Published in final edited form as:
J Phys Chem B. 2013 October 17; 117(41): . doi:10.1021/jp4050594.

Specific and Non-Specific Protein Association in Solution: Computation of Solvent Effects and Prediction of First-Encounter Modes for Efficient Configurational Bias Monte Carlo Simulations

Antonio Cardone[1,2], **Harish Pant**[3], and **Sergio A. Hassan**[4,*]

[1]Institute for Advanced Computer Science, University of Maryland, College Park, MD 20742

[2]SSD, National Institute of Standards and Technology, Gaithersburg, MD 20899

[3]Laboratory of Neurochemistry, NINDS

[4]Center for Molecular Modeling, DCB/CIT, National Institutes of Health, Bethesda, MD 20892

Abstract

Weak and ultra-weak protein-protein association play a role in molecular recognition, and can drive spontaneous self-assembly and aggregation. Such interactions are difficult to detect experimentally, and are a challenge to the force field and sampling technique. A method is proposed to identify low-population protein-protein binding modes in aqueous solution. The method is designed to identify preferential first-encounter complexes from which the final complex(es) at equilibrium evolves. A continuum model is used to represent the effects of the solvent, which accounts for short- and long-range effects of water exclusion and for liquid-structure forces at protein/liquid interfaces. These effects control the behavior of proteins in close proximity and are optimized based on binding enthalpy data and simulations. An algorithm is described to construct a biasing function for self-adaptive configurational-bias Monte Carlo of a set of interacting proteins. The function allows mixing large and local changes in the spatial distribution of proteins, thereby enhancing sampling of relevant microstates. The method is applied to three binary systems. Generalization to multiprotein complexes is discussed.

Keywords

protein-protein association; weak and ultra-weak interactions; macromolecular interfaces; aqueous interfaces; long-range solvent effects

I. Introduction

Cellular signal transduction involves networks of protein-protein interactions that transmit information.[1,2] Many of these proteins interact with more than one partner and can form stable multiprotein hetero-complexes.[2,3] A number of pathologies have been linked to disruptions of the delicate balance of forces between proteins, most commonly as a result of mutations4 or partial misfolding.[5] Understanding the physicochemical basis of macromolecular association in solution is then a requisite to understand many biological processes in the cell, from subcellular organization[3] to physiological function and disease.[6,7] To elucidate the origin of specificity and affinity structural information is often combined

*Tel: (301) 402-1382; hassan@mail.nih.gov.

with microcalorimetric and kinetic data,[8] but microscopic insight is often limited. Moreover, recent advances in paramagnetic relaxation enhancement techniques have revealed the existence of transient, ultra-weak protein self-associations that are difficult to detect with conventional biophysical methods.[9,10] Data suggest that proteins can interact at multiple sites, forming an ensemble of binding modes with very low populations.[11] These transient complexes can play a role in protein recognition, and may drive spontaneous self-assembly of higher-order architectures.[11] These studies have shown that ultra-weak association is controlled mainly by electrostatics, although hydrophobic interactions also play a role.[11] Crowded environments[12] could strengthen weak electrostatic interactions, which may explain the relatively high aggregation state of soluble proteins in living cells.[13]

The study of macromolecular complexation requires not only prediction of highly-specific binding modes, a common goal in computational biology,[14,15] but also calculation of association/dissociation rates, binding enthalpies and entropies, and detection and characterization of weak and ultra-weak association. These are major challenges for the force field, as it must describe the physics of a variety of aqueous environments and thermodynamic conditions, and the unique properties of aqueous interfaces. The protein environment is determined by several factors, including the amount of water excluded by neighboring proteins, complexes, and assemblies. The incomplete and anisotropic hydration created by these structures affect the magnitude and direction of forces induced by water.[16] The protein environment is also characterized by the properties of water close to the protein surface.[17–22] Aqueous interfaces are involved in many effects elicited by ions and cosolutes, including protein denaturation, stabilization, aggregation, and dissociation.[23–25] Aqueous interfaces display non-bulk behavior that can propagate a few hydration layers into the bulk. For example, neutron scattering and X-ray diffraction data suggest that simple ions can affect the water structure beyond their first hydration shells,[26] whereas osmotic stress experiments show that membranes and nucleic acid arrays affect the water behavior up to a few nanometers from their surfaces.27,28 Deeper interfaces have been reported in colloidal systems.[29–31] Transferring these findings to the cytosol is problematic because experiments are difficult to design and interpret, often leading to conflicting conclusions.[13,32] For example, NMR data suggest that the dynamics of cell water do not differ much from the dynamics of bulk water,[33] implying that only the first hydration shells are affected. However, neutron scattering and X-ray data indicate a larger proportion of non-bulk water,[34,35] suggesting deeper interfacial regions.

A continuum solvent model that incorporates some of these effects has been described,[16] and is reviewed in Section II. The model accounts for the effects of liquid-structure forces at aqueous interfaces, and for short- and long-range electrostatic effects of water exclusion. The latter partially determine binding free energies,[16] and is optimized here based on binding enthalpy data.

Thermodynamic calculations and prediction of binding modes also require an efficient method for sampling the configuration space. Configurational bias Monte Carlo (MC) has long been used in the condensed state,[36] including polymers[37,38] and crystals,[39] and is used here to enhance sampling of physically relevant microstates of a set of interacting proteins in solution. The configuration space generally includes both the spatial distribution of proteins and their internal conformations. The focus here is on the spatial distribution. Biased MC of internal degrees of freedom have been reported previously[40–42] and used in *ab initio* prediction of polypeptides conformations in solution.[16,40,43,44] Both methods can be combined to address the problem posed by the presence of multiple conformers and by induced fit during protein recognition and association, as discussed in Section V. A critical step in a biased scheme is the selection of the biasing function, which could hinder rather than improve sampling if not properly chosen. A function that approximates the canonical

distribution (unknown *a priori*) can greatly improve statistics and convergence, especially when large structural changes are needed to visit many configurations with statistical significance. An efficient method to construct such a function is presented in Section III. The method is applied in Section IV to three binary systems. Extension to multiprotein complexes is discussed in Section V.

II. Solvent effects: Electrostatic and liquid-structure forces

Biomolecules interact through non-covalent forces, which are strongly modulated (e.g., electrostatics) or directly elicited (e.g., hydrophobicity) by the aqueous medium. In molecular mechanics the electrostatic force $\mathbf{F}_i$ on an atom i of a system composed of N atoms is given by $\mathbf{F}_i = -\nabla_i E_e$, where E_e is the total electrostatic energy of the system in solution. The magnitude and direction of hydration forces determine the binding process. These forces are sensitive to the configuration of the system, which is determined by the N atomic coordinates $\mathbf{r} \equiv \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$. In the screened Coulomb potentials-based (SCP) model[45–47] E_e is given by[16]

$$E_e = \frac{1}{2}\sum_{i\neq j}^{N} \frac{q_i q_j}{r_{ij} D_{ij}(r_{ij};\mathbf{r})} + \frac{1}{2}\sum_{i=1}^{N} \frac{q_i^2}{R_i(q_i;\mathbf{r})} \left\{ \frac{1}{D_i[R_i(q_i;\mathbf{r});\mathbf{r}]} - 1 \right\} \quad (1)$$

where q_i is the charge of atom i. In this phenomenological partition the first sum is the interaction energy term, and the second sum is the self-energy term. The total energy in the SCP model also contains a cavity-formation term and a correction to account for the effects of liquid-structure forces (SIF) at aqueous interface (not discussed; see[16,47]). The mean-field effects of SIF are recast in R and optimized to reproduce the hydrogen-bond energies of all amino-acid pairs, as estimated from a systematic calculation of potentials of mean force in explicit water.[16] Both the screening functions D and the effective radii R depend on the system configuration. Modeling this dependence in a computationally efficient manner is a challenge, but essential to correctly represent both the magnitude and direction of hydration forces. A summary of the model follows.

II.1. Electrostatic effects of water exclusion

The screening functions in Eq. (1) are given by[16] $D_i(x;\mathbf{r}) = (1 + \varepsilon_0)/\{1 + k \exp[-\alpha_i(\mathbf{r})x]\} - 1$ and $D_{ij}(x;\mathbf{r}) = (1 + \varepsilon_0)/\{1 + k \exp[-\alpha_{ij}(\mathbf{r})x]\} - 1$, where ε_0 is the static permittivity of the solvent, and k is a constant. The dependence of D_i on the system configuration is through the screening coefficients α_i, given by[16]

$$\alpha_i \approx \alpha_{0,i} - A\sum_{J\neq I}^{M} \exp(-r_{IJ}/\sigma) \quad (2)$$

where J runs over the M residues of the proteins, and r_{IJ} is the distance between the C_α atoms of residues I and J; $A > 0$ and $\alpha_{0,i}$ determine the screening assigned to the atom i in the fully-hydrated residue I. The screening coefficients α_{ij} depend on the configuration through

$$\alpha_{ij} \approx \alpha_{0,ij} - \frac{A}{2}\sum_{K\neq I}^{M} \exp(-r_{IK}/\sigma') - \frac{A}{2}\sum_{K\neq J}^{M} \exp(-r_{JK}/\sigma') \quad (3)$$

where $\alpha_{0,ij}^2 = \alpha_{0,i}\alpha_{0,j}$. The characteristic lengths σ and σ' control the long-range decay of electrostatic water-exclusion effects.[16] Both $\alpha_{0,i}$ and ε_0 depend on the temperature, and $\alpha_{0,i}$

depends on the charge distribution as well.[47] The effective radii R_i depend on the local structure through[46]

$$R_i \approx R_{w,i} + a_i \sum_{j \neq i}^{N_c(i)} \exp(-r_{ij}/\tau_i) \quad (4)$$

where $R_{w,i}$ is a charge-dependent radius of the fully-hydrated atom i, and j runs over $N_c(i)$ atoms such that $r_{ij} < r_c$, and $a_i > 0$; r_c is a convenient threshold beyond which electrostatic interactions is said to be long ranged (according to previous theoretical estimates,[47] $r_c \sim 10$ Å; in the SCP model it is chosen as $r_c = 5.6$ Å, i.e., two hydration shells). Unlike σ and σ' in Eqs. (2) and (3), the characteristic length τ_i determines the short-range decay of the electrostatic effects of water exclusion.[16]

The summations in Eqs. (2)−(4) are suitable simplifications of general sums over the N atoms of the system,[16,45,46] and make the model highly efficient.[48,49] Figure 1 shows α, R, and the self-energy of a charge q crossing a planar interface. All the variables change smoothly with the distance, from their values in bulk water ($x \to -\infty$) to those in the interior of an infinitely large water-excluding cavity ($x \to +\infty$). The rates of changes with the distance from the surface depend on the values of σ in Eq. (2); and of τ and a in Eq. (4). For a molecule, the magnitude and direction of hydration forces depend on the values of a, τ, σ, and σ' assigned to each atom. Careful optimization is thus necessary to model the effects of water on a protein close to another protein, a membrane, or a solid surface.[50–52] The exponential functions in Eqs. (2)–(4) have been chosen for computational convenience and may need revision to better represent the decay of the electrostatic free energy with the distance from a real surface.

II.2 Model refinement

Electrostatic effects in the SCP model have been optimized previously using experimental hydration data[45] and results from dynamics simulations in explicit water.[16,53] Molecules used in the parameterization were small (amino acid and side-chain analogs), so the model better represents short-and medium-range water effects rather than long-range effects. Applications have thus been limited to peptides and small proteins at infinite dilution.[16,46,49,54] For larger systems and for processes where large amount of water are excluded from the environment (e.g., protein-protein association) consideration must be given to long-range effects. Barnase and barstar associate mainly by electrostatic forces,[55] so this complex (PDB code 1brs) is used here to optimize σ and σ' in Eqs. (2) and (3). To estimate the dissociation energy canonical MC simulations are carried out at $T = 25$ °C and fixed (standard) protonation states, using the united-atom representation (param19) of the CHARMM force field.[48] The dissociation energy ΔE_d is calculated as the energy difference between the bound and unbound states, i.e., $\Delta E_d = E_b - E_\infty$, where $E_b = Z^{-1}\Sigma_i E_i \exp(-E_i/kT) \approx \Sigma_i E_i/N_b$, and E_i and N_b in the last sum are the electrostatic energy [cf. Eq. (1)] of an accepted conformation i and the total number of accepted conformations in the bound state, respectively; E_∞ is the energy of the system with the proteins widely separated from each other. Trial moves consist of rigid-body rotations, translations, and roto-translations chosen with equal probabilities. Side-chain conformations have negligible effects on long-range electrostatics, so dihedral angle movements are not included. If long-range electrostatic effects are ignored (in practice, $\sigma \to \infty$ and $\sigma' \to \infty$) the dissociation energy is estimated at $\Delta E_d \sim 22.8$ kcal/mol (estimated sampling errors within $\sim kT$). This value changes with σ and σ' since these parameters affect the interaction and the self-energy terms in Eq. (1) independently.[16] These parameters can be adjusted to more closely reproduce the experimental binding enthalpy of the complex at the same pH and temperature, measured[56] at $\Delta H_b \sim 19.3$ kcal/mol. The optimized parameters follow a continuous line in the σ-σ' plane

(not shown), and σ = 59 Å and σ' = 37 Å are chosen here, which reproduce the experimental value within thermal energy. Two assumptions have been made. First, $\Delta H_b = \Delta U_b + p\Delta V \approx \Delta U_b \approx \Delta E_e$, i.e., changes in volume upon dissociation are neglected, and the internal energy U of the system is calculated with the continuum solvent model, and thus contains the free energy of the solvent; the SCP model also includes a standard term for the energy of cavity formation[16] (not discussed). Second, the van der Waals (vdW) contribution to the dissociation energy has been omitted. This is a common assumption[57] based on the notion that the degree of packing of atoms is similar in a protein and in water occupying the same space. However, recent ITC experiments in a number of protein-ligand complexes have shown that dispersion forces are actually quite strong and contribute significantly to the binding enthalpy when the binding pocket is sub-optimally hydrated.[58] The importance of dispersion forces in binding has long been recognized[59] and simple models have been proposed to include them in a continuum representation.[60,61] In small non-polar molecules these corrections play a measurable but modest role, especially when compared to other interfacial effects operating in heterogeneous polar molecules, such as SIF.[47,62–64] Dispersion however can no longer be ignored in larger systems and/or extended interfaces, or in cases where the interface is highly structured, and proper representation is ultimately needed to study protein association quantitatively. A simple thermodynamic cycle shows that the net vdW contribution to the binding energy between two proteins (1 and 2) can be approximated by $\Delta V_{\mathrm{vdW}} \approx -V_{12} + V_{1B} + V_{2A} - V_{A'B'}$. Here, V_{ij} is the vdW interaction energy between i and j, where A' and B' represent regions of bulk water with the same shape and volumes as proteins 1 and 2, respectively; A and B are the same regions of water but in contact with protein 2 and 1, respectively. Molecular dynamics simulations have been carried out here to estimate the relative magnitude of these terms, using the all-atom (param22) CHARMM force field and the TIP3P water model in a cubic cell of ~93 Å side lengths, with periodic boundary conditions and particle-mesh Ewald summation. For barnase (1) and barstar (2), $V_{1B} \approx 71$ kcal/mol, $V_{2A} \approx 28$ kcal/mol, and $V_{A'B'} \approx 11$ kcal/mol. The direct protein-protein vdW energy is $V_{12} \approx 118$ kcal/mol, so $\Delta V_{\mathrm{vdW}} \approx 30$ kcal/mol. Therefore, replacing a protein by water only partially offsets the direct interaction V_{12}. Although the values obtained here contain artifacts of the force field (e.g., the water model and the LJ function/parameters), the magnitude of ΔV_{vdW} should indicate that the assumption does require a closer inspection. These effects may have implications in the study of weak and ultra-weak association.

