Molecular determinants of the interactions between proteins and ssDNA

Garima Mishra and Yaakov Levy[1]

Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

ssDNA binding proteins (SSBs) protect ssDNA from chemical and enzymatic assault that can derail DNA processing machinery. Complexes between SSBs and ssDNA are often highly stable, but predicting their structures is challenging, mostly because of the inherent flexibility of ssDNA and the geometric and energetic complexity of the interfaces that it forms. Here, we report a newly developed coarse-grained model to predict the structure of SSB–ssDNA complexes. The model is successfully applied to predict the binding modes of six SSBs with ssDNA strands of lengths of 6–65 nt. In addition to charge–charge interactions (which are often central to governing protein interactions with nucleic acids by means of electrostatic complementarity), an essential energetic term to predict SSB–ssDNA complexes is the interactions between aromatic residues and DNA bases. For some systems, flexibility is required from not only the ssDNA but also, the SSB to allow it to undergo conformational changes and the penetration of the ssDNA into its binding pocket. The association mechanisms can be quite varied, and in several cases, they involve the ssDNA sliding along the protein surface. The binding mechanism suggests that coarse-grained models are appropriate to study the motion of SSBs along ssDNA, which is expected to be central to the function carried out by the SSBs.

ssDNA | protein–DNA interaction | coarse-grained model

Double-stranded DNA (dsDNA) contains all of the genetic information necessary for proper cellular functioning. The dsDNA unwinds to form ssDNA. This intermediate structure enables the stored information to be accessed; however, it is thermodynamically less stable than dsDNA, and consequently, it spontaneously forms duplex secondary structures (which inhibit subsequent DNA-processing reactions) and is more susceptible to environmental hazards (which can damage the genome). A solution to this problem adopted across many different domains of life comes from specialized ssDNA binding proteins (SSBs) that bind, protect, and stabilize the ssDNA structures required for essential genomic processes and therefore, enable these processes to proceed (1–3).

Most SSBs share a common structural motif called the *oligosaccharide/oligonucleotide* binding (OB) fold, which consists of a five-stranded β-sheet arranged as a β-barrel capped by a single α-helix to bind ssDNA (4). However, this common structural motif itself has widely varying characteristics. For example, the OB folds in different proteins can range in length from 70 to 150 aa, and the loops connecting the β-sheets that are responsible for binding specificities vary in sequence, length, and conformation (5). Furthermore, it is quite difficult to detect the OB fold domains based on sequence similarity alone, because most proteins containing this structural motif share a low degree (5–25%) of sequence similarity (4).

Most bacterial SSBs form a homotetramer and contain one OB fold per polypeptide, with each such fold capable of binding to ssDNA (6). Binding involves the ssDNA wrapping around the homotetramer, although bacterial SSBs exhibit considerable variability in their association. In contrast to bacterial SSBs, eukaryotic SSBs generally function as a heterotrimer consisting of six OB folds distributed among three subunits; however, only four of these OB folds are involved in binding to ssDNA (7), which results

in an extended arrangement for the protein–ssDNA complex. Telomere end binding proteins, which safeguard the vulnerable telomere 3′ end of ssDNA, also use different numbers of OB folds to bind with ssDNA (8). These proteins show high sequence specificities for their respective telomeric sequences, and this sequence preference may be linked to their function as telomeres (8–10). Furthermore, very few SSBs have been structurally characterized as using the K homology domain (three α-helices packed against a three-stranded β-sheet) and RNA recognition motif domains (8) to form a complex with ssDNA.

The complexes between SSBs and ssDNA are often very thermodynamically stable (11, 12), and their binding relies on the electrostatic interactions between the negatively charged phosphates of the ssDNA and the complementary positively charged residues [lysine (K), arginine (R), or histidine] of the SSBs. The same electrostatic attraction is also required for protein binding to dsDNA (13), with SSBs differentiating ssDNA from its dsDNA competitor on the basis of the intrinsic properties of each, in that dsDNA is significantly less flexible in regards to the spacing and positioning of its negative charges than ssDNA. In addition to the electrostatic interaction, detailed structural studies of crystal complexes of ssDNA–SSB complexes suggest that the aromatic residues tryptophan (W), phenylalanine (F), tyrosine (Y), and histidine (H) also play an important role in ssDNA–SSB binding (Fig. S1).

Although strong electrostatic binding of SSBs to ssDNA is necessary to stabilize the latter, the transient role of SSBs in DNA metabolic processes also requires the recycling of SSBs (i.e., their dissociation from and reassociation with ssDNA) as well as their repositioning within ssDNA complexes (14). Recent studies using single-molecule fluorescence resonance energy transfer have endeavored to shed light on the dynamic activity of different types of SSBs, which were earlier thought to form inert

complexes with ssDNAs. These studies have shown that the *Escherichia coli* SSB functions as a sliding platform that migrates on ssDNA (10, 15), whereas a homodimeric bacterial SSB from *Thermus thermophilus* is able to diffuse spontaneously along ssDNA (16). It was reported that the diffusion coefficient of the *E. coli* SSB along ssDNA is similar to that of transcription factors along dsDNA (17). Efficient sliding dynamics on ssDNA were recently elegantly shown for replication protein A (RPA) (18).

