Published online 20 February 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 7* ***3963–3972***
doi:10.1093/nar/gkt099

DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence

Simone Furini[1,*], Paolo Barbini[1] and Carmen Domene[2,3]

[1]Department of Medical Biotechnology, University of Siena, viale Mario Bracci 12, I-53100 Siena, Italy, [2]Chemistry Research Laboratory, University of Oxford, Mansfield Road, Oxford OX1 3TA, UK and [3]Department of Chemistry, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH

Received October 16, 2012; Revised January 27, 2013; Accepted January 28, 2013

ABSTRACT

The lactose repressor protein may bind DNA in two possible configurations: a specific one, if the DNA sequence corresponds to a binding site, and a non-specific one otherwise. To find its target sequences, the lactose repressor first binds non-specifically to DNA, and subsequently, it rapidly searches for a binding site. Atomic structures of non-specific and specific complexes are available from crystallographic and nuclear magnetic resonance experiments. However, what remains unknown is a detailed description of the steps that transform the non-specific complex into the specific one. Here, how the protein first recognizes its binding site has been studied using molecular dynamics simulations. The picture that emerges is that of a protein that is as mobile when interacting with non-specific DNA sequences as when free in solution. This high degree of mobility allows the protein to rapidly sample different DNA sequences. In contrast, when the protein encounters a binding site, the configuration ensemble collapses, and the protein sliding movements along the DNA sequence become scarce. The binding energies in the specific and non-specific complexes were analysed using the Molecular Mechanics Poisson Boltzmann Surface Area approach. These results represent a first step towards a throughout characterization of the DNA-recognition process.

INTRODUCTION

To perform their function, DNA-binding proteins need to find specific binding sites among an overwhelming number of non-specific DNA sequences. Experimental and theoretical evidences support the same model for the DNA recognition process, where the protein first binds non-specifically to DNA, and then it rapidly searches the sequence for the presence of binding sites (1). How a protein recognizes its binding site and how the structure of the protein–DNA complex switches from a non-specific to a specific state is still unknown. Most of the available experimental structures of protein–DNA complexes refer to proteins bound to their target sequences. Only a few DNA-binding proteins have been structurally solved in their non-specific state including, for example, BamHI (2), the λ-repressor (3) and the lactose repressor (4). The lactose repressor (LacI) controls the expression of a set of genes involved in the lactose metabolism in the bacterium *Escherichia coli*. LacI was the first gene-regulatory protein discovered, and it is still one of the most studied [see reference (5) for a review]. The abundance of structural and functional data available for LacI makes it the ideal system when aiming at characterizing the DNA-recognition process at the atomic level by means of simulation strategies. Here, we used Molecular Dynamics (MD) simulations and free energy calculations within the Molecular Mechanics Poisson Boltzmann Surface Area approach (MM/PBSA) to reveal how the lactose repressor is able to recognize its binding site while moving along a DNA molecule.

The lactose repressor is made of four identical monomers. From the point of view of the DNA recognition process, LacI can be described as a dimer of dimers (6). Each dimer binds specifically to a single DNA-binding site, and the binding of the two dimers to separate binding sites further stabilizes the protein–DNA complex (7). Considering that the DNA-binding sites are recognized by a dimer composed of identical monomers (8–11), it is not surprising that the specific sequences are symmetric, or pseudo-symmetric (12,13), and that each half-site is recognized by one of the two monomers. The DNA-binding domain of LacI is made by the first 62 residues of the N-terminal, and this 62-residue long amino-terminal fragment of LacI can still bind DNA in a

*To whom correspondence should be addressed. Tel: +39 05 775 85730; Fax: +39 05 775 86173; Email: simone.furini@unisi.it

specific way (14). Residues 1–46 form a helix-turn-helix domain, characterized by 3 α-helices: H1 (residues 5–14), H2 (residues 16–25) and H3 (residues 31–45). These secondary structural elements remain the same in specific and non-specific complexes (4), as well as in monomers unbound from DNA (15). When LacI binds to non-specific DNA, the residues of this helix-turn-helix domain interact with the backbone of DNA, and the resulting network of interactions maintains the α-helix H2, also known as the recognition helix, in close proximity to the DNA major grove. This is likely to facilitate the interaction of H2 with the edges of the DNA bases, enabling the recognition of a specific binding site. Residues 47–62 at the C-terminal of the DNA-binding domain, the hinge region, are disordered in the non-specific complex. When the protein binds to a specific DNA sequence, residues 50–58 in the hinge region fold into an α-helix, named hinge helix, and the DNA bends at the centre of symmetry of the binding site by ~37° (8,11). In the specific complex, the hinge helix of each monomer is located at the centre of symmetry of the binding site, where the helices interact with each other and with the bases of the DNA minor grove.

Interactions between the two protein monomers are crucial for specific DNA recognition and binding. The affinity of a 62-residue long amino-terminal fragment to specific DNA increases by three orders of magnitude if a disulphide bridge is engineered between the hinge helices of the two monomers (mutating Val52 to Cys) (16). This demonstrates that the helix-turn-helix domains of the two monomers are linked in the specific complex, which obviously limits their relative movements. In contrast, the hinge regions are disordered in the non-specific complex resulting in more flexible and mobile helix-turn-helix domains, as also evidenced by the lack of a crystallographic structure of LacI bound to non-specific DNA. Thus, considering that LacI finds its target site while moving along non-specific DNA, and that the movements of the helix-turn-helix domains are not correlated to each other in the non-specific complex, it should be possible to examine how LacI first recognizes its binding site by just considering a single LacI monomer aligned with half binding site. Here, this recognition process is analysed by MD simulations. MD simulations are becoming an essential tool for the study of protein–DNA complexes, as a result of better force fields and increasing computational capabilities (17). Recent noteworthy examples include the study by Seeliger *et al.* (18) that showed that MD simulations together with free-energy calculations can provide quantitative predictions of protein–DNA binding energies and the study by Yamasaki *et al.* (19) that quantified the contribution of direct in indirect readout by a statistical approach based on MD trajectories. More specifically, regarding the DNA-binding properties of the lactose repressor, the dynamics of a LacI tetramer bound to DNA was analysed by coarse-grained simulations (20), whereas all-atom MD simulations provided atomic details about the allosteric transitions of the repressor (21) and the relationship between DNA-bending and specific binding (22). In this study, we focused on the first step of the protein–DNA recognition process, i.e. the recognition of half-binding site by a LacI monomer. To this end, we compared the dynamical behaviour and the binding energies of three protein–DNA complexes. Two of these protein–DNA complexes correspond to the helix-turn-helix domain of a LacI monomer in contact with a non-specific and a specific DNA sequence, respectively. The position of the protein was defined as that observed in nuclear magnetic resonance experiments of the non-specific complex. The third model was composed by the same protein and specific DNA but with the protein rotated by ~25° on the DNA, as experimentally observed in the specific complex (4) (Figure 1). At present, it is still unknown whether a modification in the orientation of the protein in the specific complex is an essential step in the recognition of the half-binding site or whether it is a consequence of the deformation of the DNA in the specific complex. In addition, as the DNA is bent in the specific complex, and consequently distorted, the helix-turn-helix domain cannot recognize the binding site by establishing the same protein–DNA interactions present in the specific complex. Therefore, how is a binding site first identified? MD simulations have the potential to shed light into these questions revealing which driving forces characterize the recognition process and which steps drive the transformation of a non-specific complex into the specific one.

