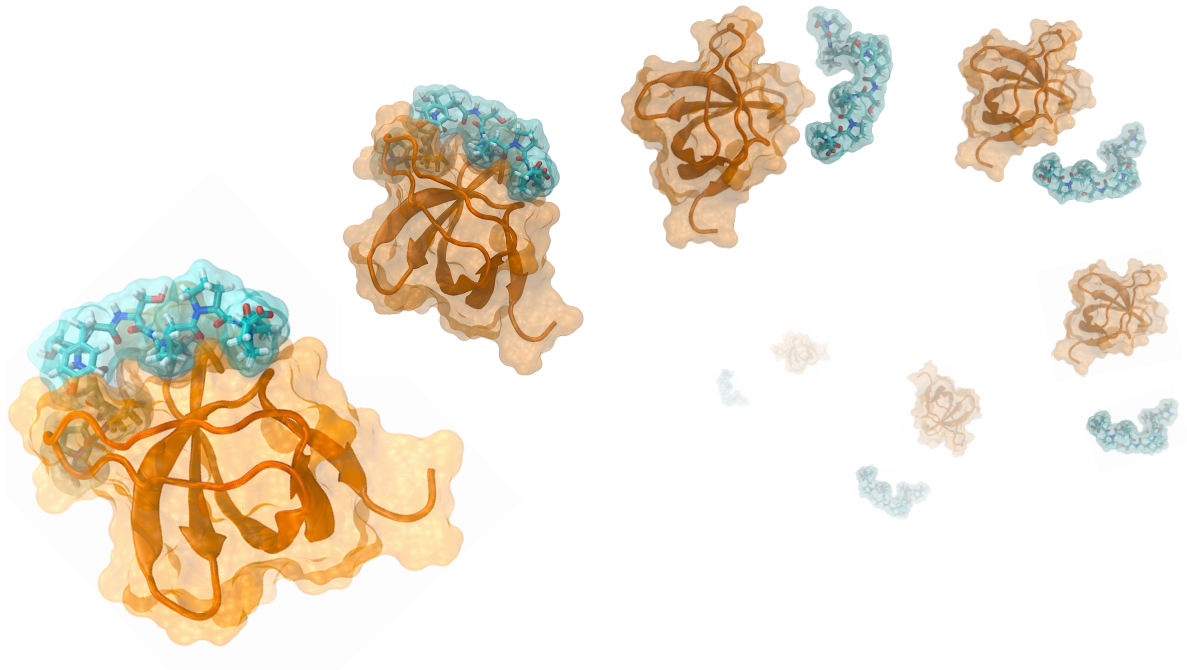
School of Physics
Georgia Institute of Technology

Department of Biochemistry and Molecular Biology
Gordon Center for Integrative Science
The University of Chicago

Centre National de la Recherche Scientifique
Laboratoire International Associé CNRS-UIUC
Université de Lorraine

University of Illinois at Urbana-Champaign
Beckman Institute for Advanced Science and Technology
Theoretical and Computational Biophysics Group

Protein:ligand standard binding free energies: A tutorial for alchemical and geometrical transformations



James Gumbart
Benoît Roux
Christophe Chipot

September 19, 2017

Please visit www.ks.uiuc.edu/Training/Tutorials/ to get the latest version of this tutorial, to obtain more tutorials like this one, or to join the `tutorial-l@ks.uiuc.edu` mailing list for additional help.

Abstract

This tutorial sets out to demonstrate the application of numerical simulations to the calculation of the standard binding free energy of a protein:ligand complex. To a large extent, computation of standard binding free energies remains a daunting theoretical challenge on account of the considerable variations in conformational, translational and orientational entropies that accompany the association of the substrate to the host protein and is not easily captured by conventional molecular-dynamics simulations. Sampling these entropic contributions is addressed here, following two distinct routes, an alchemical one and a geometrical one, wherein restraining potentials have been introduced to act on the collective variables that define the conformation of the ligand and its relative position and orientation with respect to the protein. The methodology is illustrated by the well-documented example of a small, proline-rich peptide, referred to as p41, associating to the Src homology 3 domain of a tyrosine kinase with a standard binding free energy of -7.94 kcal/mol. Considering the difficulties that current additive force fields face to describe organic ligands, e.g., drugs, the choice of an all-peptide-based molecular assembly is particularly fitting, allowing the reader to focus primarily on the methodology and the sampling.

Contents

1. Introduction

The primary objective of this tutorial is to compute the standard binding free energy of a ligand to a protein, using two distinct strategies, relying upon alchemical transformations, on the one hand, and geometrical transformations, on the other hand [1,2]. From a theoretical perspective, accurate determination of absolute free energies remains a daunting challenge, owing to the considerable variation of not only translational and rotational, but also conformational entropies underlying the binding process and which cannot be fully captured in routine, finite-length statistical simulations [3]. To address this challenge, a series of geometrical restraints is introduced to act on relevant collective variables [1,4–6], thereby alleviating the inherent sampling limitations of molecular dynamics. In this tutorial, the proposed strategy is applied to the case of the src–homology 3, or SH3, domain of tyrosine kinase Abl binding a short, proline-rich peptide referred to as p41 and of amino-acid sequence APSYSPPPPP [7,8], employing both geometrical, i.e., potential-of-mean-force, and alchemical free-energy calculations.

The reader of this tutorial is assumed to be familiarized with the use of NAMD [9] to perform standard and advanced computations, including energy minimization, molecular-dynamics simulations and free-energy calculations, both perturbative, i.e., free-energy perturbation [10, 11], and geometrical, i.e., adaptive biasing force [12, 13].

This tutorial contains advanced material. Do not attempt to tackle the problems therein if you have no preliminary experience with free-energy calculations. The neophyte reader eager to get acquainted with the computation of protein:ligand standard binding affinities is advised to complete first the introductory tutorials on alchemical, free-energy perturbation calculations and adaptive-biasing-force calculations.

The ABF algorithm is implemented as part of the "collective variable caculations" (`colvars`) module of NAMD. The `colvars` module [14, 15] is extensively documented in the NAMD user's guide. Other information about the ABF method can be found in the reference article 16. A basic working knowledge of VMD is highly recommended.

Completion of this tutorial requires

- the various files contained in the archive `tutorial-protein-ligand.tar.gz` located at `http://www.ks.uiuc.edu/Training/Tutorials/namd/PLB`;
- NAMD 2.12 or later (`http://ks.uiuc.edu/Research/namd`);
- VMD 1.9 or later (`http://ks.uiuc.edu/Research/vmd`).

1.1. Theoretical underpinnings

The equilibrium binding constant that characterizes the reversible association of the ligand with the protein, i.e., $\mathrm{protein} + \mathrm{ligand} \rightleftharpoons \mathrm{protein{:}ligand}$,

$$K_{\mathrm{eq}} = \frac{[\mathrm{protein{:}ligand}]}{[\mathrm{protein}][\mathrm{ligand}]} \tag{1}$$

can be restated as,

$$K_{\mathrm{eq}} = \frac{1}{[\mathrm{ligand}]}\frac{p_1}{p_0} \tag{2}$$

where p_0 is the fraction of free protein and p_1, the fraction of protein binding the ligand. Assuming low protein concentration, one can imagine an isolated protein in a solution of N indistinguishable ligands. Under these premises, the logarithm of the ratio p_1/p_0 can be related to the reversible work required to extract one ligand from its bulk environment and bring it to the binding site of the protein,

$$\begin{aligned} K_{\mathrm{eq}} &= \frac{1}{[\mathrm{ligand}]}\frac{N\displaystyle\int_{\mathrm{site}}\mathrm{d}\mathbf{1}\ldots\int_{\mathrm{bulk}}\mathrm{d}\mathbf{N}\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}}{\displaystyle\int_{\mathrm{bulk}}\mathrm{d}\mathbf{1}\ldots\int_{\mathrm{bulk}}\mathrm{d}\mathbf{N}\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}} \\ &= \frac{1}{[\mathrm{ligand}]}\frac{N\displaystyle\int_{\mathrm{site}}\mathrm{d}\mathbf{1}\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}}{\displaystyle\int_{\mathrm{bulk}}\mathrm{d}\mathbf{1}\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}} \end{aligned} \tag{3}$$

Here, the subscript "site" and "bulk" refer to the bound and the unbound states of the ligand, and the relevant regions of configurational space over which the integrals are evaluated. As a convention, ligand "1" occupies the binding site of the protein. Since the bulk environment is isotropic and homogenous, it follows that,

$$K_{\mathrm{eq}} = \frac{1}{[\mathrm{ligand}]}\frac{N\displaystyle\int_{\mathrm{site}}\mathrm{d}\mathbf{1}\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}}{V_{\mathrm{bulk}}\displaystyle\int_{\mathrm{bulk}}\mathrm{d}\mathbf{1}\;\delta(\mathbf{x}_1-\mathbf{x}_1^*)\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}} \tag{4}$$

V_{bulk} is the volume of the bulk medium. $\mathbf{x}_1$ is the position of the center of mass of the ligand and $\mathbf{x}_1^*$, an arbitrary location in the bulk medium, sufficiently far from the binding site of the protein. Since the concentration of the ligands is equal to N/V_{bulk}, the equilibrium binding constant can be simplified to,

$$K_{\mathrm{eq}} = \frac{\displaystyle\int_{\mathrm{site}}\mathrm{d}\mathbf{1}\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}}{\displaystyle\int_{\mathrm{bulk}}\mathrm{d}\mathbf{1}\;\delta(\mathbf{x}_1-\mathbf{x}_1^*)\int\mathrm{d}\mathbf{x}\,\mathrm{e}^{-\beta U}} \tag{5}$$

In the above expression of the equilibrium constant, the numerator reflects the final state of the binding

process and the denominator, its initial state. Per se, this expression is of limited use in the context of molecular dynamics, evaluation of the individual configurational integrals being euphemistically impractical. A viable strategy towards the computation of K_{eq} consists in decomposing the reversible association phenomenon into several stages, the corresponding contribution of which is determined in separate simulations. Two possible routes, a geometrical route and an alchemical route, can be followed to reach this objective. These two routes have a common denominator — assuming that over the timescales amenable to molecular dynamics the configurational space available to the ligand cannot be sampled adequately, a series of biasing potentials is introduced to restrain the substrate in the native conformation, position and orientation of the bound state. Introduction of properly chosen restraints [1, 4] constitutes the preamble to the separation of the ligand from the protein, either geometrically, i.e., translation, or alchemically, i.e., decoupling.