SSBs are not the only groups binding to and moving along ssDNA. For example, the complex between the protection of telomeres protein 1 (Pot1) and thiamine pyrophosphate protein 1 (TPP1; i.e., the Pot1–TPP1 complex) slides back and forth on telomeric DNA (19, 20). Similarly, RPA can bind very tightly to an ssDNA and dissociate rapidly when other SSBs are present in solution (20). Unlike studies (10, 15, 16, 19) that were performed on an infinitely dilute solution, the findings of the study in ref. 20 apply to a wide range of conditions. Thus, the study of SSB diffusion elucidates the mechanism used by SSBs to recruit other SSB-interacting enzymes onto ssDNA and give due access to the genome substrate for subsequent processing.

As described above, the experimental community has given considerable attention to probing the different aspects of SSB–ssDNA complexes. However, almost no efforts have been made on the theoretical front. To address this gap, this paper aims to develop a universal model that can predict the crystal complex of ssDNA with SSBs coming from different domains of life, which can be further extended to study other properties of these complexes in future studies.

Results and Discussion

Role of Electrostatic and Aromatic Interactions in ssDNA–SSB Binding.

The power of the developed coarse-grained model to predict the structures of SSB–ssDNA is presented in Fig. 1. Fig. 1 shows the probability of the occurrence of different states of the SSB–ssDNA complex as characterized by structural parameters that measure the similarity of the predicted ssDNA–SSB interface to the interface in the X-ray or NMR structures. For a refined comparison, the aromatic and positively charged residues that comprise the interface were split into two patches, quantified as D_1 and D_2 (Fig. S1). Areas closer to the (0,0) intersection between the D_1 and D_2 axes represent structures that more closely resemble the interface in the crystal structure. The free energy landscape for formation of each of six studied complexes where the interface between the SSB and ssDNA is modeled by electrostatic and aromatic interactions reflects that, in all six cases, near-native conformations with low values of D_1 and D_2 are reasonably populated (Fig. 1, *Left*). The importance of both the electrostatic and aromatic interactions for the specificity and stability of the interface is illustrated by plotting the interfacial aromatic residues for each system together with the electrostatic potential (Fig. 1, *Right*). We note that coarse-grained models that included only electrostatic or aromatic interactions yielded poor results, with the latter profoundly inferior to the former (Figs. S2 and S3). In the following paragraphs, we focus on the binding of several different SSBs with ssDNA in the presence of these two interactions and examine their role in binding.

The binding of the *Schizosaccharomyces pombe* Pot1p complex with homopolymeric ssDNA using both electrostatic and aromatic interactions is shown in Fig. 1*A*. Although the ssDNA sometimes found the correct protein binding pocket in the presence of electrostatic interactions only (Fig. S2), the resulting complex deviated from the crystal structure. By contrast, when aromatic interactions between the aromatic amino acids of Pot1p and the ssDNA bases were included together with the electrostatic interactions (as in Fig. 1*A*), the structure of the complex tended toward that of the crystal structure. It is noteworthy that the experimental crystal structure for the same complex was obtained with sequence-specific ssDNA (9), which can engage in base stacking. Although we did not consider sequence effects in ssDNA here, we
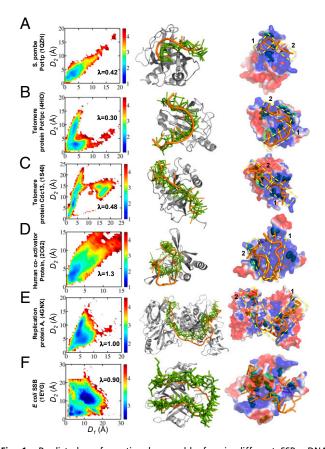

Fig. 1. Predicted conformational ensemble for six different SSB–ssDNA complexes (*A–F*). The similarity measure (*D*) quantifies the overall structural similarity between the predicted binding interface in the ssDNA–protein complex and the interface observed in the crystal structure (a lower value of *D* corresponds to a greater degree of similarity with the crystal structure, with a value of zero representing 100% concordance). For a detailed structural comparison, each SSB–ssDNA interface was divided into two regions (schematically marked by rectangles in *Right*; accurate definition in *SI Methods*) that were assessed separately as D_1 and D_2. The color bar shows the free energy of the different binding states of the complex in the presence of both electrostatic and aromatic interactions. Sample structures corresponding to high probability maps (i.e., the lowest values of D_1 and D_2) are shown in *Center*. In these structures, the ssDNA is shown in green, and for comparison, the conformation of the ssDNA in the corresponding crystal structure is shown in orange (the proteins are shown in gray). *Right* shows the interfacial aromatic residues for each system together with the electrostatic potential. The ssDNA is shown in orange. The aromatic residues (F, Y, W, and H) that interact with the ssDNA are shown as green spheres. Additional aromatic residues that are exposed to the solvent can be found on the protein surface but are not shown. In many cases, the aromatic residues are located in negatively charged regions. The values of λ (the ratio between the total electrostatic and aromatic energies of each SSB–ssDNA complex) are shown in the maps in *Left*. Additional molecular and structural details of each of the six complexes can be found in Table S1.

nevertheless found very good agreement between the conformation of the ssDNA and that of the crystal structure, which strongly suggests that aromatic interactions play a role.