MATERIALS AND METHODS

Molecular systems

Three atomic systems were defined for the protein–DNA complexes: ASPASP (residues 1–46 of LacI on a non-specific DNA sequence), SPASP and SPSP (residues 1–46 of LacI on a specific DNA sequence). Three other atomic systems were defined for the isolated protein, specific DNA and non-specific DNA. The starting structure of the ASPASP model was based on the Protein Data Bank entry 1OSL (4). Residues 1–46 of a single protein monomer were included in the model. Two nucleotides per strand in a perfect B-form were added to the DNA fragment from 1OSL. The final DNA fragment was 20-nucleotide long, with sequence 5′-GCGATAAGATATCT TATCGC-3′. An alternative DNA fragment with sequence 5′-AATTGTGAGCGCTCACAATT-3′ was defined with the software *nucgen* of the AMBER suite (23), assuming a perfect B-DNA structure. In the former DNA sequence, AATTGTGAGC corresponds to the left half site of operator O_1. This DNA molecule was superimposed on the ASPASP model, using the DNA backbone atoms for the fitting procedure. The SPASP model was defined taking the DNA structure from the specific DNA sequence and the protein structure from the ASPASP model. The crystallographic structure of the lactose repressor on the operator O_S, Protein Data Bank entry 1EFA (8), was used to define a second model for the protein attached to the specific DNA sequence. The crystallographic structure was superimposed on the SPASP model using the DNA backbone atoms of the left half binding site as a reference. Model SPSP was defined taking the DNA model from

SPASP and the protein structure from the crystallographic structure 1EFA. The starting configuration for the free protein in solution was taken from the ASPASP model removing the DNA, whereas the starting structures for specific and non-specific DNA in solution were taken from model SPASP and ASPASP, respectively, removing the protein. All the protein residues were considered in their default protonation states following the prediction of the PROPKA algorithm at neutral pH (24). Histidines were protonated in the epsilon position. N- and C-terminus were respectively acetylated and amidated. Hydrogen atoms were added with the software *psfgen* of NAMD (25). An ester bond was defined between residues 1 and 20 in each DNA strand. The systems were solvated in an orthorhombic box of ~8000 water molecules. Sodium and chloride ions were added to neutralize the systems up to a final concentration of 100 mM.

MD

The systems were equilibrated by 5000 steps of energy minimization, followed by a 250 ps MD simulation in the NVT ensemble, with harmonic restraints (20 kcal*mol^{-1}*Å^{-2}) applied to the backbone atoms of the biomolecules. The harmonic restraints were gradually reduced to zero in a 750 ps MD simulation in the NPT ensemble. Production run in the NPT ensemble followed (200 ns for each system). MD simulations were run using NAMD2.8 (25) and the CHARMM-27 all-atom force field with CMAP correction (26). The TIP3 model was used for water molecules (27). The temperature was maintained at 300 K by Langevin dynamics with damping factor equal to 5 ps^{-1}. Periodic boundary conditions were applied, and the pressure was kept at 1 atm by the Nosé-Hover Langevin method, with an oscillation period of 200 fs and a damping time of 100 fs (28,29). A smoothed cut-off (10–12 Å) was used for the van de Waals interactions. Electrostatic forces were computed by the Particle Mesh Ewald algorithm (30) with a maximum grid spacing of 1.0 Å. Bonds with hydrogen atoms were restrained by the SETTLE algorithm (31), to use a time step of 2 fs. A cumulative time of 1.2 μs was simulated.

MM/PBSA energy calculations

The free energy of binding, ΔG_{bind}, is given by:

$$\Delta G_{bind} = G_{P+D} - (G_P + G_D) \quad (1)$$

where G_{P+D}, G_P and G_D are the free energies of the complex, the isolated protein and DNA, respectively. In the MM/PBSA approach, each free energy term in Equation (1) is calculated as:

$$G = E_{bond} + E_{vdw} + E_{elec} + G_{PB} + G_{SA} - TS_s \quad (2)$$

where E_{bond} is the contribution from the molecular mechanics bond energy, i.e. the sum of the bond, angle and dihedral energies; E_{vdw} is the molecular mechanics van der Waals energy contribution; E_{elec} is the molecular mechanics electrostatic energy; G_{PB} and G_{SA} are polar and non-polar contributions to the solvation energy; T is the absolute temperature and S_s is the solute entropy. Polar and non-polar contributions to the solvation energy were calculated using the APBS software (32). The probe radius for the definition of the molecular surfaces was 1.4 Å. The relative dielectric constant of the solvent was set to 80, and three different values were adopted for the relative dielectric constant of the biomolecules (1, 2 and 4). The non-polar solvation energy was assumed proportional to the solvent accessible surface area, with proportionality constant equal to 0.0072 kcal*mol^{-1}*Å^{-2}. The solute entropy was estimated from the covariance matrix of the atom-positional fluctuations and the Schlitter's formula (33), as implemented in the CARMA software (34). The triple-trajectory paradigm was adopted, i.e. the energy terms were calculated using separate MD trajectories for the protein–DNA complex, for the isolated protein and DNA, respectively. The first 10 ns of each trajectory were considered as equilibration period and not included in the energetic analyses. Average energies were calculated for the remaining part of the MD trajectories taking snapshots every 1 ns. This sampling period corresponds to the minimal one that ensured lack of correlation between the calculated energy values as estimated by the Ljung-box (35) lack of correlation statistical test with a confidence of 95% (more details in Supplementary Material).