1.2. Geometrical transformations

From a geometrical standpoint, computation of the equilibrium constant, K_{eq}, could be conceived resorting to a simple one-dimensional potential-of-mean-force calculation, $w(r)$, wherein the ligand is separated reversibly from the protein [17],

$$K_{\mathrm{eq}} = 4\pi \int_0^R \mathrm{d}r \; r^2 \; \mathrm{e}^{-\beta w(r)} \tag{6}$$

Here, R stands for the limit of association.

This vision of the problem at hand is not only naïve, but also extremely deceitful. It assumes that over the timescale of the simulation, the ligand can sample the available configurational space. Nothing could be further from the truth, in particular for large, flexible ligands.

One possible route to access with appreciable accuracy the equilibrium constant, K_{eq}, consists in a series of geometrical transformations, wherein the substrate is progressively restrained in the native conformation, position and orientation of the bound state [1, 2]. In other words, granted that finite-length simulations cannot capture the conformational variability of the ligand in the course of the binding process, nor sample the available $8\pi^2$ of solid angle, it is preferable to tether the molecule of interest with a suitable set of restraints and subsequently evaluate the free-energy cost due to these restraints.

Introduction of geometrical restraints in the simulations can be viewed as a loss of conformational, positional and orientational entropies, the contribution to the free energy of which must be determined independently [3].

Under the assumption that the protein does not undergo perceptible conformational change in the course

of ligand association, which is a valid premise in the instance of the SH3 domain of Abl binding p41, the series of geometrical transformations involves the following steps:

(1) Determine the free-energy change for deforming the ligand in the protein:ligand complex ("site"), using as a collective variable the root mean-square deviation with respect to the conformation of the ligand in the bound state.

(2) Determine the free-energy change for reorienting the ligand in the protein:ligand complex ("site") about the first Euler angle, Θ, using as a collective variable a valence angle, restraining the conformation to that of the bound state (u_c).

(3) Determine the free-energy change for reorienting the ligand in the protein:ligand complex ("site") about the second Euler angle, Φ, using as a collective variable a dihedral angle, restraining the conformation and the orientation with respect to Θ (u_Θ) to that of the bound state.

(4) Determine the free-energy change for reorienting the ligand in the protein:ligand complex ("site") about the third Euler angle, Ψ, using as a collective variable a dihedral angle, restraining the conformation and the orientation with respect to Θ and Φ (u_Φ) to that of the bound state.

(5) Determine the free-energy change for changing the position of the ligand in the protein:ligand complex ("site") about the first polar angle, θ, using as a collective variable a valence angle, restraining the conformation and the orientation with respect to Θ, Φ and Ψ (u_Ψ) to that of the bound state.

(6) Determine the free-energy change for changing the position of the ligand in the protein:ligand complex ("site") about the second polar angle, ϕ, using as a collective variable a dihedral angle, restraining the conformation, the orientation with respect to Θ, Φ and Ψ, and the position with respect to θ (u_θ) to that of the bound state.

(7) Determine the free-energy change for changing the position of the ligand in the protein:ligand complex ("site") along the vector connecting their respective center of mass, r, using as a collective variable a Euclidian distance, restraining the conformation, the orientation with respect to Θ, Φ and Ψ, and the position with respect to θ and ϕ (u_ϕ) to that of the bound state.

(8) Determine the free-energy change for deforming the ligand in the free, unbound state ("bulk"), using as a collective variable a root mean-square deviation with respect to the conformation of the ligand in the bound state.

Once the eight, individual potentials of mean force are generated, the "geometrical" equilibrium constant is determined according to,

$$
\begin{aligned}
K_{\mathrm{eq}}^{\mathrm{geom}} = & \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta U}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c)}} \times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o+u_a)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o)}} \\
& \times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o)}} \times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c)}} \\
& \times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o+u_a)}} \times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta U}}
\end{aligned}
\tag{7}
$$

where $u_o = u_\Theta + u_\Phi + u_\Psi$ denotes the orientational potential and $u_a = u_\theta + u_\phi$, the polar-angle potential.

The three first contributions correspond, respectively, to the conformational, orientational and positional restraints acting on the ligand and represent six independent potentials of mean force. The fourth contribution corresponds to the reversible separation of the ligand from the binding site of the protein towards the bulk environment. It should be clearly understood that this potential-of-mean-force is performed, keeping all the other relevant conformational, positional and orientational degrees of freedom at their equilibrium value by means of geometrical restraints. The fifth contribution highlighted in cyan corresponds to the reorientation of a rigid body, i.e., the ligand restrained in its native conformation when bound to the protein, and can, thus, be evaluated analytically. The sixth and last contribution corresponds to free-energy cost to restrain in the bulk environment the conformation of the ligand to that in the native bound state. It constitutes the eighth and last potential-of-mean-force calculation towards the determination of K_{eq}.

1.3. Alchemical transformations

The second route for the computation of the $\mathrm{protein} : \mathrm{ligand}$ equilibrium constant consists of a series of alchemical transformations [18]. This route follows the thermodynamic cycle depicted in Figure 1, where the ligand, either in the free, unbound state, or in the bound state, is not decoupled reversibly from the environment as is, i.e., ligand^0, but rather restrained in its native conformation, position and orientation, i.e., ligand^*, characteristic of the bound state. The rationale for the introduction of a suitable set of geometric restraints in the alchemical route is rooted in the so-called wandering-ligand problem [19–21], whereby, upon decoupling of the substrate from its environment, the former becomes free to drift away from the binding site. It follows that in the backward, coupling transformation, the ligand is unlikely to form in the binding site the relevant, native network of intermolecular interactions, hence, violating the underlying principle of thermodynamic micro-reversibility.

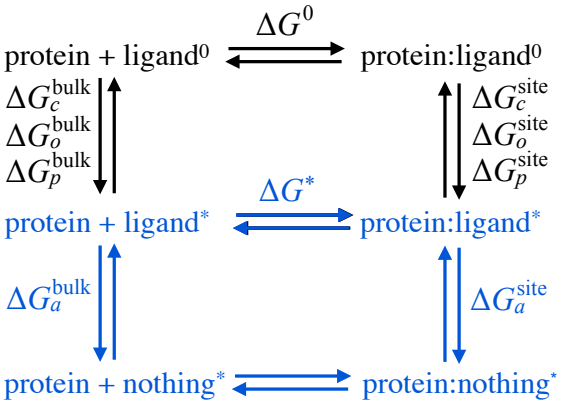


Figure 1: Complete thermodynamic cycle delineating the reversible association of a ligand to a protein and the necessary steps to determine the corresponding standard binding free energy, ΔG^0. "ligand^0" denotes an unrestrained ligand, whereas "ligand^*" refers to a ligand restrained in its native conformation, position and orientation in the protein:ligand complex.

Under the same assumption made previously that the protein does not undergo an appreciable conformational modification as the ligand binds to it, the series of alchemical transformations involves the following steps:

(1) To circumvent the wandering-ligand problem arising when the substrate is decoupled from the protein, restrain the former in the conformation, orientation and position representative of the protein:ligand complex.

(2) Determine in the bound state ("site") the alchemical free-energy change for decoupling reversibly from the protein the ligand restrained in its native conformation, orientation and position, i.e., $\mathrm{protein} : \mathrm{ligand}^*$.

(3) Determine in the free, unbound state ("bulk") the alchemical free-energy change for decoupling reversibly from the bulk water the ligand restrained in its native conformation, i.e., ligand^*.

(4) Determine in the bound state ("site") the free-energy contribution for maintaining the ligand in its native conformation, orientation and position by means of restraining potentials (u_c, u_Θ, u_Φ, u_Ψ, u_θ, u_ϕ and u_r). Towards this end, the set of collective variables utilized are a root mean-square deviation with respect to the conformation of the ligand in the bound state, the three Euler angles, Θ, Φ and Ψ, the two polar angles, θ and ϕ, and the Euclidian distance, r, separating the substrate from the protein.

