Optimization of a MM-PBSA/GBSA protocol for the prediction of binding free energy of Bcl-xL inhibitors

João Pedro do Vale Hipólito Cavalheiro
Department of Bioengineering, Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais, 1049-001, Lisbon, Portugal

1. Abstract

In this work, different protocols based on MD simulation and MM-PB/GBSA were evaluated and compared for their ability to predict relative binding energies for nine ligands of Bcl-xL, a protein involved in the inhibition of apoptosis in many types of cancer cells through a PPI mechanism.

Five of the ligands considered were derived by X-ray structures available in PDB (2YXJ, 3ZK6, 3ZLO, 3ZLN and 3ZLR), while three (3Z1B, 3ZC4, 3ZC5) were reconstructed by molecular modelling, in these cases editing the crystallographic ligand by MOE. Starting from crystallographic structures, water and not-ligand and not-protein residues were removed, missing residues were reconstructed, and the complex was energy minimized using MOE software. After, MD was performed using Amber with either ff14SB or ff99SB and either HPC or GPU in tree steps: minimization, equilibration and production. MM-PB/GBSA was then used to compute binding energies by using a hybrid approach in which a limited number of explicit water molecules (N_{wat} < 100 at an interval of 10), selected from the trajectory by the cpptraj "closest" command, were combined with implicit solvation using either the 4th, 8th or 12th nanosecond of simulation. For MM-GBSA, several GB models and atomic radii combinations were tested: mbondi and igb=1, mbondi2 and igb=5, mbondi2 and igb=8 and mbondi3 and igb=8. It was concluded that the parameters that provide the best balance between calculation time and correlation coefficient are: ff14SB in HPC, using the 4th nanosecond to perform MM-GBSA calculations with mbondi3 and igb=8.

Key words: hybrid solvation approach, MM-PBSA, MM-GBSA, Bcl-xL, closest command, N_{wat}

Abbreviations: MM-PB/GBSA - Molecular Mechanics Poisson Boltzmann Surface Area and Molecular Mechanics Generalized Born Surface Area; MM-PBSA - Molecular Mechanics Poisson Boltzmann Surface Area; MM-GBSA - Molecular Mechanics Generalized Born Surface Area; PB - Poisson Boltzmann; GB - Generalized Born; HPC - High Performance Computing cluster; N_{wat} - Number of explicit water molecules; MOE – Molecular Modelling and Simulations; PDB – Protein Data Bank; Bcl-xL - B-cell lymphoma-extra large; MD – Molecular Dynamics; Amber - Assisted Model Building with Energy Refinement

2. Introduction

Many physiological processes that modulate human life are regulated by interactions between different proteins. Consequently, any deregulation in those interactions induce a disease state which makes the production of drugs able to modulate these interactions a particularly interesting therapeutic area.[1]

A major goal of structure-based drug design is to find a ligand with sufficient affinity to a macromolecule to be interesting. Free energy of binding is an effective way to measure binding strength.[2] Finding an accurate computational method to predict free energies of binding could reduce the need of making synthesis and experiments which could save both time and money.[3] In this work, two methods were used: Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) and: Molecular Mechanics Generalized Born Surface Area (MM-GBSA). They use a combination of molecular mechanics energies and solvation energies and are able to reach a good balance between accuracy and computational cost.[2] Since free energy of binding of the solvated complex is very hard to calculate directly, the following equation is used (1): [4]

$$\Delta G^0_{bind,solv} = \Delta G^0_{bind,vacuum} + \Delta G^0_{solv,complex} - \Delta G^0_{solv,ligand} - \Delta G^0_{solv,receptor} \quad (1)$$

, where $\Delta G^0_{bind,solv}$ represents free energy of binding of the complex in solvated media, $\Delta G^0_{bind,vacuum}$ represents free energy of binding of the complex in vacuum, $\Delta G^0_{solv,complex}$ is free energy of solvation of the complex, $\Delta G^0_{solv,ligand}$ is free energy of solvation of the ligand and $\Delta G^0_{solv,receptor}$ is free energy of solvation of the receptor. It is also known that (2) [4]

$$\Delta G^0_{vacuum} = \Delta E^0_{molecular\ mechanics} - T\Delta S^0_{normal\ mode\ analysis} \quad (2)$$

, where ΔS is entropy change and $\Delta E^0_{molecular\ mechanics}$ is the potential energy. Prediction of free energies of binding is affected only by the difference between the entropy of the bound state and the entropy of the unbound state. These entropies have often comparable values and so, this difference is considered to be sufficient small to not influence

significantly the results. Furthermore, entropy calculation increases significantly standard deviation of the method. Since, entropy calculation increases substantially the standard deviation of the method, but his contribution make a better prediction of free energies of binding is small, it was decided to not make entropy calculations in the present work.[5]

