Calculation of absolute protein–ligand binding free energy from computer simulations

Hyung-June Woo[†] and Benoît Roux[‡]

Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021

Edited by Bruce J. Berne, Columbia University, New York, NY, and approved March 31, 2005 (received for review December 3, 2004)

A general methodology for calculating the equilibrium binding constant of a flexible ligand to a protein receptor is formulated on the basis of potentials of mean force. The overall process is decomposed into several stages that can be computed separately: the free ligand in the bulk is first restrained into the conformation it adopts in the bound state, position, and orientation by applying biasing potentials, then it is translated into the binding site, where it is released completely. The conformational restraining potential is based on the root-mean-square deviation of the peptide coordinates relative to its average conformation in the bound complex. Free energy contributions from each stage are calculated by means of free energy perturbation potential of mean force techniques by using appropriate order parameters. The present approach avoids the need to decouple the ligand from its surrounding (bulk solvent and receptor protein) as is traditionally performed in the double-decoupling scheme. It is believed that the present formulation will be particularly useful when the solvation free energy of the ligand is very large. As an application, the equilibrium binding constant of the phosphotyrosine peptide pYEEI to the Src homology 2 domain of human Lck has been calculated. The results are in good agreement with experimental values.

free energy perturbation | molecular dynamics | Src Homology 2 domain

A problem of central importance in computational biology is the quantitative determination of absolute binding affinities in diverse and complex systems. Predicting the binding free energy of ligands to macromolecules can have great practical values in identifying novel molecules that can bind to target receptors and act as therapeutic drugs (1). Furthermore, molecular recognition phenomena involving various kinds of binding modules linked to cell surface receptors play an essential role in a wide variety of intracellular signal transduction pathways (2–5).

Approaches at different levels of complexity and sophistication have been used to calculate binding free energies in complex biomolecular systems. Screening of large molecular databases of compounds to identify potential lead drug molecules typically relies on very simplified scoring schemes to achieve the needed efficiency (6). The binding free energy may be estimated on the basis of a continuum solvent approximation assuming quadratic fluctuations around a unique configuration (7, 8). The Molecular Mechanics/Poisson–Boltzmann (PB) and Surface Area (MM/PB-SA) method is a popular approach that relies on a mixed scheme combining configurations sampled from molecular dynamics (MD) simulations with explicit solvent, together with free energy estimators based on an implicit continuum solvent model (9). MM/PB-SA shares some similarities with the linear interaction energy method, which also uses averages calculated from explicit solvent simulations within a linear response framework (10). Despite their usefulness, such approximate schemes can be limited, and how to improve the results is unclear because they do not offer a rigorous route to compute the equilibrium binding constant.

In principle, treatments based on MD free energy perturbation (FEP) simulations with explicit solvent molecules offer the most powerful and promising approach to estimate the binding free energies of ligands to macromolecules (11). Nonetheless, although previous studies have provided many of the fundamental elements necessary for the calculation of binding free energy by means of MD (12–17), the computations so far have been limited mostly to fairly small and rigid ligands [e.g., rare gas atom (12, 15), water (14), camphor (13), benzene (15), and a single amino acid (16)], and some aspects require further considerations. The challenges that lie ahead are well illustrated by considering the association of peptide ligands to Src homology 2 (SH2) domains. SH2 domains are highly conserved noncatalytic proteins of ≈100-aa residues, which can bind phosphotyrosine-containing polypeptide sequences with high affinity and specificity. They are found in a wide variety of intracellular signal transduction pathways involving tyrosine kinases (18–22), and inappropriate cellular signaling caused by malfunctions of SH2-mediated process has been linked to many pathologic conditions (e.g., cancer, autoimmune diseases, asthma, allergies, etc.). SH2 domains are highly selective toward the sequence phosphotyrosine-Glu-Glu-Ile (pYEEI) (23), with dissociation constants ranging from micromolar to nanomolar ranges (19, 24; for a review, see ref. 25). These values correspond to an absolute binding free energy in the range of approximately −8 kcal/mol. The determinants of the phosphopeptide selectivity of SH2 domains have been characterized previously by using various empirical approaches (22, 26, 27).

Our ability to understand many aspects concerning the specificity of SH2 domains for phosphotyrosyl peptides of particular sequences would benefit from all-atom MD/FEP computations, although a straightforward application of current approaches appears to be problematic for a number of reasons. First, the sheer magnitude of the electrostatics interactions arising from the doubly charged phosphotyrosine side chain (28) suggests that the standard FEP technique used to compute absolute binding free energies, which consists of reversibly decoupling the ligand from its surrounding (12–17), is essentially impractical. Furthermore, additional difficulties are expected to arise from the significant conformational flexibility of the unbound peptide ligand in solution. Although such difficulties may at first appear to be merely technical, the truth is that available methods to compute a binding constant from simulations with explicit solvent are inadequate or prohibitive. Clearly, some significant extension to the present computational methodologies is needed to tackle the more complex situations that invariably are encountered in biological systems.

Our goal is to address these issues and design an efficient approach for calculating the binding constant of a flexible ligand to a protein. Below, the equilibrium binding constant is derived directly on the basis of configurational ensemble averages. This derivation, which differs from the traditional arguments based on chemical potentials and standard states, is particularly advantageous because of its simplicity and clarity. For instance, various computationally convenient expressions can be obtained readily

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: FEP, free energy perturbation; MD, molecular dynamics; PB, Poisson–Boltzmann; MM/PB-SA, Molecular Mechanics/PB and Surface Area; PMF, potential of mean force; rmsd, rms deviation; SH2, Src homology 2.

[†]Present address: Department of Chemistry, University of Nevada, Reno, NV 89557.

