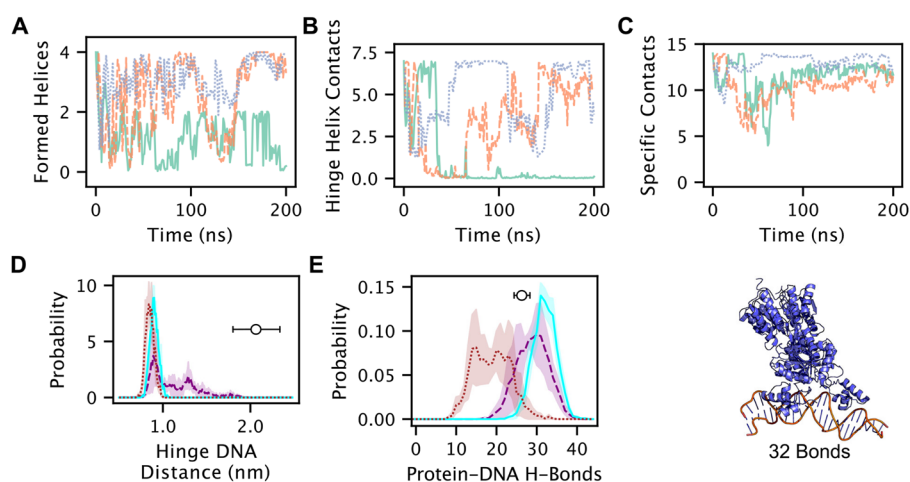
27 September 2023 14:02:52

FIG. 10. Results from metadynamics simulations that used the number of specific contacts between protein and DNA and between the two hinge regions and the ALPHARMSD for the hinge regions as collective variables. (a)–(c) show the time series for the number of helices (six residues) in the hinge region, the number of contacts between the hinge helices and the number of specific contacts between the protein and DNA for the three replicas (1-green; 2-orange, dashed; 3-blue, dotted). (d) and (e) show estimates for the distribution of configurations that are sampled in the three metadynamics simulation (purple, dashed), the unbiased (cyan), the simulations with the NMR restraints (brown, dotted) with a 68% confidence interval (shaded area) and the averages from the 20 modes in the NMR structure as a dot with whiskers. In panel (d), the histogram is shown as a function of the distance between the center of the hinge helices and the center of the DNA strand. In panel (e), the histogram is shown as a function of the number of protein–DNA hydrogen bonds that are present. In (e), a representative snapshot of the protein–DNA complex is shown.

146 kJ mol^{-1}. That such an enormous bias is not able to drive the transition from the specifically bound to the non-specifically bound state suggests that some other degree of freedom needs to be biased in our simulations.

METADYNAMICS FOR NON-SPECIFIC DNA

The previous three sections showed how difficult it is to design a metadynamics bias that can drive the transitions between the specifically and non-specifically bound states. These sections focused on illustrative examples, we have tried but not written about other combinations of CVs including the DNA bending angle, various distances between the DNA and the protein, the number of hydrogen bonds, path CVs in contact map space, and variously defined contact maps. Every combination of CVs we tried produced similarly inconclusive results.

We also tried driving the transition between the specifically and non-specifically bound states for a complex where the DNA is known to not bind specifically to the protein, the sequence known as NOD.[84] Figure 11 shows the result from one of these simulations. The metadynamics bias here acts upon the contact map and the ALPHARMSD and is similar to the one that was used to bias the dynamics in the previous section. This bias is necessary as, even though it is not possible to characterize a specifically bound complex with the non-specific sequence experimentally, simulation work has shown that the specifically bound complex with non-specific sequences remains intact over the course of a 1-μs-long simulation.[80]

Figure 11(c) shows that the metadynamics bias can drive down the number of specific contacts between the DNA and the protein. There is thus some evidence that the protein begins unbinding from the DNA. Figures 11(a) and 11(b) show that this unbinding

is accompanied by a loss of structure in the hinge–helix region. As the experiments suggest, the degree of alpha-helical structure in this region decreases as well as the number of contacts between the two hinge helices.