Several approaches have been used to mimic the role that water has in reality. Implicit solvation is one of them. When using this method solvent molecules are substituted by a continuum media with a constant dielectric constant.[6] This way, in typical implicit solvation: (3):[4]

$$\Delta G^0_{solv} = G^0_{electrostatic,\varepsilon=80} - G^0_{electrostatic,\varepsilon=1} + \Delta G^0_{hydrophobic} \quad (3)$$

where the solvation effect may be divided into polar $\left(G^0_{electrostatic,\varepsilon=80} - G^0_{electrostatic,\varepsilon=1}\right)$ and nonpolar components $\left(\Delta G^0_{hydrophobic}\right)$. [5] $\Delta G^0_{hydrophobic}$ is considered proportional to the surface area. PB or GB methods are often used to predict the polar component of solvation effect and dielectric constants are switched between a high value for the solvent $\left(G^0_{electrostatic,\varepsilon=80}\right)$ and a low value for the solute $\left(G^0_{electrostatic,\varepsilon=1}\right)$.[4, 7] The border defined by the dielectric equation that defines where dielectric constant is switched between the solvent and the solute is called Dielectric Boundary (DB). Lately, some studies have proposed different sets of atomic radii, and consequently differents DB's, optimized for PB or GB calculations (such as the modified Bondi radii mbondi, mbondi2 or mbondi3). A carefully chosen atomic radii may improve considerably the estimation accuracy of solvation free energies and forces.[7] Other internal dielectric constants ($\varepsilon_{in,\ solv}$) were also tested and evaluated the impact regarding the correlation coefficient of free energies of binding with experimental data, but unfortunately the results did not lead to the discovery of a universal internal dielectric constant.[8] This way, implicit solvation only calculates the mean interactions between solute and solvent molecules; however, it may be inadequate when there are ordered water molecules at the interface.[9]

An alternative is using explicit solvation. Explicit solvent can provide a detailed effect of solvation, however it is very expensive computationally, not only because it involves a high number of water molecules, but also because it demands conformational sampling of the referred molecules.[7]

During this work, it was used a technique that combines implicit and explicit solvation by selecting a limited number of water molecules that are, in each frame of the trajectory, closest to the ligand center mass, on several ligands of Bcl-xL, a protein involved in protection of many cancer cells from apoptosis. This process is relatively easy, does not lead to a significant increase in terms of calculation time and it has been used before with promising results.[8]

3. Scope of the thesis

The aim of this work was to optimize a protocol to predict the relative free energy of binding of several human Bcl-xL ligands: 2YXJ, 3ZK6, 3ZLO, 3ZLN, 3ZLR, 3Z1B, 3ZC4 and 3ZC5. Several variables were tested for optimization. Simulations were run for 12 nanoseconds with either ff99SB or ff14SB force-fields and using either GPU or HPC hardware. MM-PBSA and MM-GBSA calculations were run on the 4th, 8th or 12th nanoseconds of simulation and with 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 or 100 explicit water molecules (ΔN_{wat}=10). In those analysis in which MM-GBSA calculations were performed, several combinations of atomic radii and born solvation models were also tested: GB^{HCT}(igb=1) and mbondi, GB^{OBC}(igb=5) and mbondi2, GB^{Neck2}(igb=8) and mbondi2, GB^{Neck2}(igb=8) and mbondi3. Two criteria were taken into account for this optimization: correlation coefficient of predicted free energies of binding with experimental data and computational resources required. The objective of this thesis was to find, from all the possible combinations of the variables presented above, the one that presented the best trade-off between the optimization criteria.

4. Methods

4.1 Preparation of files, parametrization of charges and MD steps

All ligands were taken from Lessene *et al*.[10]. 2YXJ, 3ZK6, 3ZLN, 3ZLO and 3ZLR were taken from crystal structures presents in PDB with these accession codes. 3Z1A, 3Z1B, 3ZC4 and 3ZC5 were obtained by modifying the ligand present in crystal structure by the compound 1a of figure 1, 1b of figure 1, compound 4 of figure 2 and compound 5 of the same figure in Lessene *et al.* , respectively.[10] They may also be observed in **Figure 1** and **Figure 2** of the present article.