[‡]To whom correspondence should be addressed. E-mail: benoit.roux@med.cornell.edu.

from the fundamental expression for the equilibrium binding constant. In the present case, the absolute binding free energy is rigorously expressed as the sum of separate contributions corresponding to a step-by-step process describing the association of the ligand with the receptor. The formulation relies on potential of mean force (PMF) techniques designed to avoid the problems associated with large free energies involved in the standard FEP decoupling schemes (12–17). As an illustration of the present treatment, the absolute binding free energy of the phosphotyrosyl peptide Ace-pYEEI to the SH2 domain of Lck kinase is computed. The result of the computations is found to be in good agreement with experimental estimates.

Theoretical Developments

Let us consider a dilute solution in thermodynamic equilibrium comprising receptor proteins and flexible ligands able to associate in a bimolecular fashion. Classically, the equilibrium binding constant K_{eq} of the process $L + P \rightleftharpoons LP$ is defined as a function of the concentrations of each species, $[LP]$, $[L]$, and $[P]$, as $K_{eq} = [LP]/[L][P]$. Let p_0 and p_1 be the fraction of protein receptor with no ligand or one ligand bound, respectively. Two distinct regions of configurational space can be clearly distinguished without ambiguity: the binding "site" and the "bulk" unbound regions, and $[P] = p_0[P]_{tot}$ and $[LP] = p_1[P]_{tot}$, where $[P]_{tot}$ is the total concentration of the receptor in the system. By normalization, there can either be zero or one ligand L bound to the receptor protein P, thus, $p_0 + p_1 = 1$. It follows that the binding constant can be expressed as

$$K_{eq} = \frac{p_1[P]_{tot}}{[L]p_0[P]_{tot}} = \frac{1}{[L]} \times \frac{p_1}{p_0}. \qquad [1]$$

Assuming that the receptor concentration is sufficiently low, it is possible to consider a single one with its center of mass held fixed at the origin surrounded by a solution of ligands without loss of generality. As will be shown below, the logarithm of the ratio (p_1/p_0) is related to the reversible work needed to take one ligand molecule from the bulk and carry it to the binding site. Eq. **1** can be written as

$$\begin{aligned} K_{eq} &= \frac{1}{[L]} \times \frac{N\int_{site} d\mathbf{1}\int_{bulk} d\mathbf{2} \cdots \int_{bulk} d\mathbf{N}\int d\mathbf{X}\, e^{-\beta U}}{\int_{bulk} d\mathbf{1}\int_{bulk} d\mathbf{2} \cdots \int_{bulk} d\mathbf{N}\int d\mathbf{X}\, e^{-\beta U}} \\ &= \frac{1}{[L]} \times \frac{N\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta U}}{\int_{bulk} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta U}}, \end{aligned} \qquad [2]$$

where U is the total potential energy of the system, $1/\beta = k_B T$ is the Boltzmann constant times temperature, and $\{\mathbf{1}, \mathbf{2}, \ldots \mathbf{N}, \mathbf{X}\}$ are the degrees of freedom of the N ligand molecules and the remaining atoms (solvent or protein), respectively. The subscripts site and bulk in the integrals indicate the relevant spatial regions of the configurational space to be included in each integration, representing the bound and unbound states. In Eq. **2**, the ligand molecule "1" has been chosen arbitrarily to occupy the binding site, and the factor N accounts for the fact that any ligand could have been chosen. Because the bulk region is isotropic and homogeneous, we have

$$K_{eq} = \frac{1}{[L]} \times \frac{N\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta U}}{V_{bulk}\int_{bulk} d\mathbf{1}\delta(\mathbf{r}_1 - \mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta U}}, \qquad [3]$$

where $\mathbf{r}_1$ is the position of the center of mass of ligand 1 and $\mathbf{r}_1^*$ is some arbitrary (fixed) location in the bulk region, far away from the receptor. [The integrals over the $(N - 1)$ remaining ligands have been omitted for the sake of simplicity, assuming low concentration and absence of ligand-ligand interactions.] Because $[L] = N/V_{bulk}$, the equilibrium binding constant K_{eq} is

$$K_{eq} = \frac{\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta U}}{\int_{bulk} d\mathbf{1}\delta(\mathbf{r}_1 - \mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta U}}. \qquad [4]$$

The denominator and the numerator of Eq. **4** each represent initial and final states of the binding process: the ligand bound to the receptor and the ligand with its center of mass at $\mathbf{r}_1^*$ in the bulk, respectively (note that all coordinates are expressed relative to the center of mass of the receptor).

The fundamental expression Eq. **4** serves as the cornerstone to develop various computational strategies. In particular, it can be rewritten as

$$K_{eq} = \frac{1}{8\pi^2}\int_{site} d\mathbf{r}d\omega\, e^{-\beta W(\mathbf{r},\omega)}, \qquad [5]$$

where $W(\mathbf{r}, \omega)$ is the protein-ligand solvent-averaged PMF as a function of the relative translation and orientation $\vec{\mathbf{r}}$ and ω

$$e^{-\beta W(\mathbf{r},\omega)} = \frac{\int d\mathbf{1}\int d\mathbf{X}\delta(\mathbf{r}_1 - \mathbf{r})\delta(\omega_1 - \omega)\, e^{-\beta U}}{\int_{bulk} d\mathbf{1}\int d\mathbf{X}\delta(\mathbf{r}_1 - \mathbf{r})\delta(\omega_1 - \omega)\, e^{-\beta U}}, \qquad [6]$$

(by definition $W \to 0$ in the bulk). Although the utility of Eq. **5** is severely limited in all-atom MD simulations with explicit solvent, it is helpful to clarify the significance of approximations based on implicit solvent models such as MM/PB-SA (9). Assuming quadratic fluctuations in the bound state, the dominant contribution to the binding constant may be approximated as

$$K_{eq} \approx \frac{1}{8\pi^2}\Delta\omega\Delta V e^{-\beta W_{min}}, \qquad [7]$$

where W_{min} corresponds to the minimum of the PMF in the bound state, and $(\Delta\omega/8\pi^2)$ and ΔV represent the orientational freedom and translational volume of the bound ligand, respectively (7). In MM/PB-SA (9), a statistical ensemble of configurations of the bound ligand is generated from all-atom MD simulations, and W_{min} is estimated from the average value of the interaction free energy calculated by using an implicit continuum solvent model; the quantities $\Delta\omega$ and ΔV are estimated from the root-mean-square (rms) fluctuations by using quasiharmonic approximation (29, 30).

