Simulation of protein association: Kinetic pathways towards crystal contacts

Aaron Taudt,[1] Axel Arnold,[2] and Jürgen Pleiss[1,*]

[1]*Institute of Technical Biochemistry, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany*

[2]*Institute for Computational Physics, University of Stuttgart, Allmandring 3, 70569 Stuttgart, Germany*

(Received 16 September 2014; revised manuscript received 28 November 2014; published 30 March 2015)

We conducted molecular dynamics simulations combined with distance-based umbrella sampling and forward flux sampling to investigate the early stages of protein crystallization. Formation of contacts with long-range interactions and/or an exposed position on the protein surface was kinetically preferred over more stable hydrophobic contacts with a shorter attractive range, while the thermodynamic stability of the protein crystal was provided by hydrophobic interactions. Contacts with a large interaction area showed complex dissociation pathways that were not detected by distance-based umbrella sampling. Instead, forward flux sampling simulations of contact dissociation identified long-range attractive interactions.

DOI: 10.1103/PhysRevE.91.033311 PACS number(s): 02.70.−c, 87.14.E−

I. INTRODUCTION

Experimentally determined protein structures are the indispensable starting point of efficient strategies for protein engineering or drug design. The most important technique for protein structure determination is still diffraction crystallography, with the major bottleneck being the growth of high-quality protein crystals of sufficient size [1,2]. Methods such as cryocrystallography and synchrotron radiation decreased the necessary size of protein crystals, and strategies for enhancing crystallization propensity such as surface entropy reduction [3,4], removal of flexible termini [5,6], and the use of fusion proteins [7,8] have been successfully applied to increase the success rate of crystallization. But crystallizing a specific protein is still a trial and error approach [2]. Despite protein crystallization being used for x-ray structure determination for more than 50 years, astonishingly little is still known about the underlying mechanisms of nucleus formation and protein crystal growth. In contrast to specific interprotein contacts between naturally occurring, functional protein oligomers, crystal contacts are considered to be nonspecific, nonfunctional, unphysiological, and generally smaller [9–12]. Until recently, crystal contacts were regarded as purely stochastic and basically indistinguishable from the rest of the protein surface, because random docking of two proteins resulted in the same mean value of interface area as the contacts between monomeric proteins in a crystal [9]. In addition, the fraction of the surface which participates in crystal contacts is variable and independent of the number of contact patches [11]. By comparing crystal contacts of pancreatic ribonuclease in six different space groups, it was shown that the entire protein surface participated in crystal contacts [13]. However, there is emerging evidence that crystallization is not purely stochastic, but that surface characteristics affect the shape and strength of protein interactions. Statistical analysis of the amino acid composition of crystal contacts showed that residues with high conformational entropy such as lysine and glutamic acid are underrepresented in crystal contacts, whereas small hydrophobic amino acids like alanine enhance the crystallization propensity [3,10,14,15]. This observation was the basis of the successful surface entropy reduction approach to engineering protein variants with a high propensity of crystallization. Recent molecular dynamics simulations also showed that the potential of mean force for a particular crystal contact strongly depends on the physical chemistry of the participating interfaces [16,17]. In this work we employ molecular dynamics simulations with umbrella sampling and forward flux sampling to investigate the mechanisms of crystal contact formation and the early stage of nucleus formation. Distance-based umbrella sampling [18] is a standard technique [16,17,19] to calculate free energy differences upon binding, assuming a reaction pathway. In contrast, forward flux sampling [20] is a rare event sampling technique which yields not only the free energy of binding but also an unbiased reaction pathway. Our model system is the nonglycosylated N74S mutant of *Candida antarctica* lipase B (CALB) [21]. This enzyme catalyzes a broad range of hydrolysis and alcoholysis reactions and is widely used in industrial applications because of its high enantioselectivity, wide substrate spectrum, and stability at high temperature and in organic solvents. It is neutral from pH 4–8 [22] and consists of 317 residues.

II. METHODS

A. Molecular dynamics simulations

Molecular dynamics simulations were performed with the GROMACS package [23] using the OPLS all atom force field [24] and the TIP4P water model [25] with Ewald summation. The model protein was the ligand-free, nonglycosylated N74S mutant of CALB [21]. The structure was derived from Protein Data Bank (PDB) entry 1TCA [26] and has a resolution of 1.55 Å. The only histidine in the structure (His 224) was protonated to yield a total net charge of zero. Prior to each production run, the system was energy minimized (steepest descent, 25 000 steps) and the water around the protein was equilibrated (200 ps, position restraints on all protein heavy atoms). Simulations were carried out with periodic boundary conditions in the NPT ensemble with Parrinello-Rahman barostat at 1 bar (isotropic scaling, $\tau_p = 2.0$, compressibility $= 4.5\times10^{-5}$ bar^{-1}) and V-rescale thermostat at 310 K ($\tau_t = 0.1$) with time step 2 fs.

*Corresponding author: juergen.pleiss@itb.uni-stuttgart.de

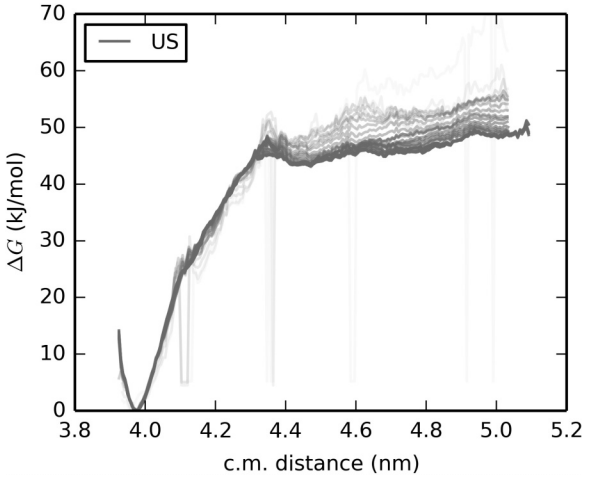

FIG. 1. Convergence of the umbrella potential of contact 1. A line is drawn every 200 ps in decreasing transparency from 200 ps until 5000 ps. The potential converges very fast for the distance when the two proteins are in contact (<4.4 nm). After dissociation (>4.4 nm), the convergence is slower due to the gained possibilities of reorientation.

B. Umbrella sampling

Umbrella sampling (US) was performed with the weighted histogram method implemented in GROMACS [27]. Starting configurations for umbrella sampling were generated from a protein dimer by pulling one protein along the "reaction coordinate" while keeping the other protein fixed. The reaction coordinate was assumed to be the principal axis of the inertia tensor which is orthogonal to the interface. Apart from constraining the translational movement in the direction of the reaction coordinate by the umbrella potential, no further restrictions were imposed on the proteins. In particular, reciprocal orientation was not controlled explicitly. However, due to the dense window spacing (every 0.5 Å) and short sampling time (5 ns), we did not observe any appreciable rotation between the proteins. The sampling time of 5 ns was sufficient to achieve convergence of the potential (Figs. 1 and 2).