RESULTS

Dynamics of specific and non-specific complexes

The ASPASP system corresponds to a 46-residue long N-terminal fragment of LacI interacting with a non-specific DNA sequence. The first and the last residues of each DNA fragment were linked together via an ester bond, thus modelling an infinite long DNA molecule by the use of periodic boundary conditions (see Figure 1 and the 'Materials and Methods' section for more details about the definition of the starting models). Modelling DNA as an infinite molecule significantly hinders bending, twisting and stretching. However, as we are interested in the first step of the recognition process, before DNA bending, this configuration was preferred to the alternative choice of a DNA molecule with free endings. As expected, the protein remained bound to the DNA for the entire course of the MD trajectory (200 ns). The average distance between the centres of mass of the protein and the DNA in the plane orthogonal to the DNA axis was found to be 16.2 ± 0.6 Å. The number of water molecules that were closer than 4 Å from both the protein and the DNA was 34 ± 4. These interfacial water molecules easily exchanged with water molecules from bulk solution in the course of the MD simulation. Salt bridges were observed between residues Arg22 and Lys2, and the DNA backbone atoms (Supplementary Figure S1 in Supplementary Material). Hydrogen bonds between the protein and the DNA backbone atoms stabilize the protein–DNA interface, with a major contribution from residues Ser16, Tyr17, Thr19, Ser31 and Thr34 (Supplementary Figure S2 in Supplementary Material).

The average number of h-bonds was found to be 4 ± 1; two atoms were considered to form a hydrogen bond if they were closer than 3.0 Å and if the donor-hydrogen-acceptor angle was lower than 30°. It was observed that the hydrogen bonds formed by residues Ser16 and Tyr17 are not always directed towards the backbone atoms of the same DNA bases. Instead, Ser16 switches between the backbone atoms of A6 and T5 of one DNA strand (blue strand in Figure 1), and Tyr17 switches between the backbone atoms of C13 and T12 of the complementary DNA strand (red strand in Figure 1). Creation and deletion of h-bonds between the protein residues and successive DNA bases is indicative of protein movements with respect to the DNA sequence. The relative movement of the protein along the DNA is confirmed by the root mean square displacement (RMSD) of the protein backbone atoms (Figure 2). The RMSD was calculated after superposition of all the frames using the DNA backbone atoms as a reference. Two positions of the protein along the DNA sequence are suggested by the RMSD shown in Figure 2, in agreement with the two possible configurations observed for the hydrogen bonds. The relative movement of the protein along the DNA can also be identified by looking at the position of residues Gln18 and Arg22 with respect to the plane of the DNA bases (Figure 3). These two residues belong to helix H2, and they are crucial for specific binding. Therefore, even if the protein remains bound to the DNA molecule for the entire MD trajectory, it is not fixed at a single DNA sequence. Conversely, the protein is moving and sampling at least two different sequences.

The protein behaves radically different if the helix-turn-helix domain is aligned to a specific binding site (left-half binding site of operator O_1) like in the SPASP and SPSP systems. The number of water molecules that were closer than 4 Å from both the protein and the DNA was 34 ± 3 in both simulations of the specific

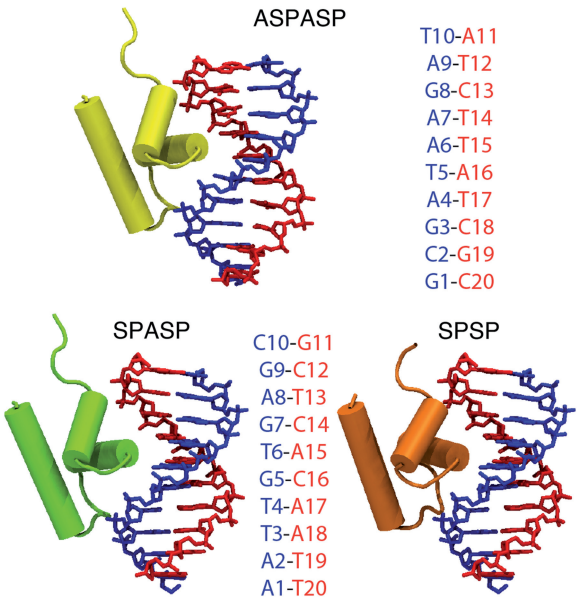

Figure 1. Atomic systems ASPASP, SPASP and SPSP. Residues 1–46 of the helix-turn-helix domain are shown in cartoon representation, together with a fragment from the 20-bp DNA molecule. The sequence of the DNA fragment is also shown, using the same colour scheme of the molecular representation. The DNA sequence is different in ASPASP and SPASP, but the position of the protein with respect to the DNA molecule is the same, and it corresponds to the experimental position of the protein bound to non-specific DNA. SPSP has the same DNA sequence of SPASP, but the position of the protein with respect to DNA corresponds to the experimental position of the protein bound to the specific DNA sequence.

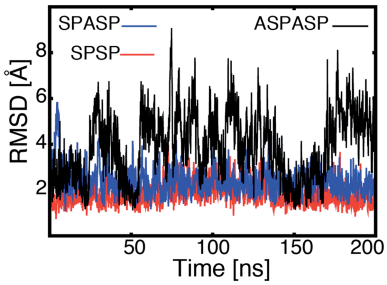

Figure 2. Protein movement along the DNA. For each trajectory, all the frames were superimposed on the first one, using the backbone atoms of the DNA molecule as reference. Then, the RMSD of the backbone atoms of residues 6–45 was calculated. Variations in the RMSD correspond to movements of the helix-turn-helix domain along the DNA molecule.

