Atomic-level characterization of protein–protein association

Albert C. Pan\textsuperscript{a,1}, Daniel Jacobson\textsuperscript{a}, Konstantin Yatsenko\textsuperscript{a}, Duluxan Sritharan\textsuperscript{a}, Thomas M. Weinreich\textsuperscript{a}, and David E. Shaw\textsuperscript{a,b,1}

\textsuperscript{a}D. E. Shaw Research, New York, NY 10036; and \textsuperscript{b}Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032

Edited by Barry Honig, Howard Hughes Medical Institute and Columbia University, New York, NY, and approved January 3, 2019 (received for review September 6, 2018)

Despite the biological importance of protein–protein complexes, determining their structures and association mechanisms remains an outstanding challenge. Here, we report the results of atomic-level simulations in which we observed five protein–protein pairs repeatedly associate to, and dissociate from, their experimentally determined native complexes using a molecular dynamics (MD)–based sampling approach that does not make use of any prior structural information about the complexes. To study association mechanisms, we performed additional, conventional MD simulations, in which we observed numerous spontaneous association events. A shared feature of native association for these five structurally and functionally diverse protein systems was that if the proteins made contact far from the native interface, the native state was reached by dissociation and eventual reassociation near the native interface, rather than by extensive interfacial exploration while the proteins remained in contact. At the transition state (the conformational ensemble from which association to the native complex and dissociation are equally likely), the protein–protein interfaces were still highly hydrated, and no more than 20% of native contacts had formed.

molecular dynamics simulations | enhanced sampling | protein–protein association

Most proteins associate with other proteins to function, forming complexes that lie at the heart of nearly all physiological processes, including signal transduction, DNA repair, enzyme inhibition, and the immune response. Determining the structures of these complexes and elucidating their association mechanisms are problems of fundamental importance. While substantial progress has been made toward the structural determination of protein–protein complexes, such structures are still relatively underrepresented in the Protein Data Bank (1), especially compared with the large number of known, functional protein–protein interactions derived from high-throughput, nonstructural approaches like yeast two-hybrid screening and affinity purification–mass spectrometry (2). Moreover, the structures of many complexes that are important drug targets for cancer and autoimmune disease remain difficult to determine experimentally (3, 4). As for protein–protein association mechanisms, powerful experimental approaches like double-mutant cycles and paramagnetic relaxation enhancement have afforded a wealth of information about potential transition states and intermediates (5, 6), but these data are often indirect or limited to, for example, metalloproteins or proteins with attached paramagnetic spin labels. Obtaining direct, atomic-level detail about association pathways for a diverse set of protein–protein complexes and developing broader insights into the common principles of protein–protein association mechanisms are still open problems.

Atomic-level molecular dynamics (MD) simulations offer a computational route toward characterizing both the structure and dynamics of protein–protein complexes. Using MD, one could in principle start a simulation with two protein monomers and “watch” them associate and dissociate reversibly during a single trajectory. Such a simulation would provide an unprecedented sampling of the possible complexes that can be formed by the protein monomers and a straightforward way to rank the stability of different complexes based on that fraction of the time during which each complex is observed. Moreover, mechanistic details like intermediate and transition states along the association pathway could be observed. In practice, however, it has proven difficult to study protein–protein association and dissociation in MD simulations: Reversible association during a single simulation trajectory has not been observed at all. Indeed, only a few examples have been reported (7–15) of simulations that successfully captured spontaneous protein–protein association to an experimentally determined complex, and these examples have been limited to examining one system, or to the association of smaller peptides (7, 11, 14). Recently, Plattner et al. (10) were the first to capture association and dissociation of a protein–protein complex (albeit in separate simulations) as part of a large-scale Markov state model study.