The similarity measures map for the telomeric protein Pot1pc with 9-mer ssDNA shows large variations in D_1 and D_2 with different probabilities in the presence of electrostatic interactions only (Fig. S2). The structure shown in Fig. S3 corresponds to the highest probability patch in the map; a similar structure also occurred at the opposite end of the binding groove with a low probability of occurrence. These extended structures occurring at either end of the Pot1pc binding groove depended solely on the electrostatic interaction. As aromatic interactions

Mishra and Levy

between Pot1pc and ssDNA came into play, a patch with a high probability developed near the crystal structure (Fig. 1B, Left), and the structure corresponding to this region (Fig. 1B, Center), was in very good agreement with the crystal structure.

Another telomere end-binding protein, namely Cdc13, produced nonsymmetric measures for D_1 and D_2 in the presence of electrostatic interactions only (Fig. S2). The small variation in D_1 suggests that the configuration of the ssDNA relative to the residues that contributed to this measure was relatively good. However, the large variation in D_2 reflects that the residues contributing to D_2 still formed a large variety of possible interfaces with the ssDNA. One reason for this may be that the ssDNA detected the negatively charged region on only one side of the binding pocket of the protein as shown in the electrostatic map (Fig. 1C, Right), which is in contrast to the case of Pot1pc, where the positive patch extended to both sides of the groove (Fig. 1B, Right). In this negatively charged region, there were also aromatic residues, which indeed, balanced out the repulsion between the phosphate of ssDNA and the negatively charged amino acid (Fig. 1C, Left) and in turn, gave rise to the crystal structure (Fig. 1C, Center). This feature is also in agreement with earlier observations that the majority of thermodynamic interactions originate from the extensive aromatic DNA interaction surface (21).

Energetics of the Interface in SSB–ssDNA Complexes. The superior ssDNA–SSB interface predictions achieved using a model that includes both electrostatic and aromatic interactions compared with a model that includes only one or the other indicate that both forces are essential to stabilize SSB–ssDNA complexes. To quantify the contribution of each term to the structural stability of the complex, we estimated the total contributions of all of the electrostatic or aromatic interactions in the conformations that are similar to the crystal structures. We then defined a ratio, λ, between the total electrostatic and aromatic energies. The values of λ for the studied systems ranged between 0.3 and 1.3 (Fig. 1). Values lower (higher) than unity indicate that the electrostatic interactions make larger (smaller) contributions than the aromatic interactions. Nevertheless, the fact that λ was close to unity for all of the systems suggests that both interactions are essential for SSB–ssDNA binding. We verified the values obtained for λ from the coarse-grained simulations by estimating the interface energetics of each SSB–ssDNA complex using an atomistic molecular dynamic (MD) force field (Table S2). We note that the systems with $\lambda < 1$ (*S. pombe* Pot1p, Telomere protein Pot1pc, and Telomere protein Cdc13) have shorter ssDNA than the systems with $\lambda \geq 1$ (human coactivator protein, replication protein A, and *E. coli* SSB). The dependence of λ on DNA length may suggest that the aromatic interactions act to anchor the ssDNA to some sites on the SSB and that these anchors are more critical for shorter stretches of ssDNA. The dependence of λ on DNA length may also suggest that, in a biological context (i.e., when the ssDNA is longer), the electrostatic and aromatic interactions are fairly equivalent in terms of their importance for stability.

Role of Flexibility in the Association of ssDNA with SSB. Although the designed transferable coarse-grained model successfully predicts SSB–ssDNA binding, its simplicity is clearly expected to limit its predictive power, especially given the high complexity of these systems. We will exemplify here how the conformational flexibility of either ssDNA or SSB can affect the predicted structures. In the case of ssDNA, its flexibility is linked to electrostatic forces and can, therefore, be modulated by salt concentration. Indeed, the persistence length of ssDNA was shown to decrease with increasing salt concentration (22, 23). To avoid modeling the ssDNA using an elaborate electrostatic model that includes ion condensation effects, we crudely represented the effect of electrostatics on the ssDNA persistence length by adding a dihedral angle between phosphate beads (*SI Methods*).

The effect of the ssDNA persistence length on binding is illustrated for the complex between human coactivator protein PC4 and the 20-mer ssDNA. The ssDNA molecule adopted an extended structure in the binding groove of protein PC4 when only electrostatic interactions were permitted between PC4 and the ssDNA (Fig. S3). As aromatic interactions are switched on, the extended conformation of the ssDNA adopted a U-shaped structure, and the predicted structures increasingly resembled the crystal structure (Fig. 1D), again supporting the importance of both electrostatic and aromatic interactions in our model. Additional improvement in the binding of ssDNA with protein PC4 was observed when the ssDNA was more flexible (i.e., when the dihedral angle potential was not applied) (Fig. S4). The better prediction obtained using flexible ssDNA may be related to the structure of this complex having been determined at a very high salt concentration (24), at which ssDNA is, indeed, expected to be more flexible because of the high screening effect of the charged phosphate groups.