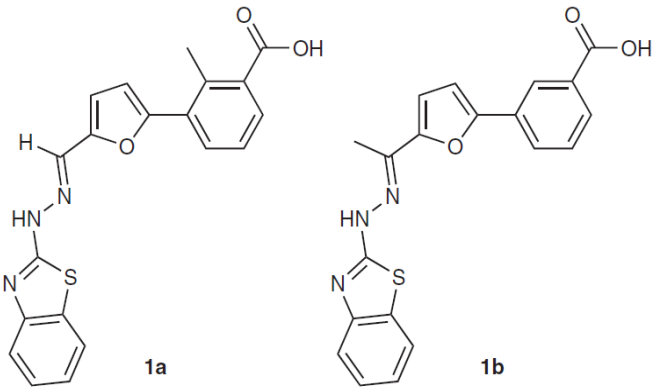

Figure 1 - Ligands of Bcl-xL:1a) 3Z1A and 1b) 3Z1B

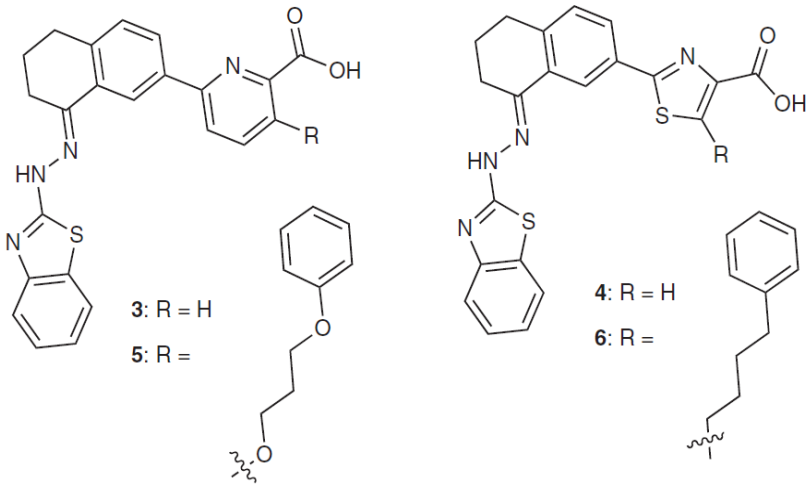

Figure 2 - Ligands of Bcl-xL:4) 3ZC4 and 5) 3ZC5

3ZK6, 3ZLN and 3ZLO had missing residues in the loop going from Ser32 to Gly44 and, so, were rebuilt using the chain A of the crystal structure of 1YSI.pdb as a template. After, all complexes were capped with an Acetyl group at the N- terminus and a NHMe at the C- terminus, followed by protonation using the Protonate3D module of MOE.[11] The geometry was minimized using AMBER12EHT force field and Born solvation model, keeping fixed the backbone and ligand atoms. Ligand geometry was minimized using the MMFF94x force field and Born solvation, keeping the receptor atoms fixed. At last, a minimization of the complex was performed using AMBER12EHT force field and Born solvation model, keeping ligand atoms fixed. Force fields such as ff99SB and ff14SB were generated to describe mainly protein systems, therefore the parameters for the ligands, which are small organic molecules, have to be derived. The General AMBER Force Field (GAFF)[4], contains most of the structural parameters for this kind of molecules, but does not contain atomic charges. Therefore, the charges of the ligands atoms were derived using Antechamber and setting BCC-AM1 as the semi-empirical QM method.

Finally, the topology and coordinate files were generated by tLeAP, by setting the desired force field and atomic radii set, uploading the ligands parameters and the previously prepared pdb files. The system was then neutralized by adding the proper amount of Na+ or Cl- ions and solvated with a TIP3P octahedric box of water with a distance of up to 10 Å from the solute.

The topology and coordinate files were used to perform molecular dynamics (MD) steps. All the steps used a non-bonded cut-off of 8.0 Å. All steps that were not a minimization, used Langevin dynamics with a collision frequency of 2.0 ps^{-1}, a seed for pseudo-random number generator of -1 and, with the exception of steps g) to m), a time step of 0.002 ps. In this work, MD was done in several steps:

a) minimization of hydrogens in constant volume (V) with a restraint of 100.0 kcal mol^{-1} for these atoms. Minimization cycles were performed until the root-mean-square (RMS) of Cartesian elements was inferior to 0.01 kcal mol^{-1} $Å^{-1}$ up to a maximum of 5000 cycles. After 1000 steps of steepest descent, it was followed a maximum of 4000 steps of conjugate gradient. The initial step size considered was 0.01 Å.

b) minimization of water and ions. Minimization cycles were performed until the RMS of the Cartesian elements of the gradient was less than 0.1 kcal mol^{-1} $Å^{-1}$ up to a maximum of 5000 cycles. After 2000 cycles, minimization method was switched from steepest descent to conjugate gradient. Protein and ligand were subjected to a restraint of 50.0 kcal mol^{-1}.

c) equilibration of the water box in conditions of constant quantity of molecules (N), constant volume (V) and a constant temperature (T) of 300 K, using 50000 steps. Bonds involving hydrogen were constrained by the SHAKE algorithmwith a geometric tolerance for coordinate resetting of 0.0003 Å and omitted in force interactions. Protein and ligand atoms were subjected to a restraint of 50.0 kcal mol^{-1}.