C. Forward flux sampling

Forward flux sampling (FFS) is a technique for sampling rare events [28]. A rare event is a transition between a (meta)stable state B and a state U, where B and U are separated by a potential energy barrier. In our case, state B is defined as the two proteins being in contact (bound), while state U is defined as the two proteins separated (unbound). An order parameter is selected which changes continuously while the system passes from B to U. This order parameter does not have to be the actual reaction coordinate, although this is most efficient. However, in most cases and in the problem studied here, the reaction coordinate is not known *a priori*. Forward flux sampling nevertheless produces correct transition rates and paths [28]. Interfaces are then introduced along the order parameter and the system is driven from B to U in a ratchetlike manner. The output of a forward flux sampling simulation is a rate constant, a potential energy profile along the order

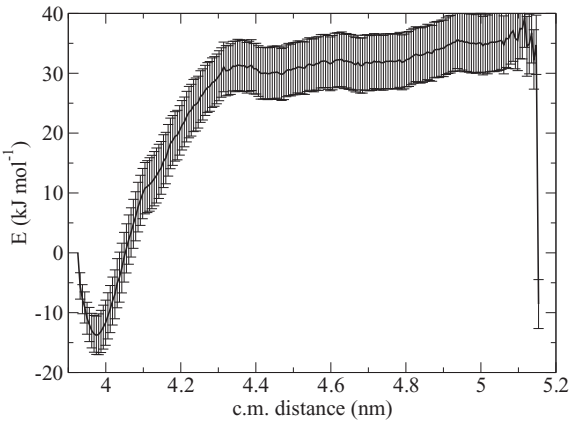

FIG. 2. Free energy profile for the dissociation of contact 1 with standard deviations calculated by the Bayesian bootstrap implemented in g_wham.

parameter [20] and the pathway of the transition. In this work, we were primarily interested in the reaction pathways and the free energy profiles.

We used the GROMACS interface of FRESHS (Flexible Rare Event Sampling Harnessing System) to conduct the forward flux sampling [29]. For the association simulation (Sec. III B), the order parameter was chosen to be the minimal surface (MS) distance between two proteins and the order parameter evaluation interval was set to 20 ps. For the dissociation simulation (Sec. III C), the center-of-mass (c.m.) distance was selected as the order parameter. The reason for the different selections is illustrated in Fig. 3: The minimal surface distance is not suited for the dissociation simulation, because it is zero until the last atoms have detached and hence forward flux sampling is useless with this order parameter (for forward flux sampling to be efficient, the order parameter has to change continuously with the reaction coordinate). Vice versa, the c.m. distance is not suited for the association simulation, since the proteins are not spherical (rather ellipsoid) and there is no single value for contact formation while the MS distance is zero.

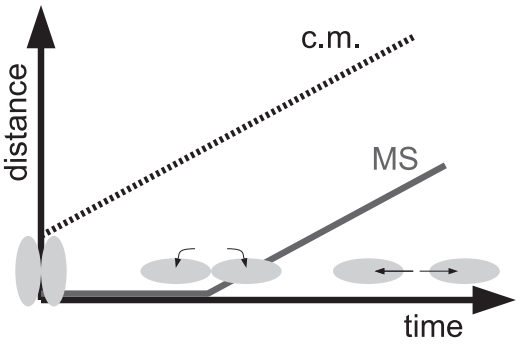

FIG. 3. Illustration of the differences of the order parameters for a fictitious dissociation of two proteins (ellipses). Dashed line, center-of-mass (c.m.) distance; solid line, minimal surface (MS) distance.

TABLE I. Forward flux sampling settings and results for the association process. Order parameter is the minimal surface distance.

λ_i	Interface position (nm)	Points per interface	Probabilities $P_{i-1,i}$	Rates $k_{A,i}$ (ps^{-1})
0	1.4	100		9.830×10^{-4}
1	1.2	100	0.299 401	2.943×10^{-4}
2	1.0	100	0.460 829	1.356×10^{-4}
3	0.8	100	0.534 759	7.253×10^{-5}
4	0.6	100	0.709 220	5.144×10^{-5}
5	0.4	100	0.719 424	3.701×10^{-5}
6	0.2	100	0.884 956	3.275×10^{-5}

1. Association simulation

Two proteins of CALB were placed at a (surface) distance of 2 nm from each other and rotated randomly. A dodecahedral box with 1 nm distance to the box boundaries and periodic boundary conditions was chosen, to ensure that the proteins eventually meet, independent of the direction in which they are moving. The forward flux sampling interface positions and rates along the order parameter are given in Table I.

2. Dissociation simulation

A pair of CALB molecules constituting a given crystal contact was aligned by its principal axes and solvated in a rectangular-shaped box of water. Box dimensions were 1.5 x-dim, 1.8 y-dim, and 1.8 z-dim, with x-, y-, and z-dim being the diameters of the pair and the x axis being the principal axis. A rectangular-shaped box was chosen to reduce the amount of water molecules needed to fill the box. However, without any constraints, this setup limits the simulated time to the time that a pair needs to rotate 90° and meet its periodic image. Therefore, all trajectories were checked for this event afterwards. Forward flux sampling interfaces were placed every 0.2 or 0.3 Å with 20 or 10 points per interface. The settings and rates are given in Table II.

Due to time and resource limitations, the forward flux sampling dissociation of contact 1 could not be completed until full detachment of the proteins. The free energy profile with c.m. distance >4.8 nm (Fig. 10) was therefore estimated with umbrella sampling. A configuration from the last forward flux sampling interface served as a starting configuration for this umbrella sampling run.

TABLE II. Forward flux sampling settings and results for the dissociation of contact 1. Order parameter is the center-of-mass distance.

λ_i	Interface position (nm)	Points per interface	Probabilities $P_{i-1,i}$	Rates $k_{A,i}$ (ps^{-1})
0	4.00	24		0.002194
1	4.02	20	0.136 054	0.000 298
2	4.04	20	0.400 000	0.000 119
⋮				
24	4.48	20	0.952 381	5.38×10^{-8}
25	4.50	20	0.909 091	4.89×10^{-8}
26	4.53	10	1.000 000	4.89×10^{-8}
27	4.56	10	0.909 091	4.45×10^{-8}
⋮				
35	4.80	10	1.000 000	3.68×10^{-8}

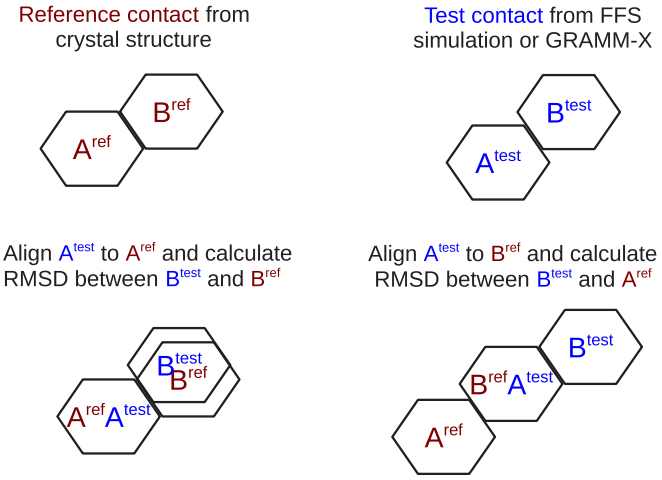

FIG. 4. (Color online) Procedure for determination of contact similarity. Protein A of the test contact is either aligned to protein A (bottom left) or protein B (bottom right) of the reference contact and the backbone RMSD is computed.

