World Research Journal of Peptide and Protein
ISSN: 2278-4586 & E-ISSN: 2278-4608, Volume 1, Issue 1, 2012, pp.-21-32.
Available online at http://www.bioinfo.in/contents.php?id=131

DETERMINING FREE ENERGIES OF PROTEIN-LIGAND BINDING AND ASSOCIATION/ DISSOCIATION PROCESSES USING COMPUTER SIMULATIONS

GAUTO D.F.[2,3], CARLOS MODENUTTI[2,3], DUMAS V.G.[1,3], LUCIA ALVAREZ[2,3], BUSTAMANTE J.P.[2,3], TURJANSKI A.G.[1,2,3] AND MARTI M.A.[1,2,3,*]

[1]Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina.
[2]Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina.
[3]Instituto de Química de los Materiales, Medio Ambiente y Energía (INQUIMAE), CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina.
*Corresponding Author: Email- marcelo@qi.fcen.uba.ar

Received: November 11, 2011; Accepted: February 14, 2012

Abstract- The development of new drugs is one of the most important research areas in the bio-sciences, and where structural bioinformatics plays a central role. Over the last thirty years rational drug design has contributed to the introduction of many new drugs in the market and computational (or in-silico based) methods are an essential part of these programs having the great advantage of the potential delivery of new drug candidates faster and at lower cost when compared to high throughput experimental methods. At the heart of these methods, lies the determination of a given drug (or ligand) affinity for a given protein receptor, which includes determination or knowledge of the protein-ligand complex. Thus, the theoretical prediction of ligand binding free energies (ΔGB), is one of the most important and yet challenging problems in computational biochemistry, and therefore the subject of the current review. The review starts describing the so called End point methods for computing ligand binding free energies which rely on performing MD simulations of the complexes with post-processing analysis, and shows recent advances and improvements on ΔGB prediction using Quantum Mechanics and explicit solvation analysis techniques. Secondly we present free energy based methods that rely on the description of the binding process itself, reviewing first the use of biased non equilibrium based methods for small ligand binding to metalo proteins and second the recent advances to the study and free energy determination of big drug like ligand binding process with biased and free diffusion methods. Finally, we perform an overall comparison of the reviewed methods, and suggest which method (or methods) should be used in different ideal cases described as examples.
Keywords- Free Energy, Protein-ligand interaction, MM-GBSA, MSMD, ligand association, ligand dissociation.

Citation: Gauto D.F. et al (2012) Determining Free Energies of Protein-Ligand Binding and Association/Dissociation Processes Using Computer Simulations. World Research Journal of Peptide and Protein, ISSN: 2278-4586 & E-ISSN: 2278-4608, Volume 1, Issue 1, pp.-21-32.

Introduction

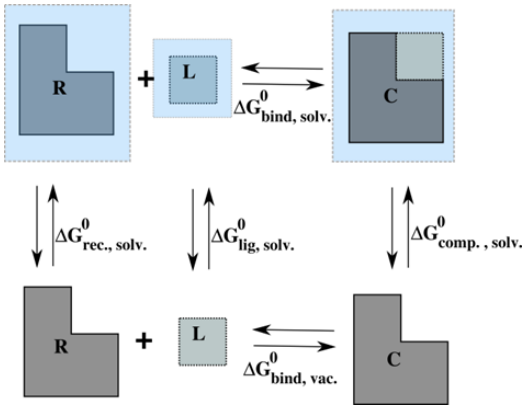
The development of new drugs is one of the most important research areas in the bio-sciences, and where structural bioinformatics plays a central role. The first pharmaceutical compound which was rationally designed starting from known receptor structure was captopril in the 80´s, the first Angiotens in Converting Enzyme selective inhibitor.[1] Over the last thirty years rational drug design has contributed to the introduction of many new drugs in the market and computational (or in-silico based) methods are an essential part of these programs.[1,2,3,4,5,6,7]The great advantage of in-silico methods is the potential delivery of new drug candidates faster and at lower cost when compared to high throughput experimental methods. At the heart of the computational methods used in drug discovery-design programs, lies the determination of a given drug (or ligand) affinity for a given protein receptor, determined by the ligand binding standard free energy (ΔG_B^0) and which includes determination or knowledge of the protein-ligand complex. On the molecular level, biological activity in many cases corresponds to the binding of a low- molecular weight (i.e drug-like) compound to a macromolecular receptor,

usually a protein. Accordingly, biological activity is intimately related with, or even can be expresses as, the affinity of both partners for each other. Under the usual equilibrium conditions the affinity is described as the corresponding thermodynamic equilibrium quantity, ΔG_B^0.[8]Thus, the theoretical prediction of ΔG_B^0, is one of the most important and yet challenging problems in computational biochemistry, and therefore the subject of the current review.[9] Precise determination of ΔG_B^0 is very important since its value is usually used to decide which of the screened compounds should be used for further experimental testing or which of a set of possible protein-inhibitor complexes actually binds with the expected affinity[1,2,6]

TheΔG^0_B can be further separated in energetic (or enthalpic) and entropic contributions. One of the key problems in estimating the $\Delta G^0_B$is the intrinsic complexity of the binding process since in order to bind; the ligand and the protein surface at the ligand binding site must be desolvated. Using a very simple thermodynamic cycle for the corresponding ligand binding and solvation processes, the ΔG^0_B in solution can first be decomposed as the combination of the gas phase association, which can be decomposed into enthalpic and enetropic terms, and the (de) solvation contribution of transferring the interacting partners from solution to the gas phase as shown in scheme 1, and leading to equation 1

$$\Delta G_B{}^0 = \Delta E_{vac} - T\,\Delta S_{vac} + \Delta G_{sv} \quad (1)$$

Where ΔE_{vac} represents the change in the system energy due to the formation of the protein-ligand complex, which is usually negative as the major contributions are the specific protein–ligand interactions; ΔS_{vac} is the change in entropy upon complex formation and involves loss of ligand rotational and translational entropy, which usually represent a ca. 15-25 kcal/mol penalty and the change in protein and ligand conformational entropy in the complex with respect to the free state in solution.[10] Finally, ΔG_{sv} represents the change in solvent free energy upon complex formation.



Scheme 1- Thermodynamic Cycle of the ligand (L) binding process to a given protein receptor (R) to yield the corresponding complex (C). The cycle allows separating the binding event, and its associated free energy ($\Delta G_{bind,\ vac.}$), in vacuum (bottom), from the solvation free energy of each sub system R, L and C ($\Delta G_{R/L/C.\ solv}$).

The contribution of ΔG_{sv} to the overall ligand affinity is very important not only for the important entropic change in the solvent involved in the ligand salvation (associated to the hydrophobic effect), but also from the displacement of tightly bound waters in the protein surface by the ligand.[11,12,13] The relevance of these surface bound waters is underscored by several studies which showed higher affinity for ligands that were designed to specifically displace them.[14]Recent developments by our group[15,16,17] and others[11,12,13,18,19] have allowed a deeper insight into this phenomenon.

Different computational methods have been developed to compute ΔG_B^0, which can be divided in those that separate the calculation of the different contributions to the free energy, and whose main representatives are the so called end-point-methods, or the methods that directly estimate the ΔG_B using a thermodynamic integration or free energy perturbation scheme. On the other hand, from a kinetically, or process dependent viewpoint the ΔG_B^0 which is directly related to the ligand dissociation equilibrium constant, depends both on the association and dissociation rates according to equation 2.

$$\Delta G_B{}^0 = -k_B T\, ln\left(K_{eq}\right) = -k_B T\, ln\,(k_{off}/k_{on}) \quad (2)$$

Where K_{eq} correspond to the ligand dissociation equilibrium constant, k_{off} is the ligand dissociation rate, k_{on} the ligand association rate and K_B the Boltzmann constant. Therefore, another way of computing the ΔG_B^0 is to determine the ligand association and dissociation rates. The association/dissociation processes are related to the corresponding free energy barriers and therefore these quantities have to be determined. To estimate the barriers the free energy profile along the association/dissociation process must be computed. Computational methods usually rely on the use of a reaction coordinate to determine the corresponding profiles; the coordinate describes the change in the system along the desired process. To compute the potential of mean force along the profiles free energy biased methods can be used, such as umbrella sampling,[10]metadynamics or the steered molecular dynamics non equilibrium based scheme, which we will described with further detail below.[20,21,22,23]The most straight forward way to determine the profiles is to start with the protein-ligand complex and gently push the ligand out from the binding site until it is completely solvated. The great advantage of the free energy profile biased methods over the end-point partition methods is that they also provide an insight on the ligand association and dissociation processes, and not only on the complex structure. Finally, as recently shown by Buch et. al.[24] the free energy profile of the ligand binding process can be determined by very extensive non biased sampling of the protein ligand system, provided enough binding and release events are observed.