d) equilibration in conditions of constant number of molecules (N), pressure (P) and temperature (T). This step was executed at a constant temperature of 300 K in 50000 MD steps and with a pressure relaxation time 2.0 ps. Bond interactions involving hydrogen were omitted. On the other hand, hydrogen bond lengths were constrained using SHAKE with a geometric tolerance for coordinate resetting of 0.0003 Å. Protein and ligand atoms were subjected to a restraint of 25.0 kcal mol^{-1}. Isotropic position scaling was used as well.

e) minimization of water, ions and side-chains; a restraint of 25.0 kcal/mol was applied on protein and ligand atoms. A maximum number of minimization cycles of 5000 was used in which after 2500 cycles, minimization method is switched from steepest descent to conjugate gradient. A stopping criteria was also used, in this way minimization stopped when the RMS of the Cartesian elements of the gradient was less than 0.1 kcal mol^{-1} $Å^{-1}$. This calculation was performed in constant volume. The initial step length was 0.01 Å.

f) total minimization; a restraint of 10.0 kcal/mol was applied on backbone atoms. Minimization stopped when RMS of Cartesian elements gradient was less than 0.1 kcal mol^{-1} $Å^{-1}$ up to a maximum of 5000 minimization cycles. After 2500 cycles, minimization method changed from steepest descent to conjugate gradient. Minimization was performed in constant volume. Initial step length of 0.01 Å was used.

g) Heating from 0 K to 50 K with a backbone restraint of 10.0 kcal mol^{-1} in constant volume. This step was performed in 10000 MD-steps with 0.0005 ps each. There was no pressure scaling. Bonds involving hydrogen were omitted and constrained by SHAKE. h) to l) steps used the same parameters as g) step except initial temperature, reference temperature and backbone restraint.

h) Heating from 50 K to 100 K with a backbone restraint of 9 kcal mol^{-1}.

i) Heating from 100 K to 150 K with a backbone restraint of 8 kcal mol^{-1}.

j) Heating from 150 K to 200 K with a backbone restraint of 7 kcal mol^{-1}.

k) Heating from 200 K to 250 K with a backbone restraint of 6 kcal mol^{-1}.

l) Heating from 250 K to 300 K with a backbone restraint of 5 kcal mol^{-1}.

m) MD simulation in NVT conditions for 100 ps, using a time step of 0.001 ps and using a backbone restraint of 5 kcal mol^{-1}. All the other parameters were set equal to those used in step g).

n) MD in NPT conditions for 200 ps and with a backbone restraint of 5 kcal mol^{-1}. This step was performed in 100000 MD steps. Bonds involving hydrogen were omitted in force evaluation and constrained by the SHAKE algorithm. Pressure relaxation time was considered to be 2.0 ps. Isotropic position scaling was used. Steps o) to r) used the same variable set as this step with exception of backbone restraint and total duration.

o) MD during 100 ps and with a backbone restraint of 4 kcal mol^{-1}.

p) MD during 100 ps and with a backbone restraint of 3 kcal mol^{-1}.

q) MD during 100 ps and with a backbone restraint of 2 kcal mol^{-1}.

r) MD during 1000 ps and with a backbone restraint of 1 kcal mol^{-1}.

s) The step of production was performed in NPT conditions: constant number of particles (N), constant pressure (P) and constant temperature (T) without any restraints. It was performed in 500000 MD steps. The rest of the parameters was set equal to those used in the n) step of minimization and equilibration steps. Since each step of production corresponds to 1 ns, production run corresponded to 12 steps (12 ns).

4.2 Trajectory Analysis and MM-PBSA/GBSA calculations

In the present work, post processing analysis included a root mean square deviation (RMSD) analysis on both the backbone of the molecule and the ligand, a root mean square fluctuation (RMSF) analysis on both the ligand and the backbone-chain and an hydrogen-bond analysis (data not shown). Finally, MM-GBSA and MM-PBSA calculations were performed. The most important programs used in this part of the work included cpptraj, that allowed to perform RMSD and RMSF, as well as to select the chosen number of water molecules nearest to the center mass of the ligand for MM-PB/GB calculations with the command “closest”, and VMD, a program used to display and analyze of MD trajectories; in this work, VMD was used to perform hydrogen-bond analysis.[12] MM-GBSA calculations were performed on the trajectories obtained using different Born solvation models: GB-HCT (igb=1), GB-OBC(II) (igb=5) e GB-Neck2 (igb=8) and analyzing different fractions of trajectory: the 4th ns of production run , the 8th ns and the 12th ns. In all these analyses, one in every 10 frames was considered, for a total of 100 frames. MM-PBSA calculations were performed as well, using the same setup conditions. Finally, the quality of prediction was evaluated by correlating computed energies with experimental binding affinities (IC_{50}).[10] In this work, both the molecular dynamics (MD) and the computation of binding energies were made using the AMBER (Assisted Model Building with Energy Refinement) package.[4]

5. Results

In the following four sections, the evolution of spearman correlation coefficients between half of the minimum inhibitory concentration and the predicted free energy of binding with the number of explicit water molecules for four different setups and for the 4th and the 12th nanoseconds of simulation is presented.