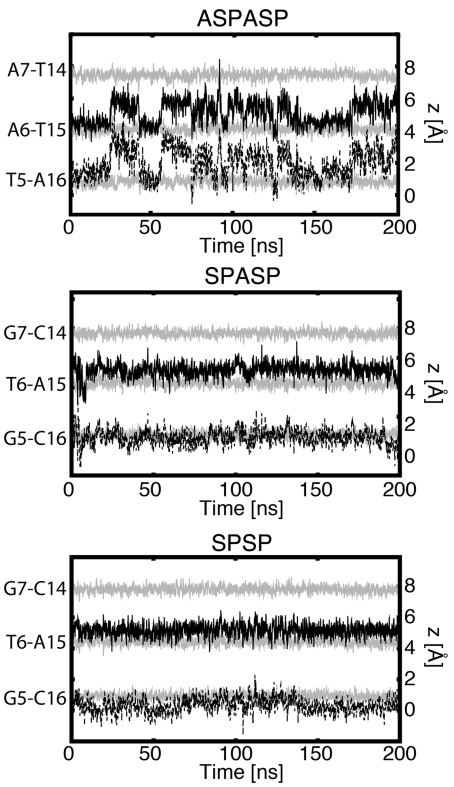

Figure 3. Movement of Gln18 and Arg22 with respect to DNA bases. The coordinate along the DNA axis of the base plane, defined as the centre of mass of the atoms belonging to the aromatic ring of the bases, is shown for three base pairs (grey lines). A black continuous line represents the axial coordinate of the centre of mass of the side chain atoms of residue Gln18. A dashed black line represents the axial coordinate of the side chain atoms of residue Arg22.

complex. Like in the simulation of the non-specific complex, these water molecules easily exchanged with bulk water. The same salt bridges observed for the non-specific complex were observed in simulations of the specific complex, with the only difference that the salt bridge formed by Lys2 appeared more stable in the trajectories SPASP and SPSP compared with the trajectory ASPASP (Supplementary Figure S1). In SPASP, the protein–DNA distance in a 200-ns MD trajectory was slightly shorter than the one observed for the non-specific complex (15.4 ± 0.5 Å). In contrast to the simulation of the non-specific complex, the protein remains anchored on the DNA molecule (Figure 2). This situation correlates well with an increase in the average number of hydrogen bonds between the protein and the DNA backbone with values ranging from 7 ± 1 in SPASP to 4 ± 1 in ASPASP. Besides, the directionality of the residues that participate in these h-bond networks in the SPASP complex is constantly targeting the same DNA bases (Supplementary Figure S3), and the key residues Gln18 and Arg22 are always aligned with the same DNA bases (Figure 3). The starting configurations of SPASP and ASPASP are identical—with the exception of the DNA sequence—and the same simulation protocol was used for the two systems. The difference between SPASP and SPSP was the starting configuration of the protein, which in SPSP resembles the experimental position of the helix-turn-helix domain in the specific complex, and in the SPASP, it resembles the experimental position of the helix-turn-helix domain in the non-specific complex (see Figure 1 and 'Materials and Methods' section). In the starting configuration of SPSP, protein and DNA are at a shorter distance, and this is maintained for the entire MD trajectory (14.6 ± 0.5 Å). Two major differences in the protein–DNA contacts can be observed between SPASP and SPSP: (i) a stable hydrogen bond between Leu6 and the DNA backbone in SPSP, which is absent in SPASP and (ii) a base-specific protein–DNA interaction in SPSP between Tyr17 and the base edges of G7 (Supplementary Figure S4). Like in the case of SPASP, in the SPSP system, the protein does not move with respect to the DNA sequence in a 200-ns MD trajectory. This is reflected in the RMSD values of the protein backbone atoms (Figure 2), the hydrogen bonds between the protein and the DNA-backbone (Supplementary Figure S4) and the alignment of Gln18 and Arg22 with the plane of the DNA bases (Figure 3). In other words, although the protein samples different DNA sequences when placed on a non-specific site (ASPASP), it remains aligned with respect to a single sequence in the case of a specific binding site (SPASP and SPSP).

To study the protein–DNA-recognition process at atomic level, it is useful to compare the dynamical features of the protein in the three complexes: ASPASP, SPASP and SPSP. The root mean square fluctuation of the alpha-carbon atoms is almost identical in the trajectories of systems ASPASP and SPASP (Figure 4). The highest mobility was observed for the N-terminal (residues 1–9) and for the loop between helices H2 and H3 (residues 25–31). The same situation was observed in the trajectory of the free protein in solution, which suggests that the mobility of the helix-turn-helix domain is analogous in the case of the protein bound to a non-specific DNA or to the specific binding site in the SPASP configuration. In contrast, a dramatic reduction in mobility characterized the SPSP system (Figure 4).

A better insight into the configurational space sampled by the protein during the simulations can be obtained from the analysis of the entropy along the MD trajectories. The entropy of the protein was calculated by the Schlitter's formula after superposition of all the frames using the backbone atoms as reference. During an MD trajectory, the entropy increases until it reaches a plateau. If two trajectories are merged, three situations may arise at the transition between the first and the second trajectory: (i) the entropy increases, regardless of the order chosen to merge the trajectories as a result of the disjoint nature of the configurational spaces sampled by the two trajectories; (ii) the entropy values remain stable, regardless of the order in which the trajectories are merged, which implies that the two trajectories share the same configurational space; and finally (iii) the entropy increases only if the trajectories are merged in one particular order, whereas the entropy stays the same if the trajectories are merged in the alternative way (36). This last situation is presented when the configurational space of one trajectory is embedded in that of the other. When the trajectories of the free protein in solution or bound to non-specific DNA are merged, the entropy behaves almost like in the second scenario (Figure 5A). Thus, the binding to non-specific DNA does not have a significant effect on the configurational space sampled by the protein. In contrast, when the trajectory of the free protein in solution is concatenated with the trajectories of the protein bound to a specific DNA sequence (either SPASP or SPSP), the situation is more similar to the third scenario. Binding to the specific DNA sequence causes the configurational space sampled by the protein to shrink, an effect that is already evident in SPASP but that it is much more pronounced in SPSP.

Energetics of specific and non-specific binding

The MM/PBSA approach was used with the purpose of revealing the driving forces responsible for the different behaviour of the LacI protein bound to a specific or a non-specific DNA sequence. It was observed that the binding energies of the protein to the specific DNA sequence were always higher than the binding energies of the protein to non-specific DNA (Table 1). The lower affinity of the protein for the specific sequence was entirely due to the high entropic cost associated with specific binding, which was not compensated by a decrease in the enthalpy of binding. In terms of enthalpy, the affinity for DNA follows the order SPSP > SPASP > ASPSP. The polar contribution to the binding energy (molecular mechanics electrostatics energy, ΔE_{elec}, plus polar contribution to the solvation energy, ΔG_{PB}) has a destabilizing effect for all the systems. The magnitude of the relative dielectric constant of biomolecules is a critical parameter in MM/PBSA calculations, with values of 1, 2 or 4 commonly used. A high dielectric constant has the