The energy of barnase and barstar along a path connecting the bound and unbound states can be calculated by gradually heating the native complex.[16] A set of relaxed structures (decoys) that includes near native and fully dissociated conformations can be generated by a MC simulation. Figure 2 shows the components of the non-bonded energy as a function of a reaction coordinate. The electrostatic interaction energy and the self-energy are shown with and without long-range water-exclusion effects included. The self-energy favors dissociation, whereas the interaction energy favors association.[16] The correct interaction results from the critical balance between these strong opposite effects. The direct vdW energy (V_{12}; above) is also shown for comparison. Inclusion of dispersion effects of water exclusion in the SCP model will be reported in a future study.

III. Prescreening of binary binding modes

Forces between macromolecules in solution operate at different length-scales and play different roles in the binding process. The method described in this section relies on the assumption that preferential first encounters are driven mainly by electrostatic interactions and by hydrophobic forces. Electrostatic forces operate at short and long range, while hydrophobicity acts only at short range (when the protein surfaces are a few hydration shells apart). Hydrogen bonds operate at even shorter distances and may determine specificity but

not first-contact modes. Surface potential complementarity can then be used to identify tentative modes of association that are most likely involved in first encounters. The final mode or modes of binding develop from these contacts and are determined by the complete force field. Surface-topography complementarity is not enforced because proteins can change conformation upon binding, a process not addressed here (see Section V).

III.1. Complementarity of surface electrostatic potential

Each protein of the complex is treated separately and at infinite dilution in pure water. For a given NMR or X-ray structure the Poisson equation is solved numerically (the problem posed by the presence of multiple conformers is discussed in Section V). The electrostatic potential φ is then mapped onto a grid of points $\mathbf{R}_n$ on the molecular surface, defined by the Lee-Richard method with a probe radius r_p = 1.4 Å, yielding $\varphi_n \equiv \varphi(\mathbf{R}_n)$.

Electrostatic (polar) interactions—Local maxima $\varphi_{M,i} \equiv \varphi(\mathbf{R}_{M,i})$ (i = 1,…, N_M) and minima $\varphi_{m,i} \equiv \varphi(\mathbf{R}_{m,i})$ (i = 1,…, N_m) of the surface potential are calculated numerically: a local maximum exists at point $\mathbf{R}_i$ if $\varphi_i > \varphi_j$ (or $\varphi_i < \varphi_j$ for a minimum) for all surface points $\mathbf{R}_j$ such that $|\mathbf{R}_j - \mathbf{R}_i| < \gamma$, where γ is a characteristic length scale of the potential variations on the surface. This value is protein-dependent and somewhat arbitrary, but enough resolution can generally be achieved with $\gamma = R_{aa} + 2R_w \sim 6.3$ Å ($R_{aa} \sim 3.5$ Å is the average radius of an amino acid in a protein, and $R_w \sim 1.4$ Å is the radius of a water molecule). This value is also computationally convenient as it leads to relatively small N_M and N_m for most proteins (see Section IV). Because of the discrete nature of the grid, φ_n shows large variations between neighboring points. Moreover, a local extremum carries no information on the spread of the potential on the local surface patch. To correct for these limitations $\mathbf{R}_{M,i}$ and $\varphi_{M,i}$ are reweighted, as

$$\mathbf{R}_{M,i} = \sum_{n=1}^{N_i} \phi_n \mathbf{R}_n \Big/ \sum_{n=1}^{N_i} \phi_n \quad (5)$$

$$\phi_{M,i} = N_i^{-1} \sum_{n=1}^{N_i} \phi_n \quad (6)$$

(likewise for a minimum) where $\mathcal{N}_i$ is the number of surface grid points such that $|\mathbf{R}_n - \mathbf{R}_{M,i}| < \gamma$ and $\varphi_n > 0$ (or $\varphi_n < 0$ for a minimum). Because $\mathbf{R}_{M,i}$ given by Eq. (5) do not generally lie on the molecular surface, they are projected onto the closest surface grid point.

With this procedure each protein p in a complex is represented by a reduced set of $N^{(p)}$ points, consisting of $N_M^{(p)}$ maxima and $N_m^{(p)}$ minima of the surface potential. Modes of electrostatic complementarity between proteins 1 and 2 are obtained upon minimization of the two-way norm,

$$e = a \sum_{i=1}^{N_M^{(1)}} \frac{\phi_{M,i}^{(1)} \phi_j^{(2)}}{r_{ij}^{(1)} + d} + a \sum_{i=1}^{N_m^{(1)}} \frac{\phi_{m,i}^{(1)} \phi_j^{(2)}}{r_{ij}^{(1)} + d} + a \sum_{i=1}^{N_M^{(2)}} \frac{\phi_{M,i}^{(2)} \phi_j^{(1)}}{r_{ij}^{(2)} + d} + a \sum_{i=1}^{N_m^{(2)}} \frac{\phi_{m,i}^{(2)} \phi_j^{(1)}}{r_{ij}^{(2)} + d} S_{12} \quad (7)$$

where the distances are given in Å and the potentials in kcal mol^{-1} C^{-1}; a is set to 1 C Å mol/kcal, so e is dimensionless. Index j in the first and second term determines the point $\mathbf{R}_j^{(2)}$ on protein 2 closest to point $\mathbf{R}_i^{(1)}$ in protein 1, i.e., $r_{ij}^{(1)} \equiv |\mathbf{R}_i^{(1)} - \mathbf{R}_j^{(2)}| = \min_k(|\mathbf{R}_i^{(1)} - \mathbf{R}_k^{(2)}|)$; a similar definition holds for j in the third and

fourth terms, after switching indices 1 and 2. The potentials $\phi_{M,i}^{(p)}$ and $\phi_{m,i}^{(p)}$ are, respectively, a maximum and a minimum on protein *p*, while $\phi_j^{(p)}$ is either a minimum or a maximum on protein *p*. The form of Eq. (7) is suggested by the electrostatic energy of two interacting charges of radii *d*/2 separated by a distance $R_{12} = r_{12} + d$; here *d* ~3 Å, about twice the average van der Waals radius that defines the molecular surface. The term S_{12} in Eq. (7) prevents structural overlaps. This is usually accounted for by the r^{-12} term of a LJ potential, but is represented here by an atom-centered hard-sphere model.

Hydrophobic (non-polar) interactions—Analogous procedure can be used to determine non-polar complementarity. A subset of surface grid points $\{\boldsymbol{R}_{n'}\} \subset \{\boldsymbol{R}_n\}$ with potentials $\{\varphi_n\}$ is first selected, such that $|\varphi_n| < \varphi_0$, where φ_0 is an appropriate threshold. Calculation on the active form of Calmodulin (PDB 1cll) and a number of small alkanes suggests that using $\varphi_0 \sim 0.1$ V may be sufficient to identify all the functionally-important non-polar regions in a protein. Local minima of the absolute value of the potential, $\psi_{m,i} \equiv |\varphi_{m,i}|$, are then calculated numerically in the new domain $\{\mathbf{R}_{n'}\}$, where $\varphi_{m,i} \equiv \varphi(\mathbf{R}_{m',i})$ and $i = 1,\ldots, N_{m'}$. The $N_{m'}$ positions and the absolute values of the potentials are adjusted according to Eq. (5) and Eq. (6), but using ψ instead of φ. Low surface potential is a necessary but insufficient condition to predict a hydrophobic region. Many points of low φ result simply from being at the boundary between regions of positive and negative fields. However, the average of $|\varphi|$ over a patch [Eq. (4)] allows discrimination of *bona fide* hydrophobic patches that could be involved in first encounters. With this procedure each protein *p* in a complex is represented by a reduced set of $N_{m'}^{(p)}$ points consisting of all the non-polar centers $\mathbf{R}_{m',i}$ on the proteins surfaces, each characterized by a degree of polarity defined by $\psi_{m,i}$. Local surface area accessibility[65,66] is used to define an appropriate norm. This is a simple but physically reasonable approximation commonly used in implicit solvation. Modes of non-polar complementarity between proteins 1 and 2 are obtained through a minimization of the two-way norm,

$$h = b^{(1)} \sum_{i=1}^{L^{(1)}} \theta(2R_w - |\mathbf{r}_i^{(1)} - \mathbf{r}_j^{(2)}|) + b^{(2)} \sum_{i=1}^{L^{(2)}} \theta(2R_w - |\mathbf{r}_i^{(2)} - \mathbf{r}_j^{(1)}|) + S_{12} \quad (8)$$

where θ is the Heaviside step function and R_w is the radius of a water molecule; the dimensionless parameters $b^{(p)} < 0$ are discussed below. Unlike the summations in Eq.(7), which covers all the points (maxima and minima) throughout the proteins surfaces, the summations in Eq. (8) are restricted to $L^{(p)}$ points $\mathbf{r}^{(p)}$ (a subset of the grid point $\mathbf{R}_{n'}$ such that $|\mathbf{r}^{(p)} - \mathbf{R}_{m'}| < \gamma$ and $|\varphi(\mathbf{r}^{(p)})| < \varphi_0$) on the local surface patch surrounding each hydrophobic center $\mathbf{R}_{m'}$; in practice $\gamma = 2R_w = 2.8$ Å. Indexes *i* and *j* are defined as in Eq. (7). The first term in Eq. (8) quantifies the degree of burial of a hydrophobic patch in protein 1 by a hydrophobic patch in protein 2; the second term yields the degree of burial of patch 2 by patch 1.

III.2. Norm optimization

Optimization of *e*—In this section a "point" refers to either a maximum or a minimum of the surface electrostatic potential. Optimization of *e* is carried out by first selecting a point *i* with coordinate $\mathbf{R}_i$ in protein 1 and a point *j* with coordinate $\mathbf{R}_j$ in protein 2 are first selected such that their potentials φ_i and φ_j have opposite signs. There are a total of $N_{tot} = N_M^{(1)} N_m^{(2)} + N_m^{(1)} N_M^{(2)}$ such (*i*, *j*) pairs. The two points are then superimposed and the proteins oriented, as follows: a vector $\boldsymbol{\nu}_i$ is defined on protein 1 as $\boldsymbol{\nu}_i = \Sigma_n (\mathbf{R}_n - \mathbf{R}_i)$, where *n* runs over all the grid points on the surface such that $|\mathbf{R}_n - \mathbf{R}_i| < s$, where *s* defines the size of a local patch of surface centered at *i*; statistics of protein/protein interfaces in the PDB

suggests s = 10 Å. A vector $\boldsymbol{\nu}_j$ is defined similarly on protein 2. If $\boldsymbol{\nu}_{i,o} = \boldsymbol{\nu}_i/|\boldsymbol{\nu}_i|$ and $\boldsymbol{\nu}_{j,o} = \boldsymbol{\nu}_j/|\boldsymbol{\nu}_j|$ are unit vectors pointing outwardly from the surfaces, the initial orientation is such that $\boldsymbol{\nu}_{i,o} = -\boldsymbol{\nu}_{j,o}$. Although this is not strictly necessary since the optimization protocol can rapidly find conformations with no structural overlaps regardless of the initial orientation, it prevents unnecessary clashes at the outset of the simulation.

