

HHS Public Access

Author manuscript *J Comput Chem.* Author manuscript; available in PMC 2016 May 15.

Published in final edited form as:

J Comput Chem. 2015 May 15; 36(13): 983–995. doi:10.1002/jcc.23883.

Detection and characterization of nonspecific, sparselypopulated binding modes in the early stages of complexation

A. Cardone^{1,2}, A. Bornstein³, H. C. Pant⁴, M. Brady¹, R. Sriram¹, and S. A. Hassan^{5,*}

¹Software and System Division, *National Institute of Standards and Technology*, Gaithersburg, MD 20899

²Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742

³Mathematics and Computer Science, Goucher College, Baltimore, MD 21204

⁴Laboratory of Neurochemistry, NINDS, *National Institutes of Health*, Bethesda, MD 20892

⁵Center for Molecular Modeling, Division of Computational Bioscience, CIT, *National Institutes of Health*, Bethesda, MD 20892

Abstract

A method is proposed to study protein-ligand binding in a system governed by specific and nonspecific interactions. Strong associations lead to narrow distributions in the proteins configuration space; weak and ultra-weak associations lead instead to broader distributions, a manifestation of non-specific, sparsely-populated binding modes with multiple interfaces. The method is based on the notion that a discrete set of preferential first-encounter modes are metastable states from which stable (pre-relaxation) complexes at equilibrium evolve. The method can be used to explore alternative pathways of complexation with statistical significance and can be integrated into a general algorithm to study protein interaction networks. The method is applied to a peptide-protein complex. The peptide adopts several low-population conformers and binds in a variety of modes with a broad range of affinities. The system is thus well suited to analyze general features of binding, including conformational selection, multiplicity of binding modes, and nonspecific interactions, and to illustrate how the method can be applied to study these problems systematically. The equilibrium distributions can be used to generate biasing functions for simulations of multiprotein systems from which bulk thermodynamic quantities can be calculated.

Keywords

protein-protein association; protein aggregation; complex formation; configurational bias Monte Carlo; nonspecific interactions; solvent effects; implicit solvent model

^{*}hassan@mail.nih.gov.

Disclaimer: Certain commercial equipment, instruments, or materials are identified in this report in order to specify the experimental or computational procedure adequately. Such identification is not intended to imply recommendation or endorsement by the NIST or NIH, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

I. Introduction

Cellular signal transduction involves complex networks of protein-protein interactions, the topological properties of which have become the focus of many experimental and theoretical studies.^{1–4} A clear understanding of networks behavior may pave the way for the development of strategies to intervene at the level of vertices, edges, and hubs to modify cellular function in pathological conditions.^{5,6} It has been recognized that, to achieve this goal, the networks need to be characterized first in terms of the physicochemical properties of the individual interaction.^{2,7,8} Early attempts in this direction related network topology to the three-dimensional structures of proteins.² Microcalorimetric, kinetic, and spectroscopic methods are commonly used to probe protein-protein interactions, but microscopic insight is often limited.9 Advanced techniques are being developed to gain atomic-resolution insight, including weak and ultra-weak association,¹⁰⁻¹² detection of sparsely-populated conformers and metastable complex intermediates,^{13,14} and tracking of real-time conformational changes upon binding.^{15,16} The importance of computer simulations in this context is selfevident, but the challenges are still numerous. These include sampling of the configuration space, as well as more fundamental questions related to the physical nature of the interactions, such as the role of the aqueous medium in mediating/inducing intermolecular forces.

In recent years, a number of experimental studies have evidenced the multifaceted, multilayered complexity that real biological media pose to a computational approach. In living cells, for example, proteins form complexes with an average of 4–5 proteins per complex.¹ The prevalence of monomers is rather low (20%), while dimers constitute a small majority (40%); the remaining proteins form higher order architectures.^{4,17} A recent largescale study^{4,18} has shown that about half of the proteins form homooligomers (homomers) and the rest heterooligomers (heteromers). Both kinds of oligomers play functional and morphological roles, and the therapeutic importance of homomers has been recognized.¹⁸ Homomers are highly symmetrical, and dihedral symmetry is more abundant than either cyclic or cubic symmetry, indicating that proteins can self-associate through multiple interfaces and with different affinities. Proteins in multimeric complexes tend to interact through highly specific contact surfaces, but recent experiments have shown that proteins can also interact at multiple sites through non-specific, ultra-weak interactions.¹⁰⁻¹² These transient associations are difficult to detect with conventional spectroscopic techniques, but they are thought to play an important role in many biological processes, including protein recognition and spontaneous self-assembly of higher-order architectures.¹⁰ Moreover, proteins can be multifunctional and interact specifically with several proteins in different cellular processes throughout the cell cycle;^{1,4,17} up to six interfaces have been identified in some cases.4

Earlier evidence^{19,20} had suggested that a large proportion of proteins in cells are in a state of high condensation, transiently bound to one another, to membranes, or to the cytoskeleton, the most extensive protein complex in eukaryote cells. Electron tomography and modeling have recently provided more direct evidence of the subcellular organization of large protein complexes in a small prokaryote cell.⁴ Data show a homogeneous distribution of complexes throughout the cytosol, punctuated by regions of higher concentration. Local

environments where proteins are in close proximity of one another may facilitate translocation (compared to simple three-dimensional diffusion) and association and may thus be the result of biological evolution. Moreover, local macromolecular crowding can change the strength of protein-protein interactions for reasons (enthalpic) unrelated to conventional (entropic) volume-exclusion effects.²¹ On the one hand, electrostatic interactions between proteins can become stronger due to the exclusion of water (highly polarizable medium) by nearby complexes,²² an effect recently probed by computer simulations in barnase-barstar.²³ In addition, changes in the structural and dynamic behavior of (non-bulk) water at interfaces and interstitial spaces can affect both electrostatic and liquid-structure forces in ways not yet fully understood.^{22,24,25} The combination of these enthalpic and entropic effects may result in the emergence of binding sites or modes not observed at higher dilution.

The summary above is concerned with the statistical properties of proteins in the subcellular environment. The organization and composition of the complexes, however, vary both in time and space. This paper does not directly address the system's dynamic behavior, which is studied once the correct protein/protein interfaces are identified (prediction of 'hot spots'²⁶ on protein surfaces is in and of itself a major goal in most predictive methods^{27–29}). The paper deals instead with the spatial characterization of the system. The method provides insight into the early stages of complex formation, including aggregation and self-assembly, and can be used to explore alternative complexation pathways with statistical significance, which has implications for the study of protein interaction networks.^{2,4,6,30,31}

Section II describes a biased Monte Carlo (MC) algorithm for the efficient sampling of the configurational space. The biasing function allows mixing large and local changes in the spatial distribution of proteins, which enhances sampling of relevant microstates. The method is based on the notion that a discrete set of preferential first-encounter modes are metastable states from which stable complexes at equilibrium evolve. Strong, specific interactions lead to a single binding mode,³² whereas weak or ultra-weak interactions lead to a distribution of sparsely-populated, non-specific modes with multiple interfaces. The solvent model used to describe protein-ligand interactions is reviewed in Section III, which represents the conditions of partial and anisotropic hydration typical in these systems. The model accounts for short- and long-range effects of water exclusion, including electrostatic and dispersion forces. Long-range electrostatic interactions are optimized using binding enthalpy data of two medium-size binary complexes. The method is used in Section IV to study binding of a flexible peptide to a protein. The peptide adopts several conformations and binds in a variety of modes with different affinities. The selected complex is well suited to analyze general features of binding, including conformational selection, and to illustrate how the method can be used to study such problems systematically. Generalization to multispecies, multiprotein systems is discussed in Section V.