Figure 11(d) shows the histogram as a function of the distance between the center of the hinge helices and the DNA. We can see that the metadynamics bias drives the system to explore configurations where this distance is large. Furthermore, Fig. 11(e) shows that the metadynamics bias also drives the system to explore configurations that have fewer hydrogen bonds between the DNA and the protein. We observed three different types of unbinding in our simulations: either H1, H2, and H3 unbind, the hinge unbinds or the whole DNA binding protein separates from the DNA. In experiments, the second of these types, the unbinding of the hinge, is observed.

Even though there are some positives in the results in Fig. 11, they are still outweighed by the negatives. Recrossing was not observed so we cannot extract relative free energies. Furthermore, any kinetic information that we extract from these simulations is likely spurious as the metadynamics bias is very large and basins are likely to be overfilled. However, there is possibly a deeper problem with this simulation than even the issues with the metadynamics. The fact that specific binding to the NOD sequence is not observed in experiments suggests that the configuration that we started the simulation from is at best metastable. If this is the case, our simulation has been started from a configuration that is far from equilibrium. Without a better sense of the free energy surface, it is unclear if the simulation would return to anything resembling the initial, specifically bound configuration even if a set of collective variables that can drive this transition for the specific sequence had been identified.

Although the barrier between the specifically bound and non-specifically bound complex is probably lower for NOD, it is better
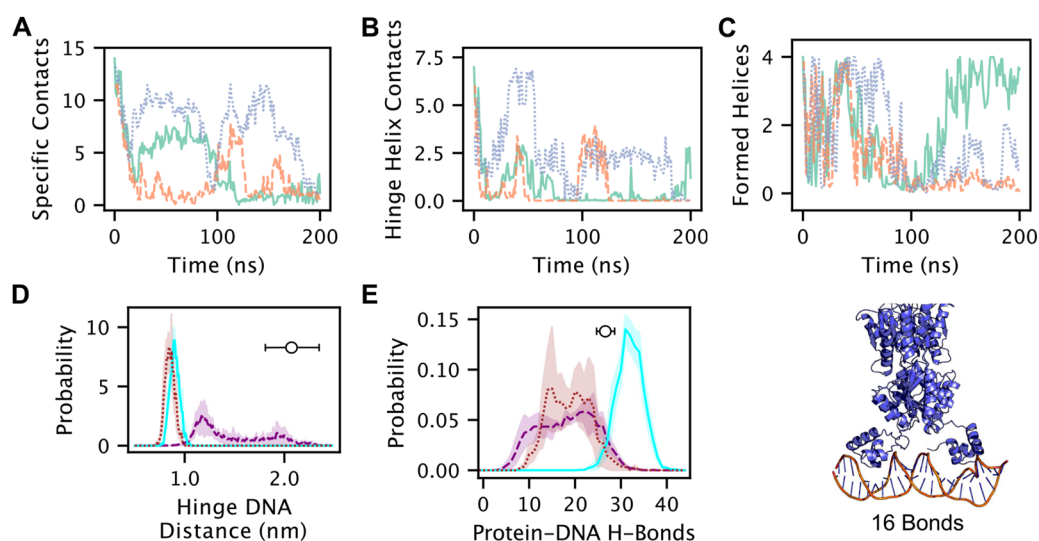
FIG. 11. Results from metadynamics simulations on the complex between LacI and the nonspecific sequence NOD using the number of specific contacts between protein and DNA and between the two hinge regions and the ALPHARMSD for the hinge regions as collective variables. (a)–(c) show the time series of the number of helices (six residues) in the hinge region, the number of contacts between the hinge helices and the number of specific contacts for the three replicas (1-green; 2-orange, dashed; 3-blue, dotted). (d) and (e) show an estimate for the distribution of configurations that are sampled in the three metadynamics simulations (purple, dashed), the unbiased (cyan), the simulations with the NMR restraints (brown, dotted) with a 68% confidence interval (shaded area) and the averages from the 20 modes in the NMR structure as a dot with whiskers. In panel (d), the histogram is shown as a function of the distance between the center of the hinge helices and the center of the DNA strand. In panel (e) the histogram is shown as a function of the number of protein–DNA hydrogen bonds that are present. In (e), a representative snapshot of the protein–DNA complex is shown.