The complex of RPA and the 30-mer ssDNA shows, once again, the importance of electrostatic and aromatic interactions for defining the interface of the complex (both in terms of the participating residues and the shape of the ssDNA) (Fig. 1E and Fig. S2). Imparting flexibility to the RPA protein in addition to both these interactions (Fig. S4) was important to achieving the crystal structure complex. This flexibility, which was introduced to residues with a high B factor, improved the interactions of the ssDNA with B domains (however, near-native assembly was obtained even without the additional flexibility). We point out that, for RPA in contrast to the human coactivator protein PC4, increasing the flexibility of the ssDNA impaired the quality of the predicted complex (Fig. S4).

Another very important and widely studied SSB is a tetramer protein from *E. coli*. The two major modes by which the *E. coli* SSB tetramer binds to ssDNA are (*i*) binding through two of four tetramer subunits, which occludes 35 nt [the (SSB)$_{35}$ mode] and (*ii*) binding through all four tetramer subunits, such that ~65 nt wrap around the tetramer [the (SSB)$_{65}$ mode] (25, 26). The binding mode adopted depends on the solution conditions, with the (SSB)$_{65}$ mode favored at high salt concentrations (>0.2 M NaCl). Raghunathan et al. (6) studied the crystal structure of the SSB tetramer bound to two 35-mer ssDNAs and proposed a model for how a continuous ssDNA can interact with all four subunits of the tetramer in (SSB)$_{65}$ binding mode when the ssDNA segments are in close proximity to each other. Computational study of the association between ssDNA and *E. coli* SSB is, thus, challenging because of not only the size of the systems (of both the protein and the ssDNA) and the high plasticity but also, the structural heterogeneity caused by different salt concentration. In the coarse-grained simulation, the salt condition is introduced by the Debye–Hückel potential that includes a screening effect for the electrostatic interactions at the interface ssDNA–SSB and modulation of the ssDNA persistence length, but obviously, the salt effect can be much richer.

The free energy plot for binding of *E. coli* SSB to ssDNA was calculated in two steps. First, we limited the analysis to those segments of the ssDNA and the dimer that interact with each other. This analysis resulted in the exclusion of 9 nt from the middle of the ssDNA (that might be viewed as a flexible linker), leaving only two segments of 28 nt in the model. Second, we calculated the D_1 and D_2 values corresponding to the first and second dimers with their respective ssDNA segments. The similarity map with electrostatic interactions only (Figs. S2 and S3) is very similar to the map with both electrostatic and aromatic interactions (Fig. 1F), although with different probabilities.

To further differentiate between these structures, we examined the conformation of the ssDNA when interacting with each dimer by measuring the average distance between the nucleotides in both the crystal structure and the simulated structures (Fig. S5). We

plotted the average internucleotide distances in our model vs. the total similarity parameter, D, for electrostatic interactions alone and electrostatic interactions combined with aromatic interactions (Fig. S5). We found that a low value for the internucleotide distance together with smaller values for D_1 and D_2 gave rise to the extended loop of ssDNA that wrapped around L_{23} as shown in Fig. S5. This structure differed from the structure obtained in the presence of electrostatic interactions only (Fig. 1 and Fig. S3). In each dimer, there are two such L_{23} loops, and because our model does not include any bias, the loop of the ssDNA could be formed using either of the L_{23} loops in either of the dimers.

In this model, the binding of the complete tetramer structure to ssDNA gave rise to one extended loop of ssDNA around L_{23} in one dimer, whereas the loop of the other dimer may widen but did not encircle L_{23}. A plausible explanation is that the binding of a dimer to the ssDNA according to the correct crystal structure may constrain the rest of the ssDNA. Consequently, the ssDNA may require more time to bind to both the loops of the tetramer, which is shown in the crystal structure model. However, there was no such constraint when studies were performed with two 35-mer ssDNA in tetramer binding. It should be mentioned that, in our model, the ssDNA linkers between the two dimers each showed different connectivities (Fig. 2), which might originate from the several possibilities of wrapping around four L_{23} loops in the tetramer. In Fig. 2A, the ssDNA took the shortest connectivity path, whereas in Fig. 2B, it took a longer route for connection. In both of these cases, the entry and exit sites of the ssDNA were in close proximity to each other, which is consistent with the proposed model (6). This difference may be attributed to the lack of ssDNA polarity in our model.