The setup described above leaves only one degree of freedom, namely, rotation by an angle ω around the axis $\boldsymbol{\nu}_{i,o}$. This way N_ω initial conformations with random ω are selected for each pair (i, j). Any of these initial conformations should converge to the same optimized structure, but this is not always the case in practice, especially for rugged interfaces, due to imperfect sampling. Equation (7) is optimized by simulated annealing MC using a Boltzmann-like distribution $f = \exp(-e/T)$, where T is a dimensionless cooling parameter. Protein 1 (chosen as the larger protein of the pair) is fixed during the optimization, while protein 2 is translated, rotated, or roto-translated randomly with equal probabilities. Rotations are defined by an angle γ about a randomly-selected axis $\boldsymbol{\Omega}$ that passes through point j. Trial moves are selected randomly from Gaussian distributions with standard deviations σ_t (translations) and σ_r (rotations) using the Box-Muller method. These are set initially at σ_t = 2.8 Å, i.e., one hydration layer allowed at the interface, and σ_r = 180°. Both distributions are adjusted on the fly during the simulation to keep the acceptance rate above 0.4 (see below). A constraint is imposed on translations such that $|\mathbf{R}_i - \mathbf{R}_j| < R_c$, which forces i to remain close to j throughout the optimization process; R_c is initially set at R_c = 2.8 Å, and trial moves that violate this distance criterion are rejected. The simulation starts at a (system-dependent) temperature $T_M = 10N_{tot}\max_{ij}(|\varphi_i\varphi_j|)/d$, which is decreased logarithmically in N_T steps up to the lowest temperature, here $T_m \sim 10^{-3}$ (in practice N_T = 20). A total of 10^4 trial moves are performed at each temperature; this limited sampling justifies the choice of N_ω initial structures. If the acceptance rate at a given temperature is less than 0.4, both σ_t and σ_r are rescaled by a factor 2/3 at the next temperature. There is no need to impose detailed balance at this stage.

Evaluation of S_{12} in Eq. (7) requires the calculation of distances d_{kl} between a surface atom k in protein 1 and a surface atom l in protein 2. A trial move is rejected if $d_{kl} < R_{vdw,k} + R_{vdw,l} + c$ for any pair of atoms; here $R_{vdw,k}$ and $R_{vdw,l}$ are the van der Waals radii of the atoms; $c \geq 0$ is a soft-core parameter that can be used to improve sampling of structures that are locally trapped due to the constrain $|\mathbf{R}_i - \mathbf{R}_j| < R_c$ imposed in the initial alignment. This problem can arise in the presence of very irregular interfaces, whereby either i or j are buried in crevices. This is the case of residues that tend to confer binding-specificity, which are often “locked” into a cavity in the host protein (see Section IV). In the protocol proposed here c = 0, and the problem posed by locally-trapped structures is circumvented by rescaling R_c by a factor 1.2 every 10^4 moves, up to a maximum of $2R_c$ (i.e., two hydration layers allowed at the interface, at most). This relaxation criterion is physically more appealing, and is applied only at the highest temperature T_M. Once a structure is accepted, the simulation at T_M continues for another 10^4 moves, and the acceptance rate is calculated over this latter period. If the rate at T_M is still zero after 10^5 moves and once the constraint reached $2R_c$, the initial alignment is discarded (this situation has been observed in few of the several tests performed).

Optimization of Eq. (7) requires finding closest neighbors to either points or atoms in each trial move. These queries are of two kinds: (1) find point i on the surface of one protein that is closest to a point j on the surface of the other protein to evaluate the electrostatic terms; (2) find atom k in one protein that is closest to an atom l in the other protein to evaluate S_{12}. In both cases a search based on Delaunay triangulation is used, which speeds computation one order of magnitude when compared to a direct search over pairs.

For each pair (i, j), the N_ω optimized structures can be grouped into conformational families. The C_α-root mean square deviations (RMSD) between all the structures are first calculated after superimposing protein 1. These values are stored in a $N_\omega \times N_\omega$ symmetrical arrangement and clustered using a hierarchical technique[67] according to the maximum intra-cluster RMSD variance (δ) desired (in practice, δ = 5 Å). The process yields N_δ d N_ω clusters (conformational families). For each of the clusters, the structure with the lowest RMSD with respect to all other members of the same cluster is selected as a representative member of the family. The optimization thus generates $\Gamma=\sum_1^{N_{tot}} N_\delta$ structures $\{s_1, s_2, \ldots, s_\Gamma\}$ as potentially relevant electrostatic-driven binding modes that warrant further scrutiny with the complete force field. The index m in $\{s_m\}$ represents a convenient array unrelated to the values of the optimized norm.

Optimization of *h*—The same algorithm is used. A "point" refers now to one of the non-polar centers. In analogy with the setup described above, point i and j are selected on the protein 1 and 2, respectively, yielding a total of $N'_{tot}=N_{m'}^{(1)} N_{m'}^{(2)}$ (i, j)-pairs. For each pair, the proteins are aligned as described above. The parameters $b^{(p)}$ in Eq. (8) is chosen as to reflect the degree of polarity of the patch, and is given by $b^{(p)}(\varphi) = A + B|\varphi|$, where $A = -b(0)$ and $B = b(0) / |\varphi_0|$. Thus, the more polar the patch is, the weaker the hydrophobic effect expected; and *vice versa*. Any positive value can be chosen for $b(0)$; here $b(0) = 4.2$ (if the summations in h had dimensions of Å^2, solubility data of alkanes suggest[45] ~4.2 kcal/mol/Å^2). The simulated annealing MC optimization is carried out with a distribution $f = \exp(-h / T)$ and a maximum temperature $T_M = 10b(0)\max_{ij}(2L)$, where L depends on i and j according to the surface area of the patch. For each (i, j)-pair clustering of the N_ω initial alignments generates N'_δ conformational families. Optimization of h yields a total of $\Gamma'=\sum_1^{N'_{tot}} N'_\delta$ structures $\{s'_1, s'_2, \ldots, s'_{\Gamma'}\}$ as potentially-relevant hydrophobicity-driven first-encounter modes.

III.3. Probability maps and biased sampling

The $\Lambda = \Gamma + \Gamma'$ conformations $\{s_m\} = \{s_1, s_2, \ldots, s_\Gamma\} \cup \{s'_1, s'_2, \ldots, s'_{\Gamma'}\}$ identified from optimization of e and h are treated on equal basis. Each mode is a potentially-relevant first encounter mode, and its relative importance is determined by a screening protocol described below. A probability distribution can be constructed from $\{s_m\}$ and used as the biasing function in the full MC simulation. In each trial move a structure s_m is first selected randomly out of the Λ potential modes. Moves consist of translations, rotations, and roto-translations of protein 2 selected with equal probabilities, while protein 1 remains fixed over the course of the simulation. Random rotations of side-chain dihedral angles are a fourth type of movement and can be applied to both proteins with equal probability.[16] At the beginning of the simulation the center of mass of protein 1 is positioned at the origin of the laboratory coordinate system, and rotated such that its primary axis of inertia is oriented in the z direction ($\mathbf{I}^{(1)} = \hat{\mathbf{k}}$). The secondary and tertiary axes of inertia are oriented in the x and y directions, respectively ($\mathbf{I}^{(2)} = \hat{\mathbf{i}}$ and $\mathbf{I}^{(3)} = \hat{\mathbf{j}}$). All movements of protein 2 are thus relative to the molecular frame of protein 1, so simple coordinates transformations can be applied to the equations derived below if protein 1 is moved, e.g., when more than two proteins are involved.

The six degrees of freedom necessary to position protein 2 relative to protein 1 are determined by six random variables u_i (i =1, …, 6) distributed uniformly in the interval [0, 1]. A translation is defined by the transformation $\mathbf{r} = \mathbf{r}_m + \Delta\mathbf{r}$, where $\mathbf{r}_m$ are the coordinates of protein 2 in the selected mode s_m and $\Delta\mathbf{r} = (x, \Delta y, \Delta z)$ is a random displacement obtained from normal distributions with zero mean and non-unit variance, according to the

transformations $\Delta x=\sigma_x\cos(2\pi u_2)\sqrt{-2\ln(u_1)};\Delta y=\sigma_y\sin(2\pi u_2)\sqrt{-2\ln(u_1)};\Delta z=\sigma_z\cos(2\pi u_4)\sqrt{-2\ln(u_3)}$, where σ_x, σ_y and σ_z are the standard deviations in each direction.

A rotation is defined by the transformation $\mathbf{r} = \mathbf{R}\bar{\mathbf{r}}_m$ where the matrix $\mathbf{R}$ represents a random rotation of protein 2 by an angle $\Delta\gamma$ around a random axis determined by the unit vector $\boldsymbol{\Omega} = (\omega_x, \omega_y, \omega_z)$ that passes through the center of mass of protein 2. In quaternion notation this matrix is given by

$$\bar{\mathbf{R}}=\begin{pmatrix} q_0^2+q_1^2-q_2^2-q_3^2 & 2q_1q_2-2q_0q_3 & 2q_1q_3+2q_0q_2 \\ 2q_1q_2+2q_0q_3 & q_0^2-q_1^2+q_2^2-q_3^2 & 2q_1q_3-2q_0q_1 \\ 2q_1q_3-2q_0q_2 & 2q_2q_3+2q_0q_1 & q_0^2-q_1^2-q_2^2+q_3^2 \end{pmatrix} \quad (9)$$

where $\mathbf{q} = (q_0, q_1, q_2, q_3) = (\cos\alpha, \omega_x \sin\alpha, \omega_y \sin\alpha, \omega_z \sin\alpha)$ and $\alpha = (\pi/360)\,\Delta\gamma$, with γ in degrees. To keep track of coordinate changes the vector $\boldsymbol{\Omega}$ is obtained from a random rotation of the primary axis of inertia of protein 2 in the mode s_m, as determined by the ortho-normal components $\mathbf{I}_m^{(1)}=(I_{x,m}^{(1)}, I_{y,m}^{(1)}, I_{z,m}^{(1)}=(\sin\varphi_m\cos\theta_m, \sin\varphi_m\sin\theta_m, \cos\varphi_m)$, where $(_m, \theta_m)$ are the angles in spherical coordinates. The rotation matrix is then defined by the transformations $\phi = \phi_m + \Delta_\phi$ and $\theta = \theta_m + \Delta\theta$, and by a rotation $\Delta\gamma$ around this new axis, where

$\Delta\varphi=\sigma_\varphi\sin(2\pi u_4)\sqrt{-2\ln(u_3)}, \Delta\theta=\sigma_\theta\cos(2\pi u_6)\sqrt{-2\ln(u_5)}, \Delta\gamma=\sigma_\gamma\sin(2\pi u_6)\sqrt{-2\ln(u_5)}$ are the corresponding Box-Muller transformations, and σ_ϕ, σ_θ and σ_γ the standard deviations. Normal distributions of φ and θ are not necessary since the main restriction is on $\Delta\gamma$, but imposed here for completeness.

In thermodynamic equilibrium strict detailed balance implies that the old (o) and the new (n) states are related through[68] $P_o\pi_{o\to n} = P_n\pi_{n\to o}$, where P is the corresponding Boltzmann occupancy probability, and π is the transition probability between the states, given by $\pi_{o\to n} = \alpha_{o\to n}\, p_{o\to n}$ and $\pi_{n\to o} = \alpha_{n\to o}\, p_{n\to o}$. Here α is the underlying matrix of the Markov process and p is the acceptance probability given by

$$p_{o\to n}=\min\left(1, \frac{\alpha_{n\to o}}{\alpha_{o\to n}}\exp(-\beta\Delta E\right) \quad (10)$$

where $\Delta E = E_n - E_o$, and E is the energy of each state, now calculated with the complete force field. The ratio of *a priori* probabilities in Eq. (10) can be estimated from a sum of Gaussian distributions over the Λ binding modes. Defining the linear array $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6) = (x, y, z, \phi\ \theta\ \gamma)$, the probability of generating a trial move within an element $\delta\boldsymbol{\eta}$ centered at $\boldsymbol{\eta}$ given that a mode m has been selected, is

$$P(\boldsymbol{\eta}|m)=\prod_{i=1}^{6} g_i(\eta_i|m)\delta\eta_i \quad (11)$$

where g_i are the normal distributions

$$g_i(\eta_i|m)=(2\pi\sigma_{i,m}^2)^{-1/2}\exp[-(\eta_i-\eta_{i,m})^2/2\sigma_{i,m}^2] \quad (12)$$

and $\eta_{i,m}$ and $\sigma_{i,m}$ are the value of η_i and its standard deviation in mode m. The total probability is

$$P(\boldsymbol{\eta})=\sum_{m=1}^{\Lambda} h_m \prod_{i=1}^{6} g_i(\eta_i|m)\delta\eta_i \quad (13)$$

where h_m is the probability of selecting mode *m*. Introducing Eq. (12) into Eq. (13) yields

$$P(\boldsymbol{\eta})=\sum_{m=1}^{\Lambda} a h_m \kappa_m \exp(-J_m) \quad (14)$$

where $a=\prod_{i=1}^{6}\delta\eta_i/8\pi^3$ and $\kappa_m=1/\prod_{i=1}^{6}\sigma_{i,m}$, with

$$J_m=\sum_{i=1}^{6}(\eta_i-\eta_{im})^2/2\sigma_{i,m}^2 \quad (15)$$

so the ratio of probabilities in Eq. (10) is given by

$$\frac{\alpha_{n\to o}}{\alpha_{o\to n}}=\frac{\sum_{m=1}^{\Lambda}\kappa_m h_m \exp(-J_m^{(o)})}{\sum_{m=1}^{\Lambda}\kappa_m h_m \exp(-J_m^{(n)})} \quad (16)$$

where κ_m can be adjusted on-the-fly through $\sigma_{i,m}$ to control the acceptance rate per mode, if needed; the same approach applies to h_m and J_m, although the latter also accommodate changes in the coordinates $\eta_{i,m}$ of the mode as new structures are accepted. If $\sigma_{i,m}$ and $\eta_{i,m}$ are kept fixed over the course of a simulation (i.e., fixed *a priori* probabilities), the biasing function is non-adaptive; if $\sigma_{i,m}$ and/or $\eta_{i,m}$ change, the function is adaptive.