The design of a computational method relying on FEP simulations with explicit solvent molecules consists of inserting intermediate states in Eq. **4** such that each individual contribution can be easily calculated. Here, the intermediate states are constructed by introducing various restraining potentials, which are designed to bias the ligand–protein complex toward the configuration it adopts in the bound state. It is useful to first establish a local frame of reference from three centers in each binding partner (Fig. 2), in which the position of the center of mass of the ligand relative to the receptor $\mathbf{r}_1$ can be specified by (r_1, θ_1, ϕ_1) in spherical coordinates, and its orientation can be specified from the three Euler angles $(\Theta_1, \Phi_1, \Psi_1)$. To restrain the ligand orientation as in the bound complex, we introduce the potential $u_o(\Theta_1, \Phi_1, \Psi_1)$. We also introduce the potential $u_a(\theta_1, \phi_1)$, designed to restrain the ligand position along a specific axis as in the bound complex. Lastly, we introduce the potential u_c, designed to restrain the conformation of the ligand around the average conformation that it adopts when it is bound to the receptor. Although other choices are possible, a simple potential can be constructed on the basis of ξ, the rms deviation (rmsd) of the ligand relative to its average conformation. With these definitions, the equilibrium binding constant K_{eq} in Eq. **4** can be written as

$$\begin{aligned} K_{eq} = &\frac{\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta U}}{\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c]}} \times \frac{\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c]}}{\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o]}} \\ &\times \frac{\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o]}}{\int_{site} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o+u_a]}} \end{aligned}$$

$$\times \frac{\int_{\text{site}} d\mathbf{1} \int d\mathbf{X}\, e^{-\beta[U+u_c+u_o+u_a]}}{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o]}}$$

$$\times \frac{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o]}}{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c]}}$$

$$\times \frac{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c]}}{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta U}}. \quad [8]$$

Most of the terms in Eq. **8** are dimensionless ratios of configurational integrals corresponding to free energy differences that can be calculated from a standard application of the FEP simulation technique,

$$e^{-\beta G_c^{\text{site}}} = \frac{\int_{\text{site}} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c]}}{\int_{\text{site}} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta U}} = \langle e^{-\beta u_c}\rangle_{(\text{site},U)} \quad [9a]$$

$$e^{-\beta G_o^{\text{site}}} = \frac{\int_{\text{site}} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o]}}{\int_{\text{site}} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c]}} = \langle e^{-\beta u_o}\rangle_{(\text{site},U+u_c)} \quad [9b]$$

$$e^{-\beta G_a^{\text{site}}} = \frac{\int_{\text{site}} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o+u_a]}}{\int_{\text{site}} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta(U+u_c+u_o)}} = \langle e^{-\beta u_a}\rangle_{(\text{site},U+u_c+u_o)} \quad [9c]$$

$$e^{-\beta G_o^{\text{bulk}}} = \frac{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o]}}{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c]}} = \langle e^{-\beta u_o}\rangle_{(\text{bulk},U+u_c)} \quad [9d]$$

$$e^{-\beta G_c^{\text{bulk}}} = \frac{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c]}}{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta U}} = \langle e^{-\beta u_c}\rangle_{(\text{bulk},U)}. \quad [9e]$$

One may note that the delta function involving $\mathbf{r}_1^*$, when it appears both in the numerator and denominator, does not affect the calculated free energies in the bulk region because it is invariant to translation. The free energy G_o^{bulk} in Eq. **9d** can be calculated directly, as an angular integral, because the bulk is isotropic. However, the fourth term in Eq. **8**, which involves a ratio of configurational integrals with the bound ligand (numerator) and the ligand held with its center of mass at $\mathbf{r}_1^*$ in the bulk by a delta function, requires special attention because it does not correspond to a free energy difference like the other terms. It can be reexpressed as (see *Supporting Text*, which is published as supporting information on the PNAS web site)

$$\frac{\int_{\text{site}} d\mathbf{1}\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o+u_a]}}{\int_{\text{bulk}} d\mathbf{1}\delta(\mathbf{r}_1-\mathbf{r}_1^*)\int d\mathbf{X}\, e^{-\beta[U+u_c+u_o]}} = S^*I^*, \quad [10]$$

where S^* is an integral over the angles θ_1 and ϕ_1,

$$S^* = (r_1^*)^2 \int_0^{\pi} \sin(\theta_1) d\theta_1 \int_0^{2\pi} d\phi_1 e^{-\beta u_a(\theta_1,\phi_1)}, \quad [11]$$

and I^* is a one-dimensional (1D) integral over r_1

$$I^* = \int_{\text{site}} dr_1 e^{-\beta[\mathcal{W}(r_1)-\mathcal{W}(r_1^*)]}, \quad [12]$$

defined in terms of the PMF $\mathcal{W}(r_1)$ calculated in the presence of the configurational and orientational restraints u_c, u_o, and u_a (see Eq. **3** in *Supporting Text*). It follows that the binding constant can be expressed as

$$K_{\text{eq}} = S^*I^* e^{-\beta[G_c^{\text{bulk}}+G_o^{\text{bulk}}-G_a^{\text{site}}-G_o^{\text{site}}-G_c^{\text{site}}]}. \quad [13]$$

By definition, K_{eq} has the units of volume (e.g., Å^3). Bimolecular equilibrium binding constants are normally expressed in inverse moles per liter, and it is customary to define an absolute binding free energy $G_{\text{bind}} \equiv -k_B T \ln[K_{\text{eq}} C^\circ]$ by assuming a standard state concentration C° of 1 mol/liter ($\equiv$ 1/1,661 Å^3).

