Ensemble-based characterization of unbound and bound states on protein energy landscape

Anatoly M. Ruvinsky,[1]* Tatsiana Kirys,[1,2] Alexander V. Tuzikov,[2] and Ilya A. Vakser[1,3]

[1]Center for Bioinformatics, The University of Kansas, Lawrence, Kansas 66047
[2]United Institute of Informatics Problems, National Academy of Sciences, 220012 Minsk, Belarus
[3]Department of Molecular Biosciences, The University of Kansas, Lawrence, Kansas 66045

Received 18 November 2012; Revised 2 February 2013; Accepted 15 March 2013
DOI: 10.1002/pro.2256
Published online 23 March 2013 proteinscience.org

Abstract: Physicochemical description of numerous cell processes is fundamentally based on the energy landscapes of protein molecules involved. Although the whole energy landscape is difficult to reconstruct, increased attention to particular targets has provided enough structures for mapping functionally important subspaces associated with the unbound and bound protein structures. The subspace mapping produces a discrete representation of the landscape, further called energy spectrum. We compiled and characterized ensembles of bound and unbound conformations of six small proteins and explored their spectra in implicit solvent. First, the analysis of the unbound-to-bound changes points to conformational selection as the binding mechanism for four proteins. Second, results show that bound and unbound spectra often significantly overlap. Moreover, the larger the overlap the smaller the root mean square deviation (RMSD) between the bound and unbound conformational ensembles. Third, the center of the unbound spectrum has a higher energy than the center of the corresponding bound spectrum of the dimeric and multimeric states for most of the proteins. This suggests that the unbound states often have larger entropy than the bound states. Fourth, the exhaustively long minimization, making small intrarotamer adjustments (all-atom RMSD $\leq$ 0.7 Å), dramatically reduces the distance between the centers of the bound and unbound spectra as well as the spectra extent. It condenses unbound and bound energy levels into a thin layer at the bottom of the energy landscape with the energy spacing that varies between 0.8–4.6 and 3.5–10.5 kcal/mol for the unbound and bound states correspondingly. Finally, the analysis of protein energy fluctuations showed that protein vibrations itself can excite the interstate transitions, including the unbound-to-bound ones.

Keywords: protein–protein interactions; energy landscape; protein ensemble; binding mechanisms

Introduction

Relationships between protein energy landscape, structure, and function have been a subject of numerous studies resulted in the development of the funnel shape energy landscape theory.[1–4] This theory has been further extended by the conformational selection paradigm to include the ensemble-based description of proteins and protein–protein interactions.[4–7] The concept suggests that bound and bound-like conformations may coexist in solution within a large ensemble of unbound conformations. By shifting equilibrium in the unbound ensemble toward the bound-like conformations, binding forces select a bound conformation corresponding to the free energy minimum. Recent studies focused on the reconstruction of the native

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: NIH; Grant number: R01GM074255.

*Correspondence to: Anatoly M. Ruvinsky, Center for Bioinformatics, The University of Kansas, Lawrence, KS 66047. E-mail: ruvinsky@ku.edu

ensembles.[8–11] An ensemble of ubiquitin structures reflecting dynamics up to the microsecond time scale was refined against residual dipolar couplings (RDCs). All crystallographically determined bound conformations of ubiquitin were found within 0.8 Å root mean square deviation (RMSD) of the C_α atoms.[10] Another RDC-optimized ensemble of ubiquitin consistent with the microsecond time scale dynamics was created by Monte Carlo sampling of the "Backrub" motions.[11] Cold denaturation and protein encapsulation were combined with nuclear magnetic resonance (NMR) to probe the ensemble.[12] Single-molecule experiments corroborated the theory of multiple interconverting conformations and revealed their relation to the fluctuating catalytic reactivity.[13] Room-temperature X-ray crystallography was able to detect such conformations in proline isomeraze.[14] Best *et al.*[8] showed that an ensemble of highly homologous X-ray structures can also reproduce structural diversity in the native ensemble probed by NMR spectroscopy in solution. A protocol combining molecular dynamics (MD) simulations of an X-ray structure with information from the NMR relaxation experiments has been suggested for studying protein conformational ensembles in solution.[15] In general MD simulations have been instrumental in mapping the conformational space.[16–19] Alternative methods for generating conformational ensembles without solving explicit equations of motion have been actively developed (see Ref. 20 for a review). Large conformational ensembles are routinely used in protein structure prediction[21,22] and studies of allosteric interactions.[23,24]

Despite the significant progress achieved in generating protein ensembles, their energetic properties and relation to the unbound-to-bound conformational changes are not well understood.[25] How to generate a bound-like structure from the unbound one is one of the main problems in structure prediction of protein complexes. Although MD simulations showed that some of the interface side chains—"anchor residues"—sample bound-like conformations,[26,27] criteria for selecting such conformations from the MD snapshots are yet to be determined. Current docking protocols are much more successful when bound conformations are used, but become less reliable in a common case when only unbound structures are known.[28,29] To advance the docking protocols, the relation between the energy landscape and conformational changes upon binding should be unraveled. Recent large-scale studies of conformational changes upon binding focused on the relationship between single bound and single unbound conformations.[30–33] However, how well the change between two selected conformations characterizes transition between the unbound and bound states within conformational ensembles as well as the transformation of the (free) energy landscapes is still unclear.

In this study, we investigate structural similarity between the ensembles of bound and unbound conformations for six proteins and characterize corresponding segments of their energy landscapes mapped by these ensembles. Wherever available, protein conformations extracted from multimers are considered separately from the conformations extracted from dimers. Within each ensemble, energies of protein conformations are considered as the corresponding energy spectrum. We investigate the impact of the energy minimization in the Generalized Born (GB) model on the distance between the ensembles of bound and unbound conformations and the ensembles' spectral properties (the energy spacing, the spectrum gap between the lowest states, and the spectrum width), which are a proxy for the folding energy landscape. Knowledge of the energy spacing allows one to infer the smoothness of the energy landscape: the larger the spacing the rougher the landscape.[34] This property is central for the principle of minimal frustration and the related theory of protein folding on a funneled energy landscape.[35,36] The spectrum gap and the spectrum width of the folded states characterize thermodynamic stability of the native and near-native states.[37] The analogy between folding and binding energy landscapes was used to quantify the specificity of protein–protein binding by the ratio of the spectrum width to the roughness of the binding energy landscape.[38]