Screening of binding modes—The probability h_m in Eq. (16) is defined over the discrete set $\{s_m\}$, and is chosen here as a Boltzmann-like distribution

$$h_m=Z^{-1}\exp(-\Delta E_m/\lambda kT) \quad (17)$$

where $Z=\sum_{i=1}^{\Lambda}\exp(-\Delta E_i/\lambda kT)$ and λ is a scaling factor discussed below. Energies are measured with respect to the fully dissociated state, $\Delta E_m = E_m - E_\infty$, where E_m is the energy of the complex in mode *m* now calculated with the complete force field. These energies are calculated as canonical averages over short MC simulations of the complex in mode *m*, $E_m \approx N_m^{-1}\sum_i E_i^{(m)}$, where $E_i^{(m)}$ and N_m are the energies and the number of accepted structures. The simulation is biased and non-adaptive, determined by $h_m = 1$ and $h_{k\neq m} = 0$, thus Eq. (16) is simplified to

$$\frac{\alpha_{n\to o}}{\alpha_{o\to n}}=\frac{\exp(-J_m^{(o)})}{\exp(-J_m^{(n)})} \quad (18)$$

with *J* given by

$$J_m^{(x)}=\sum_{i=1}^{6}(\eta_i^{(x)}-\eta_{i,m})^2/2\sigma_{i,m}^2 \quad (19)$$

where x is either o or n. The parameter $\lambda \geq 1$ in Eq. (17) is used to smooth the distribution $\{h_m\}$ over the set $\{s_m\}$. This is a safeguard measure against limitations of the prescreening protocol (including the definition of the norm) and the force field to properly identify physically relevant first-encounter modes of association. Small errors in the estimation of energies in Eq. (17) may eliminate modes (in practice, $h_m << 1$) that are worth sampling, or over emphasize sampling of less important modes, thus compromising the efficiency of the method. This problem is alleviated by using $\lambda > 1$ (see Section IV), in a process akin to high-temperature annealing.

Self-adaptive biased Monte Carlo—strict detailed balance is imposed by using Eq. (16) in the acceptance criterion established by Eq. (10). In the self-adaptive biased sampling used here both $\sigma_{i,m}$ and $\eta_{i,m}$ in Eq. (16) and Eq. (19) are allowed to change over the course of the simulation. The probability distribution $\{h_m\}$ could also change to improve efficiency by increasing/decreasing sampling of certain modes as the simulation progresses, but this adaptation is not used here. An acceptance rate b_m is calculated for each mode every 10^3 times the mode is selected, and $\sigma_{i,m}$ is then scaled up or down to keep the acceptance rate of that mode within a predetermined value. The same scaling factor applies to all the degrees of freedom, except $\sigma_{\phi,m}$ and $\sigma_{\theta,m}$. Each time a mode m is selected the coordinates $\eta_{i,m}$ are updated to the last accepted structure for that mode. This is accomplished in practice by translating the center of mass and rotating the primary axis of inertia of protein 2 to the corresponding values of the accepted structure; translations and rotations $\Delta\eta_i$ are then measured with respect to the new mode coordinates $\eta_{i,m}$. If $\Delta\eta_i$ in an accepted move is larger than $2\sigma_i$ for a given mode m, then $\sigma_{i,m}$ is reset to its original value since it is possible that a new local minimum has been identified.

IV. Results

Three binary complexes are chosen to illustrate the application of the method: Barnase/barstar (1brs) has long been used in experimental and computational studies of protein binding.[55,69,70] The complex has been used here as a guide for model refinement. The other complexes considered are: trypsin bound to a protein inhibitor (2ptc) and histidine-containing phosphocarrier protein HPr (1poh). These complexes were chosen here because they challenge different aspects of the method: in the bound state, a specificity-conferring Lys residue (K15) in the ligand of 2ptc is buried into a narrow cavity of the protein, so the complex provides a stringent test for the sampling method; 1poh has been shown to form ultra-weak self-association, with negligible dimerization in solution, so the complex provides a stringent test for the continuum model.

Figure 3 shows the electrostatic potential on the molecular surface of barnase and barstar, calculated as standard solutions of the Poisson equation. Barnase (protein 1) has 27 local maxima and 29 minima, and 36 non-polar centers, whereas barstar (protein 2) has 25 maxima and 23 minima, and 29 non-polar centers, yielding $N_{tot} = 1346$ initial (i, j) pairs to be considered for optimization of e and $N'_{tot} = 1044$ for h. For the other two complexes, $N_{tot} = 1548$ (2ptc) and 480 (1poh), and $N'_{\mathrm{tot}} = 988$ (2ptc) and 676 (1poh). Figure 4 shows the values of the potential (in Volts) at the maxima and minima; the values of $|\varphi|$ in 1poh is also shown for comparison. It is not possible to decide from these values alone which (i, j) pairs are more likely to be involved in first encounters, so all the pairs should in principle be considered in the optimization of the norms. To reduce the computational cost only the ten highest maxima and the ten lowest minima in each protein are used in the optimization of e. This simplification yields 200 (i, j) pairs; and for each of these pairs $N_\omega = 24$ initial alignments are generated. Experiments have shown that electrostatics is the main force that controls binding in the three complexes; hydrophobicity plays a role only in HPr. This knowledge *a priori* allows a convenient simplification by omitting the optimization of h in

1brs and 2ptc. This simplification applies only to the prescreening stage. The SCP model does contain[45] a simplified "hydrophobic term" (not discussed here; see Section V) which is used in both the screening stage and in the full MC simulation. Thus, bypassing the optimization of h in a particular system (here brbs and 2ptc) does not mean that hydrophobic interactions are ignored; it only means that hydrophobicity does not determine the first encounters. In contrast, for 1poh only non-polar patches with $|\varphi| < 0.03$ V (suggested by the surface potentials of calcium-loaded calmodulin) are considered [cf. Fig. 4].

After norm optimization and clustering, a total of 218 binding modes are obtained for 1brs, 474 for 2ptc, and 243 (polar) and 123 (non-polar) for 1poh. The inset of Fig. 5a shows the superposition of all the modes in 1brs, with barnase at the center. Two major first-contact regions are apparent, each containing multiple orientations of barstar; the most populated region is located in the vicinity of the native binding site of barnase. For the other two complexes there are multiple binding regions surrounding the central protein 1, with the most scattered distribution observed in 1poh (see below). Optimization of e took ~30–50 min per mode, depending on the complex; 90% of the CPU time was used to compute the electrostatic component of the norm [first four terms of Eq. (7)] and the remainder 10% for the calculation of S_{12}. Optimization of h took ~30 min per node, but this can be reduced substantially by decreasing the number of points L used to define the area of the patch in Eq. (8) (here L ~ 80–100). The optimization was performed in Matlab using standard functions from the statistic toolbox, and on a single 2.8 GHz Intel X5660 processor with 24 GB memory. The code was not parallelized.

Screening of the prescreened modes was performed with biased non-adaptive MC (10^3 steps) at 25 °C (for 1brs and 2ptc) and 35 °C (1poh), using $\sigma_x = \sigma_y = \sigma_z = 0.5$ Å; $\sigma_\Phi = \sigma_\theta = 90°$, and $\sigma_\gamma = 2.5°$. The united-atom (param19) representation of the CHARMM force field[48] was used, with the SCP model[16] implemented in the version c35 of the CHARMM program. No cutoffs were applied to the non-bonded interactions in order to account for long-range effects. Figure 5 (left panels) shows the probability distributions h_m of prescreened modes $\{m\}$ using a smoothing parameter $\lambda = 25$; only one mode stands out in 1brs, with a weight $h_m \approx 0.08$. This mode is very close to the native complex and has a C_α-rmsd of ~1.9 Å with respect to the crystal structure (Fig. 5b; blue). This shows that electrostatic pre-screening followed by screening with the complete force field is sufficient to identify a near-native conformation in the barnase/barstar complex. This is probably the case for other systems driven to association by strong electrostatic interactions. The h_m distributions in the other complexes are qualitatively different (Figs. 5c and 5e): for 2ptc the closest prescreened mode to the native complex has a C_α-rmsd of ~5 Å (Fig. 5d), and corresponds to the fifth highest weight h_m. As in 1brs, this mode is a good candidate for first contact since K15 in the ligand is near the pocket in trypsin and oriented towards it (Fig. 5d, right). For 1poh several electrostatic modes also have similar weights (Fig. 5e; black), and the highest ten modes are shown in Fig. 5f (left). These modes are clustered close to residues E5, E25, E32 and S46, which were used as labels in a recent NMR study[11] of ultra-weak self-association of HPr. The weights of the hydrophobic modes are also shown for comparison (Fig. 5e; red); the ten modes with the highest weights are displayed in Fig. 5f (right). Electrostatic and hydrophobic modes plotted in Fig. 5e are normalized independently for clarity. The scattered distribution of both types of modes (Fig. 5f) and the similarity of weights (Fig. 5e) are consistent with multiple first-encounters between the proteins and may reflect the non-specific nature of the association. The inset to Fig. 5e shows the energies ΔE_m of the modes [in Eq. (17)]. Despite the substantial energy overlap between electrostatic and hydrophobic modes it is apparent that first encounters in HPr are driven mainly by electrostatics.

The complete sets $\{h_m\}$ in Fig. 5 were used to create the initial spatial distributions for the self-adaptive MC sampling. Simulations were performed at the same temperature used for screening, and consisted of 10^6 steps with $\sigma_x = \sigma_y = \sigma_z = 2.5$ Å; $\sigma_\phi = \sigma_\theta = 90°$ and $\sigma_\gamma = 20°$. These values were chosen based on a number of combinations tested. Changing these values has no major effect in the results discussed below, but important variations in efficiency were observed due to convergence problems. In these simulations only $\eta_{i,m}$ are adapted, while $\sigma_{i,m}$ remains fixed regardless of the acceptance rate per mode. Simulations were performed in a single processor with a non-parallelized version of the SCP model, and took ~24−48 CPU hours, depending on the complex. The parallel version of the SCP model scales well up to 24 processors, and can reduce the simulation time one order of magnitude. For 1brs the native complex (Fig. 5b; green) was identified within a few thousands steps. Because there are 234 prescreened modes, the overall acceptance rate is small since all the modes are selected for trial moves, albeit with probabilities determined by h_m. The conformational distribution obtained upon convergence is very narrowly centered in a single mode identified as native. For 2ptc convergence takes much longer but the native complex was also identified correctly. The dissociation energy of the native complex is estimated at ~5.5 kcal/mol; and the association is thus strong and specific. For 1poh a single mode is also obtained (Fig. 6), but the distribution of accepted structures is much broader than in the other two complexes, which is consistent with a shallower energy surface. The predicted native complex is quite symmetrical, with residues E32 and S46 at the protein/protein interface. Dissociation from this structure requires a very small energy, only ~1.3 kcal/mol, but this is still too large and the presence of stable homodimers cannot be ruled out at 35 °C. Experiments carried out at this temperature indicate that HPr form multiple transient associations, but no specific homo-dimerization.[11] There are several possible explanations for this discrepancy that warrant further scrutiny: i) backbone flexibility may need to be included to obtain a more accurate canonical distribution. Given the transient nature of the association it is unlikely that induced fit is involved, so conformational selection may be a more important mechanism in this case (Section V); ii) current force fields are not yet accurate enough to discriminate ultra-weak modes, although progress being made, especially in the treatment of non-bonded terms (e.g., inclusion of polarizability), as these are most relevant in protein-protein interactions. Improvements and careful optimization of the solvent model, especially the treatment of the aqueous interface is an essential component and must be pursued simultaneously; iii) specific water-mediated interactions at the protein interfaces may also be important, and a continuum model cannot represent them properly unless some degree of granularity is introduced. In addition, liquid-structure forces (SIF) are non-pairwise additive and costly to compute. An algorithm has been reported to include SIF in a continuum model for use in Langevin dynamics[47]; iv) changes in protonation states upon pKa shifts have been ignored and could change the interaction energy landscape; a primitive version of the SCP model has been used to predict pH-dependent properties in proteins[71] and is well suited for on-the-fly assignment of protonation states, at the expense of CPU time. These limitations apply to all protein-protein interactions but are more problematic when dealing with weak and ultra-weak associations. These interactions thus provide a stringent benchmark for further development.

Overlooking potentially relevant modes during prescreening and/or screening (possibly due to limitations in the norm optimization protocol, the norm itself, or the force field) is of concern since success of the method hinges on having identified a mode with sufficiently large probability to be selected during sampling. To test the robustness of the method to changes in the h_m distribution, the main mode *m*' in 1rbs was removed from the set (in practice, $h_{m'} = 0$). In this case the simulation takes much longer to converge, but the native complex is also identified within a few hundred thousand steps. In this case a secondary mode with a small weight slowly moves towards the native conformation during the self-adaptive process and takes over the local distribution left unpopulated when *m*' was

removed. This drift of a distant mode towards the native mode is possible because several prescreened modes $m \neq m'$ have C_α-rmsd in the ~2 4 Å range with respect to the crystal structure. Therefore, given the chance to be selected they make important contributions to the acceptance rate once m' is removed. This also highlights the importance of smoothing h_m through λ.

V. Discussion

Weak and ultra-weak interactions can play a role in protein recognition and drive spontaneous self-assembly and aggregation of larger multimeric complexes, such as crystals, amyloid fibrils, and virus capsids. These interactions are difficult to detect experimentally. They also present a major challenge to the Hamiltonian because effects that can be ignored or treated in simplified ways in small systems at infinite dilution now require adequate treatment and optimization. These include the effects of interfaces, and long-range electrostatic and non-electrostatic effects of water exclusion. The problem posed by interfaces is complex and multifaceted, involving the dielectric[19,20,47,72] and the structural[47,73] response of the liquid, and its dynamic[19,20] and entropic[17] contributions. In particular, the entropy of an aqueous interface is difficult to capture in a mean field approximation. The entropy can be divided into an orientational and a translational contribution. The orientational component is related to the static dielectric response of the interface, and an algorithm has been proposed to estimate it self-consistently in a continuum approximation.[72] The translational behavior is more complicated and is related to the mobility of water in the hydration shells. Recent simulations have shown that water in the second shell of a DNA molecule is more mobile than water in either the first shell or the bulk phase.[17] Because of the substantial changes in surface hydration upon protein association or dissociation, different hydration shells may contribute differently to the free energy of binding. These effects need additional studies, especially in large complexes, and may eventually require proper implementation in a continuum model. The SCP model partially contains both components of the entropy, which is reflected in the sigmoidal shape of the screening functions D and in the mean field effects of SIF through R. In contrast to the entropy of water, the entropy of the molecular system under consideration can be calculated directly from the statistical distributions obtained from the biased sampling; backbone flexibility may introduce practical but not conceptual complications. Methods also exist to estimate the vibrational entropy contributions.