II. Configurational-bias self-adaptive Monte Carlo simulations

A flow chart of the general algorithm is shown in Fig. 1; detailed flowcharts of blocks A and B are presented in Fig. 2. Given two proteins, l and p, in an aqueous solution, a corresponding set of conformers, $\{l_i\}_{i=1,K}$ and $\{p_i\}_{i=1,K'}$, are first identified. All the

conformers are treated as independent structures coexisting in the solution. Any conformer of p can potentially bind any conformer of p or l (and vice versa) and trigger a dynamic response. The problem associated to conformational selection is thus reduced to a combinatorial problem and addressed at this stage. Finding a complete set of conformers, however, poses challenges of its own. For small, drug-like compounds, a plain MC search might suffice. For a protein of known three-dimensional structure, relevant conformers are expected to be structurally similar to the known structure, thus ab initio prediction is not needed. In this case, molecular Dynamics (MD)-based methods may suffice to detect relevant sub-states,³³ for example using principal component analysis³⁴ or related techniques for trajectory analysis. Local, enhanced sampling may still be needed to identify multiple conformations of unstructured segments (loops), which are known to play a role in recognition and binding.³⁵ Medium-size systems of unknown structures are the most challenging, and Section IV deals with this problem. Peptides, for example, are typically unstructured in aqueous solution or may exist in a variety of interconverting conformers. These conformers are difficult to detect with conventional NMR, so advanced methods are being developed, including paramagnetic relaxation enhancement,¹¹ chemical-exchange saturation transfer,³⁶ and CPMG relaxation dispersion.³⁷ These systems are also a challenge to computational methods, and efficient ab initio techniques have been developed to generate conformational canonical ensembles from which a reduced set of conformers can be extracted.38,39

In what follows, labels 1 and 2 refer indistinctly to two interacting proteins (or solutes, in general, including ions and small molecules) or any two conformers adopted by these proteins. The method used to predict their binding modes consists of two consecutive stages:³² a prescreening of binary interactions to identify physically meaningful first-encounter modes, followed by an adaptive configurational biased sampling to identify statistically relevant binding modes at equilibrium. Prescreening involves optimizations of two dimensionless quantities, namely, an electrostatic norm *e* and a hydrophobic norm *h* of fast computation. To define these functions, the electrostatic potential φ on the molecular surface of each solvated protein is first evaluated with a conventional Poisson equation solver.³² The *N*_M and *N*_m positions {**r**_M} and {**r**_m} of the local maxima { φ_M } and minima { φ_m } (polar centers) are then identified; likewise for the *N*_m positions {**r**_m/} of the local minima { $|\varphi_m'|$ } of $|\varphi|$ (nonpolar centers). Pairs of polar centers with potentials of opposite signs are then aligned, as discussed,³² and the two-way norm *e* is defined as (Fig. 3A)

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

where the corresponding proteins are indicated in parentheses in the upper indexes; *j* in the first and second sums denotes the point in protein 2 that is closest to point *i* in protein 1; similarly for the third and fourth sums. The distance between *i* and *j* is represented by r_{ij} , and α and *d* are constants. The term S_{12} prevents structural overlaps and is modeled as an atom-centered hard-sphere potential.

Pairs of nonpolar centers are also aligned,³² and the norm h is defined as (Fig. 3A)

$$h = b^{(1)} \sum_{i=1}^{L^{(1)}} \theta(2R_w - r_{ij}) + b^{(2)} \sum_{i=1}^{L^{(2)}} \theta(2R_w - r_{ij}) + S_{12} \quad (2)$$

where θ is the Heaviside step function, R_w is the radius of a water molecule, and *b* are system-dependent coefficients. Unlike Eq. 1 where *i* and *j* denote other maxima and minima of φ , in Eq. 2 these indexes run over *L* points defined over local surface patches centered in the aligned centers.

The functional forms of *e* and *h* are suitable simplification of the physical effects that each one intends to describe and designed specifically for computational efficiency: Eq. 1 represents local electrostatic complementarity between the surfaces, and Eq. 2 represents the degree of burial of the local hydrophobic surfaces. No assumptions are made about surface complementarity because both proteins may undergo post-binding structural relaxations (cf. Section IV.3). Norm optimizations are carried out by simulated-annealing MC simulations, from which a total of Γ electrostatics-driven modes $\{u_i\}$ and Γ' hydrophobicity-driven modes $\{v_i\}$ are obtained. The complete set of $\Lambda = \Gamma \Gamma'$ prescreened modes $\{s_i\}$ is formed by the union of both sets, i.e., $\{s_i\}_{i=1,\Lambda} = \{u_i\}_{i=1,\Gamma} \cup \{v_i\}_{i=1,\Gamma'}$, and represent preferential encounters between the proteins in their particular conformations under consideration. This set of modes is used to define an initial biasing function, *P*, for the MC simulation, which is adapted as the simulation progresses. Each mode *m* defines the center of a local Gaussian distribution in the conformation space of the complex, and the total distribution over all the modes is given by

$$P(\boldsymbol{\eta}) = \sum_{m=1}^{\Lambda} a h_m \kappa_m \exp(-J_m) \quad (3)$$

where $\mathbf{\eta} \equiv (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ is a set of generalized coordinates that determine the relative position of the proteins; *h* is the Boltzmann probability for selecting the mode *m*; $\kappa_m = 1/\prod_{i=1}^6 \sigma_{im}$ and $a = \prod_{i=1}^6 \delta \eta_i / 8\pi^3$, where σ_i is the standard deviation and $\delta \eta_i$ the unit distance along the η_i coordinate in the $\mathbf{\eta}$ space. In Eq. 3, the functions *J* are given by

$$J_m = \sum_{i=1}^{6} (\eta_i - \eta_{im})^2 / 2\sigma_{im}^2 \quad (4)$$

where η_{im} is the corresponding coordinate η_i in mode *m*.

The function *P* is the probability to generate a trial rigid-body move within an element $\delta \eta$ centered at η . In thermodynamic equilibrium, strict detailed balance implies that the old (*o*) and the new (*n*) states are related through $B_o \pi_{o \to n} = B_n \pi_{n \to o}$, where *B* 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. According to the Metropolis criterion, the canonical probability of accepting a new state is

$$p_{o \to n} = \min\left(1, \frac{\alpha_{n \to o}}{\alpha_{o \to n}} \exp(-\beta \Delta E)\right) \quad (5)$$

where $E = E_n - E_o$ is the energy difference between the states, and $\beta = 1/kT$. The ratio of a priori probabilities in Eq. (5) is chosen here as

$$\frac{\alpha_{n \to o}}{\alpha_{o \to n}} = \frac{\sum_{m=1}^{\Lambda} \kappa_m h_m \exp(-J_m^{(o)})}{\sum_{m=1}^{\Lambda} \kappa_m h_m \exp(-J_m^{(n)})} \quad (6)$$

where $J_m^{(o)}$ and $J_m^{(n)}$ are given by Eq. 4 evaluated at the old and new states, respectively. The a priori probabilities can be adjusted on the fly through the variances $\sigma_{i,m}$, the mode probabilities h_m , or the mode coordinates η_{im} . If all these quantities are kept fixed over the course of a simulation, the biasing function is said to be non-adaptive; otherwise the function is adaptive. In the applications presented here h_m and σ_{im} are kept fixed and determined at the beginning of the simulations, whereas η_{im} is updated. The weight h_m is given by $h_m \propto \exp(-\langle E_m \rangle / \lambda kT)$, where the average energy $\langle E_m \rangle$ is calculated from a short nonadaptive MC simulation using the complete force field; the significance of the smoothing parameter λ has been discussed.³²

For each trial move, the algorithm first selects a mode m with probability h_m , followed by either a rotation of a side-chain dihedral angle or one of three types of rigid-body movements, namely, translation, rotation, or rototranslation, all selected with equal probabilities. The principal axes of inertia of protein 1 define the laboratory coordinates system, as indicated in Fig. 3B. For a binary system, the position of 2 is given relative to 1; for a multiprotein system, appropriate transformations to the equations below apply. The coordinates η are defined here as $\eta = (x, y, z, \phi, \theta, \lambda)$, where (x, y, z) are the cartesian coordinates of the center of mass of protein 2, and (ϕ, θ, γ) are suitable angles that univocally determine the protein orientation. Here, the angles ϕ and θ are chosen as the spherical angles of the primary axis of inertia of protein 2, and γ is the azimuthal angle of the second moment (Fig. 3B). A translation $\mathbf{r} \rightarrow \mathbf{r} + \mathbf{r}$ is represented by a random displacement $\mathbf{r} = (x, y, y)$ z) obtained from normal distributions with zero mean and non-unit variance (σ_x , σ_y , σ_z), using the Box-Muller (BM) method. A formal rotation consists of random movements (ϕ, θ, γ) of the three independent angles, obtained from normal distributions with zero mean and variances ($\sigma_{\phi}, \sigma_{\theta}, \sigma_{\gamma}$), according to the BM method. In practice, rotations $\mathbf{r} \rightarrow \mathbf{Rr}$ are represented by a quaternion matrix **R** consisting of rotations of an angle χ around a randomly-chosen axis Ω that passes through the center of mass of protein 2, and defined by its spherical angles (Φ, Θ) in the molecular frame of 2. Rotations (Φ, Θ, γ) are thus related to (Φ,Θ, χ) by simple transformation; the restrictions on (ϕ,θ,χ) are reflected on χ .