5.1 GPU_ff99SB

The evolution of the correlation coefficient for GPU hardware and ff99SB force-field for the 4th and the 12th nanoseconds of simulation is shown in Figure 3.

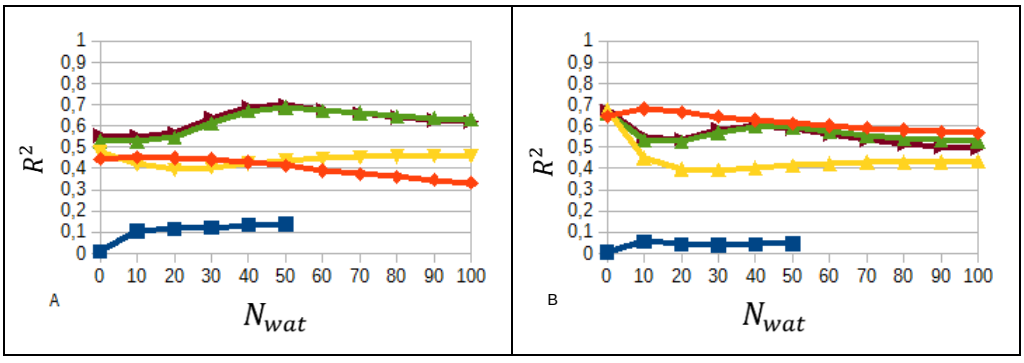

Figure 3 - Evolution of the correlation coefficient as a function of N_{wat} for simulations run on GPU, using ff99SB as force-field to perform calculations and for: MM-PBSA (—■—), MM-GBSA, igb=1 and mbondi (—◆—), MM-GBSA, igb=5 and mbondi2 (—✚—), MM-GBSA, igb=8 and mbondi2 (—▲—), MM-GBSA, igb=8 and mbondi3 (—►—) for A) 4ns and B) 12ns.

5.2 GPU_ff14SB

In Figure 4 the evolution of correlation coefficient for GPU hardware and ff14SB force-field for the 4th and the 12th nanoseconds of simulation is depicted.

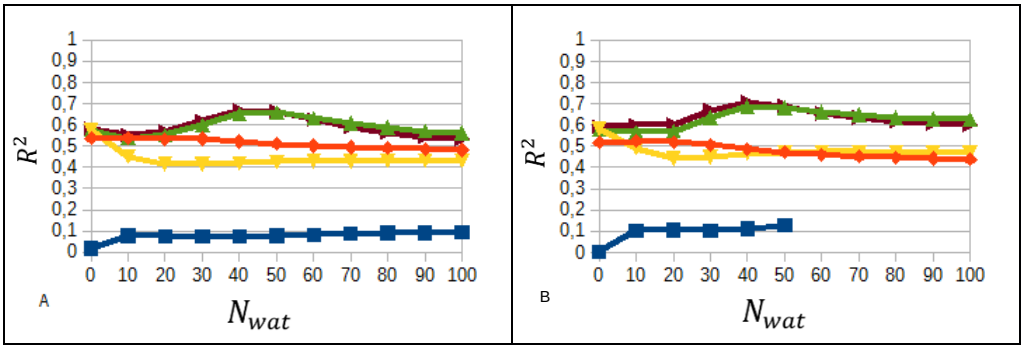

Figure 4 - Evolution of the correlation coefficient as a function of N_{wat} for simulations run on GPU, using ff14SB as force-field to perform calculations and for: MM -PBSA (—■—),MM-GBSA, igb=1 and mbondi (—●—),MM-GBSA, igb=5 and mbondi2 (—▼—),MM-GBSA, igb=8 and mbondi2 (—▲—), MM-GBSA, igb=8 and mbondi3 (—►—) for A) 4ns and B) for 12ns.

5.3 HPC_ff99SB

Figure 5 shows the results obtained on the evolution of correlation coefficient for HPC hardware and ff99SB force-field for the 4th and the 12th nanoseconds of simulation..