Difficulties encountered in protein–protein association simulations include the formation of nonnative associated states with long lifetimes compared with simulation timescales. Such kinetic traps severely hamper the sampling of other states—including states that may be thermodynamically more favorable, and thus more likely to represent the most populated complex at physiological conditions (16, 17). Even if the most thermodynamically stable complex is sampled, observing spontaneous dissociation could require simulations on the order of seconds to days (18). Some approaches have attempted to overcome this timescale problem by combining data from multiple short simulation runs. Different approaches have attempted to overcome this timescale problem by combining data from multiple short simulation runs.

Significance

Most proteins associate with other proteins to function, forming complexes that are central to almost all physiological processes. Determining the structures of these complexes and understanding how they associate are problems of fundamental importance. Using long-timescale molecular dynamics simulations, some performed using a new enhanced sampling method, we observed spontaneous association and dissociation of five protein–protein systems and from their experimentally determined native complexes. By analyzing the simulations of these five systems, which include members of diverse structural and functional classes, we are able to draw general mechanistic conclusions about protein association.

Author contributions: A.C.P., D.J., K.Y., and D.E.S. designed research; A.C.P., D.J., K.Y., D.S., and T.M.W. performed research; and A.C.P., D.J., K.Y., and D.E.S. wrote the paper. The authors declare no conflict of interest.

This is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

To whom correspondence may be addressed. Email: Albert.Pan@DEShawResearch.com or David.Shaw@DEShawResearch.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1815431116/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1815431116
trajectories (10, 12, 14) and have had success in modeling various aspects of protein–protein association. None of these methods, however, has been applied to a broad set of structurally diverse protein–protein systems, making it difficult to draw general conclusions about protein association. Moreover, such methods require an additional layer of modeling to combine the short trajectories and necessarily rest on assumptions that can bias the results toward atypical pathways or miss important conformational states (19, 20).

In this paper, we have used long-timescale MD simulations in combination with a newly developed enhanced sampling approach, which we call “tempered binding,” to simulate the reversible association of five structurally and functionally diverse protein–protein systems, and have also performed conventional MD simulations to capture many spontaneous association events. In the tempered binding simulations, we observed each of the protein–protein pairs repeatedly associate to, and dissociate from, their experimentally determined native complexes. The spontaneous association events observed in subsequent conventional MD simulations allowed us to draw general mechanistic conclusions about the association process of the five proteins we studied: We found that if the proteins made contact far from the native interface, the native state was reached by dissociation and eventual reassociation near the native interface, as opposed to extensive exploration of various protein–protein interfaces while the proteins remained in contact. The protein–protein interfaces were still highly hydrated at the transition state, and no more than 20% of native contacts had formed.

Results

Tempered Binding Simulations of Protein–Protein Association Reversibly Visited the Native Complex. We applied tempered binding to six protein–protein systems and observed reversible association in five out of the six systems (SI Appendix, Figs. S1 and S4 and Table S1). For the five reversibly associating systems, the most stable complex in the simulation agrees with the complex determined crystallographically within atomic resolution (Fig. 1). In the sixth protein–protein system, the protein dimer CLC-ec1 (21), we observed association to the experimentally determined complex, but did not observe dissociation on the timescales of our simulations (SI Appendix, Fig. S4). Some of these simulations also sampled alternative bound states that could have functional relevance, and provided quantitative estimates of the free-energy difference between the native state and these alternative states (Fig. 2 and SI Appendix, Figs. S2 and S3). Although for this initial study we have limited ourselves to proteins that do not undergo large conformational changes upon binding (SI Appendix, Table S1), we note that such proteins in themselves constitute a large class of important protein–protein complexes (22). The ribonuclease HI–SSB-Ct system could be considered an exception, since the SSB-Ct peptide folds upon binding to the ribonuclease, but the RMSD difference between the folded and disordered forms of the peptide is only about 2 Å.