Our focus on the energy minimization in implicit solvent was motivated by a recent study[39] showing that ranking protein structures by minimized GB energies can distinguish the near-native structures from decoys better than ranking based on the energy minimization either in vacuum or explicit solvent. First, our study shows that although the shortly minimized GB energies of the bound and unbound ensembles often significantly overlap, the center of the unbound spectrum tends to have a higher energy than the centers of the bound spectrum of the dimeric and multimeric states. Moreover, the larger the overlap the smaller the RMSD between bound and unbound conformational ensembles. Second, the existence of the structurally different equipotential states in both ensembles suggests that unbound states have larger entropy than the bound states. The entropy-driven modeling of the unbound-to-bound conformational changes suggests a novel direction in advancing protein–protein docking algorithms, which, in fact, commonly neglect entropy effects. Third, the results show that the bound conformations of the RNase A interface pre-exist in the unbound ensemble, indicating conformational selection as the binding mechanism. Pancreatic trypsin inhibitor (PTI), ubiquitin, and lysozyme C also have high similarity between bound and unbound interfaces as well as small deviations that can be attributed to flash cooling[40,41] or variations in the crystallization conditions. Fourth, the exhaustively long minimization (LM) by the Adopted Basis Newton–Raphson algorithm results in small mostly intrarotamer adjustments that drastically reduce the distance between the centers of the bound and unbound spectra as well as the spectra extent. It condenses unbound and bound energy levels into a thin layer at the bottom of the energy landscape. At the same time, the whole spectrum from the shortly minimized states to the bottom of the folding funnel can cover up to 40.3% of the lowest energy, indicating that the folded states may significantly differ in energy. The average energy spacing at the bottom of the energy landscape varies between 0.8–4.6 and 3.5–10.5 kcal/mol for the unbound and bound states correspondingly. The energy gap between the two lowest states varies between 0.9 and 12.1 kcal/mol. Finally, the results show that protein vibrations itself can stimulate the interstate transitions, thus supporting the conformational selection theory. We suggest an approach for estimating the number of normal modes involved in conformational transition and show that, on average, 20 low-energy normal modes are needed to describe transition between two neighboring energy states. At the same time, transitions between the two lowest states may involve an order of magnitude larger number of the modes.

Results and Discussion

To compile ensembles of bound and unbound states, we first selected a subset of 165 protein–protein complexes from the non-redundant DOCKGROUND set 3.0 of 233 protein–protein complexes.[42] This subset characterized by small unbound-to-bound conformational changes (all-atom RMSD ≤ 2 Å) represents a majority of the protein complexes (71%). Then, proteins that are monomers in the unbound state were selected from the subset and screened against the Protein Data Bank (PDB) for highly homologous proteins. Keeping PDB structures with small number of missing residues or mutations (see Methods section) resulted in a set of only six proteins (ovomucoid, PTI, ubiquitin, RNase A, CheY, and lysozyme C) that have more than five PDB structures in its unbound and bound/dimeric ensembles (Table I and Supporting Information Table SI). We found that only 1.8% of the proteins in the DOCKGROUND subset have corresponding multistate conformational ensembles. To exclude energy fluctuations related to structural disorder and mutations, a program Profix was used to build the disordered residues and missing atoms and to reverse all point mutations.

Table I. *Ensembles of Bound and Unbound Proteins*

Protein	Unbound structures	Bound structures[a]	
		Dimers	Multimers
RNase A	49	32	3
Pancreatic trypsin inhibitor (PTI)	27	18	29
Chemotaxis protein CheY	73	6	
Ubiquitin	394	8	
Ovomucoid	124	6	
Lysozyme C	45	19	24

[a] Considered separately from the other subunit(s) in the dimers/multimers.

Table I shows that, as expected, all unbound ensembles have more conformations than the corresponding dimeric ensembles. On the other hand, only three proteins were found to form multimers. Among them, lysozyme C and PTI have more conformations in their multimeric ensembles than that in the dimeric ones. The ensemble population and its heterogeneity are the key factors for the following analysis. Best *et al.*[8] showed that from 20 to 40 structures were required to reproduce side-chain heterogeneity in the unbound protein. All our unbound ensembles satisfy this requirement (Table I). Variations in crystallization conditions such as pH or ionic strength provide conformational heterogeneity in the ensembles and enable characterization of different subspaces on the landscape. An extra heterogeneity comes with the combination of X-ray and NMR structures, which represent protein dynamics at different time scales. X-ray structures carry on effects of crystal packing interactions and the flash cooling technique,[41] which may lock and optimize one of the conformations available in solution.

The GB energy of each of the selected structures was subject to a two-stage minimization by the GBMV2 module implemented in the Chemistry at HARvard Macromolecular Mechanics (CHARMM) package. The first stage consisted of 50 steps of the steepest descent minimization [short minimization (SM)] followed by 10^4 steps of the Adopted Basis Newton–Raphson minimization (LM). The GB energy included polar and nonpolar solvation energies, and an internal energy described by the CHARMM22 molecular force fields. The GBMV2 as well as other GB models enable fast calculations of the Poisson–Boltzmann (PB) electrostatic solvation energy by

$$G_{\mathrm{elec}} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\mathrm{p}}} - \frac{1}{\varepsilon_{\mathrm{s}}}\right) \sum_{ij} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i^{\mathrm{GB}} R_j^{\mathrm{GB}} \exp\left(-r_{ij}^2 / K R_i^{\mathrm{GB}} R_j^{\mathrm{GB}}\right)}}, \qquad (1)$$