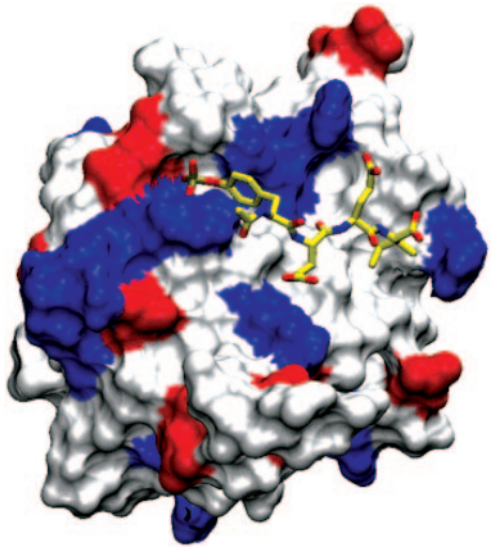

Fig. 1. Crystallographic structure of the p56^{lck} SH2 domain in complex with a pYEEI peptide (21).

Computational Methods

The crystallographic structure of the AcpYEEI peptide in complex with the human p56^{lck} SH2 domain from ref. 21. was used as the initial structure for the bound state (Fig. 1). The phosphotyrosine was assumed to be doubly charged as determined experimentally (28). The complex was solvated and equilibrated for 0.7 ns. The resulting structure was used for the PMF calculations and as the reference conformation for the rmsd restraint of the ligand. All MD simulations were generated by using the CHARMM program (31). The PARAM27 force field (32) was used with the TIP3P water potential (33). The trajectories were generated with periodic boundary conditions in the isobaric-isothermal ensemble at constant pressure of 1 atm (1 atm = 101.3 kPa) and temperature of 300 K. Electrostatic interactions were treated with a particle-mesh Ewald method (34); a grid with roughly one point per Å was used for all systems. Potassium and chloride ions were added both to neutralize the overall systems and simulate an aqueous salt solution at 150 mM.

The coordinate system for specifying the overall relative position and orientation of the ligand with respect to the protein was constructed by choosing three groups of atoms within the protein and in the ligand (see Fig. 2). The three centers, P1, P2, and P3, for the SH2 domain were given by the center of mass of (H208), (D171), and (I183, L202, L205, L165), respectively. The three centers, L1, L2, and L3, for the peptide were given by the center of mass of (pY1, E2, E3, I4), (pY), and (I4), respectively. The spherical coordinate system to establish the position of the ligand relative to the protein comprises, the P3–L1 distance r_1, the P2–P1–L1 angle θ_1, and the P3–P2–P1–L1 dihedral angle ϕ_1. The Euler angles needed to define the orientation of the ligand relative to the protein are the P3–L1–L2 angle Θ_1, the P2–P1–L1–L2 dihedral angle Φ_1, and the P1–L1–L2–L3 dihedral angle Ψ_1.

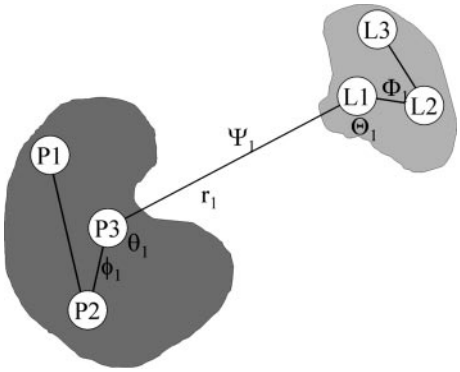

Fig. 2. Schematic representation of the local reference frame used to define the position and orientation of the ligand relative to the receptor protein and construct the restraining potentials. The spherical coordinate system establishing the position of the ligand relative to the protein are the P3–L1 distance r_1, the P2–P1–L1 angle θ_1, and the P3–P2–P1–L1 dihedral angle ϕ_1. The Euler angles needed to define the orientation of the ligand relative to the protein are the P3–L1–L2 angle Θ_1, the P2–P1–L1–L2 dihedral angle Φ_1, and the P1–L1–L2–L3 dihedral angle ψ_1.

The primary tools for the calculation of all of the components in Eq. **13** are FEP simulations (11, 35), umbrella sampling simulations (36), and the weighted histogram analysis method for unbiasing the data from multiple simulations (37–39). Harmonic biasing potentials used are the conformational restraint u_c, the orientational restraint u_o, and the axis restraint u_a. The conformation of the peptide was restrained by using the potential $u_c(\xi) = k_c(\xi[\mathbf{1};\mathbf{1}_{ref}])^2$, where ξ is the rmsd of the ligand relative to its average conformation in the bound state $\mathbf{1}_{ref}$, which is used as a reference. A force constant of $k_c = 1.192$ kcal/mol·Å^2 was used. The orientation of the ligand peptide was restrained by using the harmonic potential $u_o(\Theta_1, \Phi_1, \Psi_1) = k_o[(\Phi_1 - \Phi_1^{ref})^2 + (\Theta_1 - \Theta_1^{ref})^2 + (\Psi_1 - \Psi_1^{ref})^2]$, where $(\Theta_1^{ref}, \Phi_1^{ref}, \Psi_1^{ref})$ corresponds to the average orientation of the bound ligand. The ligand peptide also was restrained to lie along the 1D axis r_1 by using the harmonic potential $u_a(\theta_1, \phi_1) = k_a[(\theta_1 - \theta_1^{ref})^2 + (\phi_1 - \phi_1^{ref})^2]$, where $(\phi_1^{ref}, \theta_1^{ref})$ corresponds to the average values for the bound ligand. The force constants were $k_o = k_a = 100$ kcal/mol per rad^2.