In a tempered binding simulation, the strength of interactions between the protein monomer atoms, and sometimes between the protein monomer and solvent atoms, is scaled at regular time intervals using a simulated Hamiltonian tempering framework (14, 23–25). This scaling allows long-lived states to dissociate more quickly. In practice, tempered binding involves a conventional MD simulation augmented by frequent Monte Carlo moves that update the scaling strength among rungs on a ladder of values. The Monte Carlo updates are detail balanced such that, at each rung of the ladder, a Boltzmann distribution of states corresponding to that rung’s value of the scaling factor is properly sampled. In particular, the sampling at the lowest rung of the ladder (rung 0) corresponds to the completely unscalled Hamiltonian and is consistent with the distribution of states sampled in a conventional MD simulation.

Our current tempered binding protocol (which focuses on scaling the near electrostatic interactions between protein monomers, between protein monomers and water molecules, and, in the case of CLC-ec1, between protein monomers and lipid molecules) resulted in a significant increase in sampling efficiency in our simulations of the protein–protein systems studied in this work. In a tempered binding simulation of the enzyme–inhibitor system barnase–barstar, for example, the protein–protein system escaped from its native complex in hundreds of microseconds (SI Appendix, Fig. S1), whereas the lifetime of the native complex is on the order of a day (26), a speedup of almost nine orders of magnitude. It is possible, however, that other tempered binding protocols might further improve sampling.
Raf-RBD, at least three different complex. In some simulations, there were unsuccessful association events preceding a successful association event. We note that we observed several successful association events for each of the five systems (Table 1), and, as expected, we observed no examples of dissociation once the experimentally determined complex formed.

Successful association events in these five systems shared several key features. Rather than forming an encounter complex at a random interface and reaching the native interface (without dissociating) by way of an extensive search, in successful association events the encounter complexes tended to form near the native interface, at least for events observed within the timescale of our simulations (on the order of tens of microseconds). (For a given protein–protein system, the contact preceding a successful association event tended to be nonspecific, varying among different events, but typically involved interactions between charged residues in or close to the native binding interface.) In contrast, encounter complexes that formed during unsuccessful association events displayed a wide variety of relative protein–protein positions, with no pronounced preference for their positions in the native complex (Fig. 3 and SI Appendix, Fig. S5).

Successful association events in the five systems also shared similar features at the transition state (the ensemble of configurations from which association and dissociation are equally likely): No more than 20% of native contacts had formed in these configurations, and the protein–protein interfaces were still highly hydrated. We identified configurations in the transition state ensemble of association by calculating the probability of successful association, $p_{assoc}$ (also known as the committor probability), for several configurations during a successful association event (27, 28), and were able to identify configurations at or near the transition state for all five systems. A configuration is classified as a member of the transition state ensemble if $p_{assoc} = 50\%$ (that is, if additional trajectories initiated from that configuration with randomized velocities drawn from a Boltzmann distribution commit half of the time to the native complex and half of the time to the unbound state). All of the transition states characterized here occurred when <20% of native contacts had formed and while there were still a significant number of water molecules between atoms that are in contact in the native complex (Fig. 4). Given the intensive computational effort required for determining committor probabilities and identifying transition states, we only determined committor values in one successful association event for each system condition. The general features of the transition state configurations were found to be qualitatively similar among simulations of barnase–barstar with different force fields (Table 1), and even among completely different systems (Fig. 4D), however, providing strong evidence that these transition states are representative of the transition state ensembles for these five systems.