End point methods for computing ligand binding free energies

End-point free energy methods enclose a diverse set of strategies that rely on the simulation of usually only the protein-ligand (i.e. the complex) state, and sometimes the protein free or unbound state, to determine later the ΔG_B between them. Most common computational methods are usually intended to estimate the relative binding affinity between a small set of structurally related compounds rather than the absolute free energy of binding. End-point free energy methods generally involve an all atom molecular dynamics (MD) simulation of the complex and/or unbound ligand and protein and are mainly based on post-processing of the MD trajec-

tories of the complex. Therefore, they are much faster, and less accurate, than those methods based on alchemical or structural transformations, such as free energy perturbation (FEP) and/or thermodynamic integration methods (TI).[10,25] They have also the advantage of allowing the calculation of the ΔG_B values for a diverse set of ligands, while FEP and TI only allow the comparison of fairly similar compounds. As we stated above these methods do not allow the study of the ligand association or dissociation process, and require previous knowledge of the complex structure. [10]

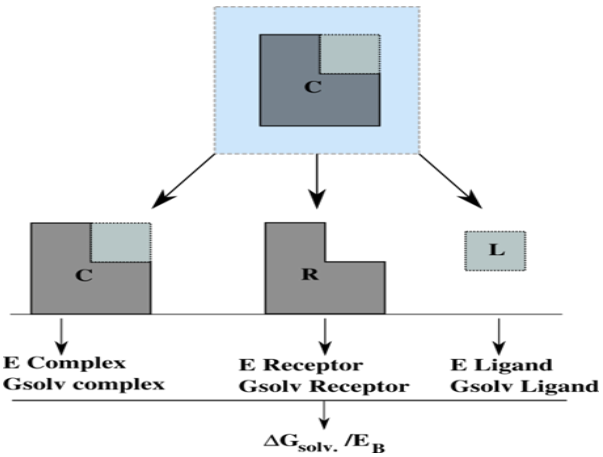
End-point methods typically rely on the partitioning of the free energy into a sum of different energetic and entropic contributions. [25,26,27]In this sense several frameworks exist that use implicit solvent approximations to reduce computational demands even further.[28] A general partition scheme is described by equation 3 shown below:[2,10]

$$\Delta G_B = \Delta E_{int} - T\Delta S_{Lig} - T\Delta S_{prot} + \Delta G_{sv} \quad (3)$$

Where, ΔE_{int} corresponds to the direct protein-ligand interaction energy in the complex, ΔS_{lig} corresponds to the ligand entropy loss due to complex formation, ΔS_{prot} corresponds to the change in the protein conformational entropy upon binding, ΔG_{sv} represents the solvent associated free energy change along the binding process and T is the system temperature, usually 300K. From these contributions, usually the most relevant are ΔE_{int} and ΔG_{sv}, and therefore there are several strategies to compute them accurately. It is important to note that an inherent approximation of all end point methods is the assumption that the conformational ensemble of the protein-ligand complex, from which each term is computed, is efficiently sampled along the MD trajectory.[9]

Classical methods with implicit solvent, MM-PB (SA) and MM-GB (SA)

Possibly the most well-known and widely used end point methods for computing ligand binding free energies are the MM-PB(SA) (Molecular Mechanics-Poisson Boltzmann Surface Area) and MM-GB(SA) (Molecular Mechanics-Generalized Born Surface Area) strategies, that combine a molecular mechanics force field with a continuum solvation model to determine the ΔE_{int} and ΔG_{SV} contributions respectively. There are several recent reviews on these methods and therefore we will describe them only briefly here.[29], [30] To determine the binding free energy with this type of methods the procedure is quite simple (as schematically shown in scheme 2).



Scheme 2- Brief scheme of the single MD approach to perform ΔGB calculations using end-point methods. C represents the complex, R the receptor and L the ligand. Light blue surrounding represents explicit solvation. E C/R/L represents the corresponding system energy. Gsolv C/R/L represents the corresponding system solvation free energy.

First, roughly 50ns long MD simulation of each protein-ligand complex is performed. Secondly, a set of evenly or randomly selected snapshots (usually 1 to 5 thousand) are selected, the explicit waters are removed, and for each, three new systems trajectories are built corresponding to the protein-ligand complex, the protein alone and the ligand alone. Once these new snapshots ensembles are built, the ΔEint contribution to equation 3 can be computed as the average potential energy difference derived from the evaluation of each system structure using the corresponding force field energy, as shown by equation 4, and averaged over all snapshots.

$$\Delta E_{int} = E_{complex} - E_{receptor} - E_{ligand} \quad (4)$$

In classical force fields the energy of each systems is usually written as a combination of bonded and non-bonded terms, for example in the classical Amber Force Field the energy (EAmber) is written as defined by equation 5

$$E_{Amber} = E_{bond} + E_{angle} + E_{torsion} + E_{vdW} + E_{elec} \quad (5)$$

Where Ebond, Eangle, Etorsion are the bonded energies associated to bond, angle and torsions respectively. EvdW and Eelec correspond to the van der Waals and electrostatic terms of the potential energy[31] Given that a single trajectory approach is used, for each snapshot, the bonded terms in the complex and isolated protein and ligand systems are exactly the same and cancel out (as well as all protein-protein vdW and electrostatic interactions) and therefore the only contributions to the ΔEint stems from the non bonded vdW and electrostatic protein ligand cross interactions. The single trajectory approach has also shown to converge faster and to avoid noise in the binding energy that could be generated due to non relevant small conformational changes during the protein dynamic evolution, that could be sampled differentially if two simulations, one for the complex and one for the free protein were used.[9]

The single trajectory derived snapshots are also used to compute the ΔGsv contribution to the binding free energy. In the PB and GB implicit solvation methods the term is further decomposed as the sum of two contributions: First, the electrostatic solvation energy which is calculated using the Generalized Born or Poisson Boltzmann model respectively,.[10]and second, a cavitation term that takes into account both the hydrophobic effect, i.e. the burying of non polar surface area upon complex formation (which is simply computed as the product of the solvent surface tension times the change in the solvent-accessible surface area (SASA) upon ligand binding); and the cost of creating the cavity which is proportional to the volume of the molecule.

Entropy Contributions

The entropy terms, cannot be computed directly from the obtained trajectory and need a separate treatment. Despite the entropic contribution to the desolvation of the drug both the protein and the

ligand modify their conformational space upon binding and these changes need to be estimated. Usually ΔSlig is computed using rigid body approximation to estimate the loss of roto-translational entropy of the ligand upon binding. Other approximations rely on the calculation of the difference in ligand vibrational entropy using harmonic approximations.[10] For ΔSprot two approaches are commonly used. In line with the single trajectory approach, one strategy relies in the calculation of the average entropy difference arising from normal model calculation of the protein in the complex and the protein alone.[31] This method is however quite time consuming since proper geometry optimization for each snapshot needs to be performed. Also important, Kongsted et. al.[28,32] showed that one of the disadvantages of the entropy calculation using normal mode analysis of harmonic frequencies is the large variation along various MD snapshots. To tackle this weak point, a practical solution was suggested, which relies on minimizing only a truncated system plus a buffer region containing all atoms (including solvent) between 8 to 12 Å from the ligand in the complex. Finally, during the minimization process, only the residues of the protein which are closer than 8 Å of the ligand are allowed to move and those belonging into the buffer region are kept frozen. [32] An alternative approach relies in the use of entropy estimation using essential modes derived from long MD. However, to compute the entropy difference, an additional long simulation of the protein in the ligand free state is required. For a recent example of how to determine entropy using essential modes for ligand binding see the recent work by Guardia et. al. [33]

Overall performance

Concerning the performance of the MM-PB/GB(SA) methods, each of them has been successfully applied to the study of several protein-ligand and protein-peptide complexes, although it should be noted that their performance is strongly system dependent.[34,35,36]In a recent study, Stoical et. al. thoroughly analyzed the convergence of the MM/PBSA method by ranking the binding affinities of the inhibitor saquinavir with the wild type (WT) and three resistant mutants of HIV-1 protease. By performing 10 ns of unrestrained dynamics for each protease-inhibitor complex the authors show that sampling of at least 4ns are necessary to obtain converged enthalpies while at least 6 ns of sampling, are necessary for the convergence of the entropy. Interestingly, converged enthalpy and entropy estimates produce ligand binding affinities within 1.5 kcal/mol of experimental values, with a remarkable level of correlation to the experimentally observed ranking of resistance levels. [37]

As described in the above mentioned example, of the main problem with these methods is the difficulty to converge the energy and entropy averages reliably. One reason for this difficulty is that energy calculations encompass fluctuations not only of the ligand and the binding site, but also of parts of the protein that are remote from the binding site, which are less relevant to the binding process, but nonetheless may contribute with considerable energy fluctuations.[9] To solve this issue usually the single-trajectory approach is used, in which only one MD simulation of the protein-ligand complex is carried out, and conformations of the nominally free ligand and free protein, are then derived simply by deleting the protein or ligand from the complex. In this single-trajectory approach, only the ligand-protein interactions contribute to the computed change in energy, clearly reducing the noise in the computed averages. [30] However, even with the single-trajectory approach and a precise force field, long MD simulations do not necessarily sample a representative ensemble of the relevant complex conformations, and therefore lead to inaccurate binding free energy calculations.