Mechanism of ssDNA Binding to SSB Protein. To obtain additional insights into the binding of ssDNA with SSB, we explored the possible mechanisms involved in the formation of the crystal structure. Fig. 3 illustrates the mechanism of ssDNA binding to RPA protein with respect to four sites colored purple (Fig. 3, *Right*). Sites I–IV are defined by the centers of mass of specific residues (Fig. 3). To study the mechanism of association of the ssDNA with SSB, we followed the time-based evolution (measured in MD steps) of the distance between each of four
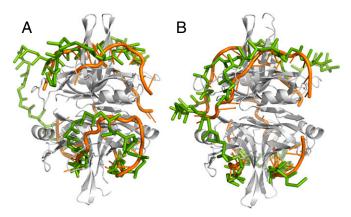

Fig. 3. Mechanism of ssDNA binding to RPA. The association between ssDNA and RPA is analyzed for a typical association trajectory by focusing on four RPA sites marked I–IV and probing the time evolution of the minimal distance between these sites and a DNA nucleotide (black lines) and the identity of that nucleotide (purple lines) (A–D, respectively). Each site is defined by the center of mass of the following residues: site I, W71 and F100; site II, Y308, K318, and Y344; site III, F209 and Y291; and site IV, R29, W31, F57, and R80. The structures in *Right* show four snapshots sampled at the different times indicated by the arrows in A, *Left*. In each snapshot, sites I–IV are shown by van der Waals spheres. The site at which binding to ssDNA takes place is shown in purple, whereas the other sites are shown in gray. The gray areas in each panel indicate the sliding events performed by the ssDNA. The fact that the distance between the RPA site and the ssDNA during sliding is close to that found in the crystal structure indicates a direct interaction. However, the identity of the interacting nucleotide changes continuously because of sliding.

Fig. 2. Two different association modes adopted by ssDNA to bind to tetrameric *E. coli* SSB. As a result of the symmetry of the tetramer, the ssDNA can interact asymmetrically with each dimer in several ways, each of which produces different connectivities (A and B). The predicted ssDNA (nucleotides 1–65) is green, and the ssDNA from the crystal structures, which is discontinued in some parts of the linker region, is shown in orange. The protein is shown in gray. The four acidic C-terminal tails (residues 113–177) were excluded in our modeling, because they are intrinsically disordered and absent in the crystal structure, even when SSB is bound to ssDNA.
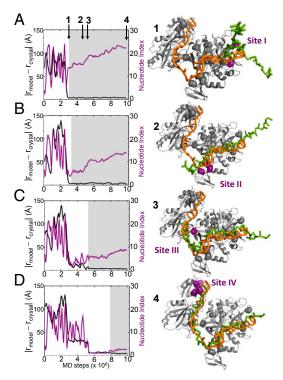
predefined sites on the protein surface and the closest DNA nucleotide (r_{model}) as well as the identity of that nucleotide.

Fig. 3, *Left* shows variations in the value of $|r_{model} - r_{crystal}|$ and the nucleotide index over time for a representative trajectory. The values of r represent the minimum distances that ssDNA nucleotides are from RPA sites I–IV (which are pictured in Fig. 3 A, *Right*, B, *Right*, C, *Right*, and D, *Right*, respectively, at four different MD time points marked by arrows in Fig. 3, *Left*). A very large value of $|r_{model} - r_{crystal}|$ indicates that none of the ssDNA nucleotides approached close to the sites. By contrast, very small near-zero values correspond to the native binding phase of ssDNA, although without giving any information regarding the identity of the nucleotide that interacted with the site, which may, therefore, differ from that used in the crystal structure.

As can be seen from Fig. 3A, *Left*, at MD step number $\sim$2.9 $\times$ 10^6, the value of $|r_{model} - r_{crystal}|$ approached zero, and the fourth nucleotide (right-hand y axis) was positioned at the minimum distance from RPA binding site I in what can be treated as the initial binding event. As the number of MD steps increased, $|r_{model} - r_{crystal}|$ still remained close to zero, but the ssDNA nucleotide index changed as the nucleotides continuously slide over the binding site until they eventually found the correct position,

in which nucleotide 22 interacts with site I, such as in the crystal structure.

Fig. 3B shows that correct binding of the ssDNA with site II occurred right after site I was occupied, indicating that sliding from site I to site II was relatively fast. The sliding of the ssDNA continued from site II toward site III with formation of the correct interface at time step $\sim 5.0 \times 10^6$ that proceeded immediately toward site IV. The sliding of the ssDNA, as suggested by this trajectory, did not take place at the same pace between the sites, presumably because of the different conformations that the ssDNA acquired at each site and the heterogeneity of the molecular properties of each site (e.g., their aromatic and electrostatic composition).

The structures shown in Fig. 3, *Right* are snapshots of four different conformations selected from the trajectory and correspond to the time frames marked by arrows 1–4 in Fig. 3, *Left*. We point out that, although site I constituted the initial association event, it is possible to find cases in which other sites engaged first. The probability of the initial binding occurring first with one of the sites depends on their molecular properties, and therefore, they are not necessarily equivalent.