The PMF along the 1D axis r_1, $\mathcal{W}(r_1)$, was calculated by using umbrella sampling simulations. The system was solvated by using a preequilibrated orthorhombic box of water molecules (approximate dimension of 76 × 56 × 56 Å) followed by 200 ps of equilibration before collecting statistics of the distance r_1. Umbrella sampling window configurations then were generated in the presence of the biasing radial potential $u_r = k_r(r_1 - r_1')^2$; 28 windows were simulated with interwindow spacing of 1 Å, and 10 additional windows separated by 0.5 Å for the short distances where the PMF was seen to vary most. Harmonic window potentials $k_r(r_1 - r_1^{(i)})^2$ were used with force constant values of 0.5 and 5 kcal/mol·Å^2. The total umbrella sampling simulation is slightly longer than 3 ns. The integral in Eq. **10** was calculated numerically by using $r_1^* = 40$ Å as a reference.

The free energy G_c^{bulk}, corresponding to the restriction on the conformation of the ligand free in solution, was obtained by means of a direct integration of the Boltzmann factor after calculating the PMF $w_c^{bulk}(\xi)$ as a function of ξ, the rmsd relative to the reference conformation $\mathbf{1}_{ref}$,

$$e^{-\beta G_c^{site}} = \frac{\int d\,\xi e^{-\beta[w_c^{bulk}(\xi)+u_c(\xi)]}}{\int d\,\xi e^{-\beta w_c^{site}(\xi)}}. \qquad [14]$$

A similar expression was used for G_c^{site} with a PMF $w^{site}(\xi)$. Harmonic window potentials, $k_c(\xi - \xi^{(i)})^2$, with a force constant $k_c = 1.0$ kcal/mol·Å^2 were used. The PMF in the bulk was calculated from 20 umbrella sampling simulations separated by 0.5 Å, for a total of 2 ns; a cubic system with 897 water molecules (30 × 30 × 30 Å) was used. The PMF in the bound state was calculated from 20 umbrella-sampling simulations separated by 0.2 Å, for a total of 0.8 ns; a cubic system with 5,490 water molecules (56 × 56 × 56 Å) was used.

The free energy terms G_o^{site} and G_a^{site}, corresponding to the orientational and axial restriction in the binding site, respectively, were calculated from Eqs. **9b** and **9c** by using FEP with 10 intermediate values of the thermodynamic coupling parameter λ between 0 and 1, for a total of 0.5 ns. A cubic system with 5,490 water molecules (56 × 56 × 56 Å) was used. The free energy term G_o^{bulk} was calculated from Eq. **9d** by direct numerical integration over the three Euler angles. The surface element S^* was calculated by direct numerical integrations of Eq. **11**. The electrostatic contribution to the solvation free energy of the peptide ligand was calculated by using FEP with 10 intermediate values of the thermodynamic coupling parameter λ between 0 and 1, for a total of 1 ns. The peptide was solvated by using the Spherical Solvent Boundary Potential (SSBP) with 421 water molecules (40).

For the MM/PB-SA approximation, the value of W_{min} in Eq. **7** is approximated by the average net interaction free energy between the protein and ligand $\Delta G_{LP} = G_{LP} - G_L - G_P + \Delta E_{LP}$, where G_{LP}, G_P, and G_L are the total electrostatic (PB) solvation free energies of the bound complex, isolated protein, and isolated ligand, respectively, and ΔE_{LP} is the direct (bare) ligand–protein electrostatic interaction energy. Each energy term was obtained by averaging >40 equilibrium configurations extracted from a MD trajectory of the bound complex. The finite-difference PB calculations were performed with a grid of 0.4 Å by using the PBEQ module of the CHARMM program (31) with the optimized atomic Born radii (41). The dielectric constants of the solvent and protein were 80 and 1, respectively. The salt concentration was taken as 0.15 mM. The binding free energy then was estimated as $G_{bind} \approx -k_BT \ln[\Delta\omega/8\pi^2] - k_BT \ln[C^\circ\Delta V] + \Delta G_{LP}$, where $\Delta\omega$ and ΔV are the typical Euler angle and volume of the orientational and translational fluctuations of the ligand in the bound complex.

Table 1. Computation of the absolute binding free energy

Component	Value
G_c^{bulk}	3.70 kcal/mol
G_c^{site}	1.43 kcal/mol
G_o^{bulk}	5.35 kcal/mol
G_o^{site}	0.04 kcal/mol
G_a^{site}	0.40 kcal/mol
S^*	22.17 Å^2
I^*	3.12×10^{13} Å
K_{eq}	4.13×10^9 Å^3
K_{eq}	2.49×10^6 M^{-1}
K_d	0.40 μM
G_{bind}	−8.8 kcal/mol

Results and Discussion

Table 1 summarizes the calculated values of the various contributions to the absolute binding free energy. In Table 2, the calculated binding free energy G_{bind} is compared with experimental values taken from refs. 24 and 42. It is observed that the result of the calculation is within ≈1 kcal/mol from experimental estimates for closely related sequences (there may be some uncertainty about the experimental value for the AcpYEEI peptide, which was estimated on the basis of a IC_{50}). Given the complexity of the system and the magnitude of the molecular interactions involved, such a good agreement is very satisfying. The dominant contribution to the error in the overall binding free energy arises from the PMF involved in the factor I^* given by Eq. **12**; all other errors are correspondingly much less important. The error on the radial PMF was estimated by using different fractions of the overall time-series data collected