(5) Determine in the free, unbound state ("bulk") the free-energy contribution for maintaining the ligand in its native conformation by means of a restraining potential (u_c). Towards this end, the chosen collective variable is the root mean-square deviation with respect to the conformation of the substrate when bound to the protein.

Once the two alchemical free-energy changes, i.e., the reversible decoupling in the bound and in the unbound states, alongside the eight free-energy contributions due to the geometrical restraints, i.e., seven terms when the ligand is in the bound state and a single one when it is in the unbound state, are computed, the "alchemical" equilibrium constant is determined according to,

$$
\begin{aligned}
K_{\mathrm{eq}}^{\mathrm{alch}} = & \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta U_1}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c)}} \\
& \times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c+u_o)}} \\
& \times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c+u_o)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c+u_o+u_a)}} \\
& \times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c+u_o+u_a)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c+u_o+u_a+u_r)}} \\
& \times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c+u_o+u_a+u_r)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_0+u_c+u_o+u_a+u_r)}} \\
& \times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_0+u_c+u_o+u_a+u_r)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_0+u_c+u_o)}} \\
& \times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_0+u_c+u_o)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_0+u_c)}} \\
& \times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_0+u_c)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c)}} \\
& \times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U_1+u_c)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta U_1}}
\end{aligned}
\tag{8}
$$

where U_0 characterizes the non-interacting ("ghost") state of the substrate, and U_1 the state in which it is coupled to the environment. The above equation follows precisely the different steps described in the thermodynamic cycle of Figure 1. Its first two contributions arise, respectively, from the conformational and the orientational restraints acting on the ligand. Its third contribution corresponds to the polar-angle term, i.e., $u_a = u_\theta + u_\phi$, of the positional restraints. Its fourth contribution corresponds to the translational term, i.e., u_r, of the positional restraints. The fifth and the eighth contributions highlighted in magenta correspond to the alchemical transformations, whereby the substrate is decoupled reversibly from its environment, respectively in the bound and in the unbound states. The sixth and the seventh contributions highlighted in cyan are analytical ones and account for the reorientation and translation of a rigid body in an homogenous bulk liquid. The ninth and last contribution represents the deformation of the ligand in the free, unbound state.

One of the main goals of the present tutorial is to demonstrate that the geometrical and the alchemical routes are overall equivalent, and that the standard binding affinity of p41 towards the SH3 domain of Abl obeys $K_{\mathrm{eq}} = K_{\mathrm{eq}}^{\mathrm{geom}} = K_{\mathrm{eq}}^{\mathrm{alch}}$.

2. Setting up the simulations

The starting point of the setups for the free-energy calculations is PDB entry `1bbz`. Generation of the structure, PSF files and the initial configurations will be performed using the `psfgen` module of VMD [22]. Because the two routes for calculating $K_{\rm eq}$, from whence the standard free energy of binding of Abl to the SH3 domain of Abl is inferred, are inherently of different nature, distinct setups will be considered for the geometrical and the alchemical transformations.

2.1. Construction of the molecular assemblies

In the geometrical route, as has been outlined above, eight different potentials of mean force ought to be determined. To optimize the computational effort, three distinct simulation cells will be devised, namely,

(1) the protein:ligand assembly solvated in a cubic box of TIP3P water with adequate padding to avoid periodicity-induced artifacts. Given the size of the dimer and the charged termini of the substrate, solvation by about 3,400 water molecules, which corresponds to a dimension of about $48 \times 48 \times 48$ Å^3, has proven appropriate. This setup will be utilized for the computation of potentials of mean force, wherein the ligand remains associated to the protein, i.e., the conformational, the Euler-angle, orientational and the polar-angle, positional terms of Equation 7.

(2) the protein:ligand assembly solvated in a rectangular box of TIP3P water with adequate padding to avoid periodicity-induced artifacts. This setup will be utilized for the computation of the separation potential of mean force. Considering the extent of the separation of the substrate from the binding site, a cell of dimension of about $48 \times 48 \times 68$ Å^3 is perfectly adapted. Choice of a rectangular cell supposes that relevant geometrical restraints are enforced to prevent the complex from tumbling as the ligand moves away from the protein, i.e., collective variables `orientation` and `distance` with respect to a dummy particle. Alternatively, the user can resort to a cubic box of adequate dimension to allow the dimer to tumble in the course of the separation. A simulation cell of dimension roughly equal to $60 \times 60 \times 60$ Å^3, i.e., about 6,300 water, has proven adapted to the determination of the separation potential of mean force.

(3) the free ligand solvated in a cubic box of TIP3P water with adequate padding to avoid periodicity-induced artifacts. For simplicity, the same simulation cell will be utilized for the solvation by about 3,400 water molecules has proven appropriate and corresponds to a dimension of about $48 \times 48 \times 48$ Å^3. This setup will be utilized to determine the potential of mean force for deforming the ligand in the unbound state.

Conversely, in the alchemical route, only two different setups need to be devised, corresponding to the

bound and unbound states of the substrate, or the lefthand– and righthand sides of the thermodynamic cycle of Figure1, namely,

(1) the protein:ligand assembly solvated in a cubic box of TIP3P water with adequate padding to avoid periodicity-induced artifacts. Given the size of the dimer, solvation by about 3,400 water molecules has proven appropriate and corresponds to a dimension of about $48 \times 48 \times 48$ Å^3. This setup will be utilized for the computation of the alchemical free-energy changes, wherein the restrained ligand is decoupled reversibly from the protein, together with the free-energy contributions arising from conformational, orientational and positional restraints.

(2) the free, unbound ligand solvated in a cubic box of TIP3P water with adequate padding to avoid periodicity-induced artifacts. For simplicity and given the size of the ligand and its charged termini, solvation by about 3,400 water molecules has proven appropriate and corresponds to a dimension of about $48 \times 48 \times 48$ Å^3. This setup will be utilized to determine the alchemical free-energy change arising from the reversible decoupling of the unbound substrate from its aqueous environment, and the free-energy contribution incurred in its deformation with respect to the reference conformation in the protein:ligand complex.

2.2. Definition of the collective variables

A fundamental aspect of the approach detailed in this tutorial, following either a geometrical route, or an alchemical one, is the introduction of a suitable set of harmonic restraints to preserve the conformation, the orientation and the position of the ligand as it is dissociated reversibly from the protein.

Since molecular dynamics, in general, cannot capture the deformation, reorientation and repositioning of the ligand as it binds to the protein, it is *preferable* to restrain the former and evaluate the free-energy contribution due to the loss of configurational entropy.

Incorporation of restraints in the free-energy calculations presupposes that the position and the orientation of the substrate with respect to the protein can be described without ambiguity. Toward this end, a frame of reference is designed, from which the position of the ligand can be defined [2]. This frame of reference is formed by three groups of atoms of the protein, $\{ \mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3 \}$. Likewise, a frame of reference formed by three groups of atoms of the ligand, $\{ \mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3 \}$, is introduced to define the orientation of the latter, as depicted in Figure 2.

A-posteriori verification utilizing alternate frames of reference for both the protein and its substrate, demonstrates that the equilibrium binding constant does not depend on the choice of the triplets $\{ \mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3 \}$ and $\{ \mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3 \}$, provided unambivalent definition of the restrained degrees of freedom,

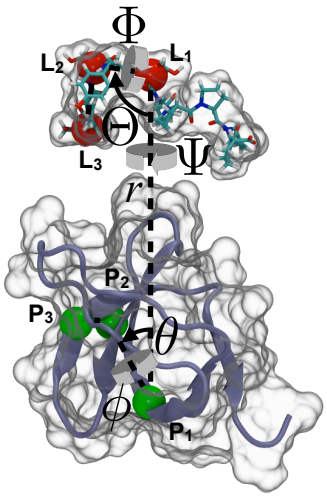


Figure 2: Binding of proline-rich ligand p41 to the SH3 domain of tyrosine kinase Abl. The position of p41 with respect to the protein is expressed by means of a Euclidian distance, r, and two polar angles, θ and ϕ, which altogether form a set of spherical coordinates. The relative orientation of the ligand is determined by the three Euler angles, Θ, Φ and Ψ. Position and orientation of the substrate rely on the definition of *nonambiguous* groups of atoms, referred here to as $\{ \mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3 \}$ and $\{ \mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3 \}$ for the protein and for the ligand, respectively. As a rule of thumb, should the substrate be separated from the protein, for instance, in the z direction of Cartesian space, it would be a good idea to align P_1 and L_1 along that direction.

notably angles θ and Θ — i.e., in other words, avoiding aligned groups of atoms is crucial to guarantee geometrical invariance of K_{eq}.

In the particular example of p41 binding to the SH3 domain of tyrosine kinase Abl, a suitable choice for $\{ \mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3 \}$ could be the backbone atoms of residue Ser5, of residue Ser3 and of residue Ala1, respectively. For $\{ \mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3 \}$, the backbone atoms of residue Leu25, of residue Glu38 and of residue Gly46 represent a reasonable choice, albeit, once again, not a unique one. The reader of this tutorial is strongly encouraged to investigate other possible selections of atoms and show that K_{eq} is independent of these selections.