A similar mechanism for back and forth sliding of ssDNA was also observed for Pot1pC (Fig. 4) by following the value of $|r_{model} - r_{crystal}|$ and the nucleotide identity to a selected site in two different trajectories. Although the ssDNA interacts with each of these sites at the beginning of the simulation (reflected by the immediate decrease of the value of $|r_{model} - r_{crystal}|$), the identity of the nucleotide interacting with the sites changes throughout the simulation. The sliding of the ssDNA is also shown pictorially in Fig. 4 at three different instances in time for each trajectory. It is clear from the electrostatic map of Pot1pC (Fig. 1) that there are two potential pathways through which ssDNA can enter into the binding pocket of the protein, unlike the situation for the telomere protein Cdc13, where the negatively charged electrostatic region on one side blocks the path of the ssDNA (Fig. 1). Fig. 4A shows the sliding behavior of ssDNA from one side, and
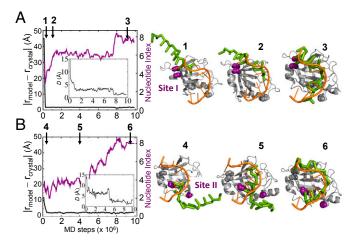
it also remains true if the ssDNA enters from the opposite side (Fig. 4B). Fig. 4, *Insets* show the time evolution of the similarity measure D, which indicates that the correct and more complete interface between the 9-nt ssDNA and Pot1pc is achieved toward the end of the simulation.

Conclusions

The complexes formed between proteins and ssDNA are not only fundamentally different from those formed between proteins and dsDNA but also, more difficult to predict. Two major factors complicate the prediction of SSB–ssDNA complexes compared with protein–dsDNA complexes: the much greater flexibility and consequent lack of definitive structure of ssDNA compared with dsDNA and the larger heterogeneity of the SSB–ssDNA interface. These two complexities are related to each other. The disordered nature of ssDNA enables it to interact with proteins mostly through the phosphate groups, which may attract charged side chains, or the bases, which may interact with aromatic side chains, with a consequent increase in interface heterogeneity. As a result of these complicating factors, SSB–ssDNA complexes have received little attention from theoreticians.

To address this gap, this paper aimed to develop a transferable model to predict the crystal complex of ssDNA with SSBs coming from different domains of life. The important molecular components of the coarse-grained model for predicting the SSB–ssDNA complexes were electrostatic and aromatic interactions between the charged and aromatic amino acids and between the DNA phosphates and bases, respectively. In addition to incorporating the inherent flexibility of ssDNA, the model incorporated the flexibility of the proteins to improve the binding predictions for some complexes. More elaborate models that include, for example, ion condensation may better describe the conformational dynamics of ssDNA and the interface that they may form with SSBs. Despite the massive computer resources that are required for these such models, they may allow researchers to explore distinct association modes at different salt conditions, which was observed experimentally for RPA (27) and *E. coli* SSB (25, 26). Our coarse-grained model successfully predicted the binding modes of six complexes of different sizes (132–694 aa and 6–65 nt). Thus, the model produces useful results using more widely available computing resources. The model can be further extended to study other properties of these complexes (such as the sliding of SSB along ssDNA) and readily combined with experimental research in future studies.

Methods

We used a coarse-grained model to explore the binding of ssDNA to SSBs at the molecular level. The model represents each protein residue by two beads placed at the Cα- and Cβ-positions. Beads representing charged amino acids (K, R, H, D, and E) have a charge at the Cβ-position. The protein is simulated by a native topology-based model and uses the Lennard–Jones (L-J) potential to represent native contact interactions. Overall, we followed an approach to protein modeling similar to that described in refs. 28 and 29. The internal energy of the protein is designated by E_{prot}. In addition to the inherent flexibility of the proteins in the coarse-grained model, which is dictated by the density of the native contacts, we incorporated enhanced flexibility for regions characterized by high B factors or residues with no electron density.

The ssDNA was modeled by three beads per nucleotide (representing the phosphate, sugar, and base) that were positioned at the geometric center of each represented group. The phosphate bead in the model bears a negative charge. The model potential for ssDNA used in our study, which follows other models (30–32), is given by $E_{ssDNA} = E_{ssDNA}^{Bond} + E_{ssDNA}^{Angle} + E_{ssDNA}^{Dihedral} + E_{ssDNA}^{Base\ pairing} + E_{ssDNA}^{Stacking} + E_{ssDNA}^{Repulsion}$. The first three terms dictate the flexibility of the ssDNA backbone, whereas the last three terms govern the global structure of the ssDNA. Two major interactions involved in the stability of ssDNA are base pairing and stacking interactions. Often, a homopolymeric ssDNA (poly T or poly C) is used in experimental studies of SSB–ssDNA complexes, which prohibits the formation of base pairing. However, the formation of stacks still remains possible depending on the type of nucleotides comprising the ssDNA. Therefore, to consider the effect of base stacking in ssDNA, we added

Fig. 4. Mechanism of ssDNA binding to Pot1pc. The association is illustrated by two different trajectories. The trajectories illustrate binding first to site (A) I or (B) II. Sites I and II are defined as the centers of mass of residues W72 and H109 and W27 and R59, respectively, and they are shown at three different time points as indicated by the arrows. The ssDNA can enter into the binding pocket of the protein from either of two sides. The initial binding to either site is followed by substantial sliding relative to the corresponding site. A pictorial representation of ssDNA sliding during its binding to the Pot1pc protein is shown for each trajectory by three snapshots (marked by 1–3 and 4–6 for the trajectories shown in A and B, respectively). *Insets* show the time evolution of the total similarity measure D for each binding trajectory, illustrating that complete binding is achieved only toward the end of the simulation (D $\sim$ 2 Å).

a short-range attraction between consecutive bases of ssDNA in the form of the L-J potential, with an energetic contribution of ε_{B-B} (*SI Methods* and Fig. S6).