### Table 1. List of conventional MD simulations of spontaneous protein–protein association

| Condition | System                  | No. of trajectories/successful association events | Aggregate time, µs | Association rate, M$^{-1}$s$^{-1}$ |
|-----------|-------------------------|-------------------------------------------------|--------------------|-----------------------------------|
| BB_1      | Barnase–barstar         | 61/24                                           | 440                | 2.3(2) × 10$^7$                  |
| BB_2      | Barnase–barstar         | 61/28                                           | 212.7              | 6(1) × 10$^7$                    |
| BB_3      | Barnase–barstar         | 51/3                                            | 225                | 4.4 × 10$^6$                     |
| IND_2     | Insulin dimer           | 61/6                                            | 294.8              | 4.1 × 10$^6$                     |
| RAS_1     | Ras–Raf-RBD             | 81/7                                            | 117                | 2.6 × 10$^6$                     |
| RNA_1     | RNase HI–SSB–Ct         | 51/5                                            | 294                | 1.0 × 10$^6$                     |
| TYK_2     | TYK2-Pseudokinase       | 151/3                                           | 461                | 4.2 × 10$^5$                     |

Simulations were initiated with different velocities from the same configuration with the two proteins unbound (different association pathways of the two proteins were sampled, providing strong evidence that the choice of starting positions does not influence the binding events; SI Appendix, Fig. S6). A successful association event is defined as a trajectory in which the I-RMSD between the two proteins and the experimentally determined complex structure became less than 1.5 Å after the RMSD time series was smoothed with a 10-ns moving average. Over 2 ms of aggregate simulation time was used for these spontaneous association trajectories. The experimental rates of association for barnase–barstar and insulin dimer are $6.0 \times 10^6$ M$^{-1}$s$^{-1}$ (32) and $1.14 \times 10^6$ M$^{-1}$s$^{-1}$ (56), respectively. For Ras–Raf-RBD, at least three different experimental association rate estimates are available: $4.5 \times 10^6$ M$^{-1}$s$^{-1}$ (57), $5.9 \times 10^6$ M$^{-1}$s$^{-1}$ (58), and $1.02 \times 10^6$ M$^{-1}$s$^{-1}$ (59). We were unable to find experimental rates of association for RNase HI–SSB–Ct or TYK2-Pseudokinase.

Pan et al.
Unsuccessful association  Successful association

Barnase – Barstar

Insulin dimer

Ras – Raf-RBD

RNase HI – SSB-Ct

TYK2 – Pseudokinase

Fig. 3. Encounter complexes visited in successful association events favored structures in which the two proteins were positioned similarly to how they are positioned in the experimentally determined complex. Simulation frames were uniformly sampled from encounter complexes and aligned to the larger protein. A single snapshot of the larger protein (blue cartoon) is shown for reference, overlaid with multiple snapshots of a Ca atom of the smaller protein near the center of the native binding interface taken from unsuccessful (red spheres) and successful (green spheres) association trajectories. The large yellow sphere indicates a region defined by a 10-Å radius around the center of mass of the binding interface of the larger protein. Kinetically trapped nonnative states, which neither dissociated nor reached the native state during a simulation, were not included in this analysis.

Discussion

For the enzyme–inhibitor complex barnase–barstar, one of the best experimentally characterized protein–protein complexes (26), our simulations are consistent with previous experimental and computational work, including extensive mutational analysis (5, 29) and Brownian dynamics simulations (30, 31). Combining the information from the tempered binding and conventional MD simulations, we estimated the binding free energy, \( \Delta G_{b} \), to be 19.2(2) kcal mol\(^{-1} \) \([ \Delta G_{b} = 1.0(2) \times 10^{-1} \text{ M J} ]\), the association rate, \( k_{\text{on}} \), to be 2.3(2) \( \times 10^{7} \) M\(^{-1} \) s\(^{-1} \), and the dissociation rate, \( k_{\text{off}} \), to be 2.3(3) \( \times 10^{-3} \) s\(^{-1} \). These simulation-derived values correspond relatively well with the known experimental values of \( \Delta G_{b} = 19 \) kcal mol\(^{-1} \), \( k_{\text{on}} = 6 \times 10^{9} \) M\(^{-1} \) s\(^{-1} \), and \( k_{\text{off}} = 8 \times 10^{-6} \) s\(^{-1} \) (32). (Further discussion of how the simulation values were calculated can be found in SI Appendix.)