D. Determination of contact similarity

Each contact comprises two identical CALB molecules, A and B, in a specific orientation. We chose the root mean square deviation (RMSD) to measure the similarity of a given contact (test contact) to a crystal contact (reference contact). The RMSD was computed as follows (Fig. 4 illustrates the procedure):

(1) Align protein A of the test contact to protein A (B) of the reference contact.

(2) Calculate the backbone RMSD between protein B of the test contact and protein B (A) of the reference contact. The smaller this value, the more similar the contacts are.

We conducted a reference simulation to determine a threshold up to which two contacts can be regarded as similar (Fig. 5). The computed RMSD value fluctuated between 0 and 1.5 nm for an intact contact and, thus, two contacts can be regarded as identical if the RMSD $\leqslant$ 1.5 nm. Furthermore, as visual inspection of the trajectory showed, up to an RMSD of 2.5 nm, two contacts can still be regarded as similar.

III. RESULTS

A. Crystal contacts of CALB

Candida antarctica lipase B has been crystallized in space group $P2_12_12_1$ where the unit cell lengths are $a = 6.21$, $b = 4.67$, $c = 9.21$ nm and the angles are $\alpha = \beta = \gamma = 90°$ (PDB entry 1TCA). Each unit cell contains a single protein molecule consisting of 317 residues, and each of these proteins forms 6 distinct crystal contacts to its 12 neighboring proteins, which cover 25.8% of the protein surface (total 12 034 $Å^2$). The number of crystal contacts and the residues that are

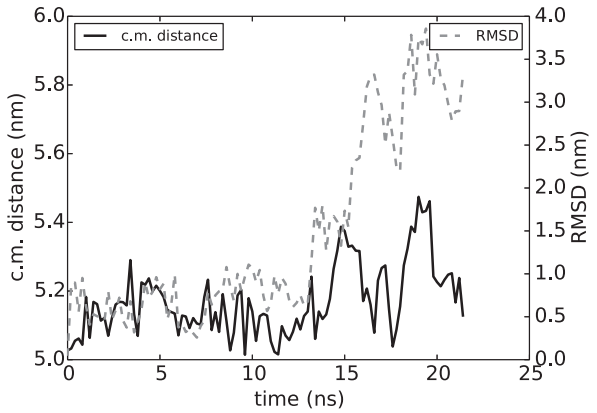

FIG. 5. Reference simulation for the RMSD that was used to determine contact similarity (contact 4). For an intact crystal contact ($t \leqslant 15$ ns, c.m. distance $\leqslant$5.4 nm), the RMSD fluctuated between 0 and 1.5 nm. The reorientation of the two proteins ($t > 15$ ns) was accompanied by a larger fluctuation in the c.m. distance and an increase in the RMSD.

involved in the contacts result from the overall decrease of free energy upon crystallization and the specific pathway of the crystallization process. The crystal contacts were classified by PISA (Protein Interfaces, Surfaces and Assemblies [30]). The factors contributing to the interaction free energy, such as contact area, solvation energy, number of H and disulfide bonds, and number of salt bridges were analyzed (Table III). PISA uses an empirically parametrized method based on chemical thermodynamics to obtain an estimate of the binding energies.

We found six crystal contacts with areas between 640 and 30 Å^2. Contact 1 (Fig. 6) had the largest contact area (640 Å^2) and was predicted to be the most stable one ($\Delta G_{\text{total}} = -68$ kJ/mol) due to its large hydrophobic surface. In addition to a large hydrophobic stretch, contact 1 also contained a small charged patch, with the two positively charged K308 and R309 in the immediate vicinity of the negatively charged E188 (see Table IV). This interaction is probably forming a salt bridge even though no salt bridge was predicted by PISA. Contacts 2, 3, and 4 had contact areas of 290, 250 and 240 Å^2, respectively. Contact 4 was predicted to be the second most stable contact ($\Delta G_{\text{total}} = -14$ kJ/mol), but in contrast to contact 1, the interactions in contact 4 were dominated by hydrogen bonds. Contacts 5 and 6 had the smallest interaction

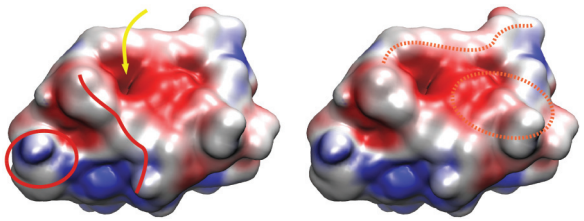

FIG. 6. (Color) Visualization of contact 1 colored by electrostatic potential [32] (red, negative; blue, positive). Lines indicate the amino acids which constitute the hydrophobic stretch and circles indicate the amino acids which form the charged patch. Solid line, partner A; dashed line, partner B. The yellow arrow indicates the entrance to the catalytic site.

area and the lowest free energy of binding. The cohesion of contact 6 was mediated by only a single hydrogen bond and probably also by electrostatic attraction. The classification by PISA was confirmed by protein-protein docking using GRAMM-X [31]. Among the predicted GRAMM-X contacts, contact 1 was identified as one of the thermodynamically most preferred contacts. However, GRAMM-X predicted 299 additional stable contacts between two CALB molecules, none of which was identified among the 6 crystal contacts (data not shown).

B. Association

To study protein association, forward flux sampling trajectories were generated and analyzed starting from two CALB proteins placed at a surface distance of 2 nm with random orientations. The forward flux sampling trajectories were started from 100 randomly oriented pairs of CALB proteins. At each interface, 100 configurations were selected, and the sampling was terminated when the minimum distance between the two proteins was less than 0.2 nm. Thus, in total 100 trajectories were generated. This approach can only find favorable pair contacts and does not consider the geometric constraints that are present in a crystal lattice. To account for those geometric constraints, we compared the 100 resulting pair contacts to the crystal contacts and found 13 trajectories that resulted in crystal contacts (Table V). To assess whether these 13 contacts resulted from specific interactions or just from the random starting configurations, we compared the contact distribution of forward flux sampling with the contact distribution of random contact formation (Fig. 7). The p value

TABLE III. Properties of the six different crystal contacts of the N74S mutant of CALB as obtained from PISA. None of the contacts involves any disulfide bonds, covalent bonds, or salt bridges. $\Delta^i G$ indicates the solvation free energy gain upon interface formation. ΔG_{polar} indicates the free energy contribution of the H bonds.