Although short-range electrostatic effects of water-exclusion [represented in the self-energy term of Eq. (1) through the conformation-dependence of R given by Eq. (4)] make important contributions to the binding energy, long-range corrections [represented in by both the interaction and the self-energy terms through the conformation-dependence of D's given by Eqs. (2) and (3)] cannot be ignored. The problem posed by long-range bulk-water electrostatics in modeling hydration forces has been discussed.[74] These effects become increasingly important as the size of the system increases, e.g., during aggregation or self-assembly, or in crowded environments. Ignoring these corrections introduces an error of ~3.5 kcal/mol (~20%) in the binding enthalpy of the barnase/barstar complex as estimated with the SCP model. Errors of this magnitude can be ignored when predicting specific (usually strong) binding modes, but are clearly unacceptable in thermodynamic calculations and for prediction of weak association for which chemical accuracy is ultimately needed. It has been shown here that long-range electrostatics can be fine-tuned to provide a better estimate of binding enthalpies. The balance between interaction and self-energy terms in Eq. (1) is critical to reproduce the correct binding energy because they oppose each other. Long-range electrostatic contributions in real systems may decay more rapidly or more slowly than the exponential decays modeled by Eqs. (2) and (3), and systematic calculations in systems of different sizes should be performed to refine the model.

There is experimental evidence that dispersion forces make important contributions to protein-ligand binding enthalpy.[58] This has long been recognized[59] and attempts have been made to include a dispersion term in implicit solvent models. Except in the case of purely non-polar solutes such corrections can be neglected since even in small polar or charged molecules other effects at the aqueous interface (e.g., the dielectric response of the liquid and liquid-structure forces) play a more important role.[75] In larger systems/interfaces, however, their contribution can be substantial and no longer be ignored. Thus, both long-range electrostatics and dispersion contribute to the cohesive energy of a macromolecular complex. The simulations discussed in Section II.2 support these findings. Non-electrostatic effects of water exclusion may actually play an important role in weak and ultra-weak association.

Developments of the SCP model have hitherto focused on electrostatics and liquid-structure forces at protein/water interfaces. These are the most important effects in a large class of biological systems, including proteins and nucleic acids, ions, osmolytes and cryoprotectants (see review in[32,75]). In other bioactive macromolecules (e.g., Ca^{2+}-loaded Calmodulin used in Section III) hydrophobic interactions are known to be a key feature of their function. A more advanced treatment of hydrophobicity in the SCP model may thus be desirable. However, modeling hydrophobic forces in molecules of arbitrary shapes and morphologies is difficult[76–80] and has not yet been addressed in a practical manner. In small non-polar molecules improvements have been reported with rather minor changes to the commonly used solvent-accessible surface-area model.[81–84] It is unclear whether more sophisticated treatments are needed in real proteins (generally characterized by sparse distributions of relatively small hydrophobic patches punctuated by regions of high polarity and local charge).[32] Recent dynamics simulations of small amphiphilic molecules have provided insight into the role of the micro-complexity of water on the hydrophobic effect in systems that more closely resemble the heterogeneity of real protein surfaces.[85] Simulations have also shown that such surfaces display a behavior in between that of an idealized hydrophobic surface (a common theoretical construct) and one that is strongly hydrophilic.[86] Unlike protein electrostatics there is a paucity of useful experimental information that can be used to validate hydrophobic models, so carefully-designed simulations may ultimately be needed to advance the field.

A method has been described to construct a biasing function for efficient configurational bias simulations that allows detection of weak and ultra-weak binding modes and populations. The method has been tested in three binary complexes, but can be extended to multiprotein systems provided that complexation occurs through a succession of binary reactions. This extension is required to simulate crowded environments or subcellular processes where multimeric complexes (averaging four or more units per complex[2,3]) are common. In a recent assessment[87] of experimental methods aimed at predicting protein-protein binding in a three-component systems only nine out of twelve participant groups were able to conclude that barnase and BiNase2 compete for binding to barstar, so that the formation of a ternary complex is not possible. Multi-component systems present a greater challenge, especially if some of the proteins interact weakly or no-specifically. Therefore, having the capability to explore efficiently (that is, rapidly and with statistical significance) the spatial distribution of many proteins simultaneously is desirable. The biasing function proposed here allows mixing large and local changes in the protein spatial distributions, which enhances sampling of microstates that may be overlooked with non-biased sampling. The method can also be used to identify regions at a protein surface that are most likely to bind ions and cosolutes since they may be attracted to multiple sites. These molecules affect almost all macromolecular properties (including protein denaturation, stabilization, aggregation, and dissociation), and can interact specifically and non-specifically with the proteins.

It has been assumed that preferential encounters in solution are driven by electrostatic and hydrophobic forces, and the norms *e* and *h* defined in Eqs. (7) and (8) reflect this assumption. The functional forms of the norms are adequate simplification of the physical effects that each intends to describe, and designed specifically for computational efficiency. Electrostatic complementarity has long been used as a strategy to predict specific binding modes,[55] but this approach alone is insufficient to predict weaker association,[88] a problem compounded in the case of non-specific and multiple binding modes. The approach has been extended and used here only to identify first-encounter modes. The binding modes obtained from norm optimization determine the spatial distribution from which the complexes evolve. The final mode (or modes) of association are obtained from the canonical distribution upon convergence of the self-adaptive biased sampling.

Proteins in aqueous solution display varying degrees of backbone flexibility. Statistics from the PDB have revealed that many proteins undergo only small changes in their overall fold upon binding (typically ~1 Å in C_α-rmsd) as their interfaces are largely pre-formed.[88] The rigid-backbone approximation is thus reasonable in many cases and has been used successfully to predict the structure of unknown complexes.[89] This approximation is usually the first stage in almost all docking algorithms,[90–92] and good estimates of binding modes in this initial stage is critical. The rigid-backbone approximation might actually suffice in the case of weak or ultra-weak binding because these interactions are short-lived, possibly lasting less than the time-scale necessary to induce backbone conformational changes (although this is a conjecture that needs experimental corroboration). Important exceptions however exist since flexibility is at the core of protein function. For example, trypsin-TPI undergoes rigid-backbone association, but the closely-related trypsinogen-TPI does not. In general, oligomeric proteins and antigen-antibody complexes tend to challenge this assumption. Moreover, some DNA- and RNA-protein complexes are known to undergo co-folding during recognition and binding.[93] Even proteins typically thought of as rigid in solution undergo localized conformational changes, usually in unstructured regions such as loops. A recent study of the dynamics of ubiquitin[94] suggests that the forty-plus crystal structures of this rather rigid protein in the PDB are likely conformers pre-selected by the ligand. Upon association of a given conformer there appear to be only small rearrangements of the backbone and the side chains. This example illustrates a general feature of macromolecular association, namely, the coexistence of induced fit and conformational selection. The method presented in this paper can be adapted to incorporate both. Because of the transient nature of weak and ultra-weak binding conformational selection is probably more important than induced fit. Induced fit is most robustly addressed molecular dynamic simulations using explicit water or by Langevin dynamics with the SCP model for consistency.[95] In this brute-force approach each binding mode identified by the method is used as a starting structure in the dynamics. An alternative is to allow backbone conformational changes over the course of the MC simulation. This is most efficiently carried out in the context of scaled collective variables,[96,97] which allows concerted movements of the backbone dihedral angles to improve the acceptance rate. This method has been used previously to study unstructured segments in globular proteins[98] and transmembrane receptors.[42] *A priori* knowledge of flexible segments, e.g., from crystallographic temperature factors or principal component analysis of a dynamic trajectory,[94,99] can reduce the computational cost by restricting collective movements to those regions only.[98] On the other hand, conformational selection can be incorporated in a straightforward manner with no additional modifications of the method presented in this paper. However, this requires identifying structural families of each molecule in solution prior to binding. Each conformation can then be treated independently. Induced fit can in turn be introduced in each sub-system as described. Identifying structural families in solution is not straightforward, and different methods should probably be used depending on the system size. Configurational bias MC simulations (e.g., conformational memories[41]) can

efficiently identify multiple conformers in peptides[41,44] and is probably the preferred method for small systems.

Acknowledgments

This study utilized the high-performance computer capabilities of the Biowulf PC/Linux cluster at the NIH. This work was supported by the NIH Intramural Research Program through the CIT and NINDS, and by the Internal NIST Research Fund.

References

1. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell. 2005; 122:957–968. [PubMed: 16169070]
2. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al. Global Landscape of Protein Complexes in the Yeast Saccharomyces Cerevisiae. Nature. 2006; 440:637–643. [PubMed: 16554755]
3. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al. The Molecular Architecture of the Nuclear Pore Complex. Nature. 2007; 450:695–701. [PubMed: 18046406]
4. Herman ML, Farasat S, Steinbach PJ, Wei MH, Toure O, Fleckman P, Blake P, Bale SJ, Toro JR. Transglutaminase-1 (TGM1) Gene Mutations in Autosomal Recessive Congenital Ichthyosis: Summary of Mutations (Including 23 Novel) and Modeling of TGase-1. Human Mutation. 2009; 30:537–547. [PubMed: 19241467]
5. Prusiner SB. Prions. Proc. Nat. Acad. Sci. (USA). 1998; 95:13363–13383. [PubMed: 9811807]
6. Colland F, Jacq X, Trouplin V, Mougin C, Groizeleau C, Hamburger A, Meil A, Wojcik J, Legrain P, Cauthier JM. Functional Proteomics Mapping of a Human Signaling Pathway. Genome Res. 2004; 14:1324–1332. [PubMed: 15231748]
7. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, et al. A Protein Interaction Network Links GIT1, an Enhancer of Huntingtin Aggregation, to Huntington's Disease. Mol. Cell. 2004; 15:853–865. [PubMed: 15383276]
8. Nienhaus, GU., editor. Protein-Ligand Interactions: Methods and Applications. Humana Press; 2005.
9. Clore GM, Tang C, Iwahara J. Elucidating Transient Macromolecular Interactions using Paramagnetic Relaxation Enhancement. Curr. Op. Struc. Biol. 2007; 17:603–616.
10. Tang C, Iwahara J, Clore GM. Visualization of Transient Encounter Complexes in Protein-Protein Association. Nature. 2006; 444:383–386. [PubMed: 17051159]
11. Tang C, Ghirlando R, Clore G. Visualization of Transient Ultra-Weak Protein Self-Association in Solution using Paramagnetic Relaxation Enhancement. J. Amer. Chem. Soc. 2008; 130:4048–4056. [PubMed: 18314985]
12. Ellis RJ. Macromolecular Crowding: Obvious but Underappreciated. TIBS. 2001; 26:597–604. [PubMed: 11590012]
13. Luby-Phelps K. Cytoarchitecture and Physical Properties of Cytoplasm: Volume, Viscosity, Diffusion, Intracellular Surface Area. Int. Rev. Cytol. 2000; 192:189–221. [PubMed: 10553280]
14. Tuffery P, Derremaux P. Flexibility and Binding Affinity in Protein-Ligand, Protein-Protein and Multi-Component Protein Interactions: Limitations of Current Computational Approaches. J. R. Soc. Interface. 2012; 9:20–33. [PubMed: 21993006]
15. Meiler J, Baker D. Rosetaligand: Protein-Small Molecule Docking with Full Side-Chain Flexibility. Proteins. 2006; 65:538–548. [PubMed: 16972285]
16. Hassan SA, Steinbach PJ. Water-Exclusion and Liquid-Structure Forces in Implicit Solvation. J. Phys. Chem. B. 2011; 115:14608–14682.
17. Pascal TA, Goddard WA III, Maiti PK, Vaidehi N. Role of Specific Cations and Water Entropy on the Stability of Branched DNA Motif Structures. J. Phys. Chem. B. 2012; 116:12159–12167. [PubMed: 22998030]