The other weak point of these methods is the estimation of the ΔGsv. In this context, Hou et. al.[36] showed that good correlations with the experimental results can only be obtained when the electrostatic interactions between the protein and the ligand are compensated by the solvation free energy contribution and when the experimental binding free energies span a wide range of affinities.[36] Moreover, in the same work, they compared the performance of both PB(SA) and GB(SA) solvation methods, for ranking the binding affinities of six different protein-ligand systems, and found that while PB(SA) performed better for the determination of absolute binding free energies, GB(SA) results allowed better relative ranking, a result which is more important in many applications such as drug design.[36]

A final important issue to take into account when using implicit solvation methods, concerns the ratio between the internal dielectric constant and the atomic radii values, as shown by Naim et. al. [38] The work nicely shows that the use of higher dielectric constant and smaller atomic radii values result in smaller errors.

In summary, although fast and easy to apply, MM-GB (SA) and MM-PB (SA) methods are not accurate enough in their binding free energy estimations. Therefore, considerable amount of work is devoted to improve them. In the following section we will describe two recent developed strategies to improve ΔGB calculations using end point methods.

Quantum mechanics based methods

An appealing idea to improve the accuracy of end point methods is to rely on a better (more accurate) calculation of both the ΔEint and ΔGsv terms, for example, using quantum mechanics (QM) instead of a molecular mechanics based force field (MM) for the whole or a relevant part of the system (as in the so called QM/MM schemes).[39,40,41,42,43,44] The QM methods usually provide a better balance between the electrostatic intermolecular interaction energy in the complex and the solvation energy calculated using a continuum solvent model since charges are allowed to change upon the change in environment (polarization effects) in contrast to classical methods. A recent example of this approach is presented in the work by Anisimov et. al.[45] who introduced the use of a MM/QM-COSMO scheme. The methodology improves both the determination of the energetic interaction term by providing an enhanced description of the protein ligand interactions using a linear-scaling full QM approach, based on a semi-empirical Hamiltonian (PM3) to reduce the computational cost, and the estimation of the salvation energy contribution, by using the condutor-like screening model (COSMO), which provides an improved description of the solvent drug electrostatic interactions. Anisimov et. al. computed the binding energy of several phosphopeptides (Ac-pYEEI, Ac-pYEEG, Ac-pYEEA, Ac-pYEAI and Ac-pYAEI) to the Src Homology 2 (SH2) domain of human Lck. Table 1, shows the results obtained with both classical MM-GB(SA) and MM-PB(SA) methods, and with the MM/QM-COSMO method, as well as the experimental determined values. While classical methods signifi-

cantly overestimate the binding free energy by more than 50 kcal/mol, the MM/QM-COSMO improves the accuracy of the absolute binding free energy calculation, with differences of around 1 kcal/mol. Moreover, what is also extremely relevant, the ranking of the different complexes is much improved.

Table 1- Summary of the main results obtained by Anisimov et. al. [45] comparing Classical and QM based methods for determining protein-ligand binding free energies.

Complex	$\Delta G_{MM\text{-}GB\ (SA)}$ (kcal/mol)	$\Delta G_{MM\text{-}PB\ (SA)}$ (kcal/mol)	$\Delta G_{MM/QM\text{-}COSMO,\ radiioptim}$ (kcal/mol)	$\Delta G_{exp.}$ (kcal/mol)
Ac-pYEEI	-61,5 (0,2)	-73,1 (0,3)	-10,0 (0,3)	-9,4
Ac-pYAEI	-66,1 (0,3)	-80,6 (0,3)	-10,6 (0,3)	-8,7
Ac-pYEAI	-58,4 (0,4)	-74,7 (0,6)	-8,5 (0,3)	-8,2
Ac-pYEEG	-58,7 (0,3)	-73,2 (0,3)	-7,5 (0,4)	-7,9
Ac-pYEEA	-62,1 (0,4)	-81,0 (0,5)	-8,2 (0,3)	-7,8

These results clearly show that QM based methods are a promising tool for ligand binding estimations as they have significantly improved the binding free energy calculations by yielding values that are in the same order as those obtained experimentally and a good relative trend between the different ligands. As the classical methods failed to approximate the ΔG values and also provided wrong tendencies, we expect to see a significant increase in QM based applications. Moreover since computational cost is not too demanding and the scalability is improving.

Explicit water based methodologies

It is well known and established that upon ligand binding, water molecules that are tightly bound to the protein surface are displaced, and that this solvent reorganization significantly contributes to the binding free energy of ligands.[11,13,15,16,17] Although the above mentioned implicit solvent approaches are widely used and sometimes provide good results, they can not deal with specific and tightly bound water molecules failing to describe the ligand binding process and producing errors in the affinity estimation. The importance of this contribution is underscored by several lead optimization strategies aimed at displacing ordered water molecules to improve affinity,[46] and by the studies showing that the change in ligand affinity is found to correlate with the ease of displacement of the ordered water molecules at the protein surface.[47]

In principle, water protein interactions can be thoroughly analyzed with molecular dynamics simulations in an explicit solvent environment. However, as the exchange of water molecules between a binding site and the bulk can be slow, specialized methodologies may be required to estimate accurately the locations and thermodynamic properties of the surface bound water molecules.[48] Li and Lazaridis used the inhomogeneous fluid solvation theory (IFST) to compute the thermodynamic properties of water molecules in protein ligand binding sites.[18,19] The IFST allows the extraction of binding enthalpies and entropies as well as their components from a plain MD simulation, what has been used to investigate the role of water molecules in ligand binding to HIV-protease, Concanavalin A, and Cyclophilin A.[11,12,49] Using the same analysis as that proposed in the IFST, recently, we were able to show that solvent structure and dynamics at protein surfaces involved in carbohydrate binding proteins (lectins) are very different as those from bulk, allowing the identification of the so called water sites (WS) or hydration sites. This WS correspond to space regions adjacent to the protein surface where the probability of finding a water molecule is significantly higher than those observed in the bulk, usually more than five times. Interestingly, our results showed that the position of the WS in the apo protein closely match the position of the carbohydrate hydroxil (-OH) groups in the protein saccaride complexes, underscoring the role played and the information that can be gained by analyzing the water behaviour from an explicit solvent all atom molecular dynamics simulations.[15,16,17]

A recent example of how the IFST and the determination of WS in the apoprotein can be used to compute the salvation energy contribution to the binding free energy of ligands, is described in the recent work by Abel et. al.[47] In the mentioned publication, the contribution of the solvent to the binding free energy for asset of small molecules inhibitors of FactorX aprotein (fXa) are estimated using an ad-hoc scoring function based on the propensity of each ligand to displace the previously determined water or hydration sites. The test set consists of 28 complexes off Xa-inhibitors, with available crystal structures and thermodynamic binding data. To estimate the free energy contribution of the solvent displacement, the authors first defined the binding site volume, as the space region that lies within 3Å of any ligand heavy atom in a multiple structural based alignments of all mentioned complex structures. Secondly, the authors employed a clustering technique to build a map of water occupancy in the fXa active site using data from only a single 10ns MD simulation trajectory, and assigned chemical potential sto43 identified water/hydration sites using the IFST method. They employed this information in order to construct a semi-empirical extension of the IFST which enables computation of the free energy differences (ΔΔG values) for the selected ligands (an example of the results obtained for five complexes are shown in Table 2 below), and compared the success of this approach with the more standard technique, in this case MM-GB (SA). The free energy differences calculated from the semi-empirical model are shown to correlate exceptionally well with experimental data with a correlation coefficient, R2, of 0.81 between the experimental and computed ΔΔG values (which is reduced to 0.80 after leave-one-out validation). The method substantially out performs the analogous MM-GB (SA) calculations (R2=0.29).