Contact[a]	Buried surface (%, Å^2)	$\Delta^i G$ (kJ/mol)	ΔG_{polar} (kJ/mol)	ΔG_{total} (kJ/mol)	$\Delta^i G$/surface [J/(mol*Å^2), cal/(mol*Å^2)]	ΔG_{polar}/surface [J/(mol*Å^2), cal/(mol*Å^2)]
1	5.3, 638.2	−66.1	–1.9	−68.0	–103.6, −24.8	−3.0, −0.7
2	2.6, 298.9	−8.4	–1.9	−10.3	–28.1, −6.7	−6.4, −1.5
3	2.1, 251.0	−4.0	–3.7	−7.7	–15.9, −3.8	−14.7, −3.5
4	1.9, 236.6	−4.7	–9.3	−14.0	–19.9, −4.7	−39.3, −9.4
5	0.7, 79.7	−5.8	0	−5.8	–72.8, −17.4	0.0, 0.0
6	0.3, 32.0	1.9	–1.9	0.0	59.4, 14.2	−59.4, −14.2

[a]Contacts are ranked by contact area.

TABLE IV. Residues participating in the six different crystal contacts of the N74S mutant of CALB.

Crystal contact	Partner A	Partner B
1	Hydrophobic stretch V139 G142 P143 A146 L147 T159 Q191 P192 V194 S195 S197 P198 L199 D200 F205 Charged patch G307 K308 R309 I314 T316	Hydrophobic stretch L144 L147 V149 V272 A275 A276 A279 A282 A283 I285 V286 A287 K290 Charged patch T186 E188 L219 V221 I222 P260
2	L1 S3 G4 S5 Y91 G95 N96 N97 K98 S120 R122 S123 K124 I176	N196 S197 P198 Y203 A206 G207 K208 Q213 A214 V215 S250 A251 Y253 I255 T256
3	S31 K32 T57 Q58 L59 G60 R242 T244	A146 A14 A287 G288 P289 K290 Q291 N292 C293 C311 S312
4	Q23 G24 A25 S26 S28 S29 V30 S31 T62 A92 G93 S94 G95	N259 L261 N264 D265 L266 T267 P268 K271
5	S5 P299 Y300 P303 P317	K13 S14 V15 D17
6	T174 F205 N206	E269

of 0.001 (χ^2 test) is highly significant and thus the contacts formed upon the association simulation are not random. Furthermore, Fig. 7(c) also shows that this result is not obtained due to a biased distribution of starting configurations towards contacts 4 and 6.

The by far most probable crystal contacts generated by forward flux sampling simulations were crystal contacts 4 and 6. Both contacts are characterized by highly polar interactions formed by patches of opposite electrostatic charges and hydrogen bonds (Fig. 8 and Table III). In contrast, contacts 1 and 5 are predominantly hydrophobic and were not observed in the forward flux sampling simulations. However, contact 1 not only consists of a large hydrophobic stretch, but also has a charged patch and formation of this charged patch was observed (Fig. 9). Thus, while contact 1 is thermodynamically preferred due to its large hydrophobic area, its formation is less probable than that of the second most stable contact 4 and even the least stable contact 6.

C. Dissociation

Interaction potentials for all six contacts were obtained by umbrella sampling. We also performed forward flux sampling simulations for contacts 1 and 2 and compared the interaction potentials with the umbrella sampling potentials. The center-of-mass distance was chosen as the order parameter. While the two simulation methods are expected to result in the same free energy difference between the protein complex and unbound proteins, the pathway from the bound to the unbound state can be expected to differ. The results are summarized in Table VI.

For contact 1 (Fig. 10), the minimum of the potential is found at 4.0 nm center-of-mass distance, which is exactly the difference of the two proteins in the crystal lattice. Beyond this minimum, the potential increases steeply for the umbrella sampling simulation. By a relative movement of the centers of mass by only 0.3 nm, the free energy increased by 45 kJ/mol. Beyond 0.3 nm, the potential changes only slightly. In contrast, forward flux sampling predicted a different pathway. The slope of the free energy increase is approximately 20% that of the umbrella sampling simulation. The forward flux sampling simulation was continued up to a center-of-mass distance of 4.8 nm where the free energy had increased by 25 kJ/mol as compared to the minimum and the two proteins are still in contact. Due to the high computational demand of the forward flux sampling approach, we used umbrella sampling to continue from this configuration

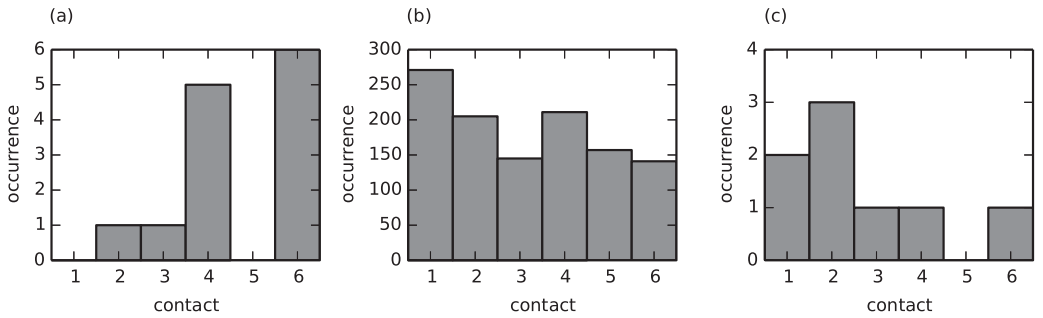

FIG. 7. Distribution of contacts obtained from (a) the forward flux sampling run, (b) random contact formation, and (c) starting configurations. The random contact formation distribution was calculated from 10 000 random contacts, 1130 of which yielded crystal contacts. The forward flux sampling distribution was calculated from 100 trajectories, 13 of which yielded crystal contacts (p value = 0.001). The similarity of the 100 starting configurations to crystal contacts was assessed by moving the two proteins towards each other until they made contact, and the closest crystal contact was determined. By this method, only 8 out of 100 starting configurations would yield crystal contacts.

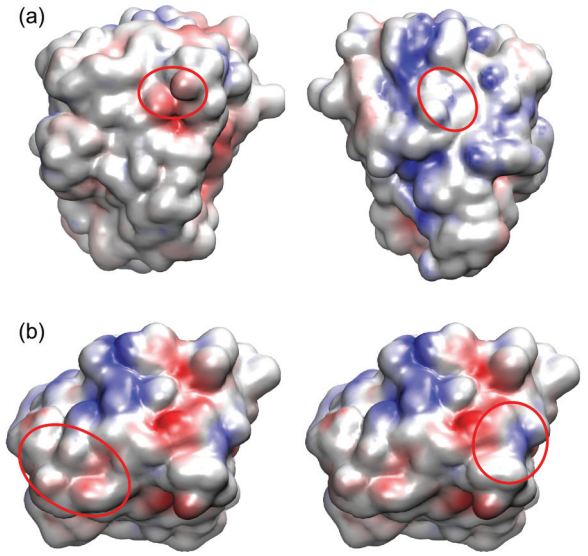

FIG. 8. (Color) Visualization of (a) contact 4 and (b) contact 6 colored by electrostatic potential [32] (red, negative; blue, positive). Circles indicate the patch on partner A (left) and partner B (right). The patches are oppositely charged.