2.3. The geometrical-route simulations

Once the position and the orientation of the ligand with respect to the protein are fully defined, the geometrical restraints outlined in section **1.2.** can be safely introduced, one after the other, just like matryoshka, or Russian nested dolls [1,2]. One-dimensional potentials of mean force are then determined to assess the contribution to the binding free energy of these geometrical restraints. To attain this objective, the adaptive biasing force, or ABF algorithm will be utilized. ABF is not the only option available here — the reader of this tutorial is suggested to try alternate approaches like umbrella-sampling-like stratification strategies [18,23], wherein the reaction pathway is broken down into multiple, overlapping, narrow windows. In the latter numerical scheme, harmonic potentials confine sampling to the region of interest of the collective variable. The unnormalized probability distribution and, hence, the free-energy landscape are recovered self-consistently by means of such algorithms as the weighting histogram analysis method, or WHAM [24] — or, alternatively, by piecewise matching of the individual potential-of-mean-force segments. Umbrella sampling and its replica-exchange variant [25, 26] are available in NAMD. In the latter instance, given the size of the molecular assembly, a significant number of computer cores

ought to be reserved to see the real benefit of a replica-based approach.

Should the reader decide to resort solely to the adaptive biasing force algorithm, a number of caveats, possibly shortcomings of the methodology, ought to be considered, namely,

Historically, ABF calculations involving geometric restraints were burdened by a stringent limitation in the `colvars` implementation of the algorithm to discriminate between thermodynamic forces arising from the potential energy function and forces arising from external harmonic potential. In recent versions of NAMD, thermodynamic forces are measured by including contributions of both the potential energy function and the enforced geometrical restraints.

The aforementioned limitation of the `colvars` implementation of the ABF algorithm in earlier versions of NAMD can be circumvented by turning to an extended, generalized coordinate, wherein the collective variable is coupled by means of a stiff spring to a fictitious particle upon which the biasing force acts. This approach is referred to as extended adaptive biasing force, or eABF method [16], and will be utilized here as an alternative to plain ABF. It is invoked by issuing the following command in the definition of the collective variable:

```
extendedLagrangian    on
```

The *true* free-energy change and its gradient are recovered through a deconvolution of the harmonic-spring contribution from that of the force field, performed on the fly within `colvars` [27,28].

Furthermore, particular care ought to be taken when defining the collective variables, specifically,

The collective variable should be fully decoupled from degrees of freedom handled by the rattle algorithm [13,14]. The latter engender constraint forces, which contaminate the measure of the instantaneous thermodynamic force acting along the model reaction coordinate.

Since turning off rattle and, hence, decreasing the integration time step, does not constitute a reasonable option, collective variables featuring heavy atoms pertaining to constrained chemical bonds should also include the hydrogen atoms thereof, e.g., atoms `CA` and `HA`.

The reader will find in the present distribution a generic `colvars` file, in which all relevant collective variables are defined [15]. The adaptive biasing force (`abf`) method is applied to the Euclidian distance, r, separating the substrate from the protein, i.e.,

```
colvar {
   name R

   width 1.0

   lowerboundary  0.0
   upperboundary 40.0

   lowerwallconstant 100.0
   upperwallconstant 100.0

   distance {
      forceNoPBC     yes
      group1 {
         atomnumbers { 367 368 369 370 384 385 }
      }
      group2 {
         atomnumbers { 949 950 951 952 958 959 }
      }
   }
}
```

whilst the other angular variables, Θ, Φ, Ψ, θ and ϕ, are set to their equilibrium value (`harmonic`), e.g.,

```
colvar {
   name Psi

   width 1.0

   lowerboundary -180.0
   upperboundary  180.0

   lowerwallconstant  0.5
   upperwallconstant  0.5

   dihedral {
      group1 {
         atomnumbers { 575 576 577 578 588 589 }
      }
      group2  {
         atomnumbers { 367 368 369 370 384 385 }
      }
      group3  {
         atomnumbers { 949 950 951 952 958 959 }
      }
      group4  {
         atomnumbers { 917 918 919 920 926 927 }
      }
   }
}
```

```
harmonic {
   colvars        Psi
   forceConstant  0.1
   centers        23.0
}
```

The conformational variable (`rmsd`) is set to zero, which corresponds to restraining the ligand in its native form, when it is bound to the protein.

2.4. The alchemical-route simulations

As has been discussed in **1.3.**, the alchemical route relies upon the introduction of geometrical restraints, the substrate being decoupled reversibly from the protein in its native conformation, position and orientation, i.e., ligand* in the thermodynamic cycle of Figure 1. For consistency reasons, use will be made here of the same groups of atoms, $\{ P_1, P_2, P_3 \}$ and $\{ L_1, L_2, L_3 \}$, introduced previously to define the position and the orientation of the ligand with respect to the protein. Just like in the geometrical route, the relevant angular variables, Θ, Φ, Ψ, θ and ϕ, are set to their equilibrium value, employing harmonic potentials (`harmonic`). However, in sharp contrast with the geometrical route, wherein the ligand is separated reversibly from the protein by means of a Euclidian distance, r, the latter will be also set here to its equilibrium value, inferred, like the other collective variables, from an adequately long thermalization simulation of the protein:ligand complex. The substrate will be restrained to its native conformation by imposing that the root mean-square deviation (`rmsd`) with respect to the structure representative of the bound state be equal to zero.

Though the final result is independent from the force constants in the conformational, positional and orientational restraints [5], their choice necessarily impacts the alchemical free-energy change. Looser restraints, allowing the ligand to rearrange spatially with respect to the protein, affect thermodynamic micro-reversibility. Conversely, setting exaggeratedly large force constants results in possible instabilities in the trajectory.

Setting force constants of 10 kcal/mol·Å^2 for the `distance` and `rmsd` collective variables, and of 0.1 kcal/mol·degrees^2 for the `angle` and `dihedral` collective variables constitutes a reasonable choice [2]. The reader is reminded that force constants in NAMD are expressed in units of the bin widths. Hint: to confirm that the force constant used in the simulation is the intended one, search the beginning of the simulation log file for a line similar to the following:

```
colvars:  The force constant for colvar "R" will be rescaled to 250
according to the specified width..
```

As a good practice, the reader is advised to perform a bidirectional transformation, wherein the substrate is decoupled from the protein, i.e., forward transformation, prior to being coupled again to the binding site, i.e., backward transformation. Convergence of alchemical free-energy calculations, in general, and free-energy perturbation, or FEP calculations, in particular, can be markedly slow [18, 29], in particular when the vanishing, or appearing compound is sizable, thus, corresponding to a significant perturbation

of the molecular assembly. In the case of p41, a fine stratification strategy is recommended, consisting of a minimum of 50 windows, in each of which at least 200,000 molecular-dynamics steps, e.g., 50,000 thermalization steps (`alchEquilSteps`) followed by 150,000 data-collection steps, are generated,

```
source               fep.tcl

alch                 on
alchType             FEP
alchFile             bound.fep
alchCol              B

alchOutFile          forward.fepout
alchOutFreq          10

alchVdwLambdaEnd     1.0
alchElecLambdaStart  0.5

alchEquilSteps       50000
set numSteps         200000

runFEP  0.0  1.0  0.02  $numSteps
```

In the above free-energy perturbation section of the NAMD configuration file, use is made of `fep.tcl`, a Tcl script for setting the range of the general-extent parameter λ over which free-energy differences are computed.

The reader is suggested to use `parseFEP` to monitor the convergence of the alchemical transformations and ultimately combine the forward and backward simulations to obtain the maximum-likelihood Bennett-acceptance-ratio, or BAR estimator of the free energy [30, 31].

Evaluating the contribution to the standard binding free energy of the geometrical restraints is appreciably less demanding from a computational perspective. The contribution of each restraint acting on a collective variable is determined alchemically, by decreasing in a stepwise fashion the force constant from its nominal value to zero — and proceeding similarly in the opposite direction. Yet, in contrast with the free-energy calculations wherein the substrate is decoupled reversibly from the protein, the present simulations will be carried out in the framework of thermodynamic integration, within the `colvars` module [15],

```
harmonic {
   colvars               R
   centers             21.3
   targetNumSteps        100000
   targetEquilSteps       25000
   lambdaSchedule         1 0.9999 0.999 0.99 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
   forceConstant        0.0
   targetForceConstant 10.0
   targetForceExponent    4
```

In this example, a harmonic potential is imposed on the Euclidian distance, r, separating the ligand from the protein, defined previously. The nominal value of the force constant, k, utilized in the course of the alchemical free-energy perturbation calculation is 10 kcal/mol·Å^2. Here, this represents the value of the force constant at λ = 1, i.e., k_1, (`targetForceConstant`). At λ = 0, the force constant, k_0, is equal to 0 kcal/mol·Å^2 (`forceConstant`). The instantaneous value of the force constant at λ is given by,

$$k = k_0 + (k_1 - k_0)\left\{\lambda^n(1-\lambda) + \lambda\left[1-(1-\lambda)^n\right]\right\} \tag{9}$$

where n is referred to as `targetForceExponent`. For each value of `lambdaSchedule` and for each collective variable, $\mathrm{d}G/\mathrm{d}\lambda$ is computed. The free-energy contribution due to the restraint of interest is obtained by integrating the gradient profile.