Table 2. Binding free energy

Peptide ligand	G_{bind}, kcal/mol
AcpYEEIP*	−9.5
pYEEIP*	−8.2
pYEEI*	−7.6
AcpYEEI†	−7.1
AcpYEEI‡	−8.8

*Ref. 24.
†Ref. 42 based on IC_{50}.
‡This work.

over the course of umbrella sampling simulations ($4 \times 1/4$ of the total). This estimate suggests that the error on the binding free energy is on the order of 2–3 kcal/mol (see also Fig. 5, which is published as supporting information on the PNAS web site). For the sake of comparison, the binding free energy also was calculated according to the MM/PB-SA scheme (9). In this approximation, an ensemble of representative configurations of the bound protein–ligand complex is generated for which the Poisson–Boltzmann continuum electrostatic calculations are performed to obtain the electrostatic free energies of the protein–ligand complex, protein, and ligand only. The overall binding free energy obtained from MM/PBSA is approximately −80 kcal/mol (see Table 3, which is published as supporting information on the PNAS web site) nearly an order of magnitude larger compared with both the experimental values and the result from the free energy simulations with explicit solvent (Table 2). Applications of MM/PB-SA often ignore the translational/orientational factor in Eq. **7** and consider exclusively the ligand–protein interaction free energy, although this treatment is incorrect (see refs. 7 and 29 for discussions). Here, those factors were estimated from the rms fluctuations of the Euler angles and the center of mass of the peptide in the bound state MD trajectories in the harmonic approximation as $\Delta\omega \simeq 0.045$ rad, and $\Delta V \simeq 5.35$ Å^3, yielding an orientational free energy contribution of $-k_BT \ln[\Delta\omega/8\pi^2]$ of 4.5 kcal/mol, and a translational free energy contribution of $-k_BT \ln[C^\circ\Delta V]$ of 3.4 kcal/mol, for a total of 7.9 kcal/mol. Clearly, the large inaccuracy arises from the continuum solvent approximation used to compute the interaction free energy and not from those smaller contributions.

One previously undescribed aspect of the present formulation is the introduction of the conformational restraint u_c based on the rmsd relative to a reference structure (here the average conformation of the bound ligand). Fig. 3 shows the PMF calculated both in the binding site and in the bulk. Together, these two PMFs help quantify the free energy "cost" arising from the loss of conformational freedom of the ligand when it must adopt a specific conformation in the binding site. Whereas a single specific average conformation of the peptide is strongly favored in the bound state (Fig. 3*A*), a wide range of conformational states differing from the bound conformation is allowed in the bulk (Fig. 3*B*). The most frequent conformations of the isolated peptide in solution are ≈4 Å in rmsd relative to the average bound conformation. A closer examination reveals that those correspond to a range of widely different conformations rather than a uniquely identifiable structure, which is indicative of the significant conformational freedom of the peptide in solution. One may note the steep rise in both PMFs at very small rmsd ($\xi < 1$ Å). This rise is related to the impossibility of quenching all of the thermal fluctuations away from a given reference structure, which is akin to an effective Jacobian associated with the rmsd volume element $d\xi$ in the multidimensional configurational space. With the PMF as a function of ξ, the free energies [G_c^{bulk} and G_c^{site}] can be calculated by direct integration of the Boltzmann factor. The ligand conformation is restrained while it is in the bulk and released once it is in the binding site, yielding the total free energy cost associated with the loss of conformational freedom of the flexible peptide, $G_c^{bulk} - G_c^{site}$.

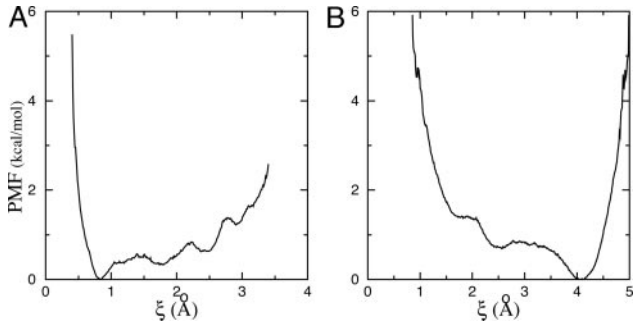

Fig. 3. Calculated PMF $w(\xi)$ of the ligand conformational degrees of freedom as function of the rmsd ξ in the bound (*A*) and bulk (*B*) states.

Fig. 4 shows the PMF $\mathcal{W}(r_1)$ as a function of the distance r_1 along the radial axis. The free energy rises steeply near the immediate vicinity of the sharp minimum at 12 Å, the stable bound state, and becomes nearly flat for $r_1 > 30$ Å (continuum electrostatic calculations based on PB confirms that the interactions between the peptide and the SH2 domain are essentially screened out at this distance). The dominant contribution to the overall binding free energy, on the order of −18 kcal/mol, clearly arises from the PMF $\mathcal{W}(r_1)$, corresponding to the interactions gained as the peptide moves from the bulk solution into the binding pocket. The integral of the PMF in Eq. **12** is dominated by the contribution from the immediate vicinity of the free energy minimum near 12 Å. For this reason, the results are not affected by the precise definition of the bound state. The strong favorable interaction in $\mathcal{W}(r_1)$ is counterbalanced by unfavorable contributions. Regrouping the various contributions according to their physical significance, one finds that the overall free energy cost associated with the loss of orientational freedom upon binding, [$G_o^{bulk} - G_o^{site}$], is on the order of 5 kcal/mol. Similarly, the free energy cost associated with the loss of internal conformational freedom, [$G_c^{bulk} - G_c^{site}$], is ≈3 kcal/mol. The magnitude of these unfavorable contributions is significant compared with the value of G_{bind}. The total cost of the conformational restriction of the peptide is, in fact, slightly overestimated because the rmsd restraining potential is not accounting for the existence of physically equivalent isomers. For example, the Tyr side chain with its phosphate group could adopt other equivalent rotameric states (3 and 2, respectively) that are treated as distinct states by the rmsd restraint. In the bulk, six equivalent rotameric states of the pTyr side chain can be visited, and the PMF $w^{bulk}(\xi)$ overestimates the free energy cost for the reference conformation by $k_BT \ln(6) = 1.07$ kcal/mol. In the bound state, these states are not easily sampled because the side chain is tightly surrounded by the protein, and there is no correction.