TABLE V. Similarity of the 100 end states of the forward flux sampling simulation to the crystal contacts. An RMSD to a crystal contact of less than 2.5 nm is regarded as similar to this contact (see Sec. II D). From 100 contacts that formed during the forward flux sampling run, 13 yield crystal contacts.

FFS trajectory number[a]	RMSD (nm)	Crystal contact
7	1.19	4
95	1.55	6
60	1.59	4
89	1.72	4
38	1.90	4
47	2.14	6
11	2.20	6
80	2.21	2
48	2.27	6
85	2.29	6
41	2.42	6
64	2.45	3
24	2.47	4

[a]The trajectory number does not imply any ranking.

until full detachment. The total free energy difference between the bound and the unbound states by this mixed approach was found to be 50 kJ/mol, which is similar to the difference found by the umbrella sampling simulation. The molecular dynamics simulation estimate of 50 kJ/mol corresponds to 74% of the free energy estimation by PISA (68 kJ/mol). While the total free energy *difference* is the same, the free energy *profiles* of the two methods deviate. Umbrella sampling predicts a range of only 0.3 nm of attractive potential, while the attractive potential by forward flux sampling is estimated to be beyond 1 nm. The reason for this considerable difference in the potential profile is the choice of reaction path for umbrella sampling. In the umbrella sampling simulation, the two proteins were moved along the direct line between the centers of mass of the two proteins. However, although this intuitive choice seems to be adequate, forward flux sampling predicts a different reaction path. Without the constraints of umbrella sampling, forward flux sampling identified a two-step reaction path for the dissociation of contact 1 as most probable: In a first step, the salt bridge of the charged patch breaks at a center-of-mass distance of 4.1 nm while the two proteins are still fully connected at the hydrophobic stretch (after 0.13 nm of displacement, corresponding to 7 kJ/mol). After

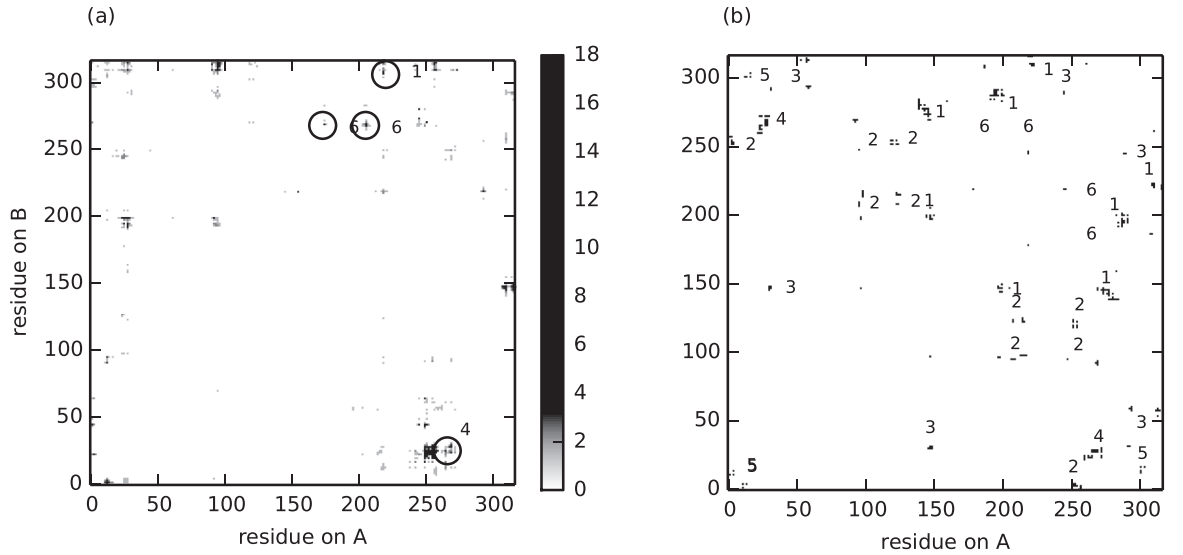

FIG. 9. Contact matrices for (a) the association simulation and (b) crystal lattice for comparison. Residues are defined as being in contact when they are closer than 7 Å. Left: Contacts at the end of the association simulations. The color indicates how often the contact was present in 100 simulations. Circles with numbers show the crystal contacts that formed entirely (4 and 6) or in part (1) during the simulation. Right: Contact matrix showing the contacts in the crystal lattice for comparison. Numbers indicate the respective crystal contact.

TABLE VI. Overview of the results for the dissociation simulations obtained by umbrella sampling. Standard deviations were obtained with the Bayesian bootstrap approach implemented in g_wham. Results for the FFS-US hybrid approach are marked by an asterisk.

Contact	ΔG (kJ/mol)	Minimum (nm)	Relative range (nm)	Absolute range (nm)	Average slope (ΔG/nm)
1	50 ± 8, 50*	4.0	0.3, 1.1*	4.3, 5.1*	166.7, 45.5*
2	22 ± 5	4.5	0.9	5.4	24.4
3	29 ± 6	4.5	0.8	5.3	36.3
4	13 ± 3	5.1	0.6	5.7	21.7
5	10 ± 4	4.8	0.4	5.2	25.0
6	9 ± 2	4.5	1.2	5.7	7.5

this short-range effect, contact 1 gradually disassembles as the center-of-mass distance further increases. This zipperlike opening of the hydrophobic stretch continues even beyond a center-of-mass distance of 4.8 nm (Fig. 11).

In contrast, for contact 2 umbrella sampling and forward flux sampling resulted in similar free energy profiles (Fig. 12). Here, the forward flux sampling simulation showed a similar path of detachment as the umbrella sampling simulation (along the connecting axis). The free energy difference is 22 kJ/mol and the attractive range 0.9 nm. This estimate for the free energy of binding is almost twice as high as the estimate from PISA of 10.3 kJ/mol. Contact 3 showed a binding free energy of 29 kJ/mol and attractive range of 0.8 nm (Fig. 13). The binding free energy of contact 4 was found to be 13 kJ/mol by umbrella sampling, which is close to the PISA estimate of 14 kJ/mol. The attractive range was found to be 0.6 nm (Fig. 14). Contact 5 had a binding free energy of 10 kJ/mol and an attractive range of 0.4 nm (Fig. 15). For contact 6, umbrella sampling predicted a binding free energy of 9 kJ/mol and an attractive range of approximately 1.2 nm (Fig. 16).

Although the binding free energies differ between the molecular dynamics estimates and the PISA estimates, the general trend is the same: Contact 1, which has the largest interface area, is the most stable contact and binding free energies decrease with decreasing interface area.