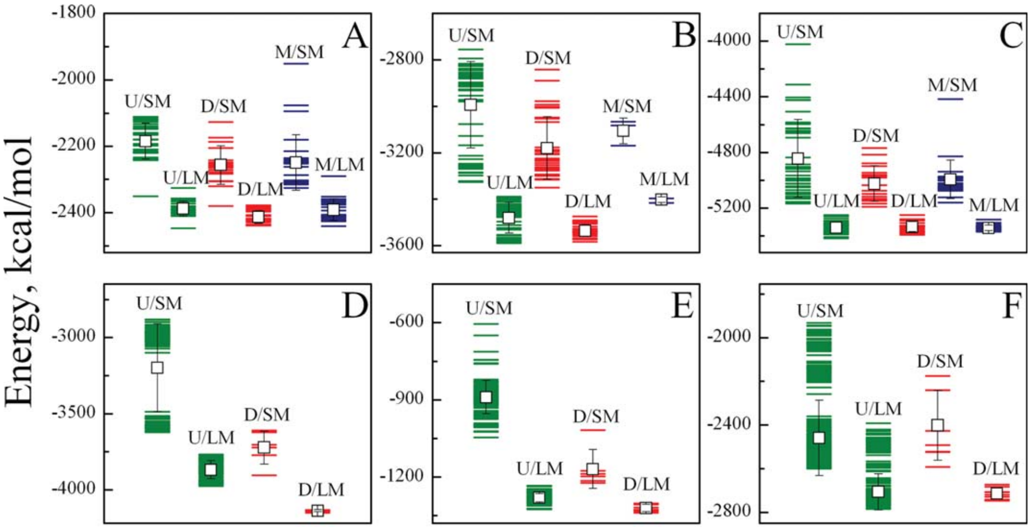

Figure 1. Energy spectrum of the unbound and bound proteins. (A) PTI, (B) RNase A, (C) lysozyme C, (D) CheY, (E) ovomucoid, and (F) ubiquitin. SM and LM indicate the spectrum after the short minimization (50 steps of the steepest descent minimization) and long minimization (SM followed by 10^4 steps of the Adopted Basis Newton–Raphson minimization). U, D, and M are unbound proteins (green), proteins crystallized as dimers (red), and multimers (blue). Open squares with error bars show the average energy (the band's center) and the standard deviation.

where r_{ij} is a distance between atoms i and j, q_i and q_j are the partial charges, ε_p and ε_s are the solute and solvent dielectric constants, R_i^{GB} is an effective Born radius of the atom i, and K is a Still factor.[43] The GBMV2 applied here was shown to be one of the best GB approaches among 23 CHARMM/AMBER/OPLS-based GB models.[44] It showed a consistently small error of less than 1% for relative solvation energies between different conformations of the same protein, making this model well suitable for the analysis of energies in protein ensembles. In addition, this model reproduced PB solvation energies within 1% error on average for large sets of proteins[44,45] and showed 0.9992 correlation between PB and GB energies.[45] The nonpolar solvation energy including the formation of a protein cavity in the solvent and protein–solvent interactions were calculated by $G_{np} = \sum_i (\sigma_i A_i + E_i)$,[46] where A_i is a solvent-accessible surface area of the atom i, σ_i is the atomic solvation parameter for the atom i, E_i is a reference solvation energy of the atom i.

Bound and unbound energy bands

Figure 1 shows minimized energy spectra of six proteins (see Methods section) represented by the ensembles of their conformations determined by X-ray crystallography and NMR (Table I and Supporting Information Table SI). SM and LM were applied to characterize the topography of the energy landscape in the GB model (see Methods section). The SM removed atom clashes inside protein structures and relaxed surface side chains involved in crystal contacts. Both energy minimizations caused small intra-rotamer readjustments of the exposed side chains resulting in a typical RMSD ≤ 0.7 Å between the minimized and the nonminimized structures, which did not substantially change neither the sizes of the bound and unbound ensembles nor the RMSD between them (Table II). Nevertheless, these changes were enough to significantly condense both spectra of the unbound and bound proteins. Figure 1 shows that the span of the spectra and the spacing between energy minima after LM are significantly smaller than that after SM. The ratio between the overall energy span (including the SM and LM bands) in the unbound ensemble and the lowest energy in the protein spectrum is 40.3% for ovomucoid, 13.7% for PTI, 26.9% for ubiquitin, 21.5% for RNase A, 22.8% for CheY, and 24.2% for lysozyme C. Excluding the lowest ratio for PTI as an outlier, the ratio for other five proteins decreases 1.7 times, with a 2.5 times increase in the number of atoms from ovomucoid to lysozyme C. The ratio decrease is expected because the lowest energy is a function of the total number of protein residues, whereas the energy extent relates mainly to the surface residues that are able to change their conformations in solution. The outlying ratio for PTI may result from the insufficient size of its unbound ensemble, which is the smallest among the proteins in our set (Table I). Ovomucoid has 51 residues, compared to 56 residues of PTI, but its unbound ensemble is 4.6 times larger than the ensemble of PTI.

The ratio between the energy span of the LM band and the lowest energy is 7.0% for ovomucoid, 4.9% for PTI, 14.0% for ubiquitin, 5.5% for RNase A, 5.0% for CheY, and 3.1% for lysozyme C (Fig. 2).

Table II. *Bound-to-Bound and Bound-to-Unbound RMSDs*

	RMSD between bound structures (Å)						RMSD between bound and unbound structures (Å)					
	All atoms			Interface			All atoms			Interface		
Protein	I[a]	SM[b]	LM[c]	I	SM	LM	I	SM	LM	I	SM	LM
RNase A												
min	0.3	0.3	0.3	0.1	0.1	0.2	0.7	0.8	0.7	0.3	0.3	0.4
max[d]	7.1	7.2	7.2	11.6	11.8	11.7	7.3	7.4	7.4	11.7	11.9	11.8
PTI												
min	0.1	0.1	0.3	0.0	0.0	0.2	1.0	1.0	0.9	1.1	1.1	1.0
max	2.3	2.4	2.3	2.7	2.7	2.7	2.5	2.5	2.4	2.7	2.7	2.8
CheY												
min	0.6	0.6	0.6	0.5	0.5	0.6	1.9	1.9	1.8	1.6	1.6	1.3
max	1.2	1.1	1.2	1.7	1.7	1.8	2.9	2.9	3.0	3.1	3.1	2.9
Ubiquitin												
min	0.3	0.1	0.4	0.4	0.1	0.5	1.0	1.0	0.9	1.1	1.1	1.0
max	1.6	1.7	1.5	2.4	2.5	2.3	4.1	4.1	4.0	5.9	5.9	6.1
Ovomucoid												
min	0.6	0.6	0.7	0.5	0.5	0.5	1.2	1.1	0.9	1.3	1.2	1.1
max	1.2	1.2	1.2	1.6	1.6	1.8	2.2	2.2	2.1	3.5	3.4	3.1
Lysozyme C												
min	0.2	0.2	0.4	0.1	0.1	0.2	0.8	0.8	0.8	0.7	0.7	0.7
max	1.8	1.8	1.9	3.0	3.2	3.1	1.9	1.9	1.9	3.1	3.1	3.3

[a] Initial (not minimized) protein structures.

[b] Structures subjected to 50 steps of the steepest descent minimization (short minimization).