Table 2. Computed and experimental ΔΔG values for a set of Factor Xa protein inhibitors taken from Abel [47]. ΔΔGexp correspond to the experimental change in binding free energy. ΔΔG3p and ΔΔG5p correspond to the estimated changes in free energy determined from the water/hydration site three and five parameter empirical function, ΔΔGMM-GB(SA) corresponds to the binding free energy difference obtained with the MM-GB(SA) method described previously. All values are in kcal/mol.

Table 2-

Protein-inhibitorcomplex	$\Delta\Delta G_{exp}$ (kcal/mol)	$\Delta\Delta G_{3p}$ (kcal/mol)	$\Delta\Delta G_{5p}$ (kcal/mol)	ΔΔGMM-GBSA (kcal/mol)
Young:38 - 2J4I:GSJ	-6.26	-4.87	-4.83	-7.27
Young:32 - Young:33	-4.11	-4.87	-4.83	-7.72
1MQ5:XCL -1MQ6:XLD	-2.94	-2.85	-2.54	-4.22
2BQ7:IID - 2BQW:IIE	-2.01	-1.73	-1.95	-8.81
1NFX:RDR - 1NFX:RRR	-0.59	+1.94	+1.53	+2.01

To estimate or compute the binding free energy using the hydration sites derived information the authors used a three and/or five-parameters coring function, consisting of a sum over all ligand heavy atoms–hydration site pairs. Each time a ligand heavy atom was found within some parameterized distance of a hydration site, a negative (or favorable) contribution to the binding energy was added based on the hydration site thermodynamic (i.e. Energetic and Entropic) properties, according to the following equation (Equation 6):

$$\Delta G_{bind} = \sum_{lig,hs} E_{rwd}\left(1-\frac{|\vec{r}_{lig}-\vec{r}_{hs}|}{R_{co}}\right)\theta(E_{hs}-E_{co}\times\theta(R_{co}-|\vec{r}_{lig}-\vec{r}_{hs}|)) - T\sum_{lig,hs} S_{rwd}\left(1-\frac{|\vec{r}_{lig}-\vec{r}_{hs}|}{R_{co}}\right)\theta(S^{e}_{hs}-S_{co}\times\theta(R_{co}-|\vec{r}_{lig}-\vec{r}_{hs}|)) \quad (6)$$

Where: ΔG_{bind} is the predicted binding free energy of the ligand, E_{hs} is the interaction energy of a given hydration/water site, S^{e}_{hs} is the excess entropy of a given hydration site, Θ is the Heavy side step function, S_{co}, E_{co}, S_{rwd} and E_{rdw} are fitted entropy and energy cutoff/normalization parameters and R_{co} is a cutoff distance used to determine whether a heavy atom of the ligand is able to displace a water molecule from the hydration site. They also constructed a three parameters coring function by fixing R_{co}, S_{rwd} and E_{rdw}.

In summary the authors have been able to show that calculation of the solvent reorganization free energy contribution to the ligand binding free energy is possible using the IFST and that even considering only this contribution it is possible to obtain good results for the prediction of relative binding free energies of a set of inhibitors. The results, beyond being a proof of concept of the relevance of solvent reorganization in the binding process, show that the analysis of solvent structure can yield a better estimation of the protein-ligand interaction energy and also of the entropic change upon binding, resulting in much better free energy calculations.[47]

Free Energy methods of the ligand association/dissociation process

In opposition to the end-point methods described above, to determine the free energy barrier for either the association or dissociation process, the whole process of ligand binding must be studied, and what is more demanding the free energy needs to be computed along the way. One of the most popular and successful methods to study such processes are the Multiple Steered Molecular Dynamics methods.[50,51,52,53,54,55,56]

Multiple steered molecular dynamics (MSMD).

In this method, a time dependent external applied force is to the system under study. This force drives the system through an arbitrary reaction path coordinate (RC), forcing the molecule to visit energetically less probable configurations. The external force applied can be expressed as:

$$F = K\,(x_0 + vt - x) \quad (7)$$

Where K is an arbitrary constant, x is the actual value of the system RC, x_0 is the initial desired position of the RC position, and v is the velocity at which the RC equilibrium position is moved to guide the system. The selection of the reaction coordinate and the velocity result critical when performing MSMD. The external work necessary to move the system along the chosen reaction coordinate can thus be easily computed by integrating over the external applied forced. Different starting configurations, extracted from an equilibrium ensemble, can be used and different non-equilibrium work profiles, for the same process along the selected reaction coordinate can be obtained. Based on these data, in 1997Jarzynski [20,21] proved that the resulting free energy for the process ($\Delta G_{A=>B}$) can be computed with the following equation:

$$\Delta G_{A\to B} = -\frac{1}{\beta}\ln(\exp(-\beta W_{A\to B}))_A = -\frac{1}{\beta}\ln\sum_{i=1}^{N}\frac{1}{N}\exp(-\beta W_{i,A\to B}) \quad (8)$$

Where: W_i is the computed work profile (for the i^{th} trajectory) when moving the system from state A to state B, and the exponential average is only done over an equilibrium ensemble of state A. This relation provides a way to obtain the free energy change of the process under study. The only two requirements for this equality to be valid are that the initial ensemble over state A be equilibrated, and that the exponential average be converged. There is no requirement as to how the switch from state A to state B should be done. In the following sections we describe the application of the presented method for the study of small and relatively large ligands association/dissociation processes in proteins.

Process 1: Heme proteins and small ligand association/ dissociation

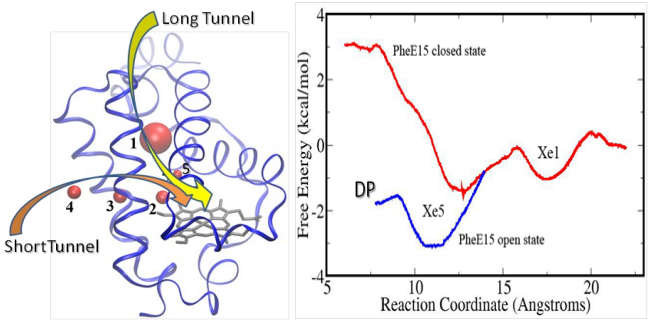
The first studies of ligand association and dissociation process using MSMD where performed using small ligands (CO, NO, O_2), since they can move fast and smoothly through the protein matrix. These small ligands bind tightly to metalo-proteins, particularly heme proteins, and there are many studies of small ligand migration process in these proteins.[50]Heme proteins are all the proteins that contain an iron-porphyrin complex as a prosthetic group, they are found in all living organisms and perform a wide variety of tasks, including sensing and transport of small gases and catalysis. Subtle regulation of the protein's affinity for these small ligands is the key issue determining a heme protein's function, as shown for different widely studied members of this group. [39,57,58,59] Usually, the heme is deeply buried inside the protein and therefore the ligand affinity is intimately related to the ligand migration process across the protein matrix, which is determined by the presence of internal cavities and tunnels [54,57,60,61,62] and/or the presence of specific residues acting as "gates". [51,63,64]

Due to their small size and the presence of tunnels and gates, the tudies using Molecular Dynamics, with a variety of strategies, see Arroyo-Mañez *et. al.* for a recent review,[50] including umbrella sampling, metadynamics, and MSMD.[54] In the following paragraphs, we present two examples where the free energy profiles for ligand entry and escape have been computed allowing to successfully explain the experimentally observed kinetic data with structural atomic detail.[54]

Free energy profiles of ligand association in the truncated hemoglobin (trHb) family of proteins.

The trHbs are a sub group of the globin protein family, displaying a conserved structural fold consisting of a two-over-two small and compact helical structure, with the heme group totally buried inside it.[58,65] One of their most salient features revealed by the

crystallographic structures[57,66] is the presence of a tunnel cavity system that connects the solvent with the heme active site, where the small ligands binds to the iron.[39,51,54,60,61,63,64] Both the main, or long tunnel (LT), and the secondary short tunnel (ST), were clearly revealed by high xenon pressure crystallographic studies of Mt-trHbN[57,66], which showed several Xe atoms located along it (Figure 1).