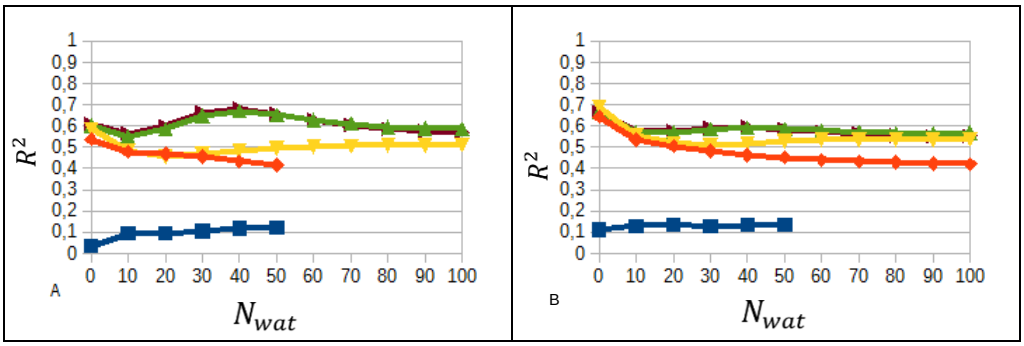

Figure 5 - Evolution of the correlation coefficient as a function of N_{wat} for simulations run on HPC, using ff99SB as force-field to perform calculations and for: MM-PBSA (—■—), MM-GBSA, igb=1 and mbondi (—●—), MM-GBSA, igb=5 and mbondi2 (—▼—), MM-GBSA, igb=8 and mbondi2 (—▲—), MM-GBSA, igb=8 and mbondi3 (—►—) for A) 4ns and B) for 12ns

5.4 HPC_ff14SB

The evolution of the correlation coefficient for HPC hardware and ff14SB force-field for the 4th and the 12th nanoseconds of simulation is shown in Figure 6..

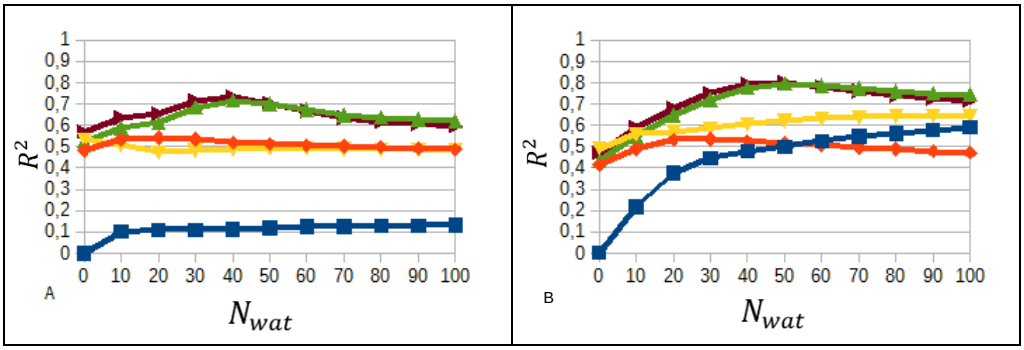

Figure 6 - Evolution of the correlation coefficient as a function of N_{wat} for simulations run on HPC, using ff14SB as force-fieldto perform calculations and for: MM-PBSA (—■—), MM-GBSA, igb=1 and mbondi (—●—), MM-GBSA, igb=5 and mbondi2 (—▼—), MM-GBSA, igb=8 and mbondi2 (—▲—), MM-GBSA, igb=8 and mbondi3 (—►—) for A) 4ns and B) for 12ns

It was also tried to extend the simulation length from 8 ns to 12 ns, just for N_{wat}=40 and N_{wat}=50 with disappointing results (data not shown).

5.5 Best results

Table 1 compiles the data for the experimental half inhibitory concentration (IC_{50}), the predicted free energy of binding (ΔG_{bind}) and the correlation between these two values (R^2) obtained for ligands 2YXJ, 3ZK6, 3ZLO, 3ZLN, 3ZLR, 3Z1B,

3ZC4 and 3ZC5, for the three best setups and for the number of water molecules that permitted the highest correlation coefficient $(N_{wat,best})$.

Table 1 - Experimental IC_{50}, predicted ΔG_{bind} and the R^2 for ligands 2YXJ, 3ZK6, 3ZLO, 3ZLN, 3ZLR, 3Z1B, 3ZC4 and 3ZC5, for the three best setups and for $N_{wat,best}$.