[c] Structures subjected to 50 steps of the steepest descent minimization, followed by 10^4 steps of the Adopted Basis Newton–Raphson minimization (long minimization).

[d] Minimum and maximum RMSDs.

Thus, LM reduces the width of the energy bands, condensing protein states into a thin layer at the bottom of the energy landscape. The energy distance (the ruggedness) between the centers of the SM and LM energy bands was calculated as an arithmetic mean of energies in a protein ensemble after SM and LM (see Methods section). Interestingly, the energies of the unbound ensembles decrease more upon minimization than the energies of the bound ensembles (Supporting Information Fig. S1) despite the fact that both ensembles have equipotential energy levels (Fig. 1). On average, the unbound proteins lose 0.6 kcal/mol per heavy atom or 4.6 kcal/mol per residue. The bound structures from dimers and multimers lose less: 0.4 and 0.3 kcal/mol per heavy atom or 3.1 and 2.5 kcal/mol per residue accordingly.

The centers of the unbound energies after SM were higher than the centers of the bound energies for all proteins, with the exception of ubiquitin. This suggests that the unbound-to-bound conformational changes guided by intermolecular interactions often follow a path that decreases the internal energy of the binding proteins. Such mechanism additionally increases binding affinity

$$\begin{aligned}\Delta G &= (E_c - TS_c) - (E_A - TS_A) - (E_B - TS_B) \\ &= V + (E'_A - E_A) + (E'_B - E_B) - TS_c + TS_A + TS_B, \qquad (2)\end{aligned}$$

where $E_c = V + E'_A + E'_B$ is the enthalpy of the complex, V is the interaction energy between proteins A and B in the complex, E'_A, E_A and E'_B, E_B are internal energies of the bound and unbound conformations of proteins A and B, S_c, S_A, and S_B are the entropies of the complex and unbound proteins A and B. The energy decrease can be achieved by improving the interface packing upon binding. The prevalence of the disorder-to-order interface transitions over the reverse transitions corroborates this hypothesis.[30] Equation (2) shows that binding forces may favor low-entropy (less flexible) conformations. Effective binding to a more-flexible high-entropy conformation

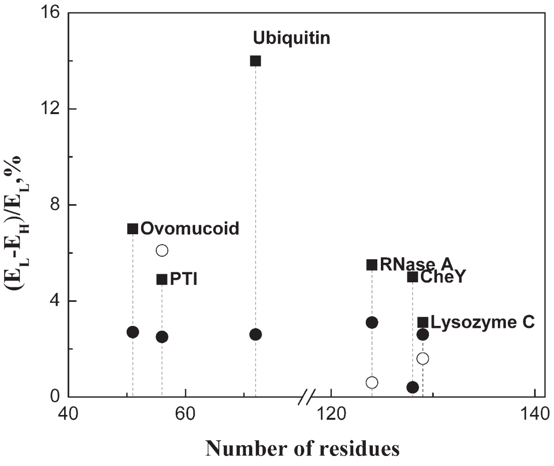

Figure 2. The ratio of the ensemble width after long minimization to the lowest energy in the joint ensemble. The data are shown for unbound (■) and bound ensembles, extracted from dimers (●) and multimers (○).

would require stronger intermolecular forces to overcome a large entropy loss upon binding. The generality of this observation for molecular association processes was discussed in Ref. 47, where a correlation was found between the increased affinities for ligand binding and the decreased structural flexibility in a series of antibody molecules. The two-sample *t*-test showed statistical significance of the difference between the bound and unbound SM-bands' centers at the 5% level for all proteins, except ubiquitin (Supporting Information Table SII). LM resulted in a significant decrease in the distance between the centers and a cancellation of the statistical significance of the difference between the centers of the unbound energies and bound energies for dimeric states of lysozyme C and ovomucoid, and multimeric states of PTI and lysozyme C. Comparison of the centers of the bound spectra of the structures extracted from dimers and multimers (Supporting Information Table SIII) shows that LM resulted in statistically significant difference between the corresponding centers for RNase and PTI. For these proteins, the center of the dimeric bound band of the dimer states is lower than the center of the multimeric bound band.

An overlap between the unbound and bound energies in the SM and LM bands, shown in Figure 1, suggests that conformational selection of a bound conformation may be dictated by entropy. Indeed, let us for simplicity consider binding of a protein A to a protein B that exists in one of two conformations only. The binding free energy of $\mathrm{A}+\mathrm{B}_i \iff \mathrm{AB}$ can be written as $\Delta G_{\mathrm{i}}=(E_{\mathrm{c}}-TS_{\mathrm{c}})-(E_{\mathrm{A}}-TS_{\mathrm{A}})-(E_{\mathrm{B_i}}-TS_{\mathrm{B_i}})$, where i equal to 1 or 2 is a number of the protein B conformation, $E_{\mathrm{B_i}}$ and $S_{\mathrm{B_i}}$ are the enthalpy and the entropy of protein B in the ith state. Therefore, if both conformations of protein B have equal energies $E_{\mathrm{B_1}} \approx E_{\mathrm{B_2}}$, then the choice of the preferred unbound conformation of protein B is dictated solely by the entropy contribution to the binding free energy. In this case, ΔG reaches its minimum for the unbound conformation with the lowest entropy. This observation can be related to the energy landscape properties by $S_{\mathrm{B_i}}=R\ln N_i$, where N_i is the number of microstates associated with the conformation i, and R is the gas constant. Since a less-populated state has a lower entropy, then the less-populated conformation is the most effective binder, in agreement with the concept of conformational selection.[48] This further suggests that conformations selected by binding forces in solution can reside in narrow (low entropy) basins formed by the intramolecular interactions. We intend to verify this hypothesis in our future study. Sampling of the low-entropy conformations may improve performance of protein–protein docking algorithms, which commonly neglect entropy effects.

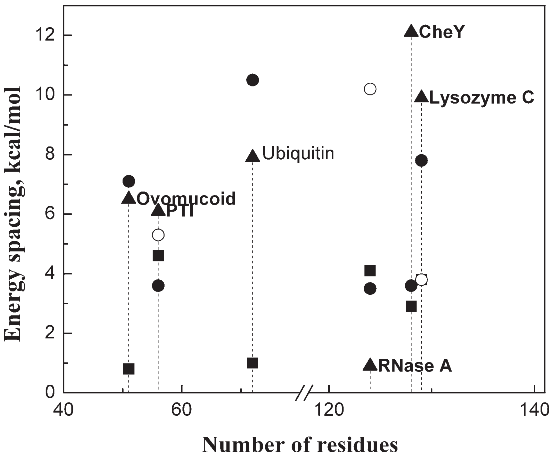

Figure 3. Energy spacing at the bottom of the folding funnel. The figure shows the energy gap between the two lowest energy minima (▲) in the joint ensemble of the bound and unbound states and the average distance between energy minima after the long minimization in the unbound (■) and bound ensembles, extracted from dimers (●) and multimers (○).