The set of coordinates η_{im} of mode *m* is continuously updated to the values corresponding to the last accepted conformation of the mode, so the probability distribution *P* itself adapts. The canonical distribution of states collected upon convergence can be post-processed by a

clustering algorithm to identify a discrete number of 'binding modes' (Fig. 1). Once the proteins associate in one of these modes they may undergo conformational changes as a result of structural relaxation and, possibly, mutually induced fit.⁴⁰. Structural relaxations are here studied by MD simulations. However, backbone conformational changes could be integrated into the MC sampling itself using, for example, scaled collective variables^{41,42} (SCV). In the SCV method, all the dihedral angles φ and ψ along the protein main chain are moved concertedly to minimize steric clashes and improve the MC sampling acceptance rate.⁴³ Although implementation of this technique is conceptually straightforward, the method requires optimization and a detailed analysis of performance, which are not pursued here.

III. Atomistic representation of the solute and implicit representation of the solvent

The effects of water are described by the screened Coulomb potentials (SCP) implicit solvent model.^{22,44} In this model, the total non-bonded energy of a hydrated solute composed of N atoms is given by

$$E = \frac{1}{2} \sum_{i \neq j}^{N} \frac{q_i q_j}{r_{ij} D_{ij}(r_{ij}; \mathbf{r})} + \frac{1}{2} \sum_{i=1}^{N} \frac{q_i^2}{R_i(q_i; \mathbf{r})} \left\{ \frac{1}{D_i [R_i(q_i; \mathbf{r}); \mathbf{r}]} - 1 \right\} + E_{vdw}^0 - \sum_{i \neq j}^{N} \varepsilon_{ij}(\mathbf{r}) (\sigma_{ij}/r_{ij})^6 + a\gamma(\mathbf{r})$$
(7)

where $\mathbf{r} \equiv {\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N}$ represents the conformation of the solute, \mathbf{r}_i is the position of atom *i*, and r_{ij} is the distance between atoms *i* and *j*. In the partition of Eq. 7 the first term is the electrostatic interaction energy and the second sum the electrostatic self-energy.^{22,45} The third term is the direct (i.e., gas phase) van der Waals energy, whereas the fourth term is the dispersive correction by the solvent.⁴⁴ The last term in Eq. 7 is the cavity formation energy, which plays the role of a hydrophobic potential; γ is the total solvent-accessible surface area of the solute, and $a \sim 5.2$ cal mol⁻¹ Å⁻¹ is obtained by fitting hydration energies of small linear alkanes.⁴⁵ Equation 7 contains four conformation-dependent functions, D_{ij} , D_i , R_i , and ε_{ij} , that capture the effects of water exclusion in conditions of partial and anisotropic hydration. These functions have been discussed in detail^{22,45} and are summarized below for completeness. The effects of liquid-structure forces (or solvent-induced forces, SIF), formally represented by a sixth term in Eq. 7 are in practice introduced empirically through a suitable modification of R, which makes the model computationally efficient.²²

The screening functions are given by^{22,46} $D_{ij}(r_{ij};\mathbf{r}) = (1+\varepsilon_0)/\{1+k \exp[-\alpha_{ij}(\mathbf{r})r_{ij}]\}-1$ and $D_i(R_i; \mathbf{r}) = (1+\varepsilon_0)/\{1+k \exp[-\alpha_i(\mathbf{r})R_i]\}-1$, where ε_0 is the static permittivity of bulk water and *k* is a constant. Structure information is embedded in the screening coefficients α_i through²²

$$\alpha_i \approx \alpha_{0,i} - A \sum_{J \neq I}^{M} \exp(-r_{IJ}/\sigma) \quad \text{(8a)}$$

where A and $\alpha_{0,i}$ are positive constants, M is the total number of residues in the system, and r_{IJ} is the distance between residues I and J. The coefficients α_{ij} are given by²²

$$\alpha_{ij} \approx \alpha_{0,ij} - \frac{A}{2} \sum_{K \neq I}^{M} \exp(-r_{IK}/\sigma^{'}) - \frac{A}{2} \sum_{K \neq J}^{M} \exp(-r_{JK}/\sigma^{'}) \quad \text{(8b)}$$

where $\alpha_{0,ij}$ is a constant. The characteristic lengths σ and σ' control the long-range decay of the electrostatic effects of water exclusion, and render the interaction and self-energy terms in Eq. 7 independent of each other. The correct balance between these two terms (discussed below) is essential for accurate quantitative results, especially as the proteins grow in size.²² Unlike long-range effects, the short-range effects of water exclusion are sensitive to the atomic details of the solute. This local-structure dependence is incorporated into R_i in the form⁴⁷

$$R_i \approx R_{w,i} + a_i \sum_{j \neq i}^{N_c(i)} \exp(-r_{ij}/\tau_i) \quad \text{(9)}$$

where a_i and $R_{w,i}$ are constants, and N_c is the total number of atoms that surround *i* within certain cutoff distance beyond which the granularity of the system becomes less critical and long-range effects take over. The atom-dependent length τ_i controls the short-range decay of electrostatic water-exclusion effects.

The function ε_{ij} quantifies the strength of dispersion interactions between atoms *i* and *j* as modulated by water, and is given by⁴⁴

$$\varepsilon_{ij}(\mathbf{r}) = a + b \sum_{l=1}^{N_k} \exp(-r_{kl}/\lambda) + b \sum_{l=1}^{N_q} \exp(-r_{ql}/\lambda) \quad (10)$$

where *a* and *b* are parameters that depend on the solute-water and water-water dispersion interactions; N_k and N_q are the number of atoms within a cutoff distance from the off-solute points *k* and *q* (both defined by *i* and *j*); r_{kl} and r_{ql} are the distances between these points and the solute atom *l*. The characteristic distance λ determines the water occupancy at any point in the system, in particular at *q* and *k*. The values of *a*, *b*, and λ have been optimized for proteins using results from dynamics simulations in explicit water.⁴⁴

All the equations that define the model are analytical and continuous (a smoothing function is used when a cutoff is introduced), and expressions for the forces corresponding to each term have been reported.^{44,45,47} Optimization of σ and σ' for the all-atom representation of the protein force field is carried out here using ITC binding enthalpy (*H*) data of two medium-size complexes at 25 °C: for barnase-barstar at pH 7⁴⁸ (1BRS) *H* was measured at ~ 19.3 kcal/mol; for Iso-1-cytochrome c bound to its peroxidase⁴⁹ (2PCC), *H* ~ 2.6 kcal/mol at pH 6. The strength of the interaction is strong in the former complex and rather weak in the latter, and in both cases the enthalpic contributions to association are favorable. Using MC simulations with the CHARMM force field,⁵⁰ binding energies *E* are estimated

in the (σ, σ') -plane as $E = E_b - E_{\infty}$, where $E_b = Z^{-1} \sum_i E_i \exp(-E_i/kT) \approx \sum_i E_i/N_b$ is the energy of the bound state; E_i and N_b in the last sum are the total non-bonded energy (Eq. 7) of an accepted conformation *i* and the total number of accepted conformations, respectively; E_{∞} is the energy of the unbound state calculated as the sum of the non-bonded energies of the two isolated proteins. Both complexes were used previously⁴⁴ in the optimization of Eq. 10, and it was shown that the total (solvent-corrected) van der Waals contributions to binding are small. Protonation states were fixed at the start of the simulations, such that Arg, Lys, and His are positively charged, whereas Glu and Asp are negatively charged; the Heme prosthetic group is taken from the default CHARMM parameters file, which carries a doubly negative charge. Trial moves consist of side-chain dihedral angle rotations, or rigid-body rotations, translations, or rototranslations, all chosen with equal probabilities. Figure 4 shows a contour plot of E for the two complexes. The intersection of the lines corresponding to the experimental H yields $\sigma = 68$ Å and $\sigma' = 22$ Å; for the extended-atom force field, the calculations yields $\sigma = 70$ Å and $\sigma' = 45$ Å, which supersedes previous estimates.^{22,32} The extreme sensitivity of the results to small changes in σ and σ' , which is apparent from the contour plots, reflects the critical balance between the interaction and the self-energy terms, as they move in opposite directions during binding.²² It also suggests that a residue-based parameterization may ultimately be needed, as the physical interpretation of the parameters indicates.²²