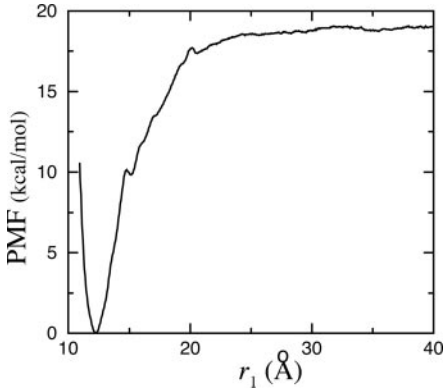

Fig. 4. Calculated PMF $W(r_1)$ as a function of the radial distance r_1 between the ligand center of mass relative to the protein.

In the present calculations, the binding constant was expressed as a step-by-step reversible process, and the ligand-receptor PMF reporting the reversible work to translate the ligand from the bulk solution to the bound state is a central quantity. The alternative "double-decoupling" scheme (12–17, 43) augmented by various restraining potentials is particularly advantageous for computing the binding constant of a ligand bound inside a pocket located deeply in the interior of a protein because it does not require any consideration of the pathway leading from the bulk to the binding site. Nonetheless, this method is nearly impractical when the absolute solvation free energy of the ligand is very large because the binding free energy then is computed as the (small) difference between two very large numbers. In such cases, the limited statistical precision of the FEP computations becomes a major hurdle, which is particularly problematic in computing the binding constant of a highly charged ligand such as the pYEEI peptide. The electrostatic contribution to the total solvation free energy of the ligand in bulk water, calculated by using FEP simulation with explicit solvent molecules, is of the order of −773 kcal/mol. This large number contrasts with the binding free energy, which is of the order of −8 kcal/mol (Table 2). Even if the statistical uncertainty of the FEP calculation was only ≈1% of the total solvation free energy, that would still translate into an error that is of the same order of magnitude as the quantity of interest itself. Attempts to compute the binding free energy of the pYEEI peptide with the SH2 domain by using a double-decoupling scheme based on Eq. **20** exhibited errors on the order of 75 kcal/mol (data not shown). To avoid such problems, the binding constant was expressed in terms of the ligand-receptor PMF, without involving the decoupled state.

The idea of estimating binding constants directly from the ligand-receptor PMF is not novel; for example, a radial bimolecular PMF was used in ref. 44. However, an approach based on the true unbiased ligand-receptor PMF would be of little use in practice because sampling over the entire range of possible motions of the ligand, receptor, and solvent would be computationally prohibitive. The present development shows that such extensive sampling is not needed and that the problem of computing the equilibrium binding constant can be rigorously formulated in terms of the PMF of the ligand restrained along the 1D axis r_1. The computational efficiency also is improved by introducing the conformational and orientational restraining potentials, u_c and u_o, which further contribute to reduce the amount of configurational space that needs to be thoroughly sampled. Those two restraining potentials help significantly in the statistical convergence of the restrained PMF calculated along the 1D radial axis r_1.

Conclusion

We have presented an efficient PMF-based computational method for calculating the equilibrium association constant between a flexible ligand and a protein receptor from MD simulations with explicit solvent. The result of the calculation is within ≈1 kcal/mol from experimental estimates. Such a good agreement may be partly fortuitous, although it is very encouraging and suggests that accurate computations of absolute binding free energies from all-atom simulations is an achievable goal.

The present PMF-based approach avoids the double-decoupling scheme (12–17), which can lead to significant difficulties in the case of highly charged ligands. Exact and computationally advantageous expressions for the equilibrium binding constant were derived directly from configurational ensemble averages. The derivation is straightforward and has the advantage of avoiding arguments based on chemical potentials invoked in the traditional framework (16, 45). Precision of the results is only limited by the computational requirements for the adequate sampling of configurations, with no uncontrollable approximations or assumptions.

One of the previously undescribed aspects of the present method is the configurational restriction potential based on the rmsd relative to the average structure of the bound state. The introduction of the rmsd into the formulation of the equilibrium binding constant permits a rational and quantitative discussion of the free energy cost associated with the loss of conformational freedom of the ligand to adopt a given bound conformation. From a practical point of view, the rmsd restraining potential essentially transforms a flexible ligand into a relatively rigid one, thereby reducing significantly the difficulties associated with the sampling of a multitude of conformations. In the present case, the configurational restriction was only applied to the flexible pYEEI ligand. Because the SH2 domain is fairly rigid, one expects that its small atomic fluctuations will be readily sampled spontaneously during unbiased MD simulations. Nonetheless, some receptor proteins are known to undergo significant conformational changes upon ligand binding, e.g., the HIV protease (46). The present approach, based on the PMF of the rmsd relative to a reference structure, could be generalized to quantify the importance of conformational flexibility of the receptor.

We thank Stefan Boresch, Jose Faraldo-Gomez, and Yuqing Deng for their comments on the manuscript. This work was supported by National Institutes of Health Grant CA-93577. The computations were carried out at the Pittsburgh Supercomputing Center with a grant from the National Resource Allocation Committee.