Energy spacing, fluctuations, and conformational changes

Figure 3 summarizes calculations of the energy spacing in the ensembles. The spacing in the unbound ensembles averaged over the protein set is 2.9 kcal/mol, which is two times less than the average spacing in the bound ensembles. On average, the dimeric and multimeric states are separated by 6.0 and 6.4 kcal/mol accordingly. Larger variations of the energy spacing in bound ensembles may be a result of a smaller size of these ensembles (Table I). Considering the proteins separately, one can see that the energy spacing between the unbound states varies between 0.8 and 4.6 kcal/mol, which includes the Hyeon and Thirumalai's[49] estimate of 0–3 kcal/mol for the energy landscape roughness or the barrier and 3.2–3.5 kcal/mol barrier measured by single-molecule dynamic force spectroscopy for a complex of GTPase Ran and the nuclear transport receptor importin-b.[50] Note that the average spacing of 2.9 kcal/mol is approximately equal to the maximum barrier found in the Hyeon and Thirumalai[49] study but less then the maximum barrier found by Nevo *et al.*[50] The energy spacing between the bound states is larger and falls in the intervals of 3.5–10.5 and 3.8–10.2 kcal/mol for the dimeric and multimeric states correspondingly. The energy gap between the two lowest energy minima in the joint ensemble of the bound and unbound states is 0.9 kcal/mol for RNase A, 6.1 kcal/mol for PTI, 12.1 kcal/mol for CheY, 7.9 kcal/mol for ubiquitin, 6.5 kcal/mol for ovomucoid, and 9.9 kcal/mol for lysozyme C. The smallest energy gap was found for RNase A, which has the smallest interface and all-atom RMSD between bound and unbound states (Table II). On the other hand, the largest gap in the

CheY spectrum corresponds to the largest RMSD between its bound and unbound ensembles.

To explore whether an interstate transition in principle can result from protein vibrations, one can estimate a number of the normal modes that needs to be involved in the protein energy fluctuation equal to the interstate barrier. The molecular energy fluctuation in a canonical ensemble can be calculated as $\delta E=\sqrt{\langle \Delta E^2\rangle}=RT\sqrt{c_V}$, where $c_V=c_T+c_R+c_{\mathrm{o}}(T)$ is a specific heat capacity per molecule at constant volume, $c_T=c_R=3/2$ are the contributions of three translational and three rotational degrees of freedom, and $c_{\mathrm{o}}(T)=\sum_i c_{\mathrm{o}}^i(T)$ is the contribution of the $3N_{\mathrm{at}}-6$ internal vibrations in a protein with N_{at} atoms. For low-energy modes $\hbar\omega_i \ll kT$ and therefore $c_o^i(T) \approx 1$ and $c_{\mathrm{o}}(T) \approx n+\sum_{i=n+1} c_{\mathrm{o}}^i(T)$, where n is a number of the low-energy modes. For high-energy modes $\hbar\omega_i \gg kT$ and $c_{\mathrm{o}}^i(T)$ exponentially goes to zero. One can find the low-bound estimate for the number of low-energy modes needed for a transition between any two states separated by the δE barrier: $n=(\delta E/RT)^2-3$. It gives 23 and 32 normal modes for the standard ambient temperature of 298.15 K and the barrier of 3.0 and 3.5 kcal/mol correspondingly. Since functional modes are often found among the lowest 20–30 modes,[51–55] we can suggest that the protein vibrations indeed can excite the interstate transitions related to protein function. Thus, an external stimulus (e.g., a ligand or a partner protein) may not be needed for changing protein conformation, which supports the conformational selection paradigm. Interestingly, many more modes are needed for an energy fluctuation covering the distance between the two lowest states in all the proteins in our set, except RNase A, which shows the smallest distance between bound and unbound ensembles. This is supported by a study of conformational changes in myosin, calmodulin, NtrC, and hemoglobin,[56] which showed that the first 20 modes contribute ≤50% of the conformational changes in these molecules. The first 30 modes of the $[\mathrm{AChE_T}]_4$–ColQ complex account for 75% of the conformational change in the tetramer.[54] For a typical 1000 atoms protein having 2994 normal modes, a fluctuation of 12.1 kcal/mol is achieved when at least 13% of the normal modes get involved. It was shown that the protein energy fluctuation can increase up to 38 kcal/mol,[57] which is more than enough for a transition over the largest gap/barrier considered in this study.

Distance between unbound and bound conformational ensembles and binding mechanisms

The size of the ensemble is controlled by ambient parameters (temperature, pH, salt concentration, etc.) and dependent on protein sequence composition (Table II). Protein–protein interactions can either select a group of bound-like conformations from the unbound ensemble or transform the whole ensemble into a new group of bound-like conformations. To find out which mechanism takes place, all-atom and interface RMSDs were calculated between all bound and unbound structures (Table II). Interface residues were defined as those losing >1 Å^2 of their surface upon binding. The smallest RMSD between the bound and unbound ensembles was found for RNase A and lysozyme C. RNase A shows the all-atom/interface RMSD of 0.7/0.3 Å. The ensembles of lysozyme C are separated by 0.8/0.7 Å of the all-atom/interface RMSD. Interestingly, the unbound ensemble of lysozyme C encompasses X-ray structures only. The unbound ensemble of RNase including both X-ray and NMR structures has the second largest share $(100\%\times N_{\text{X-ray}}/(N_{\text{X-ray}}+N_{\mathrm{NMR}})=35\%)$ of the X-ray structures ($N_{\text{X-ray}}$) in its unbound ensemble among the proteins in our set. The low bound of the unbound-to-bound all-atom and interface RMSDs varies within 0.7–1.9 and 0.3–1.6 Å. The largest RMSD between the bound and unbound ensembles of CheY corresponds to the smallest overlap between their bound and unbound SM spectra (Fig. 1), which disappears after LM. Thus, the majority of the SM and all LM bound conformations of CheY have lower energies than the unbound ones. It is likely that the entropy discussed above makes these lower energy states unfavorable for the unbound ensemble.