(A) **(B)**

Fig 1- A) Mt-trHbN xenon adducts showing several Xe atoms with different occupancy levels. The long tunnel is evidenced by the connection form Site 1, to Site 5 and from there to the distal pocket (shown by the yellow arrow). The short tunnel is evidenced by the connections form Site 4, through Site 3, Site 2 and the distal pocket. (shown by the orange arrow). **B)** Free energy profiles for ligand migration along the LT of Mt-trHbN determined using MSMD. Red line corresponds to Oxygen migration in the "closed" deoxy state. Blue line corresponds to the NO migration in the "open" oxygenated state.

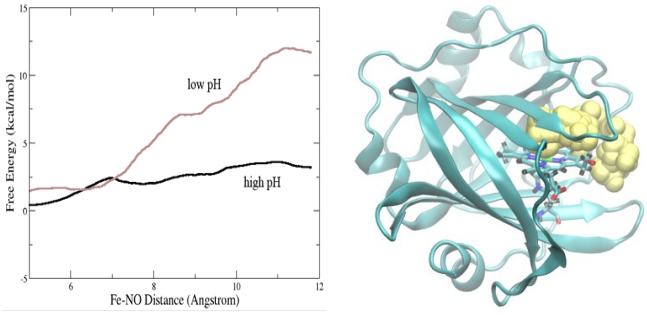
One of the salient features of Mt-trHbN lies in the fact that nitric oxide reaction with the oxygenated protein is ca. 50 times faster than oxygen or CO binding/association to the ligand free protein, indicating the presence of significant change in the ligand entry rate upon oxygenation. MD studies of the free an oxygenated protein from our group[51,64,67] showed that residue PheE15 located in the middle of the tunnel, changes its conformation depending on the coordination state of the iron. In order to analyze how ligand association rate would vary due to PheE15 conformational change, we computed the free energy profile for oxygen entry to the free protein, and NO entry to the oxygenated protein using the MSMD method described above. The resulting free energy profiles shown in Figure 1B, show that while oxygen entry to the long tunnel is sterically hindered by the closed state of PheE15 (red line in Figure 1B), and therefore oxygen must enter trough the short tunnel (data not shown). Once the protein is oxygenated, PheE15 moves to the "open" state which results in a free energy tunnel that draws NO inside the active site, as shown by the blue line in Figure 1B. As consequence, NO entry to the oxygenated protein is much faster than oxygen entry to the protein in agreement with the experimentally determined rates.[51,64,67]

Ligand migration in truncated hemoglobins from the II or O group

The LT is also evident in the structure of another truncated hemoglobin from M. tuberculosis (Mt-trHbO). Interestingly, however in this group of proteins (the O group), a conserved tryptophan occupies position G8, and partially blocks the LT access to the heme. The key role of TrpG8 for controlling ligand access in Mt-trHbO was confirmed by kinetic measurements on site directed mutants which showed that, when Trp is changed for a smaller residue, like Phe, the ligand association rate increases several times. [60,68]The results from our group show that the MSMD method correctly predicts the change in the association rate, since the barrier for small ligand access to the heme is diminished when Trp is mutated to Phe and is negligible when it is changed for alanine.[60,61] Taken together these results nicely show that the MSMD method is able to yield accurate results for the process of small ligand association to proteins. Hereafter, we will turn to an example where the same method is applied to the study of small ligand release.

pH dependent Nitric Oxide escape in the Nitrophorins.

Nitrophorins are small heme proteins that transport Nitric Oxide (NO) in a pH dependent way. At pH below 6, the protein remains loaded with NO since the escape rate is slow. When the pH increases above pH 7, as in the victims tissue, the NO dissociation rate is increased about 50 times, and NO is readily released. Previous experimental and theoretical evidence showed that the nithophorin 4 (NP4) exists in two different pH dependent states, called low and high pH, and that stable MD of each state, could be performed by selecting the corresponding appropriate initial X-ray structure and setting the differential protonation state for the key residue Asp30. [55] Based on these simulations, we used the MSMD strategy to compute the free energy barrier for NO escape from the active site to the solvent in each state. For this sake we performed 20 MSMD simulations, ending with the NO outside of NP4 for each protein conformation. The chosen reaction coordinate was the Fe-NO distance without any restraints, allowing NO to explore any possible way out of the protein. To avoid the NO escape on high-pH protein conformation, all MSMD simulations were started from a snapshot of NO located in the distal pocket. The resulting free energy profiles and escape path, shown in Figure 2, clearly demonstrate that in the low pH conformation the protein is "closed", since there is a high barrier blocking the NO escape path (~10 Kcal/mol). On the other hand in the high pH conformation the protein is clearly "open" as the NO escape barrier is only about 2 kcal/mol.[55] The same methodology and results were later obtained for nitrophorin 2, showing that modifying the free energy for ligand dissociation may be the general mechanism for regulating NO affinity in Nitrophorins.[56]



(A) **(B)**

Fig 2- A) Free energy profiles of NO release from NP4 at low-pH (gray line) and at high-pH (black line) using MSMD method. **B)** NO escape path in Np4 high-pH open conformation.

It is important to note that in order to use the SMD for the study of small ligand migration, several practical considerations need to be taken into account. First, the pulling speed or velocity needs to be carefully selected. For small (gaseous) ligands speeds between 0.001 – 0.005 Å/ps are usually adequate. For larger ligands smaller velocities may be need. A practical way of determining if the pulling rate is adequate is to perform a small (2-3) MSMD simulations with a given speed, and another set starting from the same initial conformation but using a twice slower speed. If the resulting profiles are randomly ordered, the slower speed is adequate. On the contrary if the slower MSMD yields work profiles which are significantly lower than those obtained with the fastest speed, still smaller values may be required. The number of the MSD runs must be of at least 10, and usually 10-20 simulations yield converged and statistically significant results. A practical way of determining the number of MSMD simulations and the convergence consist in performing "take-one-out" analysis. The analysis consists in computing the free energy using all "n" sets, and comparing the results with the profile obtained using all possible "n-1" work sets. If all the resulting profiles are similar to within the desired energy accuracy, i.e. 1 kcal/mol the results are converged and the number of SMD runs is fine. On the other hand if several of the "n-1" deviates significantly, particularly yielding higher values than the profile using all sets, than more SMD are needed. Another possibility to check convergence is to divide the n SMD simulations in two blocks and comparing the resulting free energy profiles.

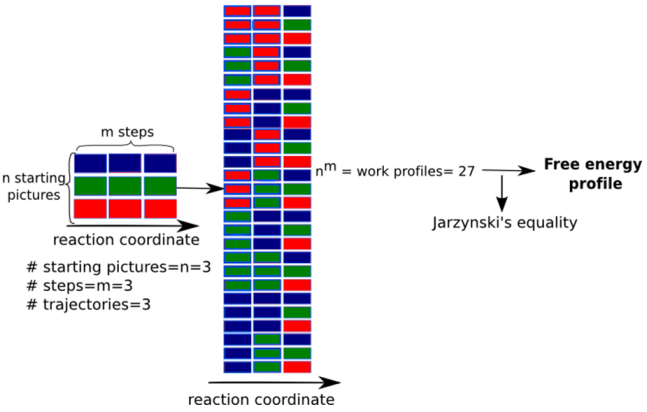
In summary, the MSMD method provides accurate estimates of the free energy profiles for small ligand migration process in proteins (for both association and dissociation), allowing therefore the comparison of ligand affinities as well as kinetic entry and escape rates. Moreover the methods explicitly sample the corresponding process yielding a structural atomic detailed picture of them that allows determination of key residues and the role that they plays on determining ligand affinity.

Process 2- Association/dissociation and affinity of drug like ligands

As for small molecules, the affinity of big drug like ligands is also associated to the free energy along the binding and release process, and can therefore be determined by computing the corresponding free energy profile. The simulation of association or binding process is a very demanding task, since if the protein ligand complex structure is unknown; it is very difficult to select a proper reaction coordinate that guides the process. If the protein ligand complex structure is known, on the other hand, it is straight forward to study the ligand release process, by pulling the ligand out of the protein, and if the free energy is measured along, the resulting profile will allow determination of the ΔG_B. However, the computational cost is high and therefore is not yet employed routinely by the scientific community.