Notably, our atomic picture of the transition state agrees with mutational and kinetic studies from Schreiber and coworkers (5, 33), which suggest that the transition state occurs before a majority of native interactions are formed, and while the protein–protein interface is still highly solvated. More recent experimental work on the coupled folding and binding of disordered protein domains also suggests that the majority of native contacts form after the transition state (34).

In addition, our observation that associating proteins already had relative positions similar to those in the native complex upon making contact during successful simulated association events is consistent with the idea of a so-called “funneled” association process, in which long-range electrostatic attraction is involved in the rapid association of barnase–barstar (31, 35) and other protein–protein pairs with oppositely charged binding sites (36). This observation is also consistent with previous atomistic MD simulation studies of protein–peptide binding, in which successful association pathways tended to be funnel-like, often due to electrostatic steering (7, 11, 14). Relatedly, in a recent simulation study of a peptide binding to the PDZ domain, multiple binding and unbinding events were observed before successful association, and the peptide did not extensively search the surface while bound (11).

The strong affinity of barstar for barnase and the hydrophilic nature of the interface make it a relatively unusual protein–protein system, so it is striking that the features of the transition state and encounter complex observed in our simulations were common to the association mechanisms of all of the protein–protein systems studied in this work (Fig. 4). This shared mechanism may thus also apply to the broader class of protein–protein complexes that—like the systems studied here (SI Appendix, Table S1)—do not undergo large conformational changes upon binding.

We have observed reversible association of a set of five protein–protein systems to their respective experimentally determined structures using an enhanced sampling method that enabled an increase in sampling efficiency of as much as nine orders of magnitude. Together with our long-timescale conventional MD simulations, which yielded many spontaneous association events, our results provide an atomic-level view of protein–protein association mechanisms. In the future, this methodology could be used to determine the structures and association mechanisms of at least some protein–protein complexes that have not yet been experimentally characterized. The ability to observe both association and dissociation events could be especially useful in this context, helping to distinguish thermodynamically stable complexes from kinetically trapped states that are sparsely populated.
To enhance the sampling in our protein simulations, we applied torsional corrections to the \( \phi \) and \( \psi \) backbone dihedrals, consisting of an additional potential term.

For the membrane–protein dimer system CLC-ec1, we truncated the dynamic N-terminal helix (residues 17–30) on both monomers, and inserted the system into a palmitoyl oleoyl phosphatidylethanolamine (POPE) lipid bilayer. The CHARMM27 force field was used for the protein, ions, and water (43, 44), and the CHARMM36 force field for lipids (45). We also adjusted the side-chain charges of aspartate, glutamate, and arginine residues (“DER” correction) to weaken the guanidinium acetate association constant (46).

Methods

**MD Simulations.** Simulations were based on the crystal structures of the protein complexes listed in SI Appendix, Table S1. Unless otherwise indicated, Lys, Arg, Asp, and Glu residues, as well as the N and C termini, were simulated in their charged states, and all His residues were neutral. For the soluble proteins, structures were solvated with water molecules, and counterions were added until the system was overall neutral. We used the Amber ff99SB*-ILDN (37–39) force field and the TIP3P (40) water model for these simulations, except for a few simulations of barnase–barstar, in which we also used the four-site water model TIP4P2005 (41) as a control. The Ras–Raf-RBD system contained the ligand GppNHp, a GTP analog, and the TYK2–pseudokinase system contained two copies of the ligand compound 7012, a pseudokinase system. For some of our protein–protein systems, we applied torsional corrections to the \( \phi \) and \( \psi \) backbone dihedrals, consisting of an additional potential term.

For the membrane–protein dimer system CLC-ec1, we truncated the dynamic N-terminal helix (residues 17–30) on both monomers, and inserted the system into a palmitoyl oleoyl phosphatidylethanolamine (POPE) lipid bilayer. The CHARMM27 force field was used for the protein, ions, and water (43, 44), and the CHARMM36 force field for lipids (45). We also adjusted the side-chain charges of aspartate, glutamate, and arginine residues (“DER” correction) to weaken the guanidinium acetate association constant (46).