Setup	Ligand	IC_{50} (μM)	ΔG_{bind}(kcal mol^{-1})	$R^2_{N_{wat}=0}$	$R^2_{N_{wat,best}}$
HPC hardware; ff14SB force-field; 12th nanosecond; MMGBSA; mbondi3; igb=8	2YXJ	0.0050	-97.5583	0,4723	0,8020
	3Z1B	2.2000	-54.7939		
	3ZC4	0.0240	-80.6065		
	3ZC5	0.0062	-103.6725		
	3ZK6	0.7700	-76.4306		
	3ZLN	0.0230	-79.9789		
	3ZLO	0.0015	-98.6334		
	3ZLR	0.0011	-124.2823		
HPC hardware; ff14SB force-field; 12th nanosecond; MMGBSA; mbondi2; igb=8	2YXJ	0.0050	-97.8756	0,4345	0,7962
	3Z1B	2.2000	-57.3337		
	3ZC4	0.0240	-82.2526		
	3ZC5	0.0062	-106.4073		
	3ZK6	0.7700	-76.9822		
	3ZLN	0.0230	-83.1865		
	3ZLO	0.0015	-101.0889		
	3ZLR	0.0011	-129.6058		
HPC hardware; ff14SB force-field; 4th nanosecond; MMGBSA; mbondi3; igb=8	2YXJ	0.0050	-96.2258	0,5715	0,7345
	3Z1B	2.2000	-73.0006		
	3ZC4	0.0240	-78.4815		
	3ZC5	0.0062	-98.6354		
	3ZK6	0.7700	-68.9888		
	3ZLN	0.0230	-79.0704		
	3ZLO	0.0015	-97.0451		
	3ZLR	0.0011	-119.8487		

6. Discussion

6.1 Number of water molecules

The choice of a proper number of explicit water molecules is important because a too high number of water molecules affects not only the velocity of the simulation but also its quality, the latter due to the increase in noise. On the other hand, a too small number of water molecules may lead to badly or insufficiently treated interactions between the complex and the solvent.[8]

Only approximately 60% of all MM-GBSA and MM-PBSA calculations were enhanced by the presence of explicit water molecules. Among those analysis in which there was improvement with the inclusion of explicit water molecules, around 58% had $N_{wat,\ best}$=40 or $N_{wat,\ best}$=50, showing that a number of molecules within this range may be a reasonable number of explicit water molecules to be considered. A previous work by Irene *et al.* also refers a comparable number of water molecules.[8] It is also important to notice that, in each case, ΔN_{wat} was set to 10. Better results, in terms of correlation to

experiments, might be obtained by finely tuning N_{wat}, even if the variation in correlation coefficient observed in the case reported here suggests that going beyond a $\Delta N_{wat} = 10$ would probably not severely affect the quality of the results.

The parameters igb=1/mbondi or igb=5/mbondi2 are seldom improved or are improved very slightly by the presence of water molecules (**Figure 3**, **Figure 4**, **Figure 5** and **Figure 6**A). By contrast, when igb=8 is used in conjunction with mbondi2 or mbondi3, an improvement is observed in many analysis with the inclusion of explicit water molecules (**Figure 3**A, **Figure 4**, **Figure 5**A and **Figure 6**), some of them in a remarkable way, for instance when simulations are run on HPC, with ff14SB using igb=8 and mbondi3 (R^2 improves more than 30% for the 12th nanosecond) (**Figure 6**B). MM-PBSA calculations always improved or at least reached comparable results with presence of explicit water molecules. However, for most analysis, correlation coefficients were low even with a large number of water molecules, with the exception of the cases where simulations are run on HPC, with ff14SB for the 12th nanosecond (Figure 6B) (R^2 improves more than 50%). As a consequence, although the best results of MM-PBSA were obtained by using the hybrid solvation model, the improvement is not statistically relevant and is obtained at the expense of an increase in computational time. Concomitantly, there was no sense in using explicit solvent molecules with MM-PBSA for this work.

The results obtained suggest that inclusion of explicit water molecules improves the prediction of free energy of binding of Bcl-Xl ligands to Bcl-Xl here analyzed, since the top three best results were obtained by including explicit water molecules in the simulations (**Table 1**). This is also consistent with other works reported in the literature[8] and is according to the theory that assumes that, since water molecules play an important role in protein-ligand interactions, they should be explicitly considered in computational simulations. However, MD is not a deterministic process, therefore multiple runs should be performed for each experiment in order to assess the true effect of N_{wat}.

6.2 Influence of duration of simulation

Molecular dynamics is the most computationally expensive step in predicting binding free energies, so a careful examination of the impact of MD duration might help to reduce this cost with a limited impact on the quality of predictions.

Except for one case, which will be discussed shortly, the results showed that the increase in simulation length did not affect significantly the quality of MM-PBSA or MM-GBSA results (**Figure 3**, **Figure 4**, **Figure 5** and **Figure 6**A).