The Mutiple-Step Trajectory Combination method

Recently, a small modification of the MSMD method was successfully applied to the study of a series of disaccharide ligand affinities in Galectin-1.[52] The Mutiple-Step Trajectory Combination (MSTC),[52,53] as it is called, is similar to the MSMD method, but with the following modification: the reaction coordinate is divided into *m* steps, and at the end of each step the system is re-equilibrated. After generating *n* different trajectories, starting from selected snapshots of the protein-ligand complex, and driving the system along the reaction coordinate in the desired m steps, all possible complete work profiles (along the whole reaction coordinate) are generated from the combination of the individual step work profiles, but taken from any possible. (As shown in Scheme 3).



Scheme 3- Schematic representation of the MSTC method. The scheme shows how by performing three trajectories divided in three steps, (Where each color block represents a trajectory step) results in 27 seven possible work profiles, resulting from the combination of blocks from different trajectories. Arrow represents progress along the reaction coordinate. The free energy profile is thus obtained from the 27 trajectories using Jarzynski´s equality.

This combination is only possible since due to the equilibration performed after each step, it can be assumed that each step trajectory is independent of the previous one. As a result, for n trajectories of m steps each, they can be combined to produce n^m different work profiles. The resulting free energy profile, is then obtained by combining the n^m work profis using Jarzynski equality already described.[53]

The MSTC method was used in order to study the binding of eight different disaccharides to bovine spleen galectin-1.[52] These animal lectins, are important modulators of several signaling processes such as apoptosis, immunoregulation and growth control in mammals, and are therefore important drug target, particularly in cancer therapy. Galectins function by binding glycan ligands of glycoprotein receptors on the cell surface and one of their key properties is their ligand, saccaride, specificity. To determine the binding free energy of each disaccharide ligand, 28 different starting snapshots were selected for each case, the pulling trajectories were performed in 8 individual long steps of 1 Å. The ligand was pulled in steps of 0.1 Å, after each small steps, the system was equilibrated for 2.5 ps. When 10 small steps were performed, a long step was done and the system was equilibrated for 50 ps. Combining the 8 individual step work profiles for all trajectories yields 10^7 work profiles that are used to determine the final free energy profile using Jarzynski equation, for each ligand. To analyze the performance of the MSTC methodology, the resulting computed binding free energies where compared with experimental Isothermal titration calorimetry (ITC) measurements for the same ligands, as shown in Table 3.

Table 3- Binding Free energy (ΔG_B) for different galectin-1/ disaccharide complexes ΔG_{Bexp} and ΔG_{Bsim} correspond to the experimental and computed (i.e. simulated) values respectively.
As can be seen from table 3, the computed ΔG_B values for the different disaccharides are in good agreement with the experimentally measured free energies. Moreover, experimental and computed free energies are linearly related with a slope of 1.1 which shows that the trend in binding affinity is correctly predicted. It is also interesting to highlight that the slope of the line between the experimental determined enthalpy contribution to binding and the ΔG_{Bsim} is only 0.7, indicating that the simulations captured additionally the entropic contribution to binding.

Table 3-

Complex	ΔG_{Bexp}(kcal/mol)	ΔG_{Bsim}(kcal/mol)
Galβ1,4GlcNAc	-5.66	-5.72
MeO-2Galβ1,4Glc	-5.84	-6.50
Galβ1,3GlcNAc	-5.79	-6.45
Galβ1,4Man	-5.39	-6.10
Galβ1,4Fruc	-5.41	-5.21
Galβ1,4Glc	-5.17	-5.62
Galβ1,3Arp	-5.15	-4.14
Galb1,4Glcβ	-5.00	-5.20

In summary the MSTC, which is a variation of the MSMD method correctly allows comparative determination of binding free energy profiles of big drug like ligands to protein. However the computational cost is roughly five to ten times that needed for conventional MSMD.

High Demanding Computational (Brute Force) Methods

As for any microscopic process the associated free energy can be determined from a plain MD simulation, provided enough sampling of the desired process is achieved. In the case of ligand binding to a protein receptor, this would mean that during the MD simulation, many (close to a hundred) ligands association and dissociation events are observed. This usually requires huge computational resources, and therefore the above mentioned biased sampling strategies are chosen. However, in a recent report and taking advantage from the recent evolution of Graphics Processing Units (GPUs) that provides the ability to reduce the running time of long MD simulations, Buch *et. al.[24]* were able to compute the ligand binding free energy for the serine protease β-trypsin inhibitor benzamidine, combining several unbiased simulations.
The goal was achieved by performing ca. 500 trajectories of free diffusion of benzamidine around trypsin, each of 100 ns length, which sums up to 50 microseconds of MD simulation. In the starting structure, the ligand was placed at 35Å from the binding pocket and allowed to move freely in the box for 50ns, but kept at a minimum distance of 20 Å from the protein. From this simulation 500 snapshots were selected as the starting structures for free diffusion production MDs. The final systems consisted of 35 thousand atoms, and 70x63x80 Å3 size box. Visual inspection of the 500 simulations showed that 187 trajectories (37%) successfully reached the bound state, that the ligand explores the entire simulation box, and that several clusters with the ligand bound to the protein surface can be observed.
To compute the binding free energy from the simulation data, proper statistical thermodynamic analysis is required. Basically, the idea is to translate the benzamidine free diffusion profile into a binding process profile. To perform this task, Markov States Models (MSMs) were constructed to describe the ligand association process in terms of structural parameters or states. [69] The requisite for using MSMs analysis is that the simulations are long enough to be in local equilibrium, which is analogous to state that the future of the system will depend only on its current state and not on its past history. In practice, first a set of microstates are defined that allow to describe the process of interest (see below), then each structure along a given MD trajectory is assigned to one of the defined microstates, and the
transition matrix is computed for each pair of microstates. Each transition matrix value (T_{ij}) is computed as the probability for the system to move from microstate *i* to microstate *j*, according to equation 9.

$$T_{ij} = \frac{number\ of\ transitions\ i \rightarrow j\ in\ time\ T}{number\ of\ starts\ in\ i} \quad (9)$$

From the first eigenvector of the transition probability matrix the potential of mean force (PMF) of the corresponding process is obtained.[69]
To analyze the data, the authors constructed three different MSMs of the process of decreasing resolution. First a high resolution three-dimensional model was built, clustering the whole box in 9700, 36 Å3 bins. The model was further coarse grained into a two-dimensional projection (50x50 bins), that better captures the binding process, and allows identification of several metastable intermediate states along the binding process. The resulting 2D PMF shows five clearly distinct states, named S0 to S4. S0 corresponds to benzamidine in the bulk (whose free energy is set to zero). S1 corresponds to the first interaction between the ligand and the protein surface, S2 and S3 are two minima located left and top from S1 in the Figure 3. Finally, S4 corresponds to the bound state, which lies -6 kcal/mol below the bulk. The analysis of the corresponding MSM eigenvectors, that provide the transition time scales between the sites, showed that transitions from S0 to S4 going through S1 and S3 have a 6-10 ns timescale, while transition from S2 to S3requires ca. 20ns. This shows that S2 is probably a secondary binding pocket not directly involved in the binding pathway, see Scheme in Figure 3B for details.

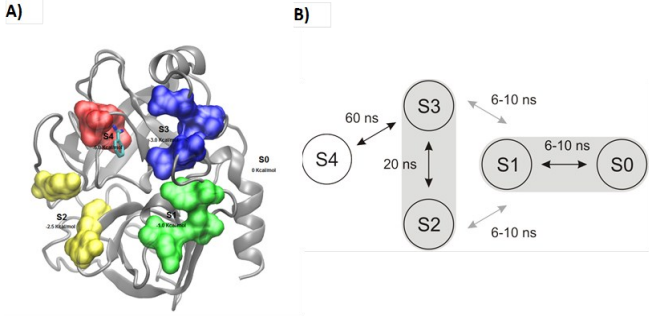


Fig 3- A) The five states resulting of 2D PMF. S0 corresponds to benzamidine in the bulk (whose free energy is set to zero). S1 (green) corresponds to the first interaction between the ligand and the protein surface, S2 (yellow) and S3 (blue) are two minima located left and top from S1. Finally, S4 (red) corresponds to the bound state, which lies -6 kcal/mol below the bulk. **B)** Scheme showing the time scales for the described transitions.