IV. Results

The method is used to study the binding of cyclin-dependent kinase 5 (cdk5) to the inhibitory peptide p6. Under normal physiological conditions, cdk5 is regulated by neuronspecific activator proteins p35 and p39.51 Physiological stress induces higher concentration of calcium, which results in the cleavage of p35 into two fragments, p25 and p10.52,53 When p25 binds cdk5 the kinase becomes hyperactive, which leads to the formation of β-amyloid plaques and neurofibrillary tangles, hallmarks of Alzheimer's disease (AD).⁵⁴ Several peptides obtained by truncation of p35 have been proposed as inhibitors of the cdk5 aberrant activity in the presence of p25, some of them with therapeutic potential.^{55,56} In particular, the 33-residue peptide p6 contains, as part of its C-terminal sequence, the 24-residue peptide p5, one of the most efficient inhibitors.^{57,58} Peptide p5 is also part of the inhibitory hybrid TFP5, which has been shown to cross the blood/brain barrier.⁵⁸ Compared to p5, however, the longer p6 poses a number of computational challenges and features that make it ideal to illustrate the method. These features include the presence of several sparsely-populated conformers, a multiplicity of binding modes with different interfaces, and a broad range of binding affinities. Other inhibitory peptides⁵⁵ (e.g., p2, p3, CIP) are generally larger than p6, hence impractical for the ab initio study intended here.

IV.1 Peptide conformations in solution

None of the structures of the inhibitory peptides have been resolved experimentally. The first stage then consists in an ab initio prediction of the p6 conformation(s) in an aqueous solution. The sequence of p6 is ¹LKPFLVESCKEAFWDRCLSVINLMSSKMLQINA³³, with uncapped NH_3^+ and $-COOH^-$ termini; residues K^+ , R^+ , E^- and D^- are all assumed to be charged throughout the simulations. To sample the peptide conformational space, replica-

exchange (REx)³⁸ dynamic simulations are used. First, a stochastic conformational optimization is carried out with the MC minimization-annealing (MCMA) method⁵⁹ to find a set of N structurally distinct conformations corresponding to N low-energy local minima in the potential energy landscape. The peptide is initially built in an extended conformation, using default topology of the all-atom CHARMM force field. In the MCMA method, residue-specific biased moves of the side-chain dihedral angles χ_1 and χ_2 or main-chain dihedral angles ϕ and ψ are generated and energy-minimized before a decision on acceptance is made. The method does not generate a canonical ensemble of conformations, but a set of N relevant structures from which the REx simulations are initiated.²² These structures are then distributed over temperatures in ascending order of energies, with the lowest energy assigned to the lowest temperature (here, 37 $^{\circ}$ C). In the present simulations N = 24, and the temperatures of the replicas are given (in Kelvin) by $T_n = int[310(1 + 1)]$ $19/310^{n-1} + 0.5$], with n = 1, ..., N; this assignment produces a uniform acceptance rate of ~0.35. Swaps of neighboring replicas are considered after 100 steps of Langevin dynamics, with a collision frequency of 2 ps^{-1} and an integration time step of 1.0 fs; the SHAKE algorithm is used for all bonds with hydrogen atoms. These parameters have been used previously,²² and an optimized REx protocol for use with the SCP continuum model has recently been reported.⁶⁰ The canonical distribution obtained at 37 °C contains a large number of structures (typically $\sim 10^4$), which are clustered to obtain a reduced set of representative conformations. To this end, the degree of similarity between all pairs of structures is first computed by Procrustes superimposition,⁶¹ and a hierarchical clustering algorithm is then applied based on the maximum intra-cluster C_{α} -rmsd variance (s) allowed.⁶² This threshold determines the number of conformational families and is set at s =1.5 Å (changing this cutoff in the 1–2 Å range has only marginal effects on the results). It is found that some of the p6 conformational families have very small populations, so a subset of K families with populations $\{g_i\}_{i=1,K}$ larger than 5% is selected. The analysis shows that p6 exists in eight main conformational families: three major ones with populations of ~15% each, one with $\sim 10\%$, and four minor families each with a population of $\sim 5\%$. The peptide is thus rather structured in solution, in the sense that only $\sim 25\%$ is random coil. Representative members $\{l_i\}_{i=1,K}$ of these families are chosen as the conformers of p6 at physiological pH and temperature (Fig. 5). Any of the conformers have the potential to bind cdk5 and elicit different biological actions. A structure l_i is said to be representative of a family if it has the smallest average C_{α} -rmsd with respect to all other members of the family. Certain elements of secondary structure are common among the conformers, including an N-terminus helix (colored in red in Fig. 5), a C-terminus helix (blue), and a third intermediate helix (green), which is present in most but not all of the families. The helices are joined by β -turns or by unstructured segments. The lengths and positions of these structural motifs vary among conformers, but are qualitatively similar to those found in p25, as observed in the crystal structure of the cdk5-p25 complex (Fig. 6). The main differences among conformers appear in the arrangements of the helices into the tertiary fold, which is also different in p25, presumably due to the partial unfolding imposed by the rest of the protein. In the absence of spectroscopic evidence these observations are presently the only way to somehow validate the structural predictions. All the conformers can in principle be detected by NMR spectroscopy, which would provide the most solid experimental validation. However, the presence of an ensemble of inter-converting, sparsely populated conformers has long posed

a challenge to traditional NMR methods, which limits their ability to provide reliable structural information. Nonetheless, recent methodological advances^{11,36,37} have allowed detection and characterization of such structures in solution. Successful applications, however, require assumptions to be made, e.g., on the number of coexisting conformers.¹³ Moreover, the choice of a particular technique depends on the kinetics of the interconversions, as they are sensitive to different exchange timescales τ . Although this study does not provide kinetics information, the conformers can be used as end points in a conformational search to estimate activation energies, hence the exchange rate constants k = $1/\tau$. A number of computational methods have been proposed to address this ubiquitous problem, including conventional umbrella sampling, replica-path,³⁸ metadynamics,⁶³ targeted molecular dynamics, ⁶⁴, and others. ⁶⁵ The complete characterization of the ensemble, i.e., populations and exchange rates, would then help to select the appropriate NMR technique and to interpret the experimental data. Based on the computational predictions, a minimum concentration of ~1 mM (threshold for detection) of the lowest population conformers would require at least ~20 mM concentration of p6 in a sample. The polarity of the peptide at the assumed pH makes aggregation unlikely even at such high concentrations, in which case the predicted average distance between the surfaces of neighboring conformers would be ~35 Å.

IV.2 Conformational selection and preferential first-encounter modes

The initial conformation of cdk5 is taken from the crystal structure of the cdk5-p25 complex (1UNL). The protein is first subjected to a 5-ns MD simulation using the protocol described in Section IV.3. The protein is here assumed to adopt a single conformer, which is taken as a representative conformation over the last nanosecond of dynamics. Although this simplification suffices for the purpose of this study, the protein actually displays a rich dynamic behavior, especially in the loops, as evidenced by principal component analysis.³⁴ Moreover, the activation loop is known to exist in two main conformations,³⁴ active and inactive, and only the active form is considered here. Identification of additional conformers, which is essential to provide a molecular basis of the binding mechanism and biological activity, would not add conceptual content to this study, and are thus omitted.