1. Wlodawer, A. (2002) *Annu. Rev. Med.* **53,** 595–614.
2. Vindigni, A. (1999) *Combinatorial Chem. High Throughput Scr.* **2,** 139–153.
3. Cheng, A. C., Calabro, V. & Frankel, A. D. (2001) *Curr. Opin. Struct. Biol.* **11,** 478–484.
4. Garvie, C. W. & Wolberger, C. (2001) *Mol. Cell.* **8,** 937–946.
5. Pawson, T. & Nash, P. (2003) *Science* **300,** 445–452.
6. Schneider, G. & Bohm, H. J. (2002) *Drug Discovery Today* **7,** 64–70.
7. Luo, H. & Sharp, K. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 10399–10404.
8. Roux, B. & Simonson, T. (1999) *Biophys. Chem.* **78,** 1–20.
9. Massova, I. & Kollman, P. A. (2000) *Perspect. Drug Discovery* **18,** 113–135.
10. Brandsdal, B. O., Osterberg, F., Almlof, M., Feierberg, I., Luzhkov, V. B. & Aqvist, J. (2003) *Adv. Protein Chem.* **66,** 123–158.
11. Kollman, P.A. (1993) *Chem. Rev.* **93,** 2395–2417.
12. Hermans, J. & Shankar, S. (1986) *Isr. J. Chem.* **27,** 225–227.
13. Helms, V. & Wade, R. C. (1998) *J. Am. Chem. Soc.* **120,** 2710–2713.
14. Roux, B., Nina, M., Pomès, R. & Smith, J. C. (1996) *Biophys. J.* **71,** 670–681.
15. Hermans, J. & Wang, L. (1997) *J. Am. Chem. Soc.* **119,** 2707–2714.
16. Boresch, S., Tettinger, F., Leitgeb, M. & Karplus, M. (2003) *J. Phys. Chem. B* **107,** 9535–9551.
17. Mann, G. & Hermans, J. (2000) *J. Mol. Biol.* **302,** 979–989.
18. Pawson, T. & Gish, G. D. (1992) *Cell* **71,** 359–362.
19. Bradshaw, J. M., Mitaxov, V. & Waksman, G. (1999) *J. Mol. Biol.* **293,** 971–985.
20. Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D. & Kuriyan, J. (1993) *Cell* **72,** 779–790.
21. Tong, L., Warren, T. C., King, J., Betageri, R., Rose, J. & Jakes, S. (1996) *J. Mol. Biol.* **256,** 601–610.
22. Sheinerman, F. B., Al-Lazikani, B. & Honig, B. (2003) *J. Mol. Biol.* **334,** 823–841.
23. Zhou, S. Y., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G., King, F., Roberts, T., Ratnofski, S., Lechleider, R. J., *et al.* (1993) *Cell* **72,** 767–778.
24. Cousins-Wasti, R. C., Ingraham, R. H., Morelock, M. M. & Grygon, C. A. (1996) *Biochemistry* **35,** 16746–16752.
25. Bradshaw, J. M. & Waksman, G. (2002) *Adv. Protein Chem.* **61,** 161–210.
26. Zvelebil, M. J., Panayotou, G., Linacre, J. & Waterfield, M. D. (1995) *Protein Eng.* **8,** 527–533.
27. Lee, J. K., Moon, T., Chi, M. W., Song, J. S., Choi, Y. S. & Yoon, C. N. (2003) *Biochem. Biophys. Res. Commun.* **306,** 225–230.
28. Bradshaw, J. M. & Waksman, G. (1998) *Biochemistry* **37,** 15400–15407.
29. Swanson, J. M. J., Henchman, R. H. & McCammon, J. A. (2004) *Biophys. J.* **86,** 67–74.
30. Sanner, M. F., Olson, A. J. & Spehner, J. C. (1996) *Biopolymers* **38,** 305–320.
31. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4,** 187–217.
32. MacKerell, A. D., Jr., Bashford, D., Bellot, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., *et al.* (1998) *J. Phys. Chem. B* **102,** 3586–3616.
33. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983) *J. Chem. Phys.* **79,** 926–935.
34. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. (1995) *J. Chem. Phys.* **103,** 8577–8593.
35. Simonson, T., Archontis, G. & Karplus, M. (1997) *J. Phys. Chem. B* **101,** 8349.
36. Torrie, G. M. & Valleau, J. P. (1977) *J. Comp. Phys.* **23,** 187–199.
37. Ferrenberg, A.M. & Swendsen, R.H. (1989) *Phys. Rev. Lett.* **63,** 1195–1198.
38. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. (1992) *J. Comp. Chem.* **13,** 1011–1021.
39. Roux, B. (1995) *Comp. Phys. Comm.* **91,** 275–282.
40. Beglov, D. & Roux, B. (1994) *J. Chem. Phys.* **100,** 9050–9063.
41. Nina, M., Beglov, D. & Roux, B. (1997) *J. Phys. Chem. B* **101,** 5239–5248.
42. Nam, N. H., Ye, G., Sun, G. & Parang, K. (2004) *J. Med. Chem.* **47,** 3131–3141.
43. Jorgensen, W. L., Buckner, J. K., Boudon, S. & Tirado-Rives, J. (1988) *J. Chem. Phys.* **89,** 3742–3746.
44. Jorgensen, W. L. (1989) *J. Am. Chem. Soc.* **111,** 3770–3772.
45. Gilson, M. K., Given, J. A., Bush, B. L. & McCammon, J. A. (1997) *Biophys. J.* **72,** 1047–1069.
46. Carlson, H. A. & McCammon, J. A. (2000) *Mol. Pharmacol.* **57,** 213–218.