To obtain the binding free energy an ever more coarse grained MSM was built, using a binary distinction of all conformations into bound (PMF ≤ 3 kcal/mol) and unbound (PMF > 3 kcal/mol), the values for each conformation PMF are taken from the full 3D PMF. The standard binding free energy was the calculated using equation 10.

$$\Delta G_B = -\Delta W_{3D} - k_B T \log (V_b/V_0) \quad (10)$$

Where: ΔW_{3D} is the depth of the PMF, k_B is the Boltzmann constant, T is the temperature, V_b is the bound volume calculated as the integral of all microstates considered to be bound, and V_0 is the standard-state volume. The result yields a binding free energy of -5.2 kcal/mol, which differs by only 1 kcal/mol from the experimentally determined value.[24,70]

Moreover, assuming a two state process, k_{on} and k_{off} directly relate to the mean times of binding and unbinding respectively. In this scenario the mean first passage time (MFPT) for the binding (on) and release (off) reactions,[71,72] which are directly determined from the binary MSM model, are directly related to association and dissociation rates according to the following equations:

$$k_{off} = 1 / MFPT_{off} \quad (11)$$

$$k_{on} = 1 / (MFPT_{on}\, C_{lig}) \quad (12)$$

where C_{Lig} is the ligand concentration, defined by the free volume. The resulting calculated MFPT are 444ns and 1.17×10^4 ns for the association and dissociation processes which yield values that deviate ca one order of magnitude from the experimental data.[73] In conclusion, this brute force approach provides accurate and quantitative information to reconstruct the complete binding process, and therefore the associated binding free energy. However, it requires a large amount of simulation time and therefore available computer power. The high computational cost and careful data analysis involved make it difficult to envisage its use for virtual screening and even ranking of a small set of different ligands, moreover when faster and similarly accurate strategies are available as those described above. On the other hand the presented methodology may be very useful to determine the ligand association process, and therefore the structure of the protein-ligand complex when it is not available. Another point worth of mention when comparing the present method with biased methods, relies in the fact that in the present case no definition of a reaction coordinate to drive the association/dissociation process is needed.

Conclusions I: A comparison between presented methods

Although a direct comparison between all presented methods is difficult, several general tendencies are evident concerning the accuracy or predictive power for the determination of a given ligand affinity as characterized by the binding free energy (ΔG_B), and the computer power required to obtain the desired values. End point methods are much faster than the association/ dissociation sampling methods, and although their accuracy is poor and system dependent when only classical approximations (MM-GB(SA) and MM-PB(SA)) are used, significant improvement is obtained when either QM (and/or QM/MM) or explicit solvent considerations are included. Given the possible combination of explicit solvation analysis and QM based methodologies, end-point methods offer a clear and direct way for improving their accuracy, that can be incorporated stepwise into a virtual screening and or rational drug design program. Moreover, the use of empirical scoring functions based on the calculation and analysis of a set of ligands whose affinities are known, allows to increase the predictive power significantly. The other advantage of end point methods is that the required calculations are relatively easy and can be automatically set-up and analyzed, while in association/dissociation process associated methods, definition of a reaction coordinate and time consuming data analysis to determine the ΔG_B is usually required.

The weak point of the end-point methods is that they require prior knowledge of the protein-ligand complex structure, and therefore if this information is not available, either molecular docking calculations need to be performed prior to the complex MD trajectories, or the methods that sample the ligand association process must be used.

Concerning the methods that sample the dissociation or association process, if the complex structure is known the MSMD, MSTC or any other enhanced sampling method that allow determination of the free energy profile (metadynamics, umbrella sampling) allows accurate calculation of the ΔG_B at a reasonable computational cost. In this cases definition of the Reaction Coordinate is also straight forward since any reaction coordinate that pulls the ligand out from the protein will probably be good enough. The gained efficiency due to the enhanced sampling, with almost no cost in accuracy makes these methods far more interesting than the brute force methods that require large computational resources and careful data analysis.

In this scenario, the brute force free diffusion methods, have the only advantage of allowing obtaining the complex structure from the structure of the free protein and the ligand separately. However, it should be noted that if careful choice of the reaction coordinate is performed the biased sampling methods may also allow the determination of the complex structure at a probably lower computational cost.

Conclusion II: What method should I choose?

The key issue for selecting the adequate method for computing the ligand binding free energy is tightly associated or dependent on at least the following three key factors:

i] What is the specific aim and context of the research involving the calculation of ΔGB?

ii] What are the characteristics of the system under study? For example what type of ligand (small, big, polar, hydrophobic, neutral, charged) and binding pocket (shallow vs deep, containing metal, hydrophobic, small or large) , and

iii] What information is available? For example, if the protein-ligand complex is known, and if experimental information on the affinity and/or association/dissociation processes is available,

Given the different possibilities for answering the above mentioned questions, several ideal cases can be analyzed.

If the project research context is the determination of several (many) different drug like ligands ΔGB to a given protein target in a screening or rational drug design program. If for some of them the protein-ligand complex structure is known and available, and for some cases also experimental affinity values are available, the best choice is to use as first approximation classical force filed

based end point methods, that can be later improved including QM based and/or explicit solvation based approaches, or even both. Ifsome or most of the protein-ligand complex structures are unknown previous molecular docking approach can be used to fill the gap. This type of approach is also useful for the study of a set of protein mutants, or set of related proteins.

If the project is based on the study of small ligand affinity/ migration to a heme or other protein, possibly the MSMD approach to compute the free energy profiles along the protein tunnel/cavity system is the best choice, since it has been successfully applied in many similar cases as shown above.

If the project aims is to study in detail the binding mode of a small set of ligands (less than five), and where relevant information is available and/or required concerning the association and/or dissociation rates, then an enhanced sampling scheme (MSMD, MSTC, metadynamics) to study each of the process is recommended.

Acknowledgements

Computer power was provided by Centro de Computación de Alto Rendimiento (C.E.C.A.R.) at the FCEN-UBA and by the cluster MCG PME No 2006-01581 at the Universidad Nacional de Córdoba, and to Grants PICT-2010-416, UBACyT 2010-2012 and Bunge y Born to MAM. DFG, CM, VGD, LA and JPB are fellowship from CONICET. AGT and MAM are staff members of CONICET.