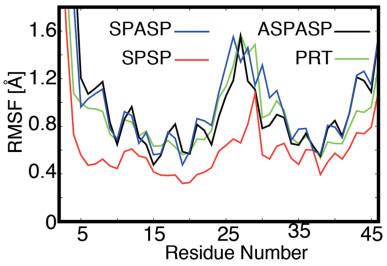

Figure 4. Root mean square fluctuation of protein alpha carbon atoms. Protein in solution (PRT, green line); Protein interacting with a non-specific DNA sequence (ASPASP, black line); Protein interacting with a specific DNA sequence with initial configuration defined as in the non-specific experimental complex (SPASP, blue line); Protein interacting with a specific DNA sequence with its initial configuration defined as in the specific experimental complex (SPSP, red line).

obvious effect of reducing the absolute value of the polar contributions to the binding energy. For the systems analysed here, this translates into a reduction of a destabilizing term. As this destabilizing effect is higher for SPSP, increasing the dielectric constant favours the specific over the non-specific binding. All the complexes are stabilized by van der Waals interactions (ΔE_{vdw}) and by the non-polar contribution to the solvation energy (ΔG_{SA}). These non-polar terms represents $\sim$−51 kcal/mol in ASPASP, $\sim$−57 kcal/mol in ASPSP and $\sim$−71 kcal/mol in SPSP (ΔG_{NP}). The binding to non-specific DNA, ASPASP, is further stabilized by the entropic term, which contributes with $\sim$−32 kcal/mol at 300 K. In contrast, the entropic term is highly destabilizing for the specific complexes SPASP and SPSP, contributing with $\sim$29 and $\sim$73 kcal/mol, respectively.

The entropy of the non-specific complex is higher than the entropy of the specific complex for two reasons: (i) the protein is more mobile when bound to the non-specific DNA than when bound to specific DNA (Figures 4 and 5) and (ii) the protein moves with respect to the DNA molecule in the MD trajectory of the non-specific complex, but not of the specific complex (Figures 2 and 3 and Supplementary Material). To separate these two contributions, the MD trajectory of the non-specific complex was divided into two sets of configurations according to the position of residue Gln18 with respect to the base pair A6-T15 (see Figure 3). The position of Gln18 is highly correlated with the network of hydrogen bonds between the protein and the backbone of DNA and with the relative movement of the protein along the DNA (compare Figure 3 with Figure 2 and Supplementary Figure S2). The distance along the axis of the DNA molecule between the side chain of Gln18 and the plane defined by the base pair A6-T15 has a bimodal distribution (Supplementary Figure S5). In the MD trajectories SPASP and SPSP, the distance between Gln18 and the base pair T6-A15 has a normal distribution. If we assume that the position of Gln18 is representative of the protein alignment along the DNA molecule, the configurations belonging to the two modes represent a protein sampling two different DNA sequences. Thus, if entropy calculation is restricted to the configurations belonging to

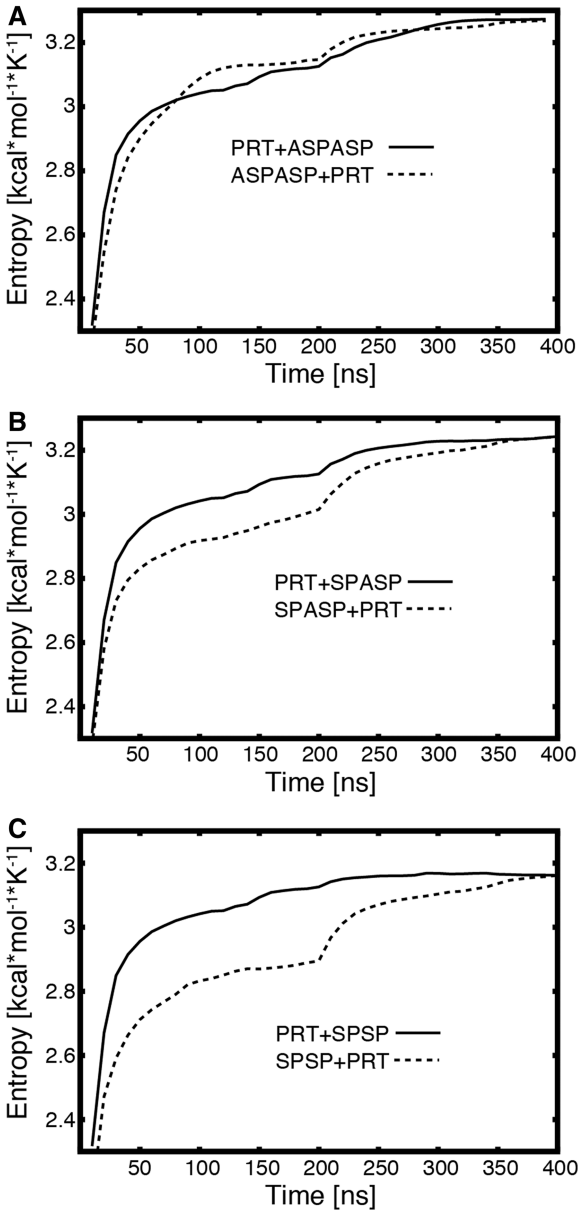

Figure 5. Protein entropy. (**A**) Trajectory of the protein bound to non-specific DNA (ASPASP) attached at the end of the trajectory of the protein in solution (continuous line) and vice versa (dashed line); (**B**) Trajectory of the protein bound to specific DNA in a non-specific orientation (SPASP) attached at the end of the trajectory of the free protein in solution (continuous line) and vice versa (dashed line); (**C**) Trajectory of the protein bound to specific DNA in a specific orientation (SPSP) attached at the end of the trajectory of the free protein in solution (continuous line) and vice versa (dashed line).

a single mode, the contribution of the relative movement of the protein with respect to the DNA is excluded from the entropy value. The bimodal distribution of the distance between Gln18 and A6-T15 was fitted by a sum of two Gaussian functions. The configurations of the MD trajectory ASPASP were divided into two sets according to the mode with higher probability, and the entropy was calculated separately for each set. The entropic cost of binding for the two sets was similar and in the range 10–20 kcal/mol at 300 K (Table 1, lines marked with a).