**Tempered Binding.** To enhance the sampling in our protein–protein association simulations, we used an approach that we call tempered binding, which is analogous to simulated tempering (47, 48), except that terms in the system’s energy function, as opposed to its temperature, are updated during the simulation. Tempered binding dynamically scales various atomic interactions during an MD simulation by a factor, \( \lambda \), that is updated among a ladder of discrete values, \( \lambda_i \). At the lowest rung of the ladder (run 0), the atomic interactions are unscaled (i.e., \( \lambda_0 = 1 \)). An update attempt was
chosen with a probability consistent with $p(\lambda i) \geq 49)$. Our tempered binding simulations scaled near electrostatic and van der Waals nonbonded interactions between different groups of atoms in the system. Additional details about tempered binding can be found in *SI Appendix*.

**ACKNOWLEDGMENTS**. We thank Michael Eastwood, Cristian Predescu, Jesus Izaguirre, and Doug Ladardi for helpful discussions; and Berkman Frank, Rebecca Bish-Cornelissen, and Jessica McGilfen for editorial assistance.

1. Lenskii MF, et al. (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPR experiment. Proteins 84:323–348.

2. Vidal M, Cusick ME, Barabasi AL (2011) Intertactome networks and human disease. Cell 144:986–998.

3. Luptadis P, et al. (2014) Structure of the pseudokinase-kinase domains from protein kinase TYK2 reveals a mechanism for Janus kinase (JAK) autoinhibition. Proc Natl Acad Sci USA 111:8025–8030.

4. Shan Y, et al. (2014) Molecular basis for pseudokinase-dependent autoinhibition of JAK2 tyrosine kinase. Nat Struct Mol Biol 21:579–584.

5. Frisch C, Fersht AR, Schreiber G (2001) Experimental assignment of the structure of the transition state for the association of barnase and barstar. J Mol Biol 308:69–77.

6. Fersht AR (2002) On the simulation of protein folding by short time scale molecular dynamics simulations. J Mol Biol 324:10549–10564.

7. Huang L, Hofer F, Martin GS, Kim SH (1998) Structural basis for the interaction of Ras and the Ras-binding domain of c-Raf-1 reveal new approach to Monte Carlo calculation of the free energy: Method of expanded ensemble simulations. J Chem Phys 108:2362–2370.

8. Saglam AS, Wang DW, Zwier MC, Chong LT (2017) Flexibility vs preorganization: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 35:1400–1415.

9. Klauuda JB, et al. (2010) Update of the CHARMM all-atom additive force field for liquids: Validation on six lipid types. J Phys Chem B 114:7830–7843.

10. Abascal JL, Vega C (2005) A general purpose model for the condensed phases of water: TIP4P/2005. J Chem Phys 123:234505.

11. Plattner N, Doerr S, De Fabritiis G, Noé F (2017) Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markovian modelling. Nat Chem 9:1005–1011.

12. Abascal JL, Vega C (2003) A general purpose model for the condensed phases of liquid water. J Phys Chem B 107:2362–2370.

13. Vreven T, et al. (2015) Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 427:3361–3383.

14. Fersht AR (2002) On the simulation of protein folding by short time scale molecular dynamics and distributed computing. Proc Natl Acad Sci USA 99:14122–14125.

15. Paci E, Cavalli A, Vendruscolo M, Cafiso A (2003) Analysis of the distributed computing approach applied to the folding of a small α-helical peptide. Proc Natl Acad Sci USA 100:13222–13227.

16. Du R, et al. (1998) On the transition coordinate for protein folding. J Chem Phys 108:334–340.

17. Schreiber G, Fersht AR (1995) Energetics of protein-protein interactions: Analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 248:478–486.