The interaction potential between a protein amino acid (AA) and ssDNA nucleotides arises from three contributions: (*i*) the electrostatic interaction between the Cβ-bead representing a charged amino acid (K, R, H, D, and E) and the negatively charged phosphate of ssDNA, (*ii*) the aromatic stacking interaction between the Cβ-bead representing an aromatic amino acid (W, F, Y, and H) and an ssDNA base, and (*iii*) the repulsive interactions between other beads of the protein and ssDNA. Thus, $E_{prot-ssDNA} = E_{prot-ssDNA}^{Elec} + E_{prot-ssDNA}^{Aromatic} + E_{prot-ssDNA}^{Repulsion}$. The electrostatic interactions acting between all of the charged beads in the system are modeled by the Debye–Hückel potential (29). The aromatic interaction, similar to base stacking, is modeled by the L-J potential, with a base–aromatic amino acid interaction strength of ε_{B-AA}. The parameter $\varepsilon_{B-AA} = 0$ for all of the nonaromatic amino acids, and $\varepsilon_{B-AA} > 0$ only for aromatic amino acids (W, F, Y, and H). Trp has a large surface area compared with other aromatic residues (F, Y, and H), and hence, ε_{T-W} may make a larger contribution than ε_{T-F}, ε_{T-Y}, and ε_{T-H} when the aromatic amino acids form stacks with the bases of ssDNA. Given the coarse-grained nature of the model, we calibrated the relative strengths of the electrostatic and aromatic contributions and selected the following values: $\varepsilon_{T-F} = \varepsilon_{T-Y} = \varepsilon_{T-H} = 2$ and $\varepsilon_{T-W} = 3$ (Fig. S7).

Using the total potential energy ($E_{prot} + E_{ssDNA} + E_{prot-ssDNA}$) of the SSB–ssDNA complex, the dynamics of the protein and ssDNA were simulated using Langevin dynamics. The ssDNA–SSB complex was studied at a salt concentration of 0.01 M using the dielectric constant of water and a temperature at which the bound state of the complex is more populated. The model was initially tested for its ability to maintain the native structure of all of the complexes when we used it as the initial input for simulation. We started the predictive simulation by placing a free ssDNA (of the same length as the ssDNA in the corresponding crystal structure) at different positions. A total of 600 simulations for each system was collected, analyzing a total of 6×10^4 conformations.

We then calculated a structural similarity parameter (*D*) to quantify the degree of deviation of the simulation from the known crystal structure. To obtain the similarity parameter, we first identified the interfacial residues as the Cβ-bead of residues (either positively charged or aromatic amino acids) with distance that was less than 9 Å from any ssDNA phosphate or any base in the crystal structure. The simulation to crystal structure similarity parameter is calculated by $D = N_{prot}^{-1} N_{ssDNA}^{-1} \sum_i^{Nprot} (\sum_j^{NssDNA} r_{ij} - \sum_j^{NssDNA} r_{ij}^0)$. Here, *i* and *j* are the *i*th bead of the selected Cβ bead of the protein and the *j*th bead of the ssDNA, respectively, and thus, r_{ij} and r_{ij}^0 are the pairwise distances between those beads in the simulated and the crystal structure, respectively. The pairwise distances were calculated either between each selected aromatic amino acids and all of the B beads of the ssDNA or between each of the selected charged amino acids and all of the phosphate beads of the ssDNA. N_{prot} is the total number of selected interfacial amino acid residues (positively charged or aromatic), and N_{ssDNA} is the number of nucleotides in the length of ssDNA examined. Thus, the term *D* quantifies the overall conformational similarity between the predicted binding interface in the ssDNA–protein complex and the crystal structure, with a *D* value = 0 indicating 100% conformational similarity. To obtain a finer structural quantification of the interface, we divided the selected interfacial residues into two groups that cover two different regions of the interface. We then calculated two order parameters, D_1 and D_2, each characterizing the accuracy of the prediction for the corresponding region of the interface (Fig. S1).

It is important to note that we used information from the crystal structure solely to model the protein, whereas transferable potentials were used to model the structure and energetics of both the ssDNA and the interface protein–ssDNA. Thus, our model does not include any bias toward the specific binding mode. In our study, we investigated six different SSB–ssDNA complexes (Table S1). A detailed description is in *SI Methods*.