Table 1. Binding energies in kcal/mol estimated with MM/PBSA

	Dielectric constant of protein and DNA	ASPASP (kcal/mol)		SPASP (kcal/mol)	SPSP (kcal/mol)
ΔE_{bond}		2.8 (4.7)		4.0 (4.2)	−9.7 (4.1)
ΔE_{vdw}		−43.0 (2.2)		−47.3 (2.2)	−60.6 (2.3)
ΔE_{elec}	$\varepsilon_{in} = 1$	−579.4 (13.0)		−652.5 (11.8)	−675.4 (11.4)
	$\varepsilon_{in} = 2$	−289.7 (6.5)		−326.2 (5.9)	−337.7 (5.7)
	$\varepsilon_{in} = 4$	−144.8 (3.2)		−163.1 (3.0)	−168.9 (2.9)
ΔG_{PB}	$\varepsilon_{in} = 1$	627.2 (11.6)		703.0 (11.0)	735.2 (10.2)
	$\varepsilon_{in} = 2$	309.7 (5.7)		346.2 (2.6)	361.2 (5.0)
	$\varepsilon_{in} = 4$	151.4 (2.8)		168.4 (2.6)	175.2 (2.4)
ΔG_{SA}		−7.7 (0.1)		−9.4 (0.1)	−10.1 (0.1)
$\Delta G_P = \Delta E_{elec} + \Delta G_{PB}$	$\varepsilon_{in} = 1$	47.9 (6.8)		50.6 (5.6)	59.8 (5.5)
	$\varepsilon_{in} = 2$	20.0 (3.3)		19.9 (2.7)	23.5 (2.7)
	$\varepsilon_{in} = 4$	6.6 (1.6)		5.3 (1.3)	6.3 (1.3)
$\Delta G_{NP} = \Delta E_{vdw} + \Delta G_{SA}$		−50.7 (0.7)		−56.6 (0.7)	−70.7 (0.7)
$\Delta H_{bind} = \Delta E_{bnd} + \Delta G_P + \Delta G_{NP}$	$\varepsilon_{in} = 1$	−0.1 (7.0)		−2.1 (5.8)	−20.6 (5.7)
	$\varepsilon_{in} = 2$	−27.9 (3.6)		−32.7 (3.1)	−56.8 (3.1)
	$\varepsilon_{in} = 4$	−41.3 (2.2)		−47.3 (2.0)	−74.0 (2.0)
$-T\Delta S_S$		−31.9		28.6	73.2
	[a]	+11.6	+17.6		
$\Delta G_{bind} = \Delta H_{bind} - T\Delta S_S$	$\varepsilon_{in} = 1$	−32.0		26.4	52.6
	[a]	11.5	17.5		
	$\varepsilon_{in} = 2$	−59.8		−4.1	16.3
	[a]	−16.3	−10.3		
	$\varepsilon_{in} = 4$	−73.2		−18.8	−0.9
	[a]	−29.7	−23.7		

Binding energies for the non-specific complex (ASPASP), and for two specific complexes (SPASP and SPSP) are shown. The specific complexes SPASP and SPSP differ in the initial position of the protein with respect to DNA, corresponding to the experimental structure of the non-specific and specific complex, respectively. The standard errors affecting the energy values are shown in parenthesis. The polar contribution to the solvation energy was calculated with a concentration of monovalent salt equal to 25 mM.
[a]Entropies and binding energies calculated dividing the MD trajectory ASPASP in two subsets according to the alignment of residues Gln18 with respect to the base pair A6-T15.

Thus, most of the entropic difference between non-specific and specific binding is ascribable to protein movements along the DNA sequence.

DISCUSSION

The helix-turn-helix domain of the lactose repressor is the first part of the protein to make contact with the target site on the DNA. Therefore, to characterize the dynamics of the recognition process, understanding how this helix-turn-helix domain identifies specific binding sites among an overwhelming number of non-specific sequences is the crucial first step. Analysis of the MD trajectories presented here show significant differences in the dynamics between the specific and the non-specific protein–DNA complexes. First and foremost, when the protein was aligned with a specific binding site, it was stable in the same position with respect to the DNA sequence for the entire MD trajectory. In contrast, a protein aligned with a non-specific sequence moved along the DNA molecule. Although this may appear to be a trivial observation, it should be noted that these results have been obtained from a comparison of two simulations, ASPASP and SPASP, where the protein initially was in the same configuration with respect to the DNA, and that the orientation of the protein was the one experimentally observed for the non-specific complex. It can be concluded that it is possible to reproduce the subtle differences between specific and non-specific protein–DNA complexes in MD simulation. Thus, the use of this computational technique represents a powerful strategy to study the recognition process and to reveal how the non-specific protein–DNA complex switches to the specific one.

The flexibility of the protein in the non-specific complex allows the fast sampling of different DNA sequences by the helix-turn-helix domain. The lactose repressor, as many other DNA-binding proteins, finds its target sites faster than the diffusion limit, and this is the result of a rapid search along the DNA molecule and inter-segmental transfer between distant DNA sequences. The experimental diffusion coefficient of LacI along DNA has an upper limit of $\sim 1*10^6$ bp^2/s (37). We previously estimated a diffusion coefficient in agreement with experiments by computing the potential of mean force for a LacI monomer along a helical trajectory around the DNA molecule (38). Here, a similar result is obtained in an unconstrained MD simulation, i.e. without forcing the protein to move along a predefined helical trajectory. The timescale accessible to classical atomistic MD simulations do not allow the sampling of a complete sliding movement of the protein along the DNA sequence. It is not possible either to exclude that the system evolves towards a non-specific complex different from the experimental one for a longer time scale, as suggested by Sun *et al.* (39) for the non-specific complex of BamHI. The MD simulations presented here show that the protein is highly mobile on the non-specific DNA sequence. The non-specific complex may exist in many different

configurations, and the ones sampled here may only correspond to the local energy minima closer to the experimental structure. However, the critical residues for the recognition process Arg22 and Gln118 showed a movement of ± 1 base pair in a 200-ns MD trajectory, which is in qualitative agreement with the average sampling time of one microsecond per base pair that can be estimated from the experimental diffusion coefficient.

The rapid movement of the protein along non-specific DNA is certainly favoured by the fact that the configurational space of the helix-turn-helix domain is the same when the protein is free in solution or bound to DNA. That is to say that the helix-turn-helix domain preserves its degrees of freedom when bound to non-specific DNA, favouring rapid DNA sampling. An analogous conclusion was reached by measuring the protection factors in hydrogen-deuterium exchange experiments (4). The configurational space collapsed in simulations of specific protein–DNA complexes, which is also in line with experimental observations. Interestingly, residues Gln18 and Arg22 in the recognition helix showed a marked decrease in mobility already in the MD trajectory of the model system SPASP, where the protein is initially in a non-specific configuration with respect to the DNA. These residues were suggested to be the first contact points to specific DNA bases by experimental measurements of hydrogen exchange rates (4).