For the electrostatic and the hydrophobic prescreening, the polar and non-polar centers are calculated numerically from solutions of the Poisson equation with parameters specified.³² In this paper, p6 and cdk5 self-associations are not considered. After norm optimization and clustering, a set of preferential first-encounter modes are identified for each of the conformers of p6. The complete set is shown in Fig. 7 (left column). These modes determine the initial biasing functions for the full MC simulations, which, upon convergence, yield a total of fifteen structurally distinct cdk5-p6 binding modes (Fig. 7, middle column). These are all plausible (pre-relaxation) modes of association in a mixture of cdk5 and p6 in an aqueous solution at physiological conditions, barring changes due to molecular crowding (cf. Introduction). All the modes involve mainly the small domain of the kinase, especially for the high-population conformers. Moreover, taking all the modes together, only a handful of residues are in direct contact with the peptide (not shown). This observation provides valuable information for systematic alanine scanning and other site-directed mutagenesis studies. Binding is generally non-specific, involving multiple sites. The recognition sites at

the cdk5/p6 interfaces contain spatially separate motifs on the surface of the kinase and include electrostatic and hydrophobic contacts that could be exploited to improve binding selectivity.

Because the binding modes are calculated from independent conformers, the relative importance of each mode needs to be assessed. From a methodological perspective, this screening is analogous to the determination of h_m (Eq. 3) and provides new weights for a generalized probability function P to be used in a configurational bias MC simulation of a multispecies multiprotein system, as shown schematically in Fig. 1 (block C). Accurate calculation of absolute binding affinities between flexible proteins is not presently feasible, and remains a challenge even for a small ligand binding to a rigid protein site.⁶⁶ Relative affinities of a set of small rigid molecules for the same protein site can be estimated more reliably, e.g., using thermodynamic integration or free energy perturbation methods.^{66,67} To obtain a reasonable estimate of the relative affinities of the cdk5-p6 complexes, a set of biased non-adaptive MC simulations are carried out, as described in Section III for the 1BRS and 2PCC complexes. The system is initially energy-minimized (100 ABNR steps) using Eq. 7 to allow the proteins to partially adjust to their new environments; this relaxation involves mainly side chain reorientations. Dissociation constants are estimated as K_d = $c^{\emptyset}exp(G/kT) \approx c^{\emptyset}exp(H/kT)$, where $c^{\emptyset} = 1$ kcal/mol is the standard concentration. These calculations neglect post-binding induced fit and configurational entropy changes, both of which are likely to contribute to negative S and H and may partially compensate each other. Rotational and translational entropy changes, previously estimated in a few kcal/ mol,^{68,69} are expected to be similar in all of the complexes and are thus ignored. Although the relative affinities calculated this way are admittedly crude estimates of their measurable values, they nonetheless provide information on the relative importance of the modes in the early stages of complexation. The calculation can also rule out modes that are probably too weak to make a meaningful contribution to the inhibitory mechanism of cdk5 hyperactivity, or single out modes strong enough to warrant further scrutiny.

The calculations show that the binding energies vary broadly among the modes, yielding dissociation constants in the mM-to-sub-pM range. These are biologically realistic values despite that non-bonded interactions were calculated with the complete forcefield (Eq. 7), i.e., no assumptions have been made regarding hydrogen-bond and dispersive interactions before and after binding. Several of the conformers bind the kinase with pM affinity, including the three most populated ones. This suggests that, even accounting for unfavorable entropic contributions, the affinities are likely to remain high. One particular high-affinity $(\sim 70 \text{ pM})$, high-population $(\sim 15\%)$ conformer (indicated by arrow *a* in Fig. 7) warrants a detailed examination, as it interacts with three structural motifs previously implicated in CIP, p25, and p35 binding and activation.³⁴ These structural motifs include the PSAALRE helix (colored blue in Fig. 6), a loop rich in acidic residues (purple), and a two-strand antiparallel β -sheet (green), which is part of an extended five-strand β -sheet arrangement; the activation loop is also shown (red), and the location of the ATP- and the substratebinding pockets are indicated. Unlike other binding sites, which may or may not be partially or completely occluded by other proteins in vivo, this site is certain to remain fully accessible to the solvent, hence to p6 and other inhibitors. This suggests that the inhibitory

action of p6 is probably due to a competition with p25 and/or p35 for cdk5 binding. A corollary of this result is that p6 would inhibit cdk5 activity in the presence of both activators. Although this prediction has yet to be corroborated experimentally for p6, it is nonetheless consistent with in vitro observations of p5: in the absence of cytoskeletal or neuron-specific molecules, such as p67, experiments show that p5 inhibits the pathological and the physiological activities similarly.⁵⁸ This study predicts that similar non-specific inhibition holds for p6, and probably for other p5-containing peptides.

Two other conformers appear to play similar inhibitory roles (arrows b and c in Fig. 7). One of these conformers (b) binds the kinase with high affinity, but its population in solution is rather small. The other conformer (c), although higher in population, binds with mM affinity, and is thus also unlikely to have a significant effect on the observed activity. The latter is a desirable result because in this mode p6 binds in the crevice formed by the activation loop and the small domain, which may lock the open conformation of the loop and stabilize the active form of the kinase.

IV.3 Structural relaxation and induced fit

A dynamic picture is ultimately needed to fully characterize association/dissociation events. The common approach, followed here, introduces dynamics once the correct protein/protein interfaces are identified, which is the critical step in most predictive methods. The prevailing view on ligand binding has evolved over the years from the lock-and-key model to more nuanced views⁷⁰ based on conformational selection and mutually induced fit.⁴⁰ These distinctions are not absolute, and the precise mechanism depends on a number of factors, including the flexibility of the molecules and the strength of the interactions, as well as the relaxation time scales. To explore structural relaxation and mutually induced fit upon binding, each of the predicted pre-relaxation binding modes is treated separately. The complexes are subjected to MD simulations in an explicit solvent representation, thereby scaling up the level of description. This sudden change in solvent representation may introduce artifacts that should be evaluated in each case. Simulations are performed in the isothermal-isobaric ensemble at a temperature of 37 °C and a pressure of 1 atm using the all-atom CHARMM protein force field with CMAP dihedral energy corrections.⁵⁰ The Langevin temperature control algorithm (damping coefficient of 5 ps^{-1}) is used in combination with a modified Nosé-Hoover constant pressure method (piston pressure control with oscillation period of 10 fs and decay of 50 fs). The complexes are embedded in cuboids with side lengths chosen so that the minimum distance between the surfaces of protein images is 40 Å; this setup yields $1-1.5 \times 10^4$ TIP3P water molecules, depending on the mode. Periodic boundary conditions and particle mesh Ewald summations are used, with parameters recommended in the NAMD documentation. All non-bonded interactions are truncated at 12 Å, using a shift function starting at 10 Å; the non-bonded list cutoff is set at 13.5 Å. The SHAKE algorithm is used, and forces are computed by Verlet integration with a 2 fs time step. Residues K^+ , R^+ , E^- and D^- are assumed to be charged at physiological pH and protonation states are kept fixed throughout the simulations. Three chloride ions are added to neutralize the system. To carry out the simulations, the system is first energy minimized; the protein atoms are then constrained and water equilibrated for 250 ps using temperature rescaling; finally, the constraints are removed, and the entire system is heated

and equilibrated for 0.5 ns at the same temperature. Production runs start at this point and continued for 5 ns, enough time for the major conformational changes to set in. Two test simulations, one of the predicted inhibitory mode (c1,a in Fig. 7) and one of the mode with the largest structural changes (c7), in implicit solvent revealed similar relaxation time scales (not shown).