1. Lohman TM, Ferrari ME (1994) Escherichia coli single-stranded DNA-binding protein: Multiple DNA-binding modes and cooperativities. *Annu Rev Biochem* 63:527–570.
2. Wold MS (1997) Replication protein A: A heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu Rev Biochem* 66:61–92.
3. Pestryakov PE, Lavrik OI (2008) Mechanisms of single-stranded DNA-binding protein functioning in cellular DNA metabolism. *Biochemistry (Mosc)* 73(13):1388–1404.
4. Theobald DL, Mitton-Fry RM, Wuttke DS (2003) Nucleic acid recognition by OB-fold proteins. *Annu Rev Biophys Biomol Struct* 32:115–133.
5. Flynn JM, Levchenko I, Sauer RT, Baker TA (2004) Modulating substrate choice: The SspB adaptor delivers a regulator of the extracytoplasmic-stress response to the AAA+ protease ClpXP for degradation. *Genes Dev* 18(18):2292–2301.
6. Raghunathan S, Kozlov AG, Lohman TM, Waksman G (2000) Structure of the DNA binding domain of E. coli SSB bound to ssDNA. *Nat Struct Biol* 7(8):648–652.
7. Fan J, Pavletich NP (2012) Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. *Genes Dev* 26(20):2337–2347.
8. Dickey TH, Altschuler SE, Wuttke DS (2013) Single-stranded DNA-binding proteins: Multiple domains for multiple functions. *Structure* 21(7):1074–1084.
9. Lei M, Podell ER, Cech TR (2004) Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nat Struct Mol Biol* 11(12):1223–1229.
10. Roy R, Kozlov AG, Lohman TM, Ha T (2009) SSB protein diffusion on single-stranded DNA stimulates RecA filament formation. *Nature* 461(7267):1092–1097.
11. Kozlov AG, Lohman TM (1998) Calorimetric studies of E. coli SSB protein-single-stranded DNA interactions. Effects of monovalent salts on binding enthalpy. *J Mol Biol* 278(5):999–1014.
12. Kozlov AG, Lohman TM (2006) Effects of monovalent anions on a temperature-dependent heat capacity change for Escherichia coli SSB tetramer binding to single-stranded DNA. *Biochemistry* 45(16):5190–5205.
13. Marcovitz A, Levy Y (2011) Frustration in protein-DNA binding influences conformational switching and target search kinetics. *Proc Natl Acad Sci USA* 108(44):17957–17962.
14. Ha T, Kozlov AG, Lohman TM (2012) Single-molecule views of protein movement on single-stranded DNA. *Annu Rev Biophys* 41:295–319.
15. Zhou R, et al. (2011) SSB functions as a sliding platform that migrates on DNA via reptation. *Cell* 146(2):222–232.
16. Zhang J, Zhou R, Inoue J, Mikawa T, Ha T (2014) Single molecule analysis of Thermus thermophilus SSB protein dynamics on single-stranded DNA. *Nucleic Acids Res* 42(6):3821–3832.

17. Lee KS, et al. (2014) Ultrafast redistribution of E. coli SSB along long single-stranded DNA via intersegment transfer. *J Mol Biol* 426(13):2413–2421.
18. Nguyen B, et al. (2014) Diffusion of human replication protein A along single-stranded DNA. *J Mol Biol* 426(19):3246–3261.
19. Hwang H, Buncher N, Opresko PL, Myong S (2012) POT1-TPP1 regulates telomeric overhang structural dynamics. *Structure* 20(11):1872–1880.
20. Gibb B, et al. (2014) Concentration-dependent exchange of replication protein A on single-stranded DNA revealed by single-molecule imaging. *PLoS ONE* 9(2):e87922.
21. Mitton-Fry RM, Anderson EM, Theobald DL, Glustrom LW, Wuttke DS (2004) Structural basis for telomeric single-stranded DNA recognition by yeast Cdc13. *J Mol Biol* 338(2):241–255.
22. Tinland B, Pluen A, Sturm J, Weill G (1997) Persistence length of single-stranded DNA. *Macromolecules* 30(19):5763–5765.
23. Murphy MC, Rasnik I, Cheng W, Lohman TM, Ha T (2004) Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy. *Biophys J* 86(4):2530–2537.
24. Werten S, Moras D (2006) A global transcription cofactor bound to juxtaposed strands of unwound DNA. *Nat Struct Mol Biol* 13(2):181–182.
25. Lohman TM, Overman LB (1985) Two binding modes in Escherichia coli single strand binding protein-single stranded DNA complexes. Modulation by NaCl concentration. *J Biol Chem* 260(6):3594–3603.
26. Bujalowski W, Lohman TM (1986) Escherichia coli single-strand binding protein forms multiple, distinct complexes with single-stranded DNA. *Biochemistry* 25(24):7799–7802.
27. Kumaran S, Kozlov AG, Lohman TM (2006) Saccharomyces cerevisiae replication protein A binds to single-stranded DNA in multiple salt-dependent modes. *Biochemistry* 45(39):11958–11973.
28. Levy Y, Wolynes PG, Onuchic JN (2004) Protein topology determines binding mechanism. *Proc Natl Acad Sci USA* 101(2):511–516.
29. Givaty O, Levy Y (2009) Protein sliding along DNA: Dynamics and structural characterization. *J Mol Biol* 385(4):1087–1097.
30. Ouldridge TE, Louis AA, Doye JP (2011) Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *J Chem Phys* 134(8):085101.
31. Morriss-Andrews A, Rottler J, Plotkin SS (2010) A systematically coarse-grained model for DNA and its predictions for persistence length, stacking, twist, and chirality. *J Chem Phys* 132(3):035105.
32. Freeman GS, Hinckley DM, de Pablo JJ (2011) A coarse-grain three-site-per-nucleotide model for DNA with explicit ions. *J Chem Phys* 135(16):165104.