to start by developing coordinates that can sample the forward and backward transitions for the complex with OSymL or weaker operators such as O1[50] where the specifically bound complex is seen in experiments.[49,95] Using these same coordinates to drive the complex with O1 between the specifically and non-specifically bound states would allow one to untangle the role kinetics and thermodynamics are playing in the experiment.

CONCLUSIONS

If it were possible to perform simulations of the reversible transition between the non-specifically bound and specifically bound complex for LacI and DNA many outstanding questions about gene regulation could be answered. Such simulations would help us to assign a structure to the non-specifically bound complex, which has been the subject of fierce debate for 60 years. Such simulations could also be used to determine how the relative free energies of the specifically bound and non-specifically bound states depend on the DNA sequence. This information would resolve arguments about whether specific binding only occurs because the kinetic barrier to the formation of the specifically bound complex is low for binding to operators and high for binding to non-operators. In other words, such simulations would tell us if there is a deep minimum in the free energy landscape for specific binding to non-operator DNA sequences that are not seen in experiments because the barrier to accessing it is inordinately high.

The simulations in the previous sections do not answer any of these questions as the CVs do not capture the essence of the transition. However, the work we have described suggests an approach

that can be used to tackle this problem. Suitable variables for biasing could be obtained by analyzing the unbiased MD simulations and by performing simulations that use the NMR restraints to identify an ensemble of structures for the non-specifically bound state. We have shown that it is difficult to come up with a single variable that describes all the differences in these two ensembles of structures. Machine learning or dimensionality reduction algorithms are perhaps better suited to tackling this problem. Furthermore, running enhanced sampling calculations on the variables that emerge from this analysis offers a method to test the efficacy of the machine-learned coordinates.

A need for benchmark, enhanced-sampling problems that are more complicated than sampling transitions in alanine dipeptide or Lennard-Jones 7[96] comes up whenever there is a discussion session at a conference on enhanced sampling. In these sessions, it is argued that more complicated problems are required to test the next generation of enhanced sampling techniques. We believe that the transition from the non-specifically bound to the specifically bound protein–DNA complex that we have described in this paper would serve as an excellent challenge to this community. Many degrees of freedom change during the transition and need to be biased for the transition to be observed on an accessible timescale. Furthermore, because the protein and DNA should not unfold during the simulation, it may be difficult to use parallel tempering[97] to enhance the sampling of unbiased degrees of freedom. Such simulations would certainly require considerably larger computational resources than we had available for our work.

The 200 ns simulations that were described in the sections on metadynamics took ~100 h on 32 CPU cores. It is thus not prohibitively expensive for those who are interested in tackling
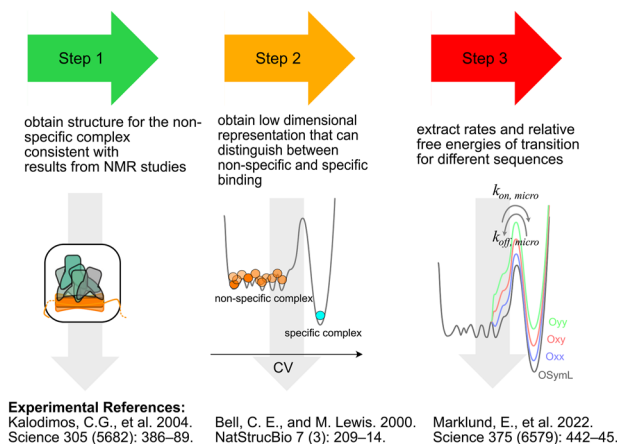
27 September 2023 14:02:52