All the binding modes undergo structural changes at varying degrees. Restructuring occurs mainly in p6 and, to a lesser extent, in unstructured segments of the small domain of cdk5 (residues 1-86); the large domain (residues 87-292) is largely unaffected. The changes in p6 result mainly from rearrangements of the helical bundle, although partial unfolding of helices is observed in some cases, usually involving a single turn at one of the helices termini. These changes in p6 are apparent in all the high-affinity modes (resulting in C_armsd as large as ~5 Å), including the main inhibitory mode discussed above (Ca-rmsd ~3 Å). The conformational changes of p6 in implicit solvent are more modest, and no partial unfolding of the termini is observed within the 6-ns simulation; repositioning of p6 relative to cdk5 are nonetheless qualitatively similar in both solvent representations, including the enhanced surface complementarity at the interface. Figures 7 (right column) shows the modes at the end of the simulations. For the main inhibitory mode, the residues in p6 that interact directly with cdk5 upon binding (L^1 , K^2 , P^3 , V^6 , E^7 , K^{10} , W^{14} , L^{23} , K^{27} , N^{32} and A³³) remain in contact with the kinase, but new hydrophobic contacts also develop (F¹³, L^{18} , I^{31}). These residues are necessary for site recognition and the inhibitory effects observed in vitro. Of particular interest are residues that also belong to p5 (from K¹⁰ upward in the sequence), which should be the focus of additional experimental studies, as changes in affinity, selectivity, or inhibitory activity can be anticipated.

The conformational changes in cdk5 depend on the particular binding mode. For the main inhibitory mode the changes are minor, with the C_{α} -rmsd of the protein and of the small and the large domains, separately, remaining below ~2 Å throughout the simulation. These are within the range of changes observed in the uncomplexed protein.³⁴ Other modes display larger conformational changes, either within the small domain (C_{α} -rmsd ~ 3 Å) and/or as a result of overall rotations of the small domain relative to the large domain. Similar conformational changes (C_{α} -rmsd ~ 2 Å) are observed in the implicit solvent simulations. Conformational changes in the ATP- and substrate-binding pockets are generally within the range observed in the uncomplexed protein.³⁴ Changes in the dynamics behavior of cdk5 upon binding, however, are more remarkable. Cross-correlation of movements can be

estimated from a normalized covariance matrix $C_{ij} = \langle \Delta \mathbf{r}_i \cdot \Delta \mathbf{r}_j \rangle / \langle \Delta r_i^2 \rangle^{1/2} \langle \Delta r_j^2 \rangle^{1/2}$, where *i* and *j* denote atoms (or subsets of atoms) in the system. In a residue-based analysis, \mathbf{r}_i and \mathbf{r}_j define the geometric centers of residue *i* and *j*, and \mathbf{r}_i and \mathbf{r}_j are the corresponding displacements between time steps; time averages are calculated over the equilibrated part of a production run. Positive ($C_{ij} > 0$) or negative ($C_{ij} < 0$) correlations indicate that residues *i* and *j* move concertedly in the same or in opposite directions respectively. Orthogonal movements yield $C_{ij} \sim 0$, so this definition may hide certain atomic correlations that could be detected by principal component analysis. A comparative analysis of the uncomplexed cdk5 and the cdk5/p25 complex evidenced a unique dynamic behavior of the ATP- and the substrate-binding sites.³⁴ In particular, the cross-correlation between the

ATP-binding site and other structural elements decreased upon association. It was conjectured³⁴ that this decoupling may be needed for catalytic efficiency, as it makes the catalytic site less sensitive to structural fluctuations of the protein. The simulations carried out here indicate the opposite behavior: the cross-correlation between the ATP-binding site and the rest of the protein appears to slightly increase upon p6 binding in almost all the modes. This is most apparent in the cross-correlation with the small domain, in particular with the five-stranded β -sheet (green and grey in Fig. 6). There is also a modest but clear expansion in the number of correlated sites. These observations suggest an independent inhibitory mechanism that may be operating in some of the modes apart from steric inhibition, namely, a long-range modulation of the kinase dynamics with potentially deleterious effect on ATP binding, hence on substrate phosphorylation.

IV. Conclusions

Predictions of protein-ligand binding modes and energies are the single most sought-after goals in computational biophysics due to their importance in diverse areas of science and technology, such as protein engineering, biomaterial design, and drug discovery. This has led to a large number of methods proposed over the years that make use of increasingly sophisticated scoring functions and innovative sampling techniques. Yet, attention to proper statistical sampling, calculations of thermodynamic quantities from the corresponding distributions, and detection and characterization of non-specific (usually weak or ultraweak) associations with multiple binding sites have not been the main driving force of such developments. Rather, the principal objective has been the prediction of the 'correct' binding modes and energies of binary systems (usually strong, highly-specific), treated at infinite dilution.^{29,71} Scoring functions are mostly empirical, with ad hoc corrections to improve estimates of binding energies from the complex coordinates. Performance is generally assessed statistically, i.e., based on the number of correct predictions over a set of known complexes. The goal of the method presented in this paper has been redirected to address protein-ligand association/dissociation within a more general framework. This development opens the possibility to calculate statistical thermodynamic quantities in a multispecies, multiprotein solution where specific and non-specific associations coexist.

This paper addressed two important challenges in the simulation of real biological media, with special consideration to computational efficiency: sampling of the configuration space and representation of long-range solvent effects. The biasing function allows mixing large and local changes in the spatial distribution of proteins, thereby enhancing sampling of relevant microstates. The method is based on the notion that a discrete set of preferential first-encounter modes are metastable states from which stable complexes at equilibrium evolve. This can be viewed as a generalization of the two-step mechanism of protein-protein association.⁷² The parallel between the funnel-like view of protein folding and protein-protein binding^{73,74} implies that the paths that proteins follow from their first encounters to their final (pre-relaxation) modes are not unique. The method can be used to explore these alternative pathways to complexation with statistical significance, and could be integrated into a more general algorithm to study protein interaction networks.

The energy function (Eq. 7) represents the anisotropic effects of water on protein-protein interactions during association or dissociation. These effects operate at different length scales and include short- and long-range electrostatic and dispersive interactions. They also include liquid structure forces, both (classical) hydrophobic forces and hydrophilic forces governed by the hydrogen bond network in the first hydration shells, which control the strength of hydrogen bonds at protein/protein interfaces. Each of these components of the force field has been parameterized separately in previous studies, using either experimental data or results from dynamics simulation in explicit water; the optimization of long-range electrostatics is presented in Section III. This description allows decomposition of the relevant forces that govern binding without the complications introduced by empirical corrections common in predictive docking algorithms. An important feature of Eq. (7) is its capability to adapt to the changing protein environment and to represent conditions of partial, anisotropic hydration typical of concentrated solutions. Macromolecular crowding is known to affect protein-protein interactions through different mechanisms, both entropic and enthalpic, and the pitfalls of studying protein function and interactions at high dilution are well documented.^{75,76} The energy function has been designed for computational efficiency and is one of the fastest all-atom force fields currently available.^{50,77}

Two major improvements still need to be incorporated to study association/dissociation in a more general framework, namely, on-the-fly backbone conformational changes and on-the-fly changes in protonation states.^{78,79} In addition, the effects of salts should be introduced through an explicit representation of ions, which in this method are treated like any other solutes. This approach is computationally more demanding than a purely continuum representation of ions (e.g., using the PB equation or DLVO-based models) but necessary to study their effects at the solute/liquid interface:⁸⁰ Solute-ion and solute-cosolute interactions are known to induce local forces that may stabilize or destabilize the structure, depending on the chemical context^{80,81} and do not lend themselves to mean-field approximations.⁸² Nonetheless, a continuum approach would still be needed to represent bulk effects of ions, for example through an empirical modification of the static dielectric permittivity (ε_0) of bulk water,⁸³ which controls the long-distance behavior of the screening functions *D* in Eq. (7).⁴⁶ Other improvements of the model and parameterization have been discussed.^{22,44}

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 an internal NIST Research Fund. AB internship at the CMM was supported by DCB/CIT at the NIH.