The entropic cost associated with the collapse of the configurational space in the transition from the non-specific to the specific complex is not compensated by an increase in binding enthalpy. Thus, the probability that the protein binds as in the non-specific complex is higher than the probability that it binds as in the specific complex. The binding energy of the non-specific complex is >50 kcal/mol higher than the binding energy of the specific complex, regardless of the dielectric constant adopted (1, 2 or 4) and of the specific complex considered (SPASP or SPSP). This difference in binding energy is entirely entropic. Entropy calculations in biological systems are inheritably problematic and affected by many approximations such as deviations from quasi-harmonic behaviour or correlation among modes. However, a difference in binding energy of 50 kcal/mol seems to be a substantial overestimation, which cannot be ascribed entirely to the approximation of the theory. An alternative estimate of the binding energies was obtained by dividing the trajectory of the non-specific complex into two sets of configurations, with Gln18 respectively aligned with the base-pair A6-T15 or A7-T14. Dividing the trajectory into two sets can be justified if the two sets correspond to different alignments of the protein on the DNA sequence. Under this hypothesis, if we want to compare the binding energy to the specific DNA sequence with the binding energy to a non-specific DNA sequence, we need to calculate the binding energy for one of the two alternative protein–DNA alignments, and not the binding energy for two non-specific sequences. When the contributions to the entropy from the relative movement of the protein along the DNA are filtered out, the difference in binding energy between the specific and the non-specific sequence decreases of $\sim$40 kcal/mol.

The difference between the binding energies of the three complexes depends on the choice of the dielectric constant. A dielectric constant equal to 1 is the most common choice for similar systems reported in the literature. However, higher values were suggested as a better approximation because of the high density of electrical charge in protein–DNA complexes (40). Together with the uncertainty in the choice of the dielectric constant, the high uncertainty in the computed energies is a further shortcoming of the MM/PBSA estimates. The standard errors quoted in this work may seem high when compared with analogous estimates reported in the literature, especially considering that the length of the MD trajectories presented here are one order of magnitude longer than most of those commonly reported in MM/PBSA calculations. Such standard errors are likely to be caused by the fact that trajectories for MM/PBSA calculations were sampled with a period of 1 ns. This choice was the minimal period that guaranteed the lack of correlation between samples (Supplementary Figure S6 in Supplementary Material). Lack of correlation is necessary if the standard error is interpreted as the uncertainty of the estimate. The uncertainty on the estimated energies is obviously a limitation if these values need to be used to rank a set of DNA sequences in terms of binding affinity, which is not the aim of this study.

Despite the shortcomings of the MM/PBSA approach, it is still possible to extract some robust results from our energetic analyses. First, the binding energy of the specific complexes is always higher than the binding energy of the non-specific complex. The difference in binding energy reaches a minimal value of $\sim$4 kcal/mol for the SPASP complex (which is close to the standard error) if the entropy contribution of protein–DNA movement is excluded from the entropy calculation, and a dielectric constant of 4 is adopted. These data suggest that the DNA-binding domain of a single LacI monomer is not able to bind specifically to DNA. Therefore, other parts of the protein (the hinge-helix, the presence of an adjacent second monomer) need to play a role even in the first step of the recognition process. Specific binding is associated with a sharp bending of the DNA molecule. In this study, we were interested in the events that immediately followed the first contact between the DNA-binding domain of a LacI monomer and a specific DNA sequence; in other words, before DNA bending induced by protein binding might occur. For this reason, DNA was modelled as an infinite molecule, thus constraining its bending, twisting and stretching motions. These geometrical constraints could be partially responsible for the high values estimated for the binding energies, suggesting that a low-energy configuration can be reached only in the presence of DNA bending and of two LacI monomers. In agreement with this hypothesis, the binding energy of the specific protein–DNA complex in a situation where the protein is initially placed as observed experimentally (SPSP) is always at least 20 kcal/mol higher than the binding energy when the protein is placed like in the non-specific complex (SPASP). Thus, the re-orientation of the protein with respect to the DNA observed experimentally, that is rotation of 25° on the DNA, is an event that happens

following recognition, and that it is likely associated also with DNA bending. Noteworthy, residue Leu6 is bound to the DNA in simulation SPSP, and not in SPASP. Experimental measurements revealed that the transition towards the specific complex is slower for Leu6 than for Gln18 and Arg22 (4). This is in agreement with our simulations, which suggests that some structural change needs to take place to account for the high entropic cost resulting from the binding of Leu6 to DNA.

Although the binding energies did not show any selectivity for the specific sequence, the enthalpy of binding was always selective for specific DNA, and the mobility of the protein on the specific DNA was always much lower than the mobility on the non-specific DNA. This has an immediate consequence on the recognition process. The dynamical behaviour of the helix-turn-helix domain of a single LacI monomer changes abruptly in the proximity of a half-binding site, without the need to establish base-specific interactions with the DNA molecule. It is this decrease in mobility that may trigger further structural changes in the protein–DNA complex, with a further reduction in enthalpy that compensate for the entropic cost. The proposed mechanism is remarkably suitable for a protein that finds its binding sites by sliding rapidly along DNA molecules.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–6.

ACKNOWLEDGEMENTS

CINECA is acknowledged for providing support and high performance computing resources through Award N. HP10BE3NY3, 2010.

FUNDING

EPSRC provided computational resources at the national GPGPU test facility hosted by the University of Edinburgh. C.D. thanks The Royal Society for a University Research Fellowship. Funding for open access charge: Departmental funds.

Conflict of interest statement. None declared.