3. Running and analyzing the simulations

Before embarking on the present tutorial, the reader ought to be reminded that the free-energy calculations proposed herein correspond to the state-of-the-art approach to address protein:ligand binding, and necessitate several hours of computer time on the parallel architecture of a commodity cluster.

Determination of K_{eq}, following either the geometrical route, or the alchemical route, is a computationally demanding endeavor and, hence, cannot be undertaken within the time usually allotted to an hands-on tutorial. The calculations described herein require a significant amount of computer time to attain fully converged free-energy estimates.

For the geometrical route, stratification of the reaction pathway is always recommended [13, 14, 23], with a number of strata as small as two for the angular free-energy profiles to as large as ten. Typical simulation times range from as low as 2 ns for the angular free-energy profiles to as much as 60 ns for the separation potential of mean force. For the alchemical route, stratification is also strongly advised, and typical simulation times can amount to as low as 6 ns to evaluate the free-energy contribution of the geometrical restraints in the bound state to as much as 40 ns for the BAR estimator. The reader is encouraged to examine alternate sampling strategies capable of offering the best precision:cost ratio.

3.1. The geometrical route

As has been discussed previously, completion of the geometrical route requires the determination of eight independent potentials of mean force, i.e., seven for the bound state and one for the unbound state. To explore the performance of an alternate algorithm to standard ABF, we have proposed that the extended adaptive biasing force, or eABF algorithm be employed. Under these premises, a post-treatment of the

data generated in the course of the simulation is necessary.

The `grad` and `pmf` files written by `colvars` in an eABF calculation reflect the average force acting on the extended, generalized coordinate — not on the actual collective variable, and should not be used as-is for analysis purposes.

Access to the *true* gradient imposes a deconvolution of the contribution due to the harmonic spring from that of the potential energy function. This deconvolution is performed on the fly in `colvars`, employing either the Zheng and Yang estimator [32], or the corrected z-averaged restraint estimator [28]. The first estimator is based on an umbrella integration (UI) [33],

$$G'(\xi') = \left(\frac{\mathrm{d}G}{\mathrm{d}\xi}\right)_{\xi'} = \frac{\sum_{\Xi'} N(\xi',\Xi') \left[\frac{(\xi' - \langle\xi_{\Xi'}\rangle)}{\beta\sigma^2_{\Xi'}} - K_\xi(\xi' - \Xi')\right]}{\sum_{\Xi'} N(\xi',\Xi')} \tag{10}$$

where $\xi(\mathbf{x})$ is the collective variable, function of the positions of the real particles of the molecular assembly and is restrained through potential $1/2\ K_\xi(\xi' - \xi_{\Xi'})^2$ to a one-dimensional fictitious particle, Ξ, moving dynamically. $N(\xi',\Xi')$ is the number of samples ξ' collected from the Ξ'-restrained ensemble, which is assumed to be Gaussian.

The second estimator is the corrected z–averaged restraint (CZAR) estimator [28],

$$G'(z') = \left(\frac{\mathrm{d}G}{\mathrm{d}z}\right)_{z'} = -\frac{1}{\beta}\frac{\mathrm{d}\ln\tilde{\rho}(z')}{\mathrm{d}z'} + k\left(\langle\lambda\rangle_{z'} - z'\right) \tag{11}$$

where $z = \xi(q)$ is the collective variable, λ is the extended variable harmonically coupled to z by means of a stiff spring of force constant k, and $\tilde{\rho}(z)$ is the observed distribution of collected variable z, upon which the time-dependent eABF bias is exerted.

The end-user may choose at the level of the `colvars` configuration file between the CZAR (`CZARestimator`) and the Zheng and Yang (`UIestimator`) estimator. The former is the default for eABF calculations. Depending upon the choice of the estimator, NAMD will generate the usual potential-of-mean-force (for one-dimensional free-energy profiles only), gradient and histogram files with a distinctive `czar` or `UI` suffix.

Once the different geometric free-energy calculations are completed, the different potentials of mean force can be post-processed numerically, using, for instance, XMGrace to evaluate the ratios of configurational integrals embodied in Equation 7, i.e., Data ⟶ Transformations ⟶ Evaluate expression/ Integration.

To illustrate how individual contributions arising from the geometrical restraints ought to be determined,

two collective variables will be chosen, namely Euler angle Ψ and Euclidian distance r. The free-energy change reflecting the reorientation about Ψ of the substrate in the binding site is depicted in Figure 3. Contribution of this Euler angle to the overall binding affinity amounts to,

$$\mathrm{e}^{+\beta\Delta G^{\mathrm{site}}_{\Psi}} = \frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi+u_\Psi)}} \tag{12}$$

In practical terms, the configurational and spatial integrals over the full potential energy function can be reduced to one over the potential of mean force for a given reaction coordinate. Thus,

$$\mathrm{e}^{+\beta\Delta G^{\mathrm{site}}_{\Psi}} = \frac{\int \mathrm{d}\Psi\, \mathrm{e}^{-\beta[w_{\mathrm{site}}(\Psi)]}}{\int \mathrm{d}\Psi\, \mathrm{e}^{-\beta[w_{\mathrm{site}}(\Psi)+u_\Psi]}} \tag{13}$$

where $w_{\mathrm{site}}(\Psi)$ is the PMF for Ψ in the bound state and u_Ψ is the harmonic restraint potential, i.e., $u_\Psi = \frac{1}{2}k_\Psi(\Psi-\Psi_0)^2$. Using a force constant of 0.1 kcal/mol·degrees^2 for the three Euler angles and an equilibrium value of Θ_0 = +102°, Φ_0 = −7° and Ψ_0 = +23°, it follows that $\Delta G^{\mathrm{site}}_{\Psi}$ = 0.4 kcal/mol.

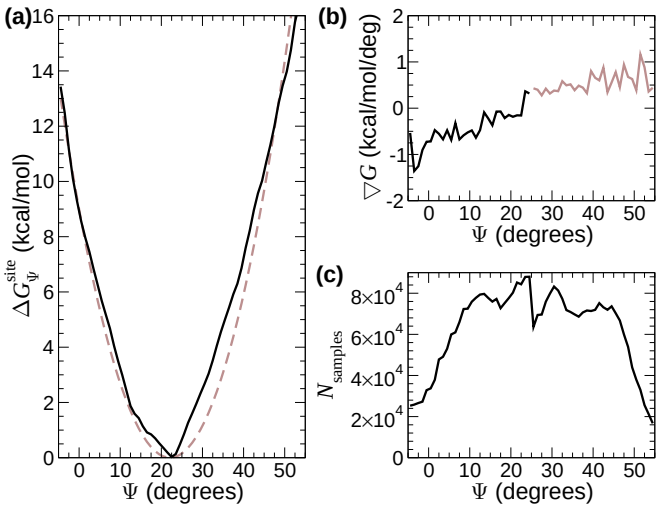


Figure 3: Free-energy change accounting for the reorientation of p41 with respect to the SH3 domain of Abl about Euler angle Ψ defined in Figure 2 (a). This potential of mean force was obtained, using two windows, approximately 40° wide. Its harmonic nature, highlighted in a dashed, light curve, is noteworthy. The reader is strongly advised to check the continuity of the gradient in adjacent windows (b). Here, two windows were utilized to cover the 80° range spanned by the collective variable. The reader is equally strongly advised to check sampling uniformity across the reaction pathway (c).

Compared with the evaluation of the free-energy contributions associated to the conformational and to the angular variables, determination of free-energy change arising from the separation of the ligand from the protein, shown in Figure 4, requires additional care. Returning to the original definition of $K^{\mathrm{geom}}_{\mathrm{eq}}$, this contribution corresponds to the fourth term of equation 7,

$$\frac{\displaystyle\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o+u_a)}}{\displaystyle\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\, \delta(\mathbf{x}_1-\mathbf{x}_1^*) \int \mathrm{d}\mathbf{x}\, \mathrm{e}^{-\beta(U+u_c+u_o)}} = S^* I^* \tag{14}$$

wherein

$$S^* = r^{*2} \int_0^{\pi} \mathrm{d}\theta \, \sin\theta \int_0^{2\pi} \mathrm{d}\phi \, \mathrm{e}^{-\beta u_a} \qquad \text{and} \qquad I^* = \int_{\text{site}} \mathrm{d}r \, \mathrm{e}^{-\beta[w(r)-w(r^*)]} \tag{15}$$

S^* is a surface term and represents the fraction of the sphere of radius r^*, centered about the binding site of the protein, accessible to the substrate. The choice of r^* should not impact the final value of K_{eq}, provided that it is sufficiently far from the binding site [2]. In the particular example of Figure 4 and for illustrative purposes, r^* will be chosen equal to 35 Å. It follows that $I^* = 3.16 \times 10^{12}$ Å and S^* = 17.15 Å^2, based on the equilibrium values of polar angles θ and ϕ appearing in the expression of u_a.