References

- 1. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006; 440:637. [PubMed: 16554755]
- Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. Science. 2006; 314:1938. [PubMed: 17185604]
- Rual JF, Venkatesan K, Hao T, Hirozane-kishikawa T, Dricot A, Li N, et al. Towards a proteomescale map of the human protein-protein interaction network. Nature. 2005; 437:1173. [PubMed: 16189514]

- Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, et al. Proteome organization in a genome-reduced bacterium. Science. 2009; 326:1235. [PubMed: 19965468]
- 5. Berger SI. Network analyses in systems pharmacology. Bioinformatics. 2009; 25:2466. [PubMed: 19648136]
- Stellzl U, Wanker EE. The value of high quality protein-protein interaction networks for systems biology. Curr Op Chem Biol. 2006; 10:551.
- 7. Iyengar I. Computational biochemistry: systems biology minireview series. J Biol Chem. 2009; 284:5425. [PubMed: 18940804]
- Aloy P, Russell RB. Structural systems biology: modelling protein interactions. Nature. 2006; 7:188.
- 9. Cooper A. Thermodynamic analysis of biomolecular interactions. Curr Op Chem Biol. 1999; 3:557.
- Tang C, Ghirlando R, Clore GM. Visualization of transient ultra-weak protein self-association in solution using paramagnetic relaxation enhancement. J Amer Chem Soc. 2008; 130:4048. [PubMed: 18314985]
- Tang C, Iwahara J, Clore GM. Visualization of transient encounter complexes in protein-protein association. Nature. 2006; 444:383. [PubMed: 17051159]
- Johansson H, Jensen MR, Gesmar H, Meier S, Vinther JM, Keeler C, Hodsdon ME, Led JJ. Specific and nonspecific interactions in ultraweak protein-protein association revealed by solvent paramagnetic relaxation enhancements. J Amer Chem Soc. 2014; 136:10277. [PubMed: 24969589]
- Sekhar A, Kay LE. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. Proc Nat Acad Sci (USA). 2013; 110:12867. [PubMed: 23868852]
- Korzhnev DM, Religa TL, Banachewicz W, Fersht AR, Kay LE. A transient and low-populated protein-folding intermediate at atomic resolution. Science. 2010; 329:1312. [PubMed: 20829478]
- Hubbel WL, Mchaourab HS, Altenbach C, Lietzow MA. Watching proteins move using sitedirected spin labeling. Structure. 1996; 4:779. [PubMed: 8805569]
- Schotte F, Lim M, Jackson TA, Smirnov AV, Soman J, Olson JS, Phillips GN, Wuff M, Anfinrud PA. Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. Science. 2003; 300:1944. [PubMed: 12817148]
- Goodsell DS, Olson AJ. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct. 2000; 29:105. [PubMed: 10940245]
- Venkatakrishnan AJ, Levy ED, Teichman SA. Homomeric protein complexes: evolution and assembly. Biochem Soc Trans. 2010; 38:879. [PubMed: 20658970]
- Luby-Phelps, K. Cytoarchitecture and Physical Properties of Cytoplasm: Volume, Viscosity, Diffusion, Intracellular Surface Area. Vol. 192. Academic Press; 1999.
- Luby-Phelps K. Cytoarchitecture and physical properties of cytoplasm: volume, viscosity, diffusion, intracellular surface area. Int Rev Cytol. 2000; 192:189. [PubMed: 10553280]
- Minton AP. Excluded volume as a determinant of macromolecular structure and reactivity. Biopolymers. 1981; 20:2093.
- Hassan SA, Steinbach PJ. Water-exclusion and liquid-structure forces in implicit solvation. J Phys Chem B. 2011; 115:14668. [PubMed: 22007697]
- 23. Qi HW, Nakka P, Chen C, Radhakrishnan ML. The effect of macromolecular crowding on the electrostatic component of barnase-barstar binding: a computational, implicit solvent-based study. PLoS ONE. 2014; 9:e98618. [PubMed: 24915485]
- 24. 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. Academic Press; 2012.
- 25. Harada R, Sugita Y, Feig M. Protein crowding affects hydration structure and dynamics. J Amer Chem Soc. 2012; 134:4842. [PubMed: 22352398]
- 26. DeLano W. Unraveling hot spots in binding interfaces: progress and challenges. Curr Op Struc Biol. 2002; 12:14.

- Camacho CJ, Vajda S. Protein-protein association kinetics and protein docking. Curr Op Struc Biol. 2002; 12:36.
- Fernandez-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci. 2002; 11:280. [PubMed: 11790838]
- 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. [PubMed: 21993006]
- Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A. Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. PLoS Comput Biol. 2013; 9:e1003369. [PubMed: 24339765]
- 31. Endy D, Brent R. Modelling cellular behaviour. Nature. 2001; 409:391. [PubMed: 11201753]
- 32. Cardone A, Pant H, Hassan SA. 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. J Phys Chem B. 2013; 117:12360. [PubMed: 24044772]
- 33. Lange OF, Lakomek NA, 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. [PubMed: 18556554]
- 34. 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. [PubMed: 20599546]
- Mehler EL, Hassan SA, Kortagere S, Weinstein H. Ab initio computer modeling of loops in Gprotein coupled receptors: lessons from the crystal structure of rhodopsin. Proteins. 2006; 64:673. [PubMed: 16729264]
- Vallurupalli P, Bouvignies G, Kay LE. Studying "invisible" excited protein states in slow exchange with a major state conformation. J Amer Chem Soc. 2012; 134:8148. [PubMed: 22554188]
- Palmer AGI, Kroenke CD, Loria JP. Nuclear magnetic resonance methods for quantifying microsecond-to-milisecond motions in biological macromolecules. Methods Enzymol. 2001; 339:204. [PubMed: 11462813]
- Mitsutake A, Sugita Y, Okamoto Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers. 2001; 60:96. [PubMed: 11455545]
- 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.
- Wang Q, Zhang P, Hoffman L, Tripathi S, Homouz D, Liu Y, Waxham MN, Cheung MS. Protein recognition and selection through conformational and mutually induced fit. Proc Nat Acad Sci (USA). 2013; 110:20545. [PubMed: 24297894]
- 41. Kidera A. Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide. Proc Natl Acad Sci USA. 1995:9886. [PubMed: 7568238]
- 42. Gabb HA, Prevost C, Bertucat G, Robert CH, Lavery R. Collective-Variable Monte Carlo Simulation of DNA. J Comp Chem. 1997; 18:2001.
- 43. Hassan, SA.; Mehler, EL.; Weinstein, H. Structure Calculations of Protein Segments Connecting Domains with Defined Secondary Structure: A Simulated Annealing Monte Carlo Combimed with Biased Scaled Collective Variables Technique. In: Hark, K.; Schlick, T., editors. Lecture Notes in Computational Science and Engineering. Vol. 24. Springer Verlag, Ag; New York: 2002. p. 197
- 44. Hassan SA. Implicit treatment of solvent dispersion forces in protein simulations. J Comput Chem. 2014; 35:1621. [PubMed: 24919463]
- 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.
- 46. Hassan SA, Mehler EL. From quantum chemistry and the classical theory of polar liquids to continuum approximations in molecular mechanics calculations. Int J Quant Chem. 2005; 102:986.
- 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. [PubMed: 12596268]