In addition to the RMSD analysis, we calculated the share of the unbound residues within 1 Å of their bound conformations for all pairs of the bound and unbound structures (Figs. 4 and 5). Each unbound structure was consequently aligned to each bound structure of the same protein by TM-align,[58] producing a number of the alignments equal to the product of populations of the bound and unbound ensemble. Then for each alignment the RMSD between the side chains in the bound and unbound structures was calculated for each protein residue. The share of the unbound residues within 1 Å of their bound conformation was calculated as $\nu_i=100\%\times N^i_{\mathrm{RMSD}\le 1}/N$, where N is the protein sequence length, and $N^i_{\mathrm{RMSD}\le 1}$ is a number of protein residues that deviate by less than 1 Å in the alignment i. This metric also showed the lowest similarity between the CheY ensembles at 0.39 level (39% of all the residues, Fig. 5). The ensembles of RNase A, lysozyme C, and PTI had the highest similarity at 0.9, 0.89, and 0.86 levels correspondingly. Comparison of the bound and unbound interfaces revealed 13 unbound structures of RNase A with *all* interface residues within 1 Å from the bound conformations (Fig. 4). Note that 2 Å is a typical size of a rotamer.[32] Thus, dimerization of RNase A with another protein can be completely described by the conformational selection mechanism.[4–7] Contrary to that, forming a multimer

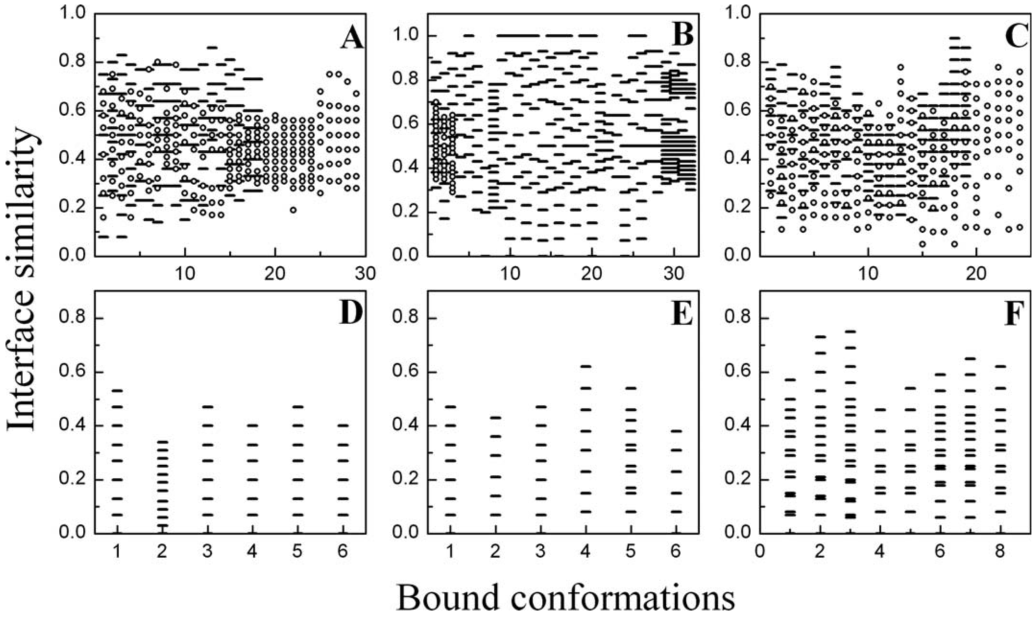

Figure 4. Similarity of bound and unbound interface conformations. (A) PTI, (B) RNase A, (C) lysozyme C, (D) CheY, (E) ovomucoid, and (F) ubiquitin. The similarity is calculated for each pair of bound and unbound structures as the share of the unbound interface residues within 1 Å RMSD from the bound interface residues. Bars and circles show the interface similarity between dimeric/multimeric and unbound conformations accordingly. The horizontal axis shows conformation in the bound ensembles.

involving RNase A invokes induced fit to expose its C-terminal (Supporting Information Fig. S2), which forms an interface β-strand that swaps with the N-terminal helix in the RNase trimer. None of the unbound structures has the exposed C-terminal. This further suggests that some proteins may employ various binding mechanisms, from the induced fit, to "lock-and-key" and conformational selection, and their combination, depending on the binding partner.[5,6]

Interestingly, flash cooling used to determine approximately 90% of macromolecular structures[40] results in a 0.2–0.8 Å backbone RMSD between the structures determined at cryogenic and room temperatures.[41] It can also change the conformational distribution of up to one-third of the protein side chains.[59] Taking this into account, we can assume that crystallographic conditions may distort the structure by 1 Å RMSD of *all* atoms. The low bound

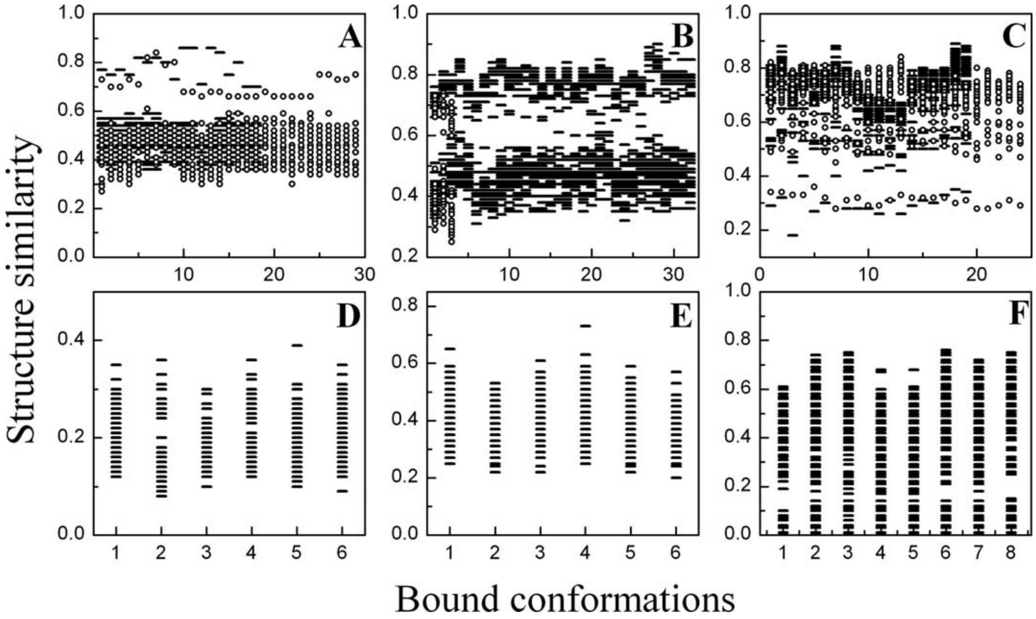

Figure 5. Similarity of bound and unbound structures. (A) PTI, (B) RNase A, (C) lysozyme C, (D) CheY, (E) ovomucoid, and (F) ubiquitin. The similarity is calculated for each pair of bound and unbound structures as the share of the unbound residues within 1 Å RMSD from the bound ones. Bars and circles show the similarity between dimeric/multimeric and unbound conformations accordingly. The horizontal axis shows conformation in the bound ensembles.