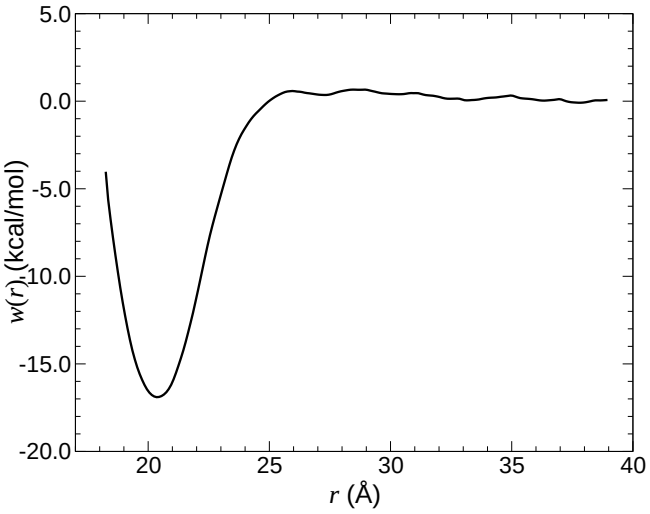


Figure 4: Free-energy change accounting for the separation of p41 from the SH3 domain of Abl, restraining conformational, orientational and positional degrees of freedom of the substrate to their equilibrium value in the bound state. It is the pièce de résistance of the geometrical route. The apparent decay of the free-energy profile at large separations reflects the probabilistic definition of the potential of mean force, wherein the geometric entropy term is not removed from the gradient. This definition of $w(r)$ is consistent with the basic expression 6 of the standard binding affinity, subsuming complete averaging of all other degrees of freedom.

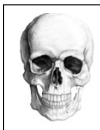
To complete the calculation of the standard binding free energy of p41 to the SH3 domain of tyrosine kinase Abl, the various contributions arising from the geometrical restraints, either in the bound state, or in the unbound state, ought to be determined in the spirit of the computation of $\mathrm{e}^{+\beta\Delta G_{\Psi}^{\text{site}}}$ detailed previously. The different pieces of the puzzle are then pasted together to recover the binding constant,

$$K_{\text{eq}} = S^* I^* \, \mathrm{e}^{-\beta(\Delta G_c^{\text{bulk}} + \Delta G_o^{\text{bulk}} - \Delta G_c^{\text{site}} - \Delta G_o^{\text{site}} - \Delta G_a^{\text{site}})} \tag{16}$$

where the second term of the exponent, ΔG_o^{bulk}, the free-energy contribution arising from the reorientation about the three Euler angles of a rigid body in a homogenous, bulk liquid, is calculated analytically,

$$\mathrm{e}^{-\beta\Delta G_o^{\text{bulk}}} = \frac{1}{8\pi^2} \int_0^{\pi} \mathrm{d}\Theta \, \sin\Theta \int_0^{2\pi} \mathrm{d}\Phi \int_0^{2\pi} \mathrm{d}\Psi \, \mathrm{e}^{-\beta u_o} \tag{17}$$

and u_o is the sum of u_Θ, u_Φ and u_Ψ.



When carrying out integrals over the potentials of mean force and restraints, pay careful attention to the units! The analytical integrals in Equation 17 are in radians but the restraints, including equilibrium values and force constants, are defined in degrees.

Following the geometrical route and assuming convergence of the different potentials of mean force of Equation 7, the final standard binding free energy, $\Delta G^0 = -1/\beta \ln K_{\rm eq}$, amounts to -7.7 kcal/mol [7, 8].

3.2. The alchemical route

The *pièce de résistance* of the alchemical route is without a doubt the alchemical transformation, wherein the restrained ligand is decoupled reversibly from its environment [21]. Even in the homogenous bulk aqueous medium, convergence of the free-energy perturbation calculation can be appreciably long, which can be understood in terms of rearrangement of the environment in response to the perturbation. The reader is strongly suggested to monitor convergence of the free energy in the different strata forming the reaction pathway that separate the coupled state from the decoupled one. Furthermore, since the transformation is run bidirectionally, the reader ought to ascertain that thermodynamic micro-reversibility is satisfied, verifying, using for instance `parseFEP` [34], that the underlying probability distributions for the various windows overlap appropriately [31].

As can be seen in Figure 5, which illustrates the reversible decoupling of p41 from the binding site of the SH3 domain of Abl, it is crucial that the bidirectional transformation reflects thermodynamic micro-reversibility, manifested here by a virtual absence of hysteresis. The shape of the two free-energy profiles ought to be commented. It mirrors the schedule chosen to extinguish, or ignite intermolecular interactions, i.e., `alchVdwLambdaEnd` set to 1.0 and `alchElecLambdaStart` set to 0.5. When the substrate is removed from the protein, electrostatic interactions are turned off faster than van der Waals interactions. For $\lambda > 0.5$, only the latter remain, which explains the point of inflection, midway on the reaction path.

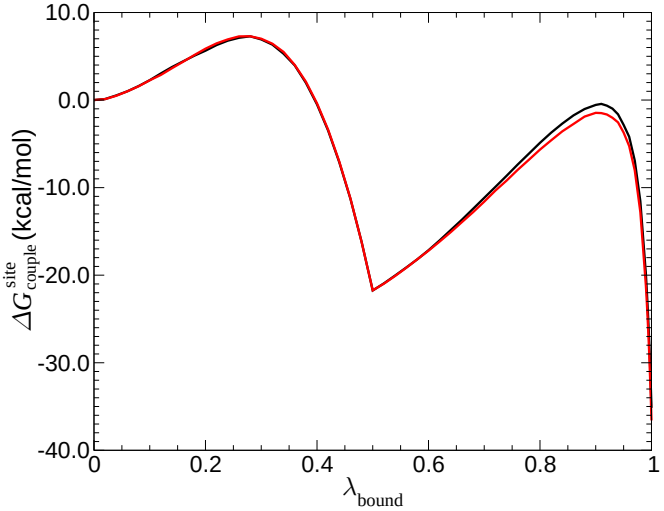


Figure 5: Free-energy change accounting for decoupling reversibly p41 from the SH3 domain of Abl, obtained using free-energy perturbation. The black profile corresponds to the forward transformation, wherein the restrained substrate is removed from the protein by turning off intermolecular interactions. A similar transformation is performed in the bulk aqueous environment, hence the popular designation of *double-annihilation* simulation [21]. The red profile characterizes the backward transformation, wherein the ligand is grown back in the binding site.

An example of the probability distributions obtained in the course of the bidirectional free-energy per-

turbation calculation is provided in Figure 6. Overlap of the histograms $P_0(\Delta U)$ and $P_1(\Delta U)$, for the forward and the backward transformations, respectively, is generally good. The reader is reminded that the relative inaccuracy of the forward simulation is related to the area under $P_1(\Delta U)$ where there is no overlap with $P_0(\Delta U)$ [18].

Visual examination of the underlying histograms of probability distributions $P_0(\Delta U)$ and $P_1(\Delta U)$, using `parseFEP` may not be sufficient to probe convergence. The reader is suggested to compare $P_0(\Delta U)$ with $P_1(\Delta U)\mathrm{e}^{+\beta\Delta U}$ and, symmetrically, $P_1(\Delta U)$ with $P_0(\Delta U)\mathrm{e}^{-\beta\Delta U}$, since $\mathrm{e}^{-\beta\Delta U}P_0(\Delta U) = \mathrm{e}^{-\beta\Delta G}P_1(\Delta U)$ [31].

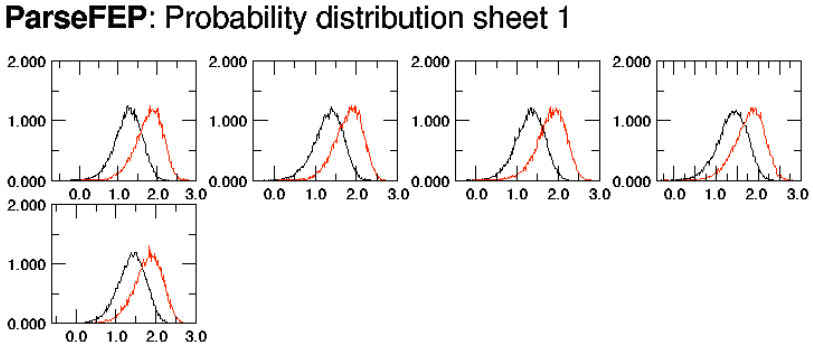


Figure 6: Histograms of the probability dstributions $P_0(\Delta U)$ and $P_1(\Delta U)$ for the forward, i.e., decoupling, and backward, i.e., coupling, transformations of p41 bound to the SH3 domain of tyrosine kinase Abl. The abscissa is ΔU, the perturbation term, expressed in kcal/mol.

In a sense, post-treatment of the free-energy calculations of the alchemical route is simpler than that of the geometrical route, requiring no tedious evaluation of ratios of configurational integrals. The contribution from the geometrical restraints is extracted from the NAMD output files,

```
> grep "dA/dLambda" rest-01.log
colvars:    Lambda= 1 dA/dLambda= 0.995394
colvars:    Lambda= 0.9999 dA/dLambda= 0.827822
colvars:    Lambda= 0.999 dA/dLambda= 0.760705
colvars:    Lambda= 0.99 dA/dLambda= 0.800472
colvars:    Lambda= 0.9 dA/dLambda= 0.783329
colvars:    Lambda= 0.8 dA/dLambda= 0.620776
colvars:    Lambda= 0.7 dA/dLambda= 1.12615
colvars:    Lambda= 0.6 dA/dLambda= 0.47743
colvars:    Lambda= 0.5 dA/dLambda= 0.305801
colvars:    Lambda= 0.4 dA/dLambda= 0.16728
colvars:    Lambda= 0.3 dA/dLambda= 0.0677748
colvars:    Lambda= 0.2 dA/dLambda= 0.0269255
colvars:    Lambda= 0.1 dA/dLambda= 0.00323781
colvars:    Lambda= 0 dA/dLambda= 0
```

It suffices then to integrate the gradients, using for instance XMGrace to recover the associated free-energy profile, from whence the different conformational, orientational and positional terms can be inferred, as illustrated in Figure 7.