- 48. 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 Biol. 1997; 267:696. [PubMed: 9126847]
- 49. Pielak GJ, Wang X. Interactions between yeast iso-1-cytochrome c and its peroxidase. Biochemistry. 2001; 40
- Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, et al. CHARMM: The Biomolecular Simulation Program. J Comp Chem. 2009; 30:1545. [PubMed: 19444816]
- 51. Tarricone C, Dhavan R, Peng J, Areces LB, Tsai LH, Musacchio A. Structure and Regulation of the CDK5-p25(nck5a) Complex. Mol Cell. 2001; 8:657. [PubMed: 11583627]
- Patrick GN, Zukerberg L, Nikolic M, de la Monte S, Dikkes P, Tsai LH. Conversion of p35 to p25 deregulates CDK5 activity and promotes neurodegeneration. Nature. 1999; 402:615. [PubMed: 10604467]
- 53. Lee MS, Kwon YT, Li M, Peng J, Friedlander RM, Tsai LH. Neurotoxicity induces cleavage of p35 to p25 by calpain. Nature. 2000; 405:360. [PubMed: 10830966]
- 54. Noble W, Olm V, Takata K, Casey E, Mary O, Meyerson J, Gaynor K, LaFrancois J, Wang L, Kondo T, Davies P, Burns M, Veeranna NR, Nixon R, Dickson D, Matsuoka Y, Ahlijanian M, Lau LF, Duff K. CDK5 is a key factor in tau aggregation and tangle formation in vivo. Neuron. 2003; 38:555. [PubMed: 12765608]
- 55. Amin ND, Albers RW, Pant HC. Cyclin-Dependent Kinase 5 (cdk5) Activation Requires Interaction with Three Domains of p35. J Neurosci Res. 2002; 67:354. [PubMed: 11813240]
- Zheng YL, Kesavapany S, Gravell M, Hamilton RS, Schubert M, Amin ND, Albers RW, Grant P, Pant HC. A Cdk5 Inhibitory Peptide reduces Tau Hyperphosphorylation and Apoptosis in Neurons. EMBO J. 2005; 24:209. [PubMed: 15592431]
- Binukumar BK, Zheng YL, Shukla V, Amin ND, Grant P, Pant HC. TFP5, a peptide derived from p35, a CDK5 neuronal activator, rescues cortical neurons from glucose toxicity. J Alzheimer's Dis. 2014; 39:899. [PubMed: 24326517]
- 58. Zheng YL, Amin ND, Hu YF, Rudrabhatla P, Shukla VK, Kanungo J, Kesavapany S, Grant P, Albers RW, Pant HC. A 24-residue peptide (p5), derived from p35, the cdk5 neuronal activator, specifically inhibits cdk5-p25 hyperactivity and tau hyperphosphorylation. J Biol Chem. 2010; 285:34202. [PubMed: 20720012]
- 59. Steinbach PJ. Exploring peptide energy landscapes: a test of force fields and implicit solvent models. Proteins. 2004; 57:2004.
- 60. Li XF, Latour RA. The Temperature Intervals with Global Exchange of replicas Empirical Accelerated Sampling Method: Parameter Sensitivity and Extension to a Complex Molecular System. J Comp Chem. 2011; 32:1091. [PubMed: 20949510]
- 61. Gower, JC.; Dijksterhuis, GB. Procrustes Problems. Oxford University Press; New York: 2004.
- 62. Szekely GJ, Rizzo ML. Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method. J Classif. 2005; 22:151.
- 63. Laio A, Parrinello M. Escaping free-energy minima. Proc Nat Acad Sci (USA). 2002; 99
- 64. Schlitter J, Engels M, Kruger P, Jacoby E, Wollmer A. Targeted molecular dynamics simulations of conformational change. Application to the T to R transition in insulin. Mol Sim. 1993; 10:291.
- Baba T, Harada R, Nakano M, Shigeta Y. On the induced-fit mechanism of substrate-enzyme binding structure of nylon-oligomer hydrolase. J Comput Chem. 2014; 35:1240. [PubMed: 24752461]
- 66. Hansen N, van Gunsteren WF. Practical aspects of free-energy calculations: a review. J Chem Theory Comput. 2014; 10:2632.
- Hassan, SA.; Gracia, L.; Vasudevan, G.; Steinbach, PJ. Computer simulation of protein-ligand interactions: challenges and applications. In: Nienhaus, GU., editor. Protein-Ligand Interactions. Vol. 305. Humana Press; New Jersey: 2005.
- 68. Hermans J, Wang L. Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. Application to a complex of benzene and mutant T4 lysozyme. J Amer Chem Soc. 1997; 119:2707.
- Yu YB, Privalov PL, Hodges RS. Contribution of translational and rotational motions to molecular association in aqueous solution. Biophys J. 2001; 81:1632. [PubMed: 11509376]

- Csermely P, Palotai R, Nussinov R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci. 2010; 35:539. [PubMed: 20541943]
- 71. Kastritis PL, Bonvin AMJJ. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. J R Soc Interface. 2013; 10:20120835. [PubMed: 23235262]
- 72. Selzer T, Schreiber G. New insights into the mechanism of protein-protein association. Proteins. 2001; 45:190. [PubMed: 11599022]
- 73. Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. Protein Sci. 1999; 8:1181. [PubMed: 10386868]
- 74. Zhang C, Chen J, DeLisi C. Protein-protein recognition: exploring the energy funnels near the binding sites. Proteins. 1999; 34:255. [PubMed: 10022360]
- Zimmermann SB, Minton AP. Macromolecular crowding: biochemical, biophysical, and physiological consequences. Annu Rev Biophys Biomol Struct. 1993; 22:27. [PubMed: 7688609]
- 76. Ellis RJ. Macromolecular Crowding: Obvious but Underappreciated. TIBS. 2001; 26:597. [PubMed: 11590012]
- 77. Juneja A, Ito M, Nilsson L. Implicit solvent models and stabilizing effects of mutations and ligand on the unfolding of the Amyloid b-peptide central helix. J Chem Theory Comput. 2013; 9:834.
- Dullweber F, Stubbs MT, Musil D, Sturzebecher J, Klebe G. Factorising ligand affinity: a combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. J Mol Bio. 2001; 313:593. [PubMed: 11676542]
- 79. Stubbs MT, Reyda S, Dullweber F, Moller M, Klebe G, Dorsch D, Mederski WWKR, Wurziger H. pH-dependent binding modes observed in trypsin crystals: lessons for structure-based drug design. ChemBio Chem. 2002; 3:246.
- Hassan SA. Amino acid side chain interactions in the presence of salts. J Phys Chem B. 2005; 109:21989. [PubMed: 16479276]
- Parsegian VA, Rand RP, Rau DC. Osmotic Stress, Crowding, Preferential Hydration, and Binding: a Comparison of Perspectives. Proc Nat Acad Sci (USA). 2000; 97:3897.
- 82. French RH, et al. Long Range Interactions in Nanoscale Science. Rev Mod Phys. 2010; 82:1887.
- Hasted JB, Ritson DM, Collie CH. Dielectric Properties of Aqueous Ionic Solutions. Parts I and II. J Chem Phys. 1948; 16:1.



Figure 1.

Flowchart of the general algorithm described in this study. Blocks A and B are detailed in Fig. 2; block C (not discussed here) represents a canonical self-adaptive configurational-bias Monte Carlo subroutine for the multispecies-multiprotein system, and is a generalization of block B for binary systems.

Author Manuscript





Flowcharts of block A (single-molecule conformational search) and block B (canonical selfadaptive configurational-bias Monte Carlo of binary interactions).

Author Manuscript



Figure 3.

(A) Norm optimization: schematic representation of the variables that define the electrostatic norm e (upper right; Eq. 1) and hydrophobic norm h (lower left; Eq. 2); (B) schematic representation of variables that define the relative orientations of proteins during the biased MC simulations.



Figure 4.

Contour map of the binding energy E in the (σ, σ') -plane (both in Å) used to calibrate longrange electrostatic effects of water exclusion (Eqs. 8a,b) on barnase-barstar (solid line; PDB id: 1BRS) and Iso-1-cytochrome c peroxidase-cytochrome c (dashed; 2PCC). Thicker lines correspond to contours with E equal to the ITC-determined binding enthalpies H; the optimal values of the parameters correspond to the intersection of the two lines.



Figure 5.

Representative structures (conformers c_1-c_8) of the main conformational families of peptide p6 in aqueous solution, as determined by replica-exchange molecular dynamics simulations using the all-atom CHARMM force field with the SCP implicit solvent model. Populations are: $g_1-g_3 \sim 15\%$; $g_4 \sim 10\%$; $g_5-g_8 \sim 5\%$. Colors are the same as in Fig. 6 for protein p25; same color corresponds to the same amino acid sequence (see text).



Figure 6.

Cyclic-dependent kinase 5 (cdk5) and pathological activator (p25) with key structural motifs involved in binding and regulation shown in color (see text); ATP- and substrate-binding pockets indicated by arrows; structures taken from the crystal structure of the cdk5-p25 complex (inset).



Figure 7.

Characterization of cdk5-p6 complex formation in solution: from conformational selection to mutually induced fit. Metastable preferential first-encounter modes (left column); pre-relaxation binding modes (middle column); post-binding dynamics and structural relaxation (right column). First encounters are characterized by hundreds of conformations for each pair of conformers (shown with cdk5 backbone superimposed). The number of pre-relaxation binding modes is reduced to 1–3 modes per pair, yielding a total of 15 binding modes with affinities in the μ M to sub-pM range (p6 shown in different colors for clarity). The dynamic behavior of the complexes upon association is explored with a fully-atomistic representation of the solvent (water and counter-ions not shown). Arrows *a*, *b*, and *c* indicate the predicted inhibitory modes.