of the *all-atom* RMSD between bound and unbound ensembles (Table II) suggests that, in addition to RNase A, the conformational selection likely guides binding processes of PTI, ubiquitin (also see Ref. 10), and lysozyme C.

Materials and Methods

Generation of the protein set

To compile a set of proteins with multiple bound and unbound conformations, a subset of protein complexes with small changes in the backbone upon binding (all-atom RMSD ≤ 2 Å) was selected from the nonredundant DOCKGROUND set 3.0.[42] The subset covers 71% of the DOCKGROUND set of 233 complexes. The subset was narrowed down to proteins that are monomers in the unbound state of the biological assembly. Their sequences were used to identify homologous proteins in PDB (sequence identity $> 98\%$ by BLAST[60]). The unbound protein structures with small ligands were excluded. All PDB entries found for each query protein were put into three ensembles: unbound monomers, dimers, and multimers. Only proteins with more than five unbound and bound structures were retained. Selected structures were analyzed for disordered residues and mutations. If some of the structures had a disordered terminal, it was deleted in all members of the ensemble. All fragments with ≤ 3 disordered residues at the interface and ≤ 5 at the non-interface were reconstructed by a program Profix from the Jackal package (http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software). Structures with disordered fragments longer than five residues were discarded. Point mutations were reversed by Profix. The resulting set consisted of six proteins (Table I and Supporting Information Table SI) with multiple X-ray and NMR-derived bound and unbound conformational states and 100% sequence identity between the states.

Minimization protocol

The MMTSB tool set[61] and the GB method that calculates Born radii by analytic volume integration (CHARMM: GBMV method 2) were used to minimize solvation free energy of the proteins.[44,62] The method was parameterized to accurately reproduce electrostatic solvation energies from standard Poisson theory. A nonpolar contribution to the solvation free energy was calculated by the ASP model considering the exposed surface area.[46] Each protein was subjected to 50 steps of the steepest descent minimization (SM) followed by 10^4 steps of the Adopted Basis Newton–Raphson minimization (LM). The CHARMM22 force field was used. The dielectric constant was set to 1 for protein and 80 for solvent. Each bound protein was minimized within its complex to keep interface unchanged. The analysis showed that protein energy changed $\leq 1.5\%$ between 500 and 10^4 steps of LM. The average RMSD between all heavy atoms of the initial and minimized structures after SM was 0.1 Å. LM produced the average all-atom RMSD at 0.7 Å between the initial and the minimized structures. As can be seen from Table II, LM did not change substantially the RMSD-based size of the conformational ensembles and the distance between the unbound and bound ensembles.

Characterization of the energy spectrum

The ratio of the spectrum width in the ensemble of the SM and LM unbound structures to the lowest energy was calculated as the absolute value of $100\%\cdot(\Delta E_1+\Delta E_2)/E_{\mathrm{L}}$, where E_{L} is the lowest protein energy in the joint ensemble of bound and unbound structures, and $\Delta E_{1,2}$ are the energy span in the unbound ensemble after the SM and LM correspondingly. If the energy spans overlap in the SM and LM ensembles, then the ratio was calculated as $100\%\cdot(E_{\min}-E_{\max})/E_{\mathrm{L}}$, where $E_{\min,\max}$ are the lowest and the highest energies in the unbound spectrum.

The ruggedness of the energy landscape was calculated as $\bar{E}_{\mathrm{SM}}-\bar{E}_{\mathrm{LM}}$, where $\bar{E}_{\mathrm{SM}},\bar{E}_{\mathrm{LM}}$ are the average energies in a protein ensemble after SM and LM accordingly. The energy spacing was calculated as the average distance between energy levels: $\sum_{i=1}^{N}\left(E_i^{\mathrm{LM}}-E_{i+1}^{\mathrm{LM}}\right)/(N-1)=\left(E_H^{\mathrm{LM}}-E_L^{\mathrm{LM}}\right)/(N-1)$, where $\left\{E_i^{\mathrm{LM}}\right\}$ is an ordered set of the LM energies, N is the number of structures in the LM ensemble.

Acknowledgments

We would like to thank Prof. Ken A. Dill, Stony Brook University, for reading the manuscript and providing valuable suggestions.

References

1. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. Science 254:1598–1603.
2. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4:10–19.
3. Wolynes PG, Onuchic JN, Thirumalai D (1995) Navigating the folding routes. Science 267:1619–1620.
4. Tsai CJ, Kumar S, Ma B, Nussinov R (1999) Folding funnels, binding funnels, and protein function. Protein Sci 8:1181–1190.
5. Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 5:789–796.
6. Csermely P, Palotai R, Nussinov R (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci 35:539–546.
7. Tsai CJ, Ma B, Nussinov R (1999) Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci USA 96:9970–9972.
8. Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M (2006) Relation between native ensembles and

experimental structures of proteins. Proc Natl Acad Sci USA 103:10901–10906.