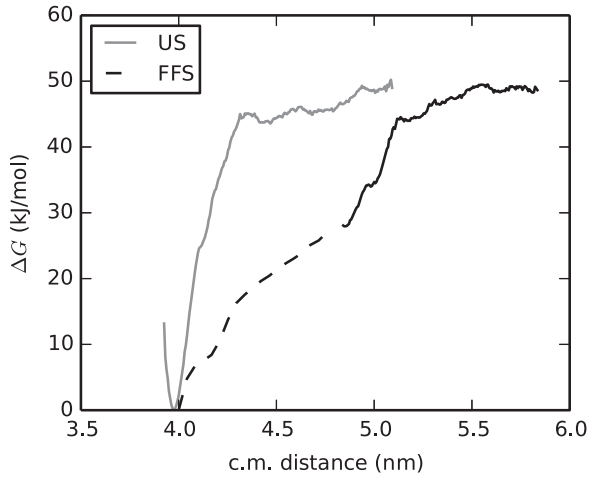

FIG. 10. Free energy profiles for the dissociation of contact 1 obtained with forward flux sampling (FFS) and umbrella sampling (US). The continuous line in the forward flux sampling profile is an estimate of the continued forward flux sampling profile obtained with umbrella sampling, due to the high computational demand of the forward flux sampling approach and time and resource limitations.

IV. DISCUSSION

A. Kinetics versus thermodynamics

Constrained sampling methods like umbrella sampling simulations are a widely used method to evaluate protein-protein interactions [16,17,19]. In most simulations, the distance between the centers of mass of two interacting protein molecules is selected as a reaction coordinate. Thus, the center-of-mass distance is gradually increased during the simulation, while the relative orientation of the two proteins is constrained. While the choice of the reaction coordinate has no impact on the free energy difference between the bound and the unbound states, it could influence the distance dependency of the interaction potential, as shown for two protein-protein contacts investigated here. For contact 2, with a contact area of 300 $Å^2$, the distance dependency of the interaction potential was found to be the same if calculated by umbrella sampling or forward flux sampling. The interaction potential is long-range and shallow and rises to 22 kJ/mol in 0.8 nm. However, for contact 1, with a contact area of 640 $Å^2$, the distance dependency of the interaction potential differed between the two methods. As calculated by umbrella sampling, the interaction potential is short-range and deep and rises to 45 kJ/mol in only 0.3 nm. In contrast, forward flux sampling predicts an interaction potential which rises to 25 kJ/mol in 0.8 nm and 45 kJ/mol in more than 1 nm. The reason for this difference is the more complex reaction pathway predicted by forward flux sampling, which allows for reorientation and conformational changes of the two proteins.

We conclude that the reaction pathway of dissociation is straightforward for small contacts such as contact 2 and becomes more complex with increasing contact area. For most of the crystal contacts that have been investigated by umbrella sampling [16,17,19], a small contact area is typical and therefore the interaction was reliably described by an umbrella sampling approach. However, for crystal contacts with a large contact area such as contact 1, a more realistic evaluation of the reaction pathway which allows for reorientation of the proteins coupled to local conformational changes seems to be necessary. This is especially true for specific, biologically

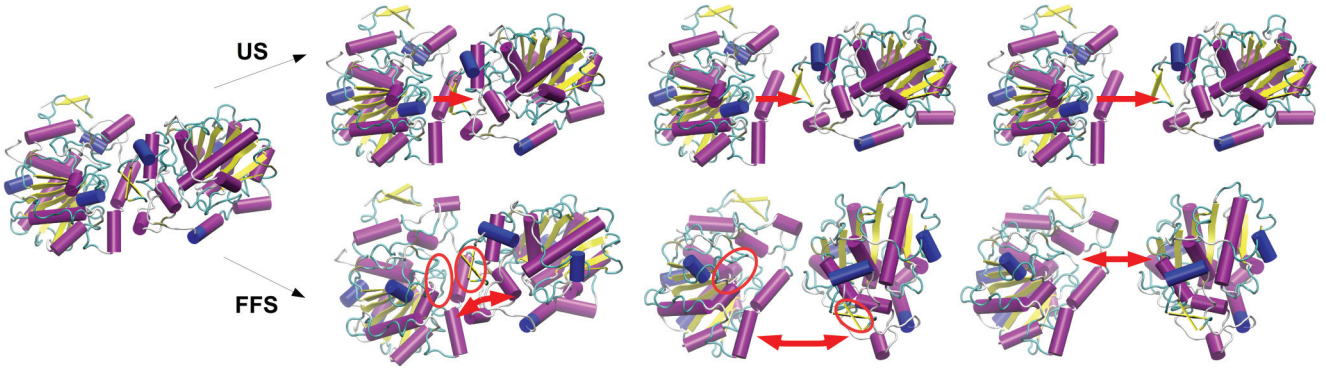

FIG. 11. (Color) Dissociation pathways of contact 1 using umbrella sampling (US) and forward flux sampling (FFS). Red arrows indicate the reaction pathway. The umbrella sampling reaction pathway is predefined along the axis connecting the centers of mass of the two proteins. In contrast, the forward flux sampling reaction pathway is a result of the simulation and is more complex: In a first step, the two proteins detach at the charged patch (red circles) while they are still connected at the hydrophobic stretch. In a second step, the hydrophobic stretch gradually disassembles like a zipper until the proteins are finally fully separated.

relevant protein-protein contacts which generally have a much larger contact area than crystal contacts [10].

It is noteworthy that contact 1 and 2, despite their different contact areas (640 $Å^2$ and 300 $Å^2$), have a similar binding free energy per contact area, 73–78 J $Å^{-2}$ mol^{-1}, which is near the estimation of 100 J $Å^{-2}$ mol^{-1} (24 cal $Å^{-2}$ mol^{-1}) of buried hydrophobic surface of amino acids [33], a value that has been confirmed for folding [34], protein-protein contacts [35], and binding of small molecules [36], suggesting that a large part of the binding energy is caused by hydrophobic interactions. In addition, the slopes of the interaction potentials are also comparable (24–31 kJ nm^{-1} mol^{-1}, FFS potential for contact 1). In contrast, contact 6 seems to have very different properties than the other investigated contacts: The slope of 7.5 kJ nm^{-1} mol^{-1} might be caused by the high density of electrostatic interactions at this contact. The binding free energy per contact area of 281 J $Å^{-2}$ mol^{-1} is probably an artifact of the pairwise simulation approach, because the two proteins are free to come closer and form a bigger surface than in the geometrically constrained crystal lattice (see Fig. 9).

In contrast to umbrella sampling with its constrained reaction path, forward flux sampling is free of explicit constraints and results in the most probable reaction path. Thus, forward flux sampling provides information about the kinetics of the dissociation and association process and provides a tool to identify whether kinetics or thermodynamics decides about the favored orientation of protein-protein contacts. Although contact 1 is thermodynamically much more favorable, contacts 4 and 6 were almost exclusively observed in association simulations using forward flux sampling. This result can be explained by the different attractive ranges of the contacts and their position on the protein surface. As the umbrella sampling and the forward flux simulations showed, the attractive range of contacts 4 and 6 reaches a center-of-mass distance of approximately 5.7 nm, while contacts 1 and 2 are attractive until a center-of-mass distance of 5.1 and 5.4 nm, respectively. Furthermore, contacts 1 and 6 are direct neighbors sharing some of the surface involved in contact formation

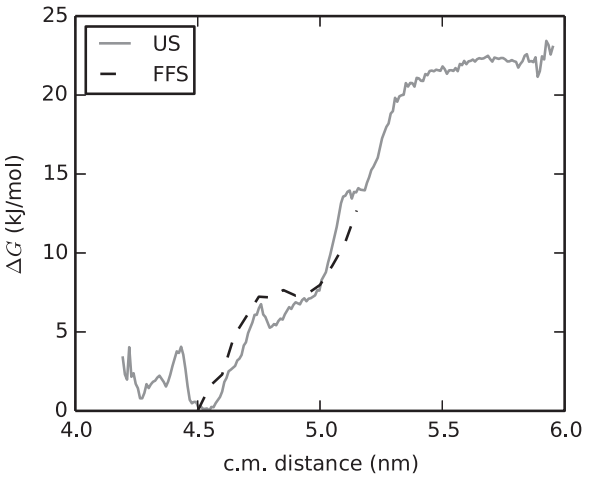

FIG. 12. Free energy profiles for the dissociation of contact 2 obtained with forward flux sampling (FFS) and umbrella sampling (US). The free energy profiles are very similar for both methods.