To complete the thermodynamic cycle of Figure 1, the analytical terms appearing in Equation 8 ought to be added, namely, for the positional contribution,

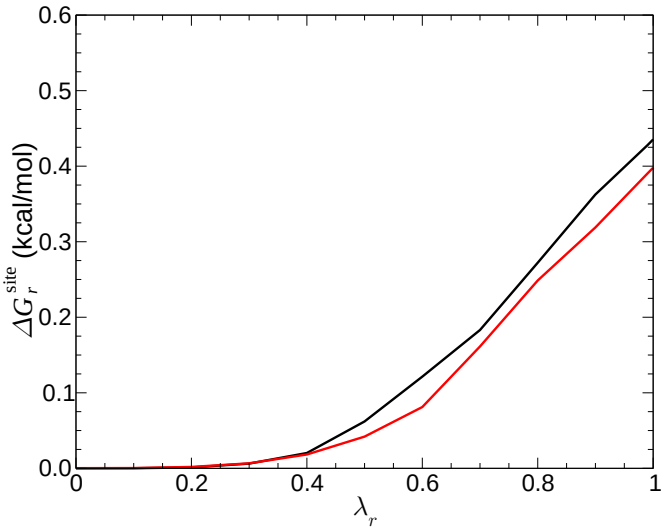


Figure 7: Free-energy contribution due to the restraining potential acting on the Euclidian distance between p41 from the SH3 domain of Abl, based on the equilibrium value in the bound state. The black and red curves delineate the forward and backward transformations, wherein, respectively, the force constant of the geometrical restraint is decreased from its nominal value to 0, and back. The hysteresis between the two free-energy profiles provides a rough estimate of the inaccuracy of the calculation.

$$F_{\text{t}}\,\mathrm{e}^{-\beta\Delta G_a^{\text{bulk}}} = \int_0^{\infty} \mathrm{d}r\; r^2\,\mathrm{e}^{-\beta u_r} \int_0^{\pi} \mathrm{d}\theta\;\sin\theta \int_0^{2\pi} \mathrm{d}\phi\,\mathrm{e}^{-\beta u_a} \tag{18}$$

where $u_a = u_\theta + u_\phi$ and F_{t} is a translational factor with units of volume [1]. For the orientational contribution,

$$\mathrm{e}^{-\beta\Delta G_o^{\text{bulk}}} = \frac{1}{8\pi^2} \int_0^{\pi} \mathrm{d}\Theta\;\sin\Theta \int_0^{2\pi} \mathrm{d}\Phi \int_0^{2\pi} \mathrm{d}\Psi\;\mathrm{e}^{-\beta u_o} \tag{19}$$

where $u_o = u_\Theta + u_\Phi + u_\Psi$. Pasting all the contributions together,

$$K_{\text{eq}} = F_{\text{t}}\,\mathrm{e}^{-\beta(\Delta G_c^{\text{bulk}} - \Delta G_{\text{couple}}^{\text{bulk}} + \Delta G_o^{\text{bulk}} + \Delta G_a^{\text{bulk}} - \Delta G_{\text{decouple}}^{\text{site}} - \Delta G_a^{\text{site}} - \Delta G_r^{\text{site}} - \Delta G_o^{\text{site}} - \Delta G_c^{\text{site}})}. \tag{20}$$

Following the alchemical route and assuming convergence of the individual free-energy calculations embodied in Equation 8, the final standard binding free energy, $\Delta G^0 = -1/\beta\;\ln K_{\text{eq}}$, amounts to -7.7 kcal/mol [7, 8].

4. Concluding remarks and extensions of the tutorial

It is not superfluous to reemphasize that this document contains advanced material, which, in glaring contrast with introductory tutorials, requires that the reader has grasped the theoretical underpinnings of the methodology to perform the proposed free-energy calculations with utmost efficiency. Such calculations are computationally demanding and still cannot be run in a routine fashion. Aside from the slowly-converging nature of the simulations [18, 29], the size of the molecular assembly imposes adapted computational resource, in the form of commodity clusters, chiefly making use of either central processing units, or graphical processing units. Contrary to introductory tutorials, the archive does not include all the files, input and output, but only the bare necessities to complete the study gracefully.

The methodology described herein raises, however, a number of questions and comments. Central to both the alchemical and the geometrical routes is the choice of atoms, or groups thereof, utilized to define the position and the orientation of the ligand with respect to the protein. It should be clearly understood that ΔG^0 is an invariant of this choice [1,2]. The reader is invited to explore alternate selections of atoms for the triplets $\{ \mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3 \}$ and $\{ \mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3 \}$, and show that the computed standard binding affinity is, indeed, independent from these selections.

Closely related to the choice of atoms forming $\{ \mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3 \}$ and $\{ \mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3 \}$ is the choice of the reference three-dimensional structure, central to the computation of the conformational terms, and from whence the equilibrium values of the collective variables are inferred. *What is a good reference structure?* A natural choice is the energy-minimized crystallographic structure of the protein:ligand complex solvated in the aqueous environment. Another relevant option consists in running an adequately long molecular-dynamics simulations past the thermalization stage and compute an average structure, which ought to be energy-minimized to remove spurious averaging artifacts, notably from freely rotating methyl groups.

A third comment that comes naturally to mind concerns the necessity to impose geometrical restraints on the protein. In the particular example of p41 binding to the SH3 domain of Abl, comparison of the protein conformation in the native complex and in the free state suggests that the spatial rearrangement that accompanies the association process is extremely limited and, therefore, obviates the need for additional harmonic potentials. If such were not the case, it would become necessary to restrain the conformation of the protein to that of the bound state, incorporating in Equation 7, and its alchemical counterpart, the relevant, supplementary contribution.

One of the virtues of this tutorial, and the study therein, is to reveal and illuminate the pros and cons of free-energy methods, and how well suited the latter are for a given problem. Whereas the pseudo-harmonic nature of the orientational contributions to $K_{\mathrm{eq}}^{\mathrm{geom}}$ suggests that the extended adaptive biasing force algorithm is well adapted for such collective variables, the same cannot be said for the separation potential of mean force, plagued by slow orthogonal degrees of freedom and hidden barriers. The beauty of free-energy methods is their ability to be combined with other algorithms targeted at improving ergodic sampling [25,26,35–38]. The reader is invited to consider alternate options to eABF, or standard stratified umbrella-sampling approaches, notably multiple-walker schemes, wherein replicas of the reaction coordinate models are spawned and handled by the different cores of a parallel computer architecture [37].

Another question eminently relevant to the problem at hand is the provision of a measure of the reliability of the free-energy estimate. As a matter of principle, just like publishing experimental binding free energies bereft of an error bar would be inconceivable, a theoretical estimate ought to be reported with

its reliability, i.e., the combination of the accuracy, or systematic error, and of the precision, or statistical error [39]. While the `parseFEP` plugin [34] of VMD supplies an estimate of both the precision and the accuracy of the Bennett acceptance ratio free-energy difference, the reader is invited to refer to the specialized literature [29, 40–42] to adopt the best strategy towards the computation of these quantities in the framework of the adaptive biasing force, or the stratified umbrella-sampling method.

The example detailed in this tutorial, p41, a proline-rich peptide binding to the SH3 domain of tyrosine kinase Abl, however intricate, still represents one standard binding free energy, that is one number. An interesting extension of the present work would consist in investigating other, structurally related substrates, thereby demonstrating that the proposed methodology is predictive. For instance, replacing the decapeptide APSYSPPPPP by APTYHPPLPP, i.e., three point mutations, results in an experimental absolute decrease of the standard binding free energy from -7.94 to -5.30 kcal/mol, whilst a single point mutation like APSYSPPPPP by APTYSPPPPP yields a more modest increase from -7.94 to -8.72 kcal/mol [43]. The reader is invited to explore these mutants, using the methodology described herein, and verify that the computed standard binding free energies are consistent with the results of relative free-energy calculations.

Appendix: Archive

This tutorial is provided with all the files necessary towards the calculation of the standard binding constants $K_{\mathrm{eq}}^{\mathrm{alch}}$ and $K_{\mathrm{eq}}^{\mathrm{geom}}$. The archive is organized in two directories, `AlchemicalRoute` and `GeometricalRoute`. `AlchemicalRoute` contains four main subdirectories, which correspond to the alchemical transformations in the bound and in the unbound states, as well as the evaluation of the free-energy contribution due to the geometrical restraints in these two states. `GeometricalRoute` contains eight main subdirectories, which correspond to the eight potential-of-mean-force calculations required to determine the standard binding free energy.