18. Oleinikova A, Sasisanker P, Weingartner H. What can really be Learned from Dielectric Spectroscopy of Protein Solutions? A Case Study of Ribonuclease A. J. Phys. Chem. 2004; 108:8467–8474.
19. Schroder C, Rudas T, Boresch S, Steinhauser O. Simulation Studies of the Protein-Water Interface: I. Properties at the Molecular Resolution. J. Chem. Phys. 2006; 124:234907. [PubMed: 16821953]
20. Rudas T, Schroder C, Boresch S, Steinhauser O. Simulation Studies of the Protein-Water Interface. II. Properties at the Mesoscopic Resolution. J. Chem. Phys. 2006; 124:234908. [PubMed: 16821954]
21. Merzel F, Smith JC. Is the First Hydration Shell of Lysozyme of Higher Density than Bulk Water? Proc. Nat. Acad. Sci. (USA). 2002; 99:5378–5383. [PubMed: 11959992]
22. Loffler G, Schreiber H, Steinhauser O. Calculation of the Dielectric Properties of a Protein and its Solvent: Theory and a Case Study. J. Mol. Biol. 1997; 270:520–534. [PubMed: 9237916]
23. Schellman JA. Fifty Years of Solvent Denaturation. Biophys. Chem. 2002; 96:91–101. [PubMed: 12034431]
24. Timasheff SM. The Control of Protein Stability and Association by Weak Interactions with Water: How Do Solvents Affect These Processes? Annu. Rev. Biophys. Biomol. Struct. 1993; 22:67–97. [PubMed: 8347999]
25. Arakawa K, Timasheff SM. The Stability of Proteins by Osmolytes. Biophys. J. 1985; 47:411–414. [PubMed: 3978211]
26. Mancinelli R, Botti A, Bruni F, Ricci MA, Soper AK. Perturbation of Water Structure due to Monovalent Ions in Solution. Phys. Chem. Chem. Phys. 2007; 9:2959–2967. [PubMed: 17551619]
27. Parsegian VA. Protein-Water Interactions. Int. Rev. Cytol. 2002; 215:1–31. [PubMed: 11952225]
28. Parsegian VA, Rau DC. Water near Intracellular Surfaces. J. Cell Biol. 1984; 99:196–200.
29. Zheng J-M, Pollack GH. Long-range Forces Extending from Polymer-Gel Surfaces. Phys. Rev. E. 2003; 68:031408.
30. Larsen AE, Grier DG. Like-Charge Attractions in Metastable Colloidal Crystallites. Nature. 1997; 385:230–233.
31. Crocker JC, Grier DG. When Like Charges Attract: The Effect of Geometrical Confinement on Long-Range Colloidal Interactions. Phys. Rev. Lett. 1996; 77:1897–1900. [PubMed: 10063199]
32. Hassan, SA.; Mehler, EL. In Silico Approaches to Structure and Function of Cell Components and their Assemblies: Molecular Electrostatics and Solvent Effects. In: Egelman, E., editor. Comprehensive Biophysics. Vol. Vol. 9. Oxford: Academic Press; 2012. p. 190-228.
33. Halle B. Protein Hydration Dynamics in Solution: a Critical Survey. Phil. Trans. R. Soc. Lond. B. 2004; 359:1207–1223. [PubMed: 15306377]
34. Frolich A, Gabel F, Jasnin M, Lehnert U, Oesterhelt D, Stadler M, Tehei M, Weik M, Wood K, Zaccai G. From Shell to Cell: Neutron Scattering Studies of Biological Water Dynamics and Coupling to Activity. Faraday Disc. 2009; 141:117–130.
35. Tehei M, Franzetti B, Wood K, Gabel F, Fabiani E, Jasnin M, Zamponi D, Oesterhelt D, Zaccai G. Neutron Scattering Reveals Extremely Slow Cell Water in Dead Sea Organism. Proc. Nat. Acad. Sci. (USA). 2007; 104:766–771. [PubMed: 17215355]
36. Siepmann, JI. Configurational-bias Monte Carlo: Background and Selected Applications. In: van Gunsteren, WF.; Weiner, PK.; Wilkinson, AJ., editors. Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications. Vol. Vol. 2. Leiden: ESCOM; 1993. p. 249-264.
37. Siepmann JI, Frenkel D. Configurational Bias Monte Carlo: A New Sampling Scheme for Flexible Chains. Molecular Physics. 1992; 75:59–70.
38. de Pablo JJ, Jain TS. A biased Monte Carlo Technique for Calculation of the Density of States of Polymer Films. J. Chem. Phys. 2002; 116:7238–7244.
39. Falcioni M, Deem MW. A biased Monte Carlo Scheme for Zeolite Structure Solution. J. Chem. Phys. 1999; 110:1754–1767.
40. Steinbach PJ. Exploring Peptide Energy Landscapes: A Test of Force Fields and Implicit Solvent Models. Proteins. 2004; 57:665–677. [PubMed: 15390266]

41. Guarnieri F, Weinstein H. Conformational Memories and the Exploration of Biologically Relevant Peptide Conformations: An Illustration for the Gonadotropin-releasing Hormone. J Amer. Chem. Soc. 1996; 118:5580–5589.
42. Mehler EL, Hassan SA, Kortagere S, Weinstein H. Ab initio Computer Modeling of Loops in G-Protein Coupled Receptors: Lessons from the Crystal Structure of Rhodopsin. Proteins. 2006; 64:673–690. [PubMed: 16729264]
43. Hassan SA, Mehler EL. A General Screened Coulomb Potential Based Implicit Solvent Model: Calculation of Secondary Structure of Small Peptides. Int. J. Quant. Chem. 2001; 83:193–202.
44. Hassan SA, Guarnieri F, Mehler EL. Characterization of Hydrogen Bonding in a Continuum Solvent Model. J. Phys. Chem. B. 2000; 104:6490–6498.
45. Hassan SA, Guarnieri F, Mehler EL. A General Treatment of Solvent Effects Based on Screened Coulomb Potentials. J. Phys. Chem. B. 2000; 104:6478–6489.
46. Hassan SA, Mehler EL, Zhang D, Weinstein H. Molecular Dynamics Simulations of Peptides and Proteins with a Continuum Electrostatic Model Based on Screened Coulomb Potentials. Proteins. 2003; 51:109–125. [PubMed: 12596268]
47. Hassan SA. Liquid-structure Forces and Electrostatic Modulation of Biomolecular Interactions in Solution. J. Phys. Chem. B. 2007; 111:227–241. [PubMed: 17201447]
48. Brooks BR, Brooks CL III, MacKerrel ADM Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al. CHARMM: The Biomolecular Simulation Program. Comp. Chem. 2009; 30:1545–1614.
49. Juneja A, Ito M, Nilsson L. Implicit Solvent Models and Stabilizing Effects of Mutations and Ligand on the Unfolding of the Amyloid β-Peptide Central Helix. J. Chem. Theory Comput. 2013; 9:834–846.
50. Schreiber G, Fersht AR. Rapid Electrostatically Assisted Association of Proteins. Nature. 1996; 3:427–431.
51. Xu X-HN, Yeung ES. Long-range Electrostatic Trapping of Single-Protein Molecules at a Liquid-Solid Interface. Science. 1998; 281:1650–1653. [PubMed: 9733506]
52. Gray JJ. The Interaction of Proteins with Solid Surfaces. Curr. Op. Struc. Biol. 2004; 14:110–115.
53. Hassan SA. Intermolecular Potentials of Mean Force of Amino Acid Side Chain Interactions in Aqueous Medium. J. Phys. Chem. B. 2004; 108:19501–19509.
54. Okur A, Miller BT, Joo K, Lee JA, Brooks BR. Generating Reservoir Conformations for Replica Exchange through the Use of the Conformational Space Annealing Method. J. Chem. Theory Comput. 2013; 9:1115–1124. [PubMed: 23585739]
55. Lee LP, Tidor B. Barstar is Electrostatically Optimized for Tight Binding to Barnase. Nature. 2001; 8:73–76.
56. Frisch C, Schreiber G, Johnson CM, Fersht AR. Thermodynamics of the Interaction of Barnase and Barstar: Changes in Free Energy versus Changes in Enthalpy on Mutation. J. Mol. Bio. 1997; 267:696–706. [PubMed: 9126847]
57. Vajda S, Weng ZP, Rosenfeld R, DeLisi C. Effect of Conformational Flexibility and Solvation on Receptor-Ligand Binding Free Energies. Biochemistry. 1994; 33:13977–13988. [PubMed: 7947806]
58. Malham R, Johnstone S, Bingham RJ, Barratt E, Phillips SEV, Laughton CA, Homans SW. Strong Solute-Solute Dispersive Interactions in a Protein-Ligand Complex. J. Amer. Chem. Soc. 2005; 127:17061–17067. [PubMed: 16316253]
59. Floris F, Tomasi J. Evaluation of the Dispersion Contribution to the Solvation Energy: A Simple Computational Model in the Continuum Approximation. J. Comput. Chem. 1989; 10:616–627.
60. Zacharias M. Continuum Solvent Modeling of Nonpolar Solvation: Improvement by Separating Surface Area dependent Cavity and Dispersion Contributions. J. Phys. Chem. A. 2003; 107:3000–3004.
61. Wagoner JA, Baker NA. Assessing Implicit Models for Nonpolar Mean Solvation Forces: The Importance of Dispersion and Volume Terms. Proc. Nat. Acad. Sci. (USA). 2006; 103:8331–8336. [PubMed: 16709675]
62. Durell SR, Brooks BR, Ben-Naim A. Solvent-Induced Forces Between Two Hydrophilic Groups. J. Phys. Chem. 1994; 98:2198–2202.

63. Ben-Naim A. Solvent-Induced Forces in Protein Folding. J. Phys. Chem. 1990; 94:6893–6895.
64. Bruge F, Fornilli SL, Malenkov GG, Palma-Vittorelli MB, Palma MU. Solvent-Induced Forces on a Molecular Scale: Non-Additivity, Modulation and Causal Relation to Hydration. Chem. Phys. Lett. 1996; 254:283–291.
65. Tanford C. Interfacial Free Energy and the Hydrophobic Effect. Proc. Nat. Acad. Sci. (USA). 1979; 76:4175–4176. [PubMed: 16592699]
66. Hermann RB. Theory of Hydrophobic Bonding. II. Correlation of Hydrocarbon Solubility in Water with Solvent Cavity Surface-Area. J. Phys. Chem. 1972; 76:2754–2759.
67. Szekely GJ, Rizzo ML. Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. J. Classif. 2005; 22:151–183.
68. Allen, MP.; Tildesley, DJ. Computer Simulation of Liquids. Oxford: Clarendon Press; 1987.
69. Gabdoulline RR, Wade RC. Protein-Protein Association: Investigation of Factors Influencing Association Rates by Brownian Dynamics Simulations. J Mol Biol. 2001; 306:1139–1155. [PubMed: 11237623]
70. Hoefling M, Gottschalk KE. Barnase-Barstar: From First Encounter to Final Complex. J. Struct. Biol. 2010; 171:52–63. [PubMed: 20211732]
71. Shan J, Mehler EL. Calculation of pKa in Proteins with the Microenvironment Modulated-Screened Coulomb Potential (MM-SCP). Proteins. 2011; 79:3346–3355. [PubMed: 21748803]
72. Hassan SA. Self-Consistent Treatment of the Local Dielectric Permittivity and Electrostatic Potential in Solution for Polarizable Macromolecular Force Fields. J. Chem. Phys. 2012; 137:074102. [PubMed: 22920098]
73. Hassan SA. Amino Acid Side Chain Interactions in the Presence of Salts. J. Phys. Chem. B. 2005; 109:21989–21996. [PubMed: 16479276]
74. Masella M, Borgis D, Cuniasse P. A Multiscale Coarse-Grained Polarizable Solvent Model for Handling Long Tail Bulk Electrostatics. J. Comput. Chem. 2013; 34:1112–1124. [PubMed: 23382002]
75. Hassan, SA.; Mehler, EL. Modeling Aqueous Solvent Effects through Local Properties of Water. In: Feig, M., editor. Modeling Solvent Environments: Applications to Simulation of Biomolecules. Weinheim: Wiley-VCH; 2010.
76. Chandler D. Interfaces and the Driving Force of Hydrophobic Assembly. Nature. 2005; 437:640–647. [PubMed: 16193038]
77. Jensen TR, Ostergaard M, Reitzel N, Balashev K, Peters GH, Kjaer K, Bjornholm T. Water in Contact with Extended Hydrophobic Surfaces: Direct Evidence of Weak Dewetting. Phys. Rev. Lett. 2003; 90:086101. [PubMed: 12633443]
78. Pratt LR. Molecular theory of Hydrophobic Effects: She is too Mean to have her Name Repeated. Annu. Rev. Phys. Chem. 2002; 53:409–436. [PubMed: 11972014]
79. Hummer G, Garde S, Garcia AE, Pratt EA. New Perspectives on Hydrophobic Effects. Chem. Phys. 2000; 258:349–370.
80. Lum K, Chandler D, Weeks JD. Hydrophobicity at Small and Large Length Scales. J. Phys. Chem. B. 1999; 103:4570–4577.
81. Ashbaugh HS, Kaler EW, Paulaitis ME. A "Universal" Surface Area Correlation for Molecular Hydrophobic Phenomena. J. Am. Chem. Soc. 1999; 121:9243–9244.
82. Wallqvist A, Gallicchio E, Levy RM. A Model for Studying Drying at Hydrophobic Interfaces: Structural and Thermodynamic Properties. J. Phys. Chem. B. 2001; 105:6745–6753.
83. Cramer CJ, Truhlar DG. An SCF Solvation Model for the Hydrophobic Effect and Absolute Free Energies of Aqueous Solvation. Science. 1992; 256:213–217. [PubMed: 17744720]
84. Wagner F, Simonson T. Implicit Solvent Models: Combining an Analytical Formulation of Continuum Electrostatics with Simple Models of the Hydrophobic Effect. J. Comp. Chem. 1999; 20:322–335.
85. Tan ML, Cendagorta JR, Ichiye T. Effects of Microcomplexity on Hydrophobic hydration in Amphiphiles. J. Amer. Chem Soc. 2013; 135:4918–4921.