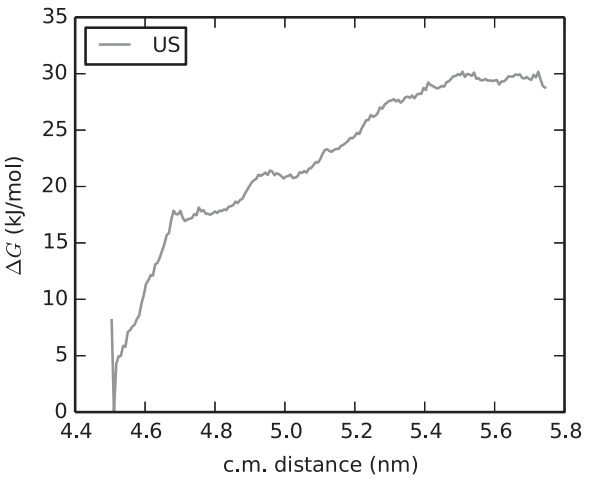

FIG. 13. Free energy profile for the dissociation of contact 3 obtained with umbrella sampling (US).

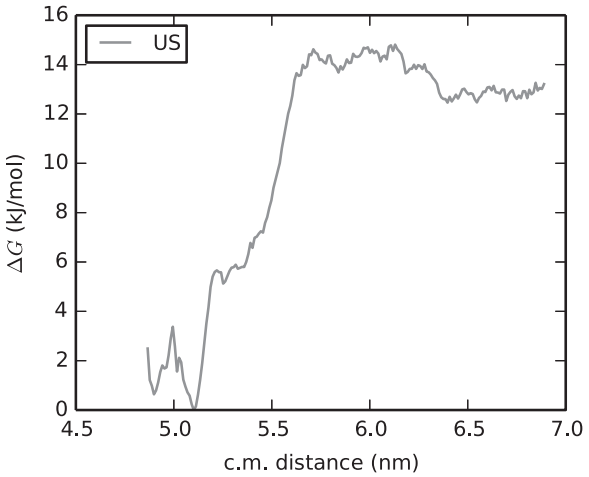

FIG. 14. Free energy profiles for the dissociation of contact 4 obtained with umbrella sampling (US).

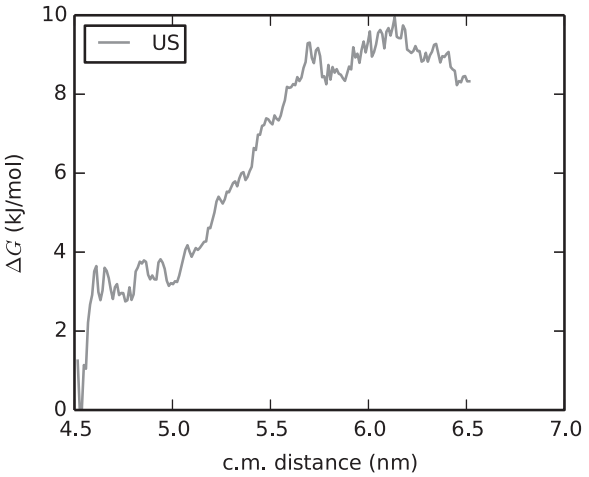

FIG. 16. Free energy profiles for the dissociation of contact 6 obtained with umbrella sampling (US).

(Table III, F205). This provides a plausible mechanism for the kinetic disfavor of contact 1: As two proteins approach in the correct orientation to form either contact 1 or contact 6, contact 6 is most likely to form because of its long-range potential. This results in a lower free energy at larger distances of the paths leading to contacts 4 and 6 and, therefore, most association paths are directed towards contacts 4 and 6, which are thus kinetically favored. Hence, upon nucleation the thermodynamically most stable contacts might not form first, but only after the kinetically preferred contacts were formed.

It is important to emphasize that our approach only compares crystal contacts. Conclusions about nucleation or crystallization which involves many proteins and additional geometric constraints are inferred by comparing crystal contacts which differ considerably in their dissociation energy and the range of the attractive potential.

B. Relevance for crystallization

There are two contradictory observations regarding crystal contacts of proteins. On one hand, they seem to be random and unspecific [9,11,13], but on the other hand, changes of the protein surface might sensitively influence the crystallization behavior and contact interactions [3,10,14–17]. These contradictory observations can be explained by the delicate balance of kinetics (preferring contact 4 and 6) and thermodynamics (preferring contact 1) during the nucleation process. We propose that nucleation is driven by association of the kinetically preferred contacts, which are characterized by a long-range potential and/or an exposed position on the protein surface. A long-range potential is characteristic for electrostatic interactions, and indeed, electrostatics was identified as the dominant driving force for the reorientation of proteins at a charged surface [37–39], driven by the heterogeneous distribution of charges on the protein surface. Also, it has been shown that two protein molecules in a crystal are oriented such that they compensate their charges [40]. However, the final thermodynamic stability of a crystal is provided by short-range hydrophobic contacts, which are formed later in the nucleation process and are a consequence of the initially formed contacts. This view explains why the whole protein surface can participate in crystal contacts [13], while at the same time specificity for the different contacts can be observed [16,17]. Our results are in accordance with the two-step model of protein crystallization [41]. First, a high-density liquid droplet without long-range order is formed, which then reorganizes into an ordered crystallization nucleus. In the high-density liquid droplet we expect the kinetically favored contacts to dominate. During the reorganization most of these initial contacts would disappear, and the most stable hydrophobic contacts form.

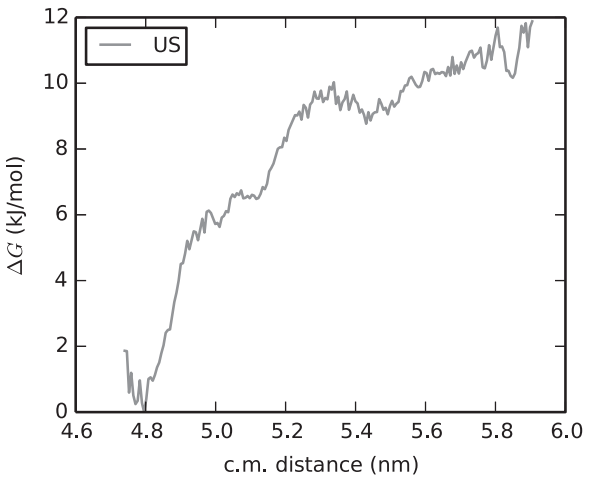

FIG. 15. Free energy profile for the dissociation of contact 5 obtained with umbrella sampling (US).