Caveat emptor. Though the tutorial includes all the relevant files needed to perform the different free-energy calculations described herein, the reader is strongly advised to not use these files blindly, without checking first their contents.

As stated previously, the transformations carried out in this tutorial, either of geometrical or alchemical nature, ought to be stratified for statistical efficiency. The various subdirectories of the archive only represent one stratum of the full reaction pathway, thus, requiring duplication of the NAMD and `colvars` files.

Caveat emptor. The reader may decide to choose triplets { P_1, P_2, P_3 } and { L_1, L_2, L_3} distinct from those proposed in the archive, in which case the `colvars` files should be adapted accordingly. Different choices of atoms can result in different sampling and convergence properties.

Amongst the many possible adaptations of the input files for the geometrical route, modification of `FullSamples`, the threshold beyond which the time-dependent bias is applied on the collective variable, may impact the overall efficiency of the adaptive biasing force simulations. In addition, the reader might find necessary to alter the stiffness of the spring connecting the collective variable to the fictitious particle, i.e., `extendedFluctuation`. The reader is also suggested to write information in the `colvars` trajectory files, or `traj` files, at an appropriate frequency for the post-treatment of the extended adaptive biasing force, or eABF simulations.

In the case of the alchemical route, similar adaptations ought to be considered. For the alchemical transformations, the stratification strategy has proven crucial and can be tailored to increase both efficiency and reliability. Stratification is also important when zeroing reversibly the force constant of the geometrical restraints utilized to preserve the conformation, the orientation and the position of the ligand representative of the bound state, i.e., `lambdaSchedule`. The number of steps, i.e., `targetNumSteps`, including the thermalization stage, i.e., `targetEquilSteps`, is of paramount importance to insure both convergence and thermodynamic micro-reversibility of the simulations.

Acknowledgments

Jérôme Hénin and Giacomo Fiorin are gratefully acknowledged for helpful discussions about the `colvars` module.

References

[1] Woo, H. J.; Roux, B., Calculation of absolute protein–ligand binding free energy from computer simulations, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6825–6830.

[2] Gumbart, J. C.; Roux, B.; Chipot, C., Standard binding free energies from computer simulations: What is the best strategy?, *J. Chem. Theor. Comput.* **2013**, *9*, 794–802.

[3] 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. Am. Chem. Soc.* **1997**, *119*, 2707–2714.

[4] Dixit, S. B.; Chipot, C., Can absolute free energies of association be estimated from molecular mechanical simulations? The biotin–streptavidin system revisited, *J. Phys. Chem. A* **2001**, *105*, 9795–9799.

[5] Deng, Y.; Roux, B., Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant, *J. Chem. Theor. Comp.* **2006**, *2*, 1255–1273.

[6] Wang, J.; Deng, Y.; Roux, B., Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials, *Biophys. J.* **2006**, *91*, 2798–2814.

[7] Pisabarro, M. T.; Serrano, L., Rational design of specific high-affinity peptide ligands for Abl-SH3 domain, *Biochemistry* **1996**, *35*, 10634–10640.

[8] Pisabarro, M. T.; Serrano, L.; Wilmanns, M., Crystal structure of the abl-SH3 domain complexed with a designed high-affinity peptide ligand: implications for SH3-ligand interactions., *J. Mol. Biol.* **1998**, *281*, 513–521.

[9] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, L.; Schulten, K., Scalable molecular dynamics with NAMD, *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[10] Landau, L. D., *Statistical physics*, The Clarendon Press: Oxford, 1938.

[11] Zwanzig, R. W., High–temperature equation of state by a perturbation method. I. Nonpolar gases, *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[12] Darve, E.; Pohorille, A., Calculating free energies using average force, *J. Chem. Phys.* **2001**, *115*, 9169–9183.

[13] Hénin, J.; Chipot, C., Overcoming free energy barriers using unconstrained molecular dynamics simulations, *J. Chem. Phys.* **2004**, *121*, 2904–2914.

[14] Hénin, J.; Fiorin, G.; Chipot, C.; Klein, M. L., Exploring multidimensional free energy landscapes using time-dependent biases on collective variables, *J. Chem. Theor. Comput.* **2010**, *6*, 35–47.

[15] Fiorin, G.; Klein, M. L.; Hénin, J., Using collective variables to drive molecular dynamics simulations, *Mol. Phys.* **2013**.

[16] Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C., The adaptive biasing force method: Everything you always wanted to know, but were afraid to ask, *J. Phys. Chem. B* **2015**, *119*, 1129–1151.

[17] Shoup, D.; Szabo, A., Role of diffusion in ligand binding to macromolecules and cell–bound receptors, *Biophys. J.* **1982**, *40*, 33–39.

[18] Chipot, C.; Pohorille, A., Eds., *Free energy calculations. Theory and applications in chemistry and biology*, Springer Verlag, 2007.

[19] Hermans, J.; Shankar, S., The free energy of xenon binding to myoglobin from molecular–dynamics simulation, *Isr. J. Chem.* **1986**, *27*, 225–227.

[20] Roux, B.; Nina, M.; Pomès, R.; Smith, J. C., Thermodynamic stability of water molecules in the Bacteriorhodopsin proton channel: A molecular dynamics and free energy perturbation study, *Biophys. J.* **1996**, *71*, 670–681.

[21] Gilson, M. K.; Given, J. A..; Bush, B. L.; McCammon, J. A., The statistical–thermodynamic basis for computation of binding affinities: A critical review, *Biophys. J.* **1997**, *72*, 1047–1069.

[22] Humphrey, W.; Dalke, A.; Schulten, K., VMD — Visual molecular dynamics, *J. Molec. Graphics* **1996**, *14*, 33–38.

[23] Roux, B., The calculation of the potential of mean force using computer simulations, *Comput. Phys. Comm.* **1995**, *91*, 275–282.

[24] Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M., The weighted histogram analysis method for free energy calculations on biomolecules. I. The method, *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[25] Sugita, Y.; Okamoto, Y., Replica–exchange molecular dynamics method for protein folding, *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[26] Sugita, Y.; Kitao, A.; Okamoto, Y., Multidimensional replica–exchange method for free–energy calculations, *J. Chem. Phys.* **2000**, *113*, 6042–6051.

[27] Fu, H.; Shao, X.; Chipot, C.; Cai, W., Extended adaptive biasing force algorithm. An on–the–fly implementation for accurate free–energy calculations, *J. Chem. Theory Comput.* **2016**, *12*, 3506–3513.

[28] Lesage, A.; Lelièvre, T.; Stoltz, G.; Hénin, J., Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method., *J. Phys. Chem. B* **2017**, *121*, 3676–3685.

[29] Lelièvre, T.; Stoltz, G.; Rousset, M., *Free energy computations: A mathematical perspective*, Imperial College Press, 2010.

[30] Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data, *J. Comp. Phys.* **1976**, *22*, 245–268.

[31] Pohorille, A.; Jarzynski, C.; Chipot, C., Good practices in free-energy calculations, *J. Phys. Chem. B* **2010**, *114*, 10235–10253.

[32] Zheng, L.; Yang, W., Practically efficient and robust free energy calculations: Double-integration orthogonal space tempering, *J. Chem. Theor. Compt.* **2012**, *8*, 810–823.

[33] Kästner, J.; Thiel, W., Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration", *J. Chem. Phys.* **2005**, *123*, 144104.

[34] Liu, P.; Dehez, F.; Cai, W.; Chipot, C., A toolkit for the analysis of free-energy perturbation calculations, *J. Chem. Theor. Comput.* **2012**, *8*, 2606–2616.

[35] Woods, C. J.; Essex, J. W.; King, M. A., The development of replica-exchange-based free-energy methods, *J. Phys. Chem. B* **2003**, *107*, 13703–13710.

[36] Jiang, W.; Hodoscek, M.; Roux, B., Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics , *J. Chem. Theory Comput.* **2009**, *5*, 2583–2588.

[37] Minoukadeh, K.; Chipot, C.; Lelièvre, T., Potential of mean force calculations: A multiple-walker adaptive biasing force approach, *J. Chem. Theor. Comput.* **2010**, *6*, 1008–1017.

[38] Jiang, W.; Roux, B., Free energy perturbation Hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations, *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.

[39] Kofke, D.A.; Cummings, P.T., Precision and accuracy of staged free-energy perturbation methods for computing the chemical potential by molecular simulation, *Fluid Phase Equil.* **1998**, *150*, 41–49.

[40] Straatsma, T. P.; Berendsen, H. J. C.; Stam, A. J., Estimation of statistical errors in molecular simulation calculations, *Mol. Phys.* **1986**, *57*, 89–95.

[41] Flyvbjerg, H.; Petersen, H. G., Error estimates on averages of correlated data, *J. Chem. Phys.* **1989**, *91*, 461–466.

[42] Rodriguez-Gomez, D.; Darve, E.; Pohorille, A., Assessing the efficiency of free energy calculation methods, *J. Chem. Phys.* **2004**, *120*, 3563–3578.

[43] Hou, T.; Chen, K.; McLaughlin, W. A.; Lu, B.; Wang, W., Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain., *PLoS Comput. Biol.* **2006**, *2*, e1.