FIG. 12. Schematic illustration and key literature that illustrates the problem that we put forth to the enhanced sampling community. The figure shows the three steps to solving the problem of understanding DNA binding and recognition. The first of these steps is to identify an ensemble of structures for the non-specifically bound state that is consistent with the experimental data from NMR. The second step is to find a CV that can drive the system between the non-specifically bound and specifically bound states. The final step is then to calculate the free energy change for this transition and the kinetic rates of unbinding. The rates and relative free energies that emerge from these simulations can then be compared with and used to interpret the ample data that is available from experiments.

this sampling challenge to reproduce the calculations we have performed. We provide example input files on GitHub (https://doi.org/10.5281/zenodo.7794684) to help interested researchers get their simulations started quickly.

Figure 12 shows a graphical summary of the challenge that we propose to the community and highlights the key experimental papers. As the figure shows, there are three parts to the problem that we think can be tackled separately. The first problem is to identify the structure of the non-specifically bound complex. In this paper, we endeavored to identify this structure by running simulations with restraints that were informed by the results from NMR experiments and obtained results that were inconclusive. We have provided data on the ensemble of structures that was explored in these simulations, but other researchers may find that they can obtain better structures by performing experiment directed simulations in some other manner.

The next step of the challenge in Fig. 12 is to identify some low dimensional coordinate that can distinguish between configurations of the protein that are specifically bound to DNA from those that are non-specifically bound. This coordinate could be a physically inspired CV like the ones we have discussed in this paper or it could be a machine-learned coordinate that is generated by combining many physically inspired CVs. Ideally, this CV could then be used to drive the protein between the specifically and non-specifically bound states, which, we have shown, is difficult to do using established methods. Such simulations would allow one to extract the relative free energies of the two bound states and the rate constants for the transitions. As the third panel of Fig. 12 indicates the results from such simulations could be compared with the results from experiments. These simulations may even provide a rationale for better

understanding the kinetic rate information that emerges from the experiments. At the very least, we hope that if the sampling challenge we have laid out in the preceding sections is accepted widely, there will be a lively discussion of the strengths and weaknesses of these various new methods and the mechanism for protein–DNA binding.

SUPPLEMENTARY MATERIAL

See the supplementary material for details on the characterization of the ion parameters, the preparation of NMR restraint files, a supplementary figure from metadynamics simulations, and a detailed description of the collective variables used for biasing and visualization.

ACKNOWLEDGMENTS

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

G.A.T. conceived the idea of the paper. J.E. conceived the idea of using molecular dynamics simulations to study microscopic binding and unbinding rates for the LacI-operator system. D.v.d.S. conceived the idea of using NMR-restraint to sample the non-specific state. M.L. carried out the simulations, analysis, and produced all figures with input from all authors. G.A.T. and M.L. wrote the manuscript with input from all authors.

Malin Lüking: Data curation (equal); Formal analysis (equal); Investigation (equal); Project administration (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **David van der Spoel**: Conceptualization (equal); Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Writing – review & editing (equal). **Johan Elf**: Conceptualization (equal); Funding acquisition (lead); Project administration (equal); Resources (equal); Supervision (equal); Writing – review & editing (equal). **Gareth A. Tribello**: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal);

Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

Starting structures, input files for the simulations as well as scripts to prepare NMR restraints and PLUMED input files for different starting structures are openly available at GitHub https://doi.org/10.5281/zenodo.7794684.[98] Trajectories for the different simulations can be found at the SciLifeLab data repository under the following DOI: https://doi.org/10.17044/scilifelab.22348714.v1.[99]

REFERENCES

[1] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).

[2] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science **330**, 341 (2010).

[3] D. P. Landau, S.-H. Tsai, and M. Exler, Am. J. Phys. **72**, 1294 (2004).

[4] D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).