86. Giovambattista N, Lopez CF, Rossky PJ, Debenedetti PG. Hydrophobicity of Protein Surfaces: Separating Geometry from Chemistry. Proc. Nat. Acad. Sci. (USA). 2008; 105:2274–2279. [PubMed: 18268339]
87. Yamniuk AP, Edavettal SC, Bergqvist S, Yadav SP, Doyle ML, Calabrese K, Parsons JF, Eisenstein E. ABRF-MIRG Benchmark Study: Molecular Interactions in a Three-Component System. J. Biomol. Tech. 2012; 23:101–114. [PubMed: 22942790]
88. Kleanthous, C., editor. Protein-Protein Recognition. New York: Oxford University Press; 2000.
89. Strynadka NCJ, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kunts I, Abagyan R, Totrov R, Janin J, Cherfils J, Zimmermann F, et al. Molecular Docking Programs Successfully determine the Binding of a β-lactamase Inhibitory Protein to term-1 β-Lactamase. Nature Struct. Biol. 1996; 3:233–239. [PubMed: 8605624]
90. Lensink MF, Mendez R, Wodak SJ. Docking and scoring protein complexes: Capri 3rd Edition. Proteins. 2007; 69:704–718. [PubMed: 17918726]
91. Ritchie DW. Recent Progress and Future Directions in Protein-Protein Docking. Curr. Protein Pept. Sci. 2008; 9:1–15. [PubMed: 18336319]
92. Vakser JA, Kundrotas P. Predicting 3D Structures of Protein-Protein Complexes. Curr. Pharm. Biotechnol. 2008; 9:57–66. [PubMed: 18393862]
93. Chen Y, Varani G. Protein Families and RNA Recognition. FEBS J. 2005; 272:2088–2097. [PubMed: 15853794]
94. Lange OF, Lakomek N-A, Faris C, Schroder GF, Walter KFA, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL. Recognition Dynamics up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution. Science. 2008; 320:1471–1475. [PubMed: 18556554]
95. Li X, Hassan SA, Mehler EL. Long Dynamics Simulations of Proteins using Atomistic Force Fields and a Continuum Representation of Solvent Effects: Calculation of Structural and Dynamic Properties. Proteins. 2005; 60:464–484. [PubMed: 15959866]
96. Go N, Noguti T, Nishikawa T. Dynamics of a Small Globular Protein in terms of Low-Frequency Vibrational Modes. Proc. Nat. Acad. Sci. (USA). 1983; 80:3696–3700. [PubMed: 6574507]
97. Noguti T, Go N. Efficient Monte Carlo Method for Simulation of Fluctuating Conformations of Native Proteins. Biopolymers. 1985; 24:527–546. [PubMed: 3986295]
98. Hassan, SA.; Mehler, EL.; Weinstein, H. Structure Calculations of Protein Segments Connecting Domains with Defined Secondary Structure: A Simulated Annealing Monte Carlo Combined with Biased Scaled Collective Variables Technique. In: Hark, K.; Schlick, T., editors. Lecture Notes in Computational Science and Engineering. Vol. Vol. 24. New York: Springer; 2002. p. 197-231.
99. Cardone A, Hassan SA, Albers RW, Sriram RD, Pant HC. Structural and Dynamic Determinants of Ligand Binding and Regulation of Cyclin-Dependent Kinase 5 by Pathological Activator p25 and Inhibitory Peptide CIP. J. Mol. Bio. 2010; 401:478–492. [PubMed: 20599546]

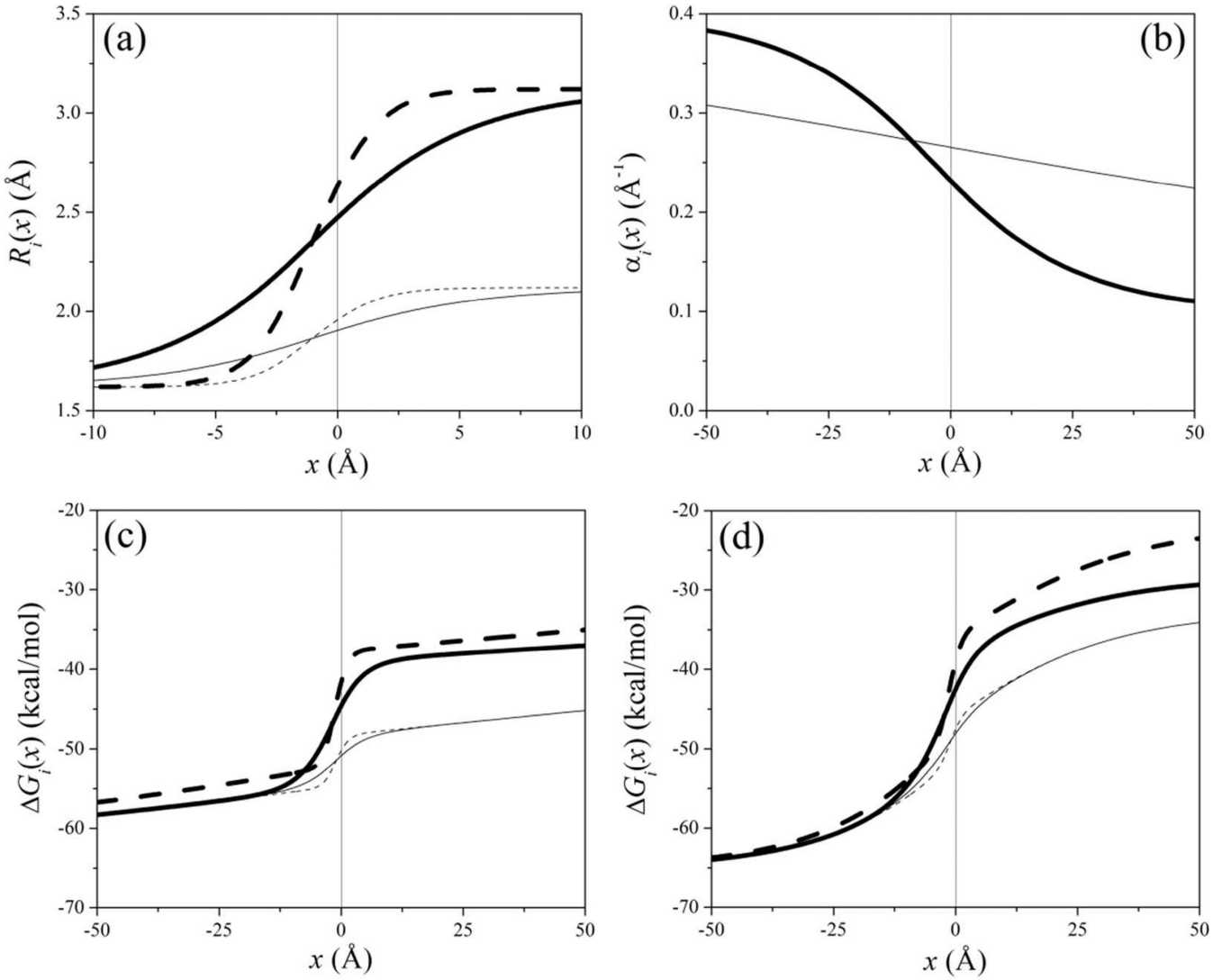

Figure 1.

Behavior of (a) the effective radius R, (b) the screening parameter α, and (c, d) the self-energy ΔG of a point charge q = +1 close to a planar aqueous interface, as described by the SCP continuum solvent model [Eqs. (1–4)]. The interface is located at x = 0, with water filling the space x < 0 and an idealized protein occupying the region x > 0. (a) From Eq. (4), using $a = R_w$ + 1.5 Å (thick lines) and $a = R_w$ + 0.5 Å (thin), for τ = 3.125 Å (solid) and τ = 1.0 Å (dashed). The parameter a determines the total change in effective radius between bulk water ($x \to -\infty$) and bulk protein (complete dehydration, $x \to \infty$); τ determines the rate of change as the particle crosses the interface. (b) From Eq. (2), using σ = 15 Å (thick); σ = 75 Å (thin). (c) From Eq. (1) with σ = 75 Å and effective radii plotted in panel (a). (d) Same as in (c), using σ = 15 Å. The protein, which determines the planar aqueous interface, was modeled as two superimposed three-dimensional cubic lattices, one representing the positions of C_α atoms in Eq. (2), with a side length of 7 Å; and the other one representing the position of all atoms in Eq. (4), with a side length of 2 Å (assuming an average volume of ~180 Å^3 per amino acid,[16] and ~20 atoms per amino acid).

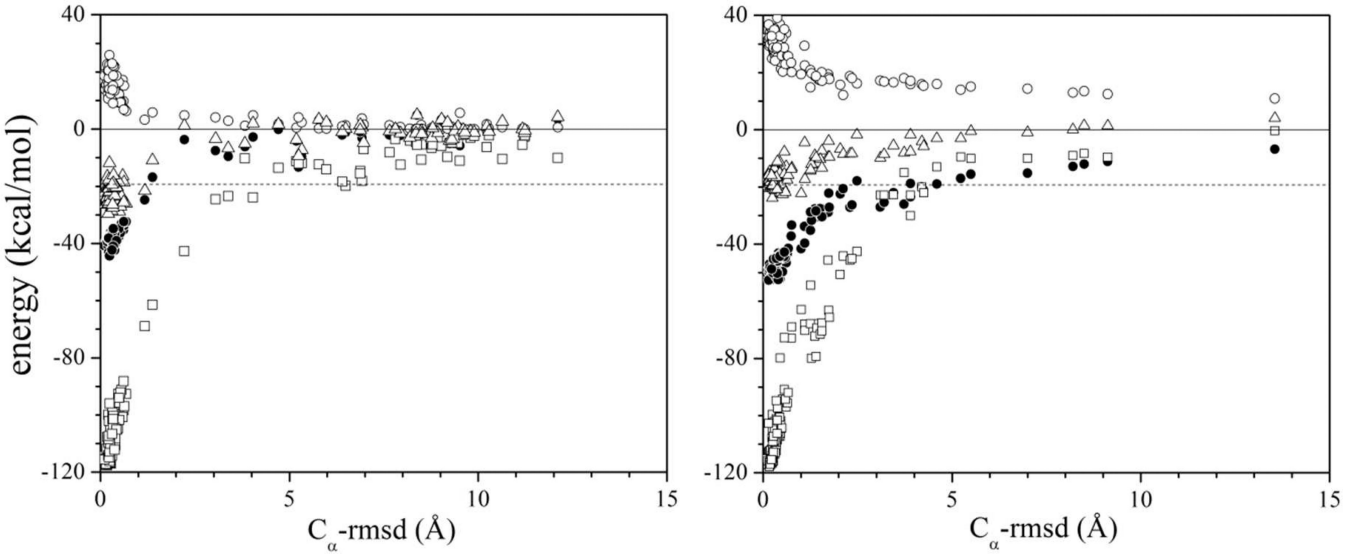

Figure 2.

Non-bonded energy decomposition of the barnase-barstar complex during dissociation by heating, calculated with the SCP continuum solvent model as implemented in the CHARMM program (version c35b4), with (a) and without (b) long-range water-exclusion effects: protein-protein van der Waals energy (squares), electrostatic interaction energy [black circles; first term in Eq. (1)], and self-energy [open circles; second term in Eq. (1)]. The total electrostatic energy E_e [Eq. (1)] of the system is also shown (triangles) and determines the dissociation enthalpy. The reaction coordinate is the C_α-rmsd with respect to the crystal structure of the complex (PDB 1brs). The limits $\sigma \rightarrow \infty$ and $\sigma' \rightarrow \infty$ in Eqs. (2) and (3) lead to over-stabilization of the complex by ~3.5 kcal/mol. Optimized values $\sigma = 59$ Å and $\sigma' = 37$ Å lead to a dissociation energy equal to the measured $\Delta H_b = 19.3$ kcal/mol of the complex.

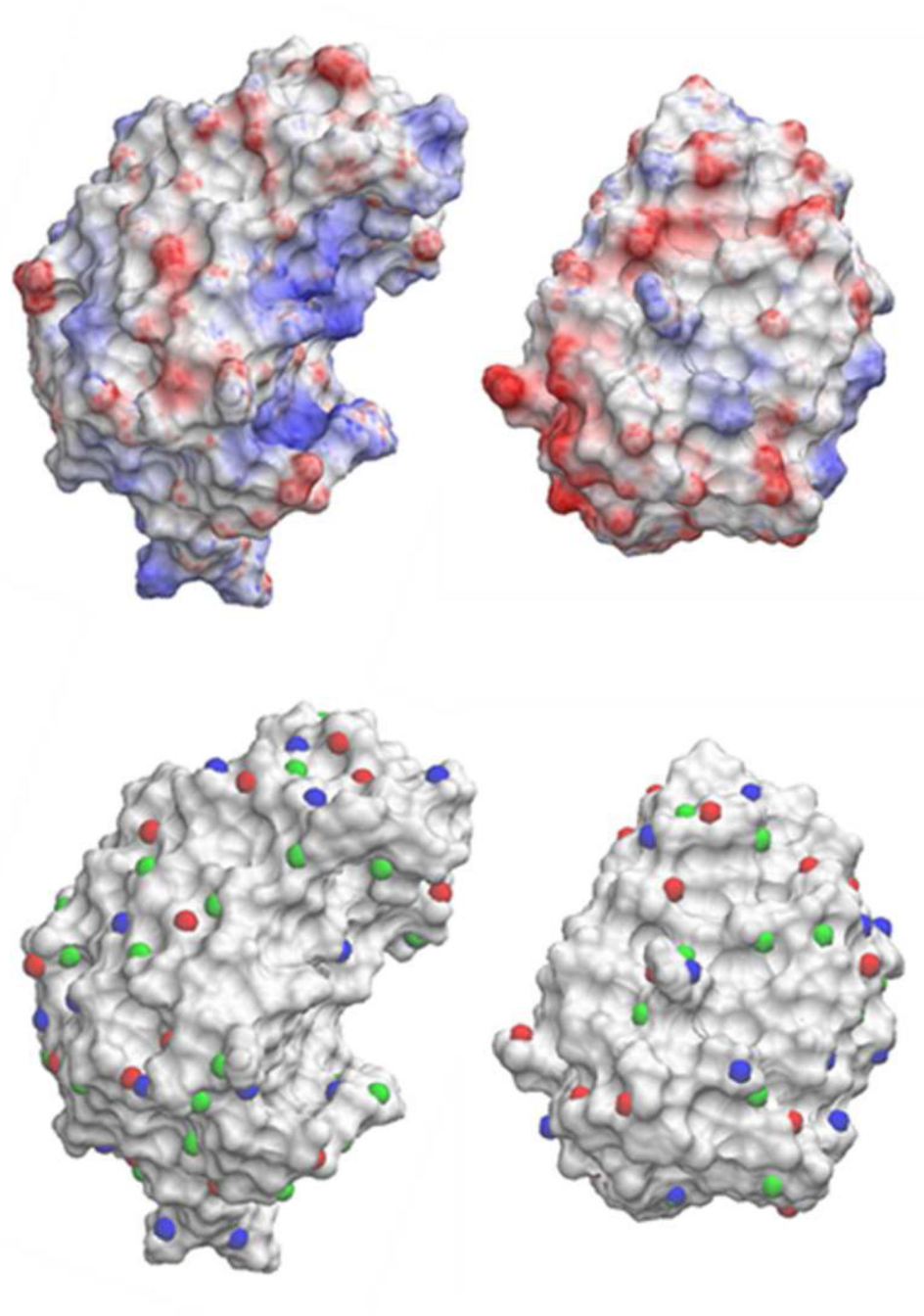

Figure 3.
Electrostatic potential (upper panel) on the molecular surface of barnase (left) and barstar (right) calculated from conventional numerical solutions of the Poisson equation ($\varepsilon_p = 2$; $\varepsilon_w = 78$). Positions of maxima (blue) and minima (red) and non-polar centers (green) are calculated from Eq. (5) and (6) (lower panel).