9. Wlodarski T, Zagrovic B (2009) Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. Proc Natl Acad Sci USA 106:19346–19351.
10. Lange OF, Lakomek N-A, Farès C, Schröder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C, de Groot BL (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320:1471–1475.
11. Friedland GD, Lakomek NA, Griesinger C, Meiler J, Kortemme T (2009) A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. PLoS Comp Biol 5:e1000393.
12. Babu CR, Hilser VJ, Wand AJ (2004) Direct access to the cooperative substructure of proteins and the protein ensemble via cold denaturation. Nat Struct Mol Biol 11:352–357.
13. Yang H, Luo G, Karnchanaphanurach P, Louie T-M, Rech I, Cova S, Xun L, Xie XS (2003) Protein conformational dynamics probed by single-molecule electron transfer. Science 302:262–266.
14. Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T (2009) Hidden alternative structures of proline isomerase essential for catalysis. Nature 462:669–673.
15. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433:128–132.
16. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652.
17. Meinhold L, Smith JC, Kitao A, Zewail AH (2007) Picosecond fluctuating protein energy landscape mapped by pressure temperature molecular dynamics simulation. Proc Natl Acad Sci USA 104:17261–17265.
18. Kohn JE, Afonine PV, Ruscio JZ, Adams PD, Head-Gordon T (2010) Evidence of functional protein dynamics from X-ray crystallographic ensembles. PLoS Comp Biol 6:e1000911.
19. Andrec M, Felts AK, Gallicchio E, Levy RM (2005) Protein folding pathways from replica exchange simulations and a kinetic network model. Proc Natl Acad Sci USA 102:6801–6806.
20. Jacobs DJ (2010) Ensemble-based methods for describing protein dynamics. Curr Opin Pharmacol 10:760–769.
21. Shortle D, Simons KT, Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 95:11158–11162.
22. Zhang Y, Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 102:1029–1034.
23. Kidd BA, Baker D, Thomas WE (2009) Computation of conformational coupling in allosteric proteins. PLoS Comput Biol 5:e1000484.
24. Hilser VJ, Dowdy D, Oas TG, Freire E (1998) The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble. Proc Natl Acad Sci USA 95:9903–9908.
25. Hegler JA, Weinkam P, Wolynes PG (2008) The spectrum of biomolecular states and motions. HFSP J 2:307–313.
26. Rajamani D, Thiel S, Vajda S, Camacho CJ (2004) Anchor residues in protein–protein interactions. Proc Natl Acad Sci USA 101:11287–11292.
27. Smith GR, Sternberg MJ, Bates PA (2005) The relationship between the flexibility of proteins and their conformational states on forming protein–protein complexes with an application to protein–protein docking. J Mol Biol 347:1077–1101.
28. Norel R, Petrey D, Wolfson HJ, Nussinov R (1999) Examination of shape complementarity in docking of unbound proteins. Proteins 36:307–317.
29. Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084.
30. Ruvinsky AM, Kirys T, Tuzikov AV, Vakser IA (2011) Side-chain conformational changes upon protein–protein association. J Mol Biol 408:356–365.
31. Guharoy M, Janin J, Robert CH (2010) Side-chain rotamer transitions at protein–protein interfaces. Proteins 78:3219–3225.
32. Kirys T, Ruvinsky AM, Tuzikov AV, Vakser IA (2012) Rotamer libraries and probabilities of transition between rotamers for the side chains in protein–protein binding. Proteins 80:2089–2098.
33. Stein A, Rueda M, Panjkovich A, Orozco M, Aloy P (2011) A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. Structure 19:881–889.
34. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND (1995) Toward an outline of the topography of a realistic protein-folding funnel. Proc Natl Acad Sci USA 92:3626–3630.
35. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 21:167–195.
36. Dill KA (1999) Polymer principles and protein folding. Protein Sci 8:1166–1180.
37. Shakhnovich E (2006) Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 106:1559–1588.
38. Wang J, Verkhivker GM (2003) Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. Phys Rev Lett 90:e1000911.
39. Chopra G, Summa CM, Levitt M (2008) Solvent dramatically affects protein structure refinement. Proc Natl Acad Sci USA 105:20239–20244.
40. Garman E (2003) 'Cool' crystals: macromolecular cryocrystallography and radiation damage. Curr Opi Struct Biol 13:545–551.
41. Juers DH, Matthews BW (2001) Reversible lattice repacking illustrates the temperature dependence of macromolecular interactions. J Mol Biol 311:851–862.
42. Gao Y, Douguet D, Tovchigrechko A, Vakser IA (2007) DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. Proteins 69:845–851.
43. Still WC, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112:6127–6129.
44. Feig M, Onufriev A, Lee MS, Im W, Case DA, Brooks CL 3rd (2004) Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 25:265–284.
45. Lee MS, Feig M, Salsbury FR Jr, Brooks CL III (2003) New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J Comput Chem 24:1348–1356.
46. Wesson L, Eisenberg D (1992) Atomic solvation parameters applied to molecular-dynamics of proteins in solution. Protein Sci 1:227–235.
47. Thorpe IF, Brooks CL III (2007) Molecular evolution of affinity and flexibility in the immune system. Proc Natl Acad Sci USA 104:8821–8826.

48. Boehr DD, Wright PE (2008) Biochemistry. How do proteins interact? Science 320:1429–1430.
49. Hyeon C, Thirumalai D (2003) Can energy landscape roughness of proteins and RNA be measured by using mechanical unfolding experiments? Proc Natl Acad Sci USA 100:10249–10253.
50. Nevo R, Brumfeld V, Kapon R, Hinterdorfer P, Reich Z (2005) Direct measurement of protein energy landscape roughness. EMBO Rep 6:482–486.
51. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. Protein Eng 14:1–6.
52. Dobbins SE, Lesk VI, Sternberg MJE (2008) Insights into protein flexibility: the relationship between normal modes and conformational change upon protein–protein docking. Proc Natl Acad Sci USA 105:10390–10395.
53. Cui Q, Li G, Ma J, Karplus M (2004) A normal mode analysis of structural plasticity in the biomolecular motor F(1)-ATPase. J Mol Biol 340:345–372.
54. Zhang D, McCammon JA (2005) The association of tetrameric acetylcholinesterase with ColQ tail: a block normal mode analysis. PLoS Comp Biol 1:e62.
55. Park J, Kahng B, Kamm RD, Hwang W (2006) Atomistic simulation approach to a continuum description of self-assembled beta-sheet filaments. Biophys J 90:2510–2524.
56. Petrone P, Pande VS (2006) Can conformational change be described by only a few normal modes? Biophys J 90:1583–1593.
57. Cooper A (1976) Thermodynamic fluctuations in protein molecules. Proc Natl Acad Sci USA 73:2740–2741.
58. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309.
59. Fraser JS, van den Bedemb H, Samelsona AJ, Langa PT, Holtonc JM, Echolsd N, Albera T (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proc Natl Acad Sci USA 108:16247–16252.
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410.
61. Feig M, Karanicolas J, Brooks CL III (2004) MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graphics Model 22:377–395.
62. Chocholousova J, Feig M (2006) Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. J Comput Chem 27:719–729.