ACKNOWLEDGMENTS

The authors acknowledge the Deutsche Forschungsgemeinschaft (SFB716) for financial support and the High Performance Computing Center Stuttgart for kindly providing computational resources.

[1] A. McPherson, Methods **34**, 254 (2004).
[2] N. E. Chayen and E. Saridakis, Nat. Methods **5**, 147 (2008).
[3] Z. S. Derewenda and P. G. Vekilov, Acta Crystallogr. Sect. D: Biol. Crystallogr. **62**, 116 (2006).
[4] D. R. Cooper, T. Boczek, K. Grelewska, M. Pinkowska, M. Sikorska, M. Zawadzki, and Z. Derewenda, Acta Crystallogr. Sect. D: Biol. Crystallogr. **63**, 636 (2007).
[5] A. Dong, X. Xu, A. M. Edwards, C. Chang, M. Chruszcz, M. Cuff, M. Cymborowski, R. Di Leo, O. Egorova, E. Evdokimova, E. Filippova, J. Gu, J. Guthrie, A. Ignatchenko, A. Joachimiak, N. Klostermann, Y. Kim, Y. Korniyenko, W. Minor, Q. Que, A. Savchenko, T. Skarina, K. Tan, A. Yakunin, A. Yee, V. Yim, R. Zhang, H. Zheng, M. Akutsu, C. Arrowsmith, G. V. Avvakumov, A. Bochkarev, L.-G. Dahlgren, S. Dhe-Paganon, S. Dimov, L. Dombrovski, P. Finerty, S. Flodin, A. Flores, S. Gräslund, M. Hammerström, M. D. Herman, B.-S. Hong, R. Hui, I. Johansson, Y. Liu, M. Nilsson, L. Nedyalkova, P. Nordlund, T. Nyman, J. Min, H. Ouyang, H.-w. Park, C. Qi, W. Rabeh, L. Shen, Y. Shen, D. Sukumard, W. Tempel, Y. Tong, L. Tresagues, M. Vedadi, J. R. Walker, J. Weigelt, M. Welin, H. Wu, T. Xiao, H. Zeng, and H. Zhu, Nat. Methods **4**, 1019 (2007).
[6] A. Wernimont and A. Edwards, PLoS ONE **4**, e5094 (2009).
[7] M. Kuge, Y. Fujii, T. Shimizu, F. Hirose, A. Matsukage, and T. Hakoshima, Protein Sci. **6**, 1783 (1997).
[8] L. Corsini, M. Hothorn, K. Scheffzek, M. Sattler, and G. Stier, Protein Sci. **17**, 2070 (2008).
[9] J. Janin, Nat. Struct. Biol. **4**, 973 (1997).
[10] S. Dasgupta, G. H. Iyer, S. H. Bryant, C. E. Lawrence, and J. A. Bell, Proteins **28**, 494 (1997).
[11] O. Carugo and P. Argos, Protein Sci. **6**, 2261 (1997).
[12] R. P. Saha, R. P. Bahadur, and P. Chakrabarti, J. Proteome Res. **4**, 1600 (2005).
[13] M. P. Crosio, J. Janin, and M. Jullien, J. Mol. Biol. **228**, 243 (1992).
[14] J. P. K. Doye, A. A. Louis, and M. Vendruscolo, Phys. Biol. **1**, P9 (2004).
[15] W. N. Price, Y. Chen, S. K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran, J. Everett, S. N. Tong, F. Forouhar, S. S. Swaminathan, T. Acton, R. Xiao, J. R. Luft, A. Lauricella, G. T. DeTitta, B. Rost, G. T. Montelione, and J. F. Hunt, Nat. Biotechnol. **27**, 51 (2009).
[16] G. Pellicane, G. Smith, and L. Sarkisov, Phys. Rev. Lett. **101**, 248102 (2008).
[17] D. Fusco, J. J. Headd, A. De Simone, J. Wang, and P. Charbonneau, Soft Matter **10**, 290 (2014).
[18] G. Torrie and J. Valleau, J. Comput. Phys. **23**, 187 (1977).
[19] J. A. Lemkul and D. R. Bevan, J. Phys. Chem. B **114**, 1652 (2010).
[20] R. J. Allen, C. Valeriani, and P. ten Rein Wolde, J. Phys.: Condens. Matter **21**, 463102 (2009).
[21] M. W. Larsen, U. T. Bornscheuer, and K. Hult, Protein Expression Purif. **62**, 90 (2008).
[22] P. Trodler and J. Pleiss, BMC Struct. Biol. **8**, 9 (2008).
[23] H. J. C. Berendsen, D. V. D. Spoel, and R. V. Drunen, Comput. Phys. Commun. **91**, 43 (1995).
[24] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).
[25] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[26] J. Uppenberg, M. T. Hansen, S. Patkar, and T. Jones, Structure **2**, 293 (1994).
[27] J. S. Hub, B. L. de Groot, and D. van der Spoel, J. Chem. Theor. Comput. **6**, 3713 (2010).
[28] R. J. Allen, D. Frenkel, and P. R. ten Wolde, J. Chem. Phys. **124**, 024102 (2006).
[29] K. Kratzer, J. T. Berryman, A. Taudt, J. Zeman, and A. Arnold, Comput. Phys. Commun. **185**, 1875 (2014).
[30] E. Krissinel and K. Henrick, J. Mol. Biol. **372**, 774 (2007).
[31] A. Tovchigrechko and I. A. Vakser, Nucl. Acids Res. **34**, W310 (2006).
[32] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker, Nucl. Acids Res. **32**, W665 (2004).
[33] C. Chothia, Nature (London) **248**, 338 (1974).
[34] M. Matsumura, W. J. Becktel, and B. W. Matthews, Nature (London) **334**, 406 (1988).
[35] E. J. Sundberg, M. Urrutia, B. C. Braden, J. Isern, D. Tsuchiya, B. A. Fields, E. L. Malchiodi, J. Tormo, F. P. Schwarz, and R. A. Mariuzza, Biochemistry **39**, 15375 (2000).
[36] S. W. Rick, I. A. Topol, J. W. Erickson, and S. K. Burt, Protein Sci. **7**, 1750 (1998).
[37] B. Yoon and A. Lenhoff, J. Phys. Chem. **96**, 3130 (1992).
[38] M. L. Grant and D. A. Saville, J. Phys. Chem. **98**, 10358 (1994).
[39] A. Steudle and J. Pleiss, Biophys. J. **100**, 3016 (2011).
[40] T. Takahashi, S. Endo, and K. Nagayama, J. Mol. Biol. **234**, 421 (1993).
[41] D. Erdemir, A. Lee, and A. Myerson, Acc. Chem. Res. **42**(5), 621 (2009).