References

[1] Jorgensen W.L. (2004) *The many roles of computation in drug discovery, Science* 303, 1813-1818.
[2] Jorgensen W.L. (2009) *Efficient drug lead discovery and optimization, Acc. Chem. Res.,* 42, 724-733.
[3] Maryanoff B.E. (2004*) J. Med. Chem.,* 47, 769-787.
[4] Amzel L.M. (1998) *Curr. Opin. Biotechnol.*, 9, 366-369.
[5] Shoichet B.K., McGovern S.L., Wei B., Irwin J.J. (2002) *Curr. Opin. Chem. Biol*., 6, 439-446.
[6] Shoichet B.K. (2004) *Nature*, 432, 862-865.
[7] Blake J.F., Laird E.R. (2003) *Annual Reports in Medicinal Chemistry,* 38, 305-314.
[8] Gohlke H. , Klebe G. (2002) *Angewandte Chemie - International Edition,* 41, 2644-2676.
[9] Swanson J.M., Henchman R.H., McCammon J.A. (2004) *Biophys. J.,* 86, 67-74.
[10] Leach A.R. (2001) *Molecular Modelling: Principles and Applications, Pearson Education EMA.*
[11] Li Z., Lazaridis T. (2005) *J. Phys. Chem. B.,* 109, 662-670.
[12] Li Z. , Lazaridis T. (2006) *J. Phys. Chem. B.,* 110, 1464-1475.
[13] Li Z., Lazaridis T. (2007) *Physical Chemistry Chemical Physics,* 9, 573-581.
[14] Ladbury J.E. (1996) *Chem. Biol.,* 3, 973-980.
[15] Gauto D.F., Di Lella S., Guardia C.M.a., Estrin D.A., Marti M.A. (2009) *Journal of physical chemistry B.,* 113, 8717-8724.
[16] Gauto D.F., Di Lella S., Estrin D.A., Monaco H.L., Marti M.A. (2011) *Carbohydr. Res.,* 346, 939-948.
[17] Di Lella S., Marti M.A., Alvarez R.M., Estrin D.A., Ricci J.C. (2007) *J. Phys. Chem. B.,* 111, 7360-7366.
[18] Lazaridis T. (1998) *Journal of Physical Chemistry,* B., 102, 3542-3550.
[19] Lazaridis T. (1998) *Journal of Physical Chemistry B,* 102, 3531-3541.
[20] Jarzynski C. (1997) *Physical Review E,* 56, 5018-5035.
[21] Jarzynski C. (1997) *Physical Review Letters,* 78, 2690-2693.
[22] Schmidtke P., Luque F.J., Murray J.B., Barril X., *Journal of the American Chemical Society,* 133, 18903-18910.
[23] Colizzi F., Perozzo R., Scapozza L., Recanatini M., Cavalli A., *Journal of the American Chemical Society,* 132, 7361-7371.
[24] Buch I., Giorgino T., De Fabritiis G. (2011) *The National Academy of Sciences of the United States of America,* 108, 10184-10189.
[25] Aqvist J., Luzhkov V.B., Brandsdal B.O. (2002) *Acc. Chem. Res.,* 35, 358-365.
[26] Srinivasan J., Miller J., Kollman P.A., Case D.A. (1998) *J. Biomol. Struct. Dyn.,* 16, 671-682.
[27] Vorobjev Y.N., Hermans J. (1999) *Biophys. Chem.,* 78, 195-205.
[28] Kongsted J., Soderhjelm P., Ryde U. *J.* (2009) *Comput. Aided Mol. Des.,* 23, 395-409.
[29] Adcock S.A., McCammon J.A. (2006) *Chem. Rev.,* 106, 1589-1615.
[30] Gilson M.K., Zhou H.X. (2007) *Annu. Rev. Biophys. Biomol. Struct.* 36, 21-42.
[31] Kollman P.A., Massova I., Reyes C., Kuhn B., Huo S., Chong L., Lee M., Lee T., Duan Y., Wang W., Donini O., Cieplak P., Srinivasan J., Case D.A., Cheatham T.E. (2000) *Acc. Chem. Res.,* 33, 889-897.
[32] Kongsted J., Ryde U. (2009) *J. Comput. Aided Mol. Des.,* 23, 63-71.
[33] Guardia D.F., Gauto S., Di Lella, Rabinovich G.A., Marti M.A., Estrin D.A. (2011) *Journal of Chemical Information and Modeling,* 51, 1918–1930.
[34] Kuhn B., Gerber P., Schulz-Gasch T., Stahl M. (2005) *Journal of Medicinal Chemistry,* 48, 4040-4048.
[35] Pearlman D.A. (2005) *Journal of Medicinal Chemistry,* 48, 7796-7807.
[36] Hou T., Wang J., Li Y., Wang W. (2010) *Journal of Chemical Information and Modeling,* 51, 69-82.
[37] Stoica I., Sadiq S.K., Coveney P.V. (2008) *Journal of the American Chemical Society,* 130, 2639-2648.
[38] Naim M., Bhat S., Rankin K.N., Dennis S., Chowdhury S.F., Siddiqi I., Drabik P., Sulea T., Bayly C.I., Jakalian A., Purisima E.O. (2007) *Journal of Chemical Information and Modeling,* 47, 122-133.
[39] Marti M.A. Crespo A., Capece L., Boechi L., Bikiel D.E., Scherlis D.A., Estrin D.A. (2006) *Journal of Inorganic Biochemistry,* 100, 761-770.
[40] Crespo A., Scherlis D.A., MartÃ- M.A., OrdejÃ³n P., Roitberg A.E., Estrin D.A. (2003) *Journal of Physical Chemistry B,* 107, 13728-13736.
[41] Antony J., Grimme S., Liakos D.G., Neese F. *Journal of Physical Chemistry A,* 115, 11210-11220.
[42] Fu Z., Li X., Merz Jr K.M., *Journal of Computational Chemistry,* 32, 2587-2597.
[43] Ciancetta A., Genheden S., Ryde U., *Journal of Computer-Aided Molecular Design,* 25, 729-742.
[44] Beierlein F.R., Michel J., Essex J.W., *Journal of Physical Chemistry B,* 115, 4911-4926.

[45]Anisimov J.W., Cavasotto C.N. (2011) *Journal of Computational Chemistry,* 32, 2254-2263.
[46]Lam P.Y., Jadhav P.K., Eyermann C.J., Hodge C.N., Ru Y., Bacheler L.T., Meek J.L., Otto M.J., Rayner M.M., Wong Y.N., (1994) *Science,* 263, 380-384.
[47]Abel R., Young T., Farid R., Berne B.J., Friesner R.a., (2008) *Journal of the American Chemical Society* 130, 2817-2831.
[48]Monecke P., Borosch T., Brickmann J., Kast S.M. (2006) *Biophysical Journal,* 90, 841-850.
[49]Li Z., Lazaridis T. (2003) *J. Am. Chem. Soc.,* 125, 6636-6637.
[50]Arroyo-Mañez P., Bikiel D.E., Boechi L., Capece L., Di Lella S., Estrin D.A., Marti M.A., Moreno D.M., Nadra A.D., Petruk A.A. (2011) *Biochimica et Biophysica Acta - Proteins and Proteomics,* 1814, 1054-1064.
[51]Bidon-Chanal A., Marti M.A., Crespo A., Milani M., Orozco M., Bolognesi M., Luque F.J., Estrin D.A. (2006) *Proteins: Structure, Function and Bioinformatics,* 64, 457-464.
[52]Echeverria I., Amzel L.M. (2011) *Biophysical Journal,* 100, 2283-2292.
[53]Echeverria I., Amzel L.M. (2011) *Proteins Structure Function and Bioinformatics,* 78, 1302-1310.
[54]Forti F., Boechi L., Estrin D.A., Marti M.A. (2011) *Journal of Computational Chemistry,* 32, 2219-2231.
[55]Marti M.A., GonzÃ¡lez Lebrero M.C., Roitberg A.E., Estrin D.A. (2008) *Journal of the American Chemical Society,* 130, 1611-1618.
[56] Swails J.M., Meng Y., Walker F.A., Marti M.A., Estrin D.A., Roitberg A.E. (2009) *Journal of Physical Chemistry B.,* 113, 1192-1201.
[57]Milani M., Pesce A., Ouellet Y., Dewilde S., Friedman J., Ascenzi P., Guertin M., Bolognesi M. (2004) *Journal of Biological Chemistry,* 279, 21520-21525.
[58]Wittenberg J.B., Bolognesi M., Wittenberg B.A., Guertin M. (2002) *Journal of Biological Chemistry,* 277, 871-874.
[59]Ghosh A. (2008) *Elsevier*.
[60]Boechi L., Marti M.A., Milani M., Bolognesi M., Luque F.J., Estrin D.A. (2008) *Proteins: Structure, Function and Bioinformatics,* 73, 372-379.
[61]Boechi L., Manez P.A., Luque F.J., Marti M.A., Estrin D.A. (2010) *Proteins Structure Function and Bioinformatics,* 78, 962-970.
[62]Milani M., Pesce A., Nardini M., Ouellet H., Ouellet Y., Dewilde S., Bocedi A., Ascenzi P., Guertin M., Moens L., Friedman J.M., Wittenberg J.B., Bolognesi M. (2005) *Journal of Inorganic Biochemistry,* 99, 97-109.
[63]Lama A., Pawaria S., Bidon-Chanal A., Anand A., Gelpi J.L., Arya S., Marti M., Estrin D.A., Luque F.J., Dikshit K.L. (2009) *Journal of Biological Chemistry,* 284, 14457-14468.
[64] Bidon-Chanal A., MartÃ- M.A., Estrin D.A., Luque F.J. (2007) Journal of the American Chemical Society 129, 6782-6788.
[65]Vuletich D.A., Lecomte J.T. (2006) *Journal of molecular evolution,* 62, 196-210.
[66]Pesce A., Milani M., Nardini M., Bolognesi M. (2008) *Methods in Enzymology,* 303-315.
[67] Crespo A., MartÃ M.A., Kalko S.G., Morreale A., Orozco M., Gelpi J.L., Luque F.J., Estrin D.A. (2005) *Journal of the American Chemical Society,* 127, 4433-4444.
[68] Guallar V., Lu C., Borrelli K., Egawa T., Yeh S.R. (2009) *J. Biol. Chem.* 284, 3106-3116.
[69] Noe F., Fischer S.(2008) *Current Opinion in Structural Biology,* 18, 154-162.
[70]Mares-Guia M., Shaw E. (1965) *Journal of biological chemistry,* 240, 1579-1585.
[71] Singhal N., Snow C.D., Pande V.S. (2004) *Journal of Chemical Physics,* 121, 415-425.
[72] Huang D., Caflisch A. (2011) *PLoS Computational Biology,* 7.
[73] Guillain F., Thusius D. (1970) *Journal of the American Chemical Society,* 92, 5534-5536.