[5] L. Wang, R. A. Friesner, and B. J. Berne, J. Phys. Chem. B **115**, 9431 (2011).

[6] G. Ashton, N. Bernstein, J. Buchner, X. Chen, G. Csányi, A. Fowlie, F. Feroz, M. Griffiths, W. Handley, M. Habeck, E. Higson, M. Hobson, A. Lasenby, D. Parkinson, L. B. Pártay, M. Pitkin, D. Schneider, J. S. Speagle, L. South, J. Veitch, P. Wacker, D. J. Wales, and D. Yallup, Nat. Rev. Methods Primers **2**, 39 (2022).

[7] R. Elber, Annu. Rev. Biophys. **49**, 69 (2020).

[8] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[9] W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).

[10] R. J. Allen, C. Valeriani, and P. Rein Ten Wolde, J. Phys.: Condens. Matter **21**, 463102 (2009).

[11] J. B. Abrams and M. E. Tuckerman, J. Phys. Chem. B **112**, 15742 (2008).

[12] G. Bussi and A. Laio, Nat. Rev. Phys. **2**, 200 (2020).

[13] J. Kästner, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 932 (2011).

[14] W. Chen and A. L. Ferguson, J. Comput. Chem. **39**, 2079 (2018).

[15] F. Giberti, G. A. Tribello, and M. Ceriotti, J. Chem. Theory Comput. **17**, 3292 (2021).

[16] S.-T. Tsai, Z. Smith, and P. Tiwary, J. Chem. Theory Comput. **17**, 6757 (2021).

[17] G. A. Tribello and P. Gasparotto, Front. Mol. Biosci. **6**, 46 (2019).

[18] G. A. Tribello, F. Giberti, G. C. Sosso, M. Salvalaglio, and M. Parrinello, J. Chem. Theory Comput. **13**, 1317 (2017).

[19] B. Cheng, G. A. Tribello, and M. Ceriotti, Phys. Rev. B **92**, 180102 (2015).

[20] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, Chem. Rev. **116**, 7078 (2016).

[21] H. Ehrlich, E. Bailey, M. Wysokowski, and T. Jesionowski, Biomimetics **6**, 46 (2021).

[22] F. Jacob and J. Monod, J. Mol. Biol. **3**, 318 (1961).

[23] F. Crick, Nature **227**, 561 (1970).

[24] A. E. Osbourn and B. Field, Cell. Mol. Life Sci. **66**, 3755 (2009).

[25] A. Jobe and S. Bourgeois, J. Mol. Biol. **69**, 397 (1972).

[26] J. Koubek, J. Schmitt, C. V. Galmozzi, and G. Kramer, Front. Mol. Biosci. **8**, 689755 (2021).

[27] W. Gilbert and B. Müller-Hill, Proc. Natl. Acad. Sci. U. S. A. **58**, 2415 (1967).

[28] F. R. Blattner, G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao, Science **277**, 1453 (1997).

[29] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf, Science **336**, 1595 (2012).

[30] A. D. Riggs, S. Bourgeois, and M. Cohn, J. Mol. Biol. **53**, 401 (1970).

[31] R. B. Winter, O. G. Berg, and P. H. von Hippel, Biochemistry **20**, 6961 (1981).

[32] P. H. Richter and M. Eigen, Biophys. Chem. **2**, 255 (1974).

[33] M. D. Barkley, Biochemistry **20**, 3833 (1981).

[34] Y. M. Wang, R. H. Austin, and E. C. Cox, Phys. Rev. Lett. **97**, 048302 (2006).

[35] A. Mahmutovic, O. G. Berg, and J. Elf, Nucleic Acids Res. **43**, 3454 (2015).

[36] E. Marklund, B. van Oosten, G. Mao, E. Amselem, K. Kipper, A. Sabantsev, A. Emmerich, D. Globisch, X. Zheng, L. C. Lehmann, O. G. Berg, M. Johansson, J. Elf, and S. Deindl, Nature **583**, 858 (2020).