REFERENCES

1. Berg,O.G., Winter,R.B. and Von Hippel,P.H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, **20**, 6929–6948.
2. Viadiu,H. and Aggarwal,A.K. (2000) Structure of BamHI bound to nonspecific DNA: a model for DNA sliding. *Mol. Cell*, **5**, 889–895.
3. Albright,R.A., Mossing,M.C. and Matthews,B.W. (1998) Crystal structure of an engineered Cro monomer bound nonspecifically to DNA: possible implications for nonspecific binding by the wild-type protein. *Protein Sci.*, **7**, 1485–1494.
4. Kalodimos,C.G., Biris,N., Bonvin,A.M., Levandoski,M.M., Guennuegues,M., Boelens,R. and Kaptein,R. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, **305**, 386–389.
5. Lewis,M. (2005) The lac repressor. *C R Biol.*, **328**, 521–547.
6. Friedman,A., Fischmann,T. and Steitz,T. (1995) Crystal structure of lac repressor core tetramer and its implications for DNA looping. *Science*, **268**, 1721–1727.
7. Oehler,S., Eismann,E.R., Krämer,H. and Müller-Hill,B. (1990) The three operators of the lac operon cooperate in repression. *EMBO J.*, **9**, 973–979.
8. Bell,C. and Lewis,M. (2000) A closer view of the conformation of the Lac repressor bound to operator. *Nat. Struct. Biol.*, **7**, 209–214.
9. Bell,C.E. and Lewis,M. (2001) Crystallographic analysis of Lac repressor bound to natural operator O1. *J. Mol. Biol.*, **312**, 921–926.
10. Spronk,C.A., Bonvin,A.M., Radha,P.K., Melacini,G., Boelens,R. and Kaptein,R. (1999) The solution structure of Lac repressor headpiece 62 complexed to a symmetrical lac operator. *Structure*, **7**, 1483–1492.
11. Kalodimos,C., Bonvin,A., Salinas,R., Wechselberger,R., Boelens,R. and Kaptein,R. (2002) Plasticity in protein-DNA recognition: lac repressor interacts with its natural operator 01 through alternative conformations of its DNA-binding domain. *EMBO J.*, **21**, 2866–2876.
12. Gilbert,W. and Maxam,A. (1973) The nucleotide sequence of the lac operator. *Proc. Natl Acad. Sci. USA*, **70**, 3581–3584.
13. Sadler,J.R., Sasmor,H. and Betz,J.L. (1983) A perfectly symmetric lac operator binds the lac repressor very tightly. *Proc. Natl Acad. Sci. USA*, **80**, 6785–6789.
14. Ogata,R.T. and Gilbert,W. (1978) An amino-terminal fragment of lac repressor binds specifically to lac operator. *Proc. Natl Acad. Sci. USA*, **75**, 5851–5854.
15. Kaptein,R., Zuiderweg,E., Scheek,R., Boelens,R. and van Gunsteren,W. (1985) A protein structure from nuclear magnetic resonance data. lac repressor headpiece. *J. Mol. Biol*, **182**, 179–182.
16. Kalodimos,C.G., Folkers,G.E., Boelens,R. and Kaptein,R. (2001) Strong DNA binding by covalently linked dimeric Lac headpiece: evidence for the crucial role of the hinge helices. *Proc. Natl Acad. Sci. USA*, **98**, 6039–6044.
17. Mackerell,A.D. Jr and Nilsson,L. (2008) Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.*, **18**, 194–199.
18. Seeliger,D., Buelens,F.P., Goette,M., de Groot,B.L. and Grubmuller,H. (2011) Towards computational specificity screening of DNA-binding proteins. *Nucleic Acids Res.*, **39**, 8281–8290.
19. Yamasaki,S., Terada,T., Kono,H., Shimizu,K. and Sarai,A. (2012) A new method for evaluating the specificity of indirect readout in protein-DNA recognition. *Nucleic Acids Res.*, **40**, e129.
20. Villa,E., Balaeff,A. and Schulten,K. (2005) Structural dynamics of the lac repressor-DNA complex revealed by a multiscale simulation. *Proc. Natl Acad. Sci. USA*, **102**, 6783–6788.
21. Swint-Kruse,L., Zhan,H. and Matthews,K.S. (2005) Integrated Insights from Simulation, Experiment, and Mutational Analysis Yield New Details of LacI Function. *Biochemistry*, **44**, 11201–11213.
22. Barr,D. and van der Vaart,A. (2012) The natural DNA bending angle in the lac repressor headpiece-O1 operator complex is determined by protein-DNA contacts and water release. *Phys. Chem. Chem. Phys.*, **14**, 2070–2077.
23. Pearlman,D.A., Case,D.A., Caldwell,J.W., Ross,W.S., Cheatham,T.E., Debolt,S., Ferguson,D., Seibel,G. and Kollman,P. (1995) Amber, a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.*, **91**, 1–41.
24. Li,H., Robertson,A.D. and Jensen,J.H. (2005) Very fast empirical prediction and rationalization of protein pKa values. *Proteins*, **61**, 704–721.
25. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kale,L. and Schulten,K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.
26. MacKerell,A.D., Bashford,D., Bellott,M., Dunbrack,R.L., Evanseck,J.D., Field,M.J., Fischer,S., Gao,J., Guo,H., Ha,S. *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.

27. Jorgensen,W.L., Chandresekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
28. Feller,S.E., Zhang,Y.H., Pastor,R.W. and Brooks,B.R. (1995) Constant-pressure molecular dynamics simulation- the langevin piston method. *J. Chem. Phys.*, **103**, 4613–4621.
29. Martyna,G.J., Tobias,D.J. and Klein,M.L. (1994) Constant-pressure molecular-dynamics algorithms. *J. Chem. Phys.*, **101**, 4177–4189.
30. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. and Pedersen,L.G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.*, **103**, 8577–8593.
31. Miyamoto,S. and Kollman,P.A. (1992) Settle—an analytical version of the Shake and Rattle algorithm for wigid water molecules. *J. Comput. Chem.*, **13**, 952–962.
32. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Acad. Sci. USA*, **98**, 10037–10041.
33. Schlitter,J. (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.*, **215**, 617–621.
34. Glykos,N.M. (2006) Software news and updates. Carma: a molecular dynamics analysis program. *J. Comput. Chem.*, **27**, 1765–1768.
35. Ljung,G.M. and Box,G.E.P. (1978) On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.
36. Hsu,S.T., Peter,C., van Gunsteren,W.F. and Bonvin,A.M. (2005) Entropy calculation of HIV-1 Env gp120, its receptor CD4, and their complex: an analysis of configurational entropy changes upon complexation. *Biophys. J.*, **88**, 15–24.
37. Wang,Y.M., Austin,R.H. and Cox,E.C. (2006) Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys. Rev. Lett.*, **97**, 048302.
38. Furini,S., Domene,C. and Cavalcanti,S. (2010) Insights into the sliding movement of the lac repressor nonspecifically bound to DNA. *J. Phys. Chem. B*, **114**, 2238–2245.
39. Sun,J., Viadiu,H., Aggarwal,A.K. and Weinstein,H. (2003) Energetic and structural considerations for the mechanism of protein sliding along DNA in the nonspecific BamHI-DNA complex. *Biophys. J.*, **84**, 3317–3325.
40. Hou,T., Wang,J., Li,Y. and Wang,W. (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.*, **51**, 69–82.