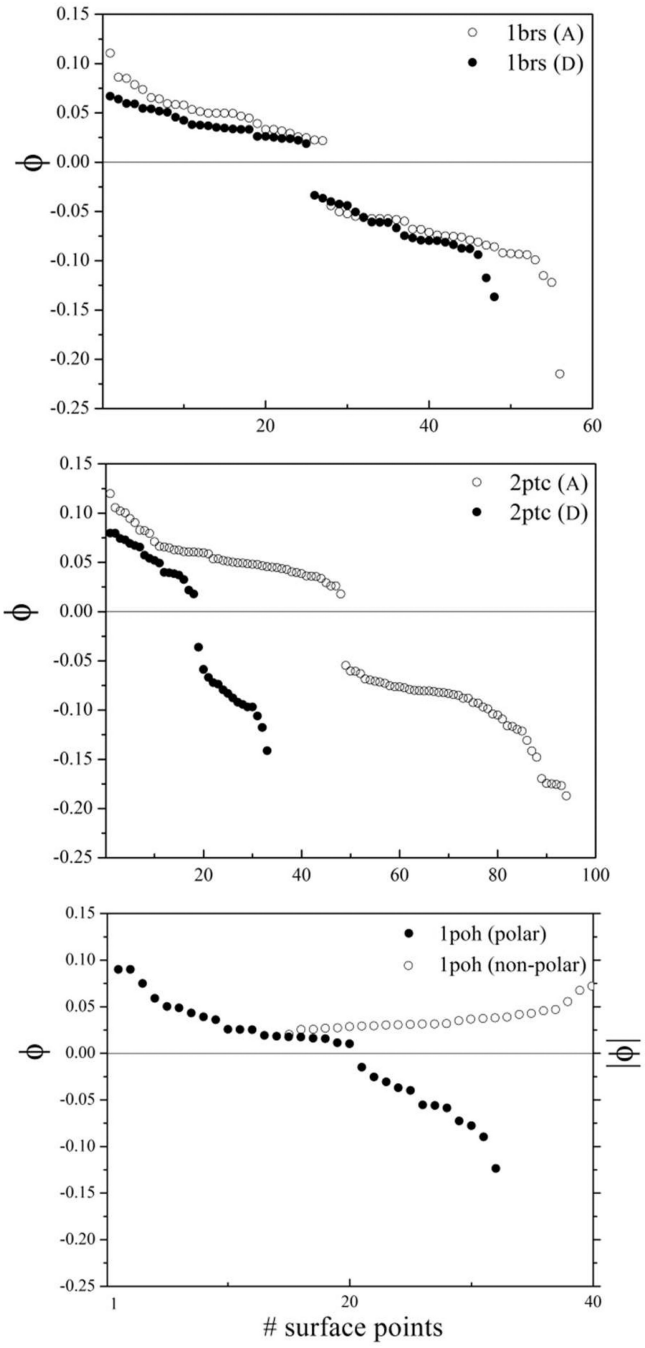

Figure 4.

Values (in Volt) of the maxima and minima of the surface potentials φ in the three systems studied (chain A in 1brs is barnase and trypsin in 2ptc). The absolute values ψ = | φ | of the surface potential at the non-polar centers in 1poh are also shown.

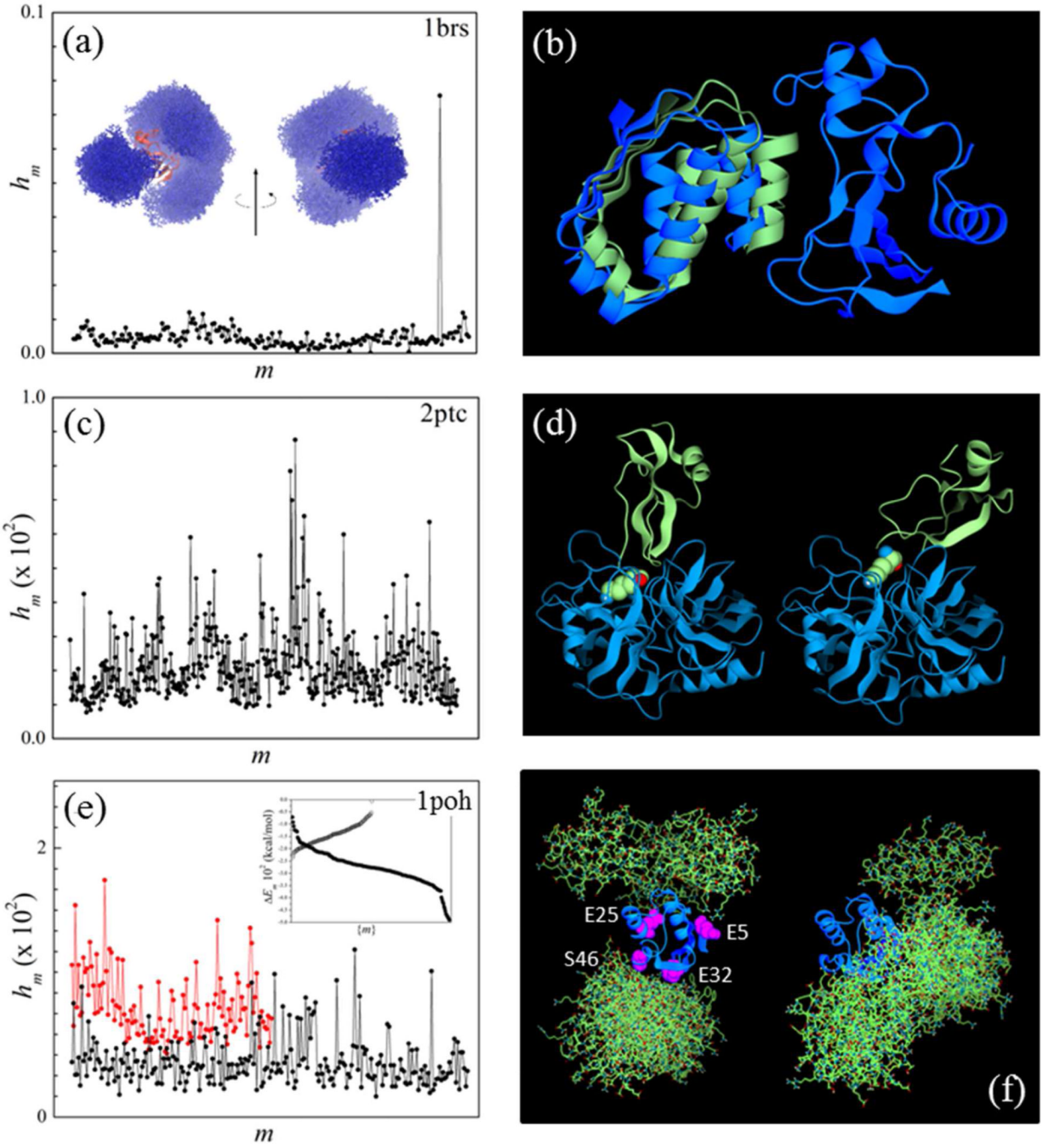

Figure 5.

(a) Probability distribution of prescreened modes of the barnase/barstar complex calculated from Eq. (17) with λ = 25. The mode *m*' with the highest weight $h_{m'}$ is near-native. Inset: prescreened modes of barstar (atom representation; blue) and barnase (ribbon; red) obtained upon optimization of the electrostatic norm *e* [Eq. (7)]. These putative electrostatic-driven first-contact modes determine the biasing function for the self-adaptive conformational bias MC sampling. (b): Mode *m*' (blue) and crystal structure (green) of barnase bound to barstar. (c) Same as in (a) for 2ptc. (d) Trypsin/inhibitor complex (2ptc): crystal structure (left) and prescreened mode *m*' with the smallest C_α-rmsd with respect to the crystal structure (right);

m' has the fifth highest weight in (c); K15 of the inhibitor protein is shown (purple). (e): Same as in (a) 1poh (black); probability distribution obtained upon optimization of the hydrophobic norm *h* [Eq. (8)] is also shown (red). Inset: energy of pre-screened modes (polar: solid circles; non-polar: open circles); (f) Histidine-containing phosphocarrier protein HPr (1poh): conformations of the ten highest h_m modes obtained upon optimization of *e* (*electrostatic modes*; left) and *h* (*hydrophobic modes*; right) superimposed to a central HPr protein; amino acids used as labels in a recent NMR study of ultra-weak self-association are shown.

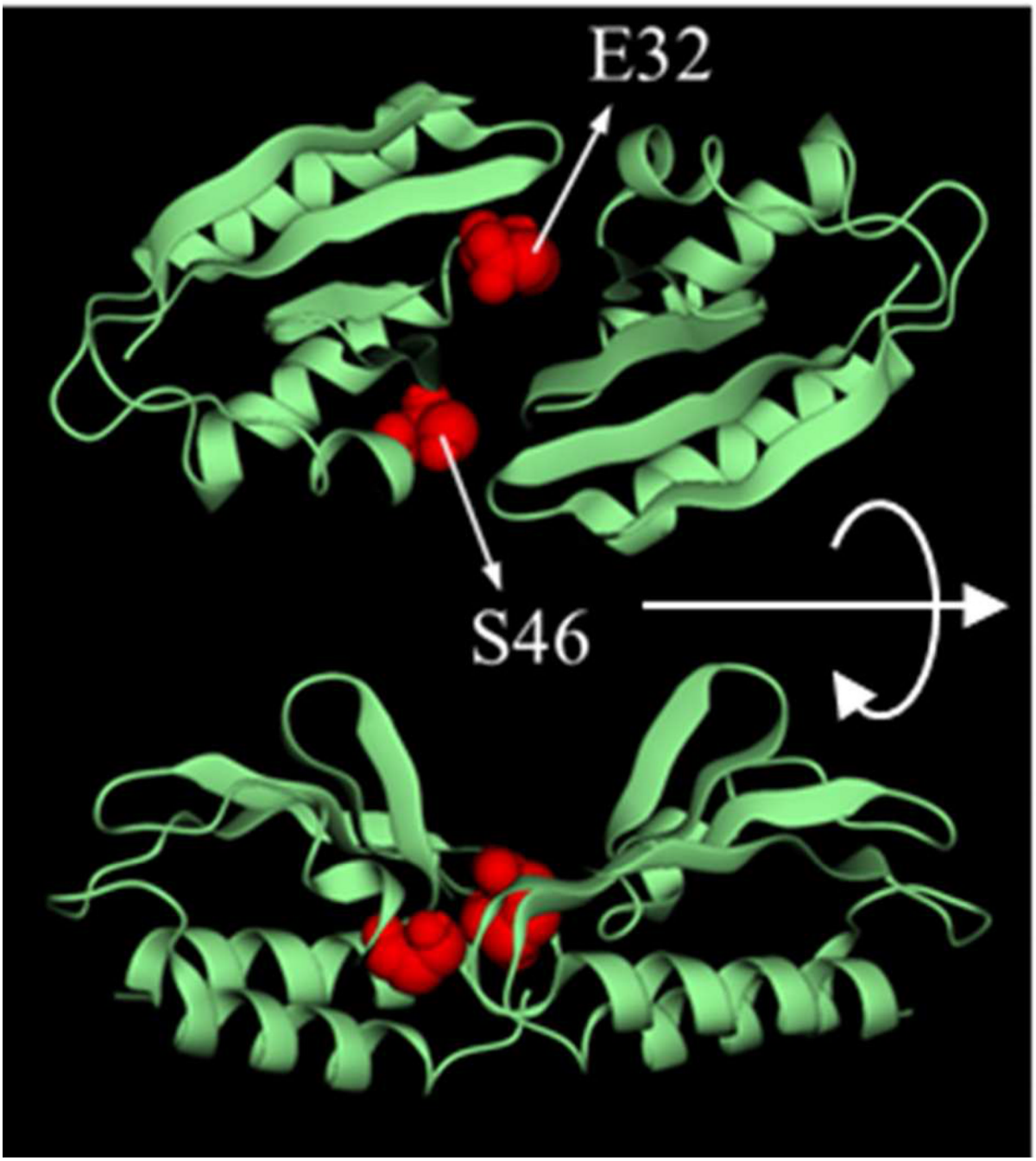

Figure 6.
Symmetrical homodimer (representative member of the ensemble) obtained upon convergence of the self-adaptive biased MC simulation. The binding energy of this state is estimated at ~1.3 kcal/mol.