[37] M. Lüking, J. Elf, and Y. Levy, J. Phys. Chem. B **126**, 9971 (2022).

[38] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'Conner, D. W. Noble, and P. H. Von Hippel, Proc. Natl. Acad. Sci. U. S. A. **74**, 4228 (1977).

[39] T. Hu, A. Y. Grosberg, and B. I. Shklovskii, Biophys. J. **90**, 2731 (2006).

[40] E. Marklund, G. Mao, J. Yuan, S. Zikrin, E. Abdurakhmanov, S. Deindl, and J. Elf, Science **375**, 442 (2022).

[41] S. Banerjee, R. Biswas, K. Seki, and B. Bagchi, J. Chem. Phys. **141**, 124105 (2014).

[42] M. Slutsky and L. A. Mirny, Biophys. J. **87**, 4021 (2004).

[43] E. G. Marklund, A. Mahmutovic, O. G. Berg, P. Hammar, D. van der Spoel, D. Fange, and J. Elf, Proc. Natl. Acad. Sci. U. S. A. **110**, 19796 (2013).

[44] I. Leven and Y. Levy, Nucleic Acids Res. **47**, 5530 (2019).

[45] H.-X. Zhou, Proc. Natl. Acad. Sci. U. S. A. **108**, 8651 (2011).

[46] A. Marcovitz and Y. Levy, J. Phys. Chem. B **117**, 13005 (2013).

[47] C. A. Brackley, M. E. Cates, and D. Marenduzzo, Phys. Rev. Lett. **109**, 168103 (2012).

[48] M. Castellanos, N. Mothi, and V. Muñoz, Nat. Commun. **11**, 540 (2020).

[49] C. E. Bell and M. Lewis, Nat. Struct. Biol. **7**, 209 (2000).

[50] S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, EMBO J. **13**, 3348 (1994).

[51] A. D. Riggs, R. F. Newby, and S. Bourgeois, J. Mol. Biol. **51**, 303 (1970).

[52] V. P. Chuprina, J. A. C. Rullmann, R. M. J. N. Lamerichs, J. H. van Boom, R. Boelens, and R. Kaptein, J. Mol. Biol. **234**, 446 (1993).

[53] C. G. Kalodimos, A. M. J. J. Bonvin, R. K. Salinas, R. Wechselberger, R. Boelens, and R. Kaptein, EMBO J. **21**, 2866 (2002).

[54] M. Barbi and F. Paillusson, Adv. Protein Chem. Struct. Biol. **92**, 253 (2013).

[55] X. Chen, M.-Y. Tsai, and P. G. Wolynes, J. Am. Chem. Soc. **144**, 1835 (2022).

[56] H. M. Sasmor and J. L. Betz, Gene **89**, 1 (1990).

[57] M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu, Science **271**, 1247 (1996).

[58] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).

[59] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1–2**, 19 (2015).

[60] D. M. Krüger and S. C. L. Kamerlin, DNA Base Mutator, 2018.

[61] L. L. C. Schrödinger and W. DeLano (2020). "PyMOL," http://www.pymol.org/pymol.

[62] M. O. Kim, S. E. Nichols, Y. Wang, and J. A. McCammon (2020). J. Comput.-Aided Mol. Des. **27**, 235 (2013).

[63] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen, J. Chem. Theory Comput. **7**, 525 (2011).

[64] R. Milo and R. Philips, "What are the concentrations of different ions in cells?" in *Cell Biology by the Numbers* ((eLife, 2020), available at http://book.bionumbers.org/what-are-the-concentrations-of-different-ions-in-cells.

[65] M. P. Doyle and J. L. Schoeni, Appl. Environ. Microbiol. **48**, 855 (1984).

[66] S. Mukherjee and D. Bhattacharyya, J. Biomol. Struct. Dyn. **31**, 896 (2013).

[67] L. Oliveira Bortot, Z. Bashardanesh, and D. van der Spoel, J. Chem. Inf. Model. **60**, 322 (2020).

[68] S.-y. Lin and A. D. Riggs, Cell **4**, 107 (1975).

[69] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**, 712 (2006).

[70] R. B. Best, W. Zheng, and J. Mittal, J. Chem. Theory Comput. **10**, 5113 (2014).

[71] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case, and M. Orozco, Nat. Methods **13**, 55 (2016).

[72] J. L. F. Abascal and C. Vega, J. Chem. Phys. **123**, 234505 (2005).

[73] L. X. Dang and P. A. Kollman, J. Phys. Chem. **99**, 55 (1995).

[74] P. Auffinger, T. E. Cheatham, and A. C. Vaiana, J. Chem. Theory Comput. **3**, 1851 (2007).

[75] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).

[76] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).

[77] G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).

[78] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).

[79] M. Zgarbová, M. Otyepka, J. Šponer, F. Lankaš, and P. Jurečka, J. Chem. Theory Comput. **10**, 3177 (2014).

[80] Q. Liao, M. Lüking, D. M. Krüger, S. Deindl, J. Elf, P. M. Kasson, and S. C. Lynn Kamerlin, J. Phys. Chem. B **123**, 3576 (2019).

[81] C. G. Kalodimos, N. Biris, A. M. J. J. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein, Science **305**, 386 (2004).

[82] E. L. Ulrich, K. Baskaran, H. Dashti, Y. E. Ioannidis, M. Livny, P. R. Romero, D. Maziuk, J. R. Wedell, H. Yao, H. R. Eghbalnia, J. C. Hoch, and J. L. Markley, J. Biomol. NMR **73**, 5 (2019).

[83] A. Sinelnikova and D. van der Spoel, J. Biomol. NMR **75**, 143 (2021).

[84] C. A. E. M. Spronk, G. E. Folkers, A.-M. G. W. Noordman, R. Wechselberger, N. van den Brink, R. Boelens, and R. Kaptein, EMBO J. **18**, 6472 (1999).

[85] C. A. E. M. Spronk, M. Slijper, J. H. van Boom, R. Kaptein, and R. Boelens, Nat. Struct. Biol. **3**, 916 (1996).

[86] A. van der Vaart, Biochim. Biophys. Acta **1850**, 1091 (2015).

[87] F. Pietrucci and A. Laio, J. Chem. Theory Comput. **5**, 2197 (2009).

[88] C. A. Spronk, A. M. Bonvin, P. K. Radha, G. Melacini, R. Boelens, and R. Kaptein, Structure **7**, 1483 (1999).

[89] C. G. Kalodimos, G. E. Folkers, R. Boelens, and R. Kaptein, Proc. Natl. Acad. Sci. U. S. A. **98**, 6039 (2001).

[90] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).

[91] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 020603 (2008).

[92] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, Comput. Phys. Commun. **185**, 604 (2014).

[93] D. Branduardi, G. Bussi, and M. Parrinello, J. Chem. Theory Comput. **8**, 2247 (2012).

[94] M. Bonomi, D. Branduardi, F. L. Gervasio, and M. Parrinello, J. Am. Chem. Soc. **130**, 13938 (2008).

[95] C. E. Bell and M. Lewis, J. Mol. Biol. **312**, 921 (2001).

[96] G. A. Tribello, M. Ceriotti, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **107**, 17509 (2010).

[97] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, J. Am. Chem. Soc. **128**, 13435 (2006).

[98] M. Lüking (2023). "Transcriptionfactor_unbinding: v.0.2.0," Zenodo, V.0.2.0, Dataset. https://doi.org/10.5281/zenodo.7794684

[99] M. Lüking, D. van der Spoel, J. Elf, and G. A. Tribello (2023). "Can we use molecular dynamics to simulate biomolecular recognition?," SciLifeLab. https://doi.org/10.17044/scilifelab.22348714.v1.

27 September 2023 14:02:52