It should be noted that, for simulations run on HPC hardware, the combination between the ff14SB force-field (**Figure 6**) provided maximum correlation coefficients in both MM-PBSA and MM-GBSA (this latter conducted using igb=8 and mbondi2 or 3 set of radii), especially when carried on the 12th nanosecond (**Figure 6**B). It should also be stressed, however, that results of analysis conducted on the 12th nanosecond for MM-GBSA calculations (**Figure 6**B) are only marginally better that those conducted on the 4th nanosecond of MD trajectory (**Figure 6**A) (HPC, ff14SB, igb=8/mbondi3, 12th ns: R^2=0.80; HPC, ff14SB, igb=8/mbondi3, 4th ns: R^2 =0,73). In my opinion, although more trials should be carried out before reaching out a valid conclusion, most likely there is no need to push simulation beyond the 4th nanosecond.

6.3 Influence of calculation method chosen (MM-PBSA or MM-GBSA)

In this work, a great difference in results was observed by using MM-PBSA or MM-GBSA as the calculation method. It was observed that MM-PBSA had always worse results than MM-GBSA (in general, $R^2_{MM\text{-}PBSA,\ max} < 0.15$ and $R^2_{MM\text{-}GBSA,\ max} > 0.4$), except when simulations were run with ff14SB force-field on HPC hardware with MM-GBSA calculations made for the 12th nanosecond (**Figure 6**B). There are many reasons which might be behind this difference. Given that PB is less approximated than GB, MM-PBSA calculations might be much more sensitive to system instabilities than MM-GBSA. This might be verified by extending simulation time further, but other approximations might play an important role. Furthermore, these prediction are in accordance with a previous work which indicated that, often, MM-PBSA calculations provided worse predictions, in terms of ranking binding free energies, than MM-GBSA; however, this study also indicated that MM-PBSA might be better in terms of predicting absolute binding energies.[13] Considering that in medicinal chemistry a correct ranking of candidate compounds is probably more desired that computing absolute binding free energies and that MM-PBSA has

a higher cost in terms of computational time, MM-GBSA can be better suited than MM-PBSA in most applications. Moreover, it should be noted that MM-GBSA calculations were tested with different kinds of GB models and different atomic radii, but no technique was used to optimize MM-PBSA calculations and so the simulations made are likely insufficient to conclude that MM-GBSA is better than MM-PBSA to calculate relative free binding energies.

6.4 Influence of the force-field chosen

With exception of when HPC hardware, ff14SB force-field for the 12th nanosecond of simulation was used (**Figure 6**B), there was no statistically relevant difference between results obtained with ff14SB and ff99SB force-fields. However, since the best results (**Table 1**), in terms of correlation between prediction and experiments, were obtained with ff14SB, results suggest that this force-field may be a better choice than ff99SB for these kind of calculations.

6.5 Influence of the set of atomic radii and kind of generalized Born model

The combination between GB-Neck2 (igb=8) and mbondi3 provided the highest correlation coefficient between predictions and experiments in most analysis, with a few exceptions: correlations obtained by GPU, ff99SB for the 12th nanosecond (**Figure 3**B), correlations obtained by HPC, ff99SB for the 8th nanosecond of simulation (data not shown) and correlations obtained by HPC, ff99SB for the 12th nanosecond of simulation (**Figure 5**B). However, these exceptions are not statistically relevant. The combinations igb=1/mbondi and igb=5/mbondi2 often gave origin to similar results, but worse than those obtained with igb=8/mbondi3 (**Figure 3**A, **Figure 4**, **Figure 5**A, **Figure 6**A).

6.6 Influence of calculation performed on GPU or HPC

Statistically, there was no significant difference between analysis performed with GPU from those performed with HPC, except in the case of simulations run with ff14SB force-field and where the 12th ns was analyzed by MM-PBSA (**Figure 6**B); only in this case, a remarkable improvement was observed by switching from GPU (R^2 = 0.15) to HPC (R^2 = 0.6). However, in order to obtain a statistically meaningful comparison between HPC and GPU hardware, the very same simulation should be repeated at least 3 times on both hardware. It should be noted, however, that the differences between the algorithms used in both hardwares are not very significant, so using both hardwares allowed to compare results since generation of the trajectories is not deterministic.

7. Conclusions

With this work, it was possible to conclude that, within the tested protocols, the best protocol to rank free binding energies of Bcl-xL with the following ligands 2YXJ, 3Z1B, 3ZC4, 3ZC5, 3ZK6, 3ZLN, 3ZLO and 3ZLR was obtained by running simulations on HPC, using the ff14SB force-field, using the Born solvation model GB-Neck2 coupled with the set of atomic radii “mbondi3” and performing MM-GBSA calculations on the 4th nanosecond.

It should be noted that the complexes considered are very similar since the receptor is always Bcl-xL, and the only thing that changes is the ligand. Consequently it was possible to obtain quite acceptable maximum correlation coefficients in most of the experiences performed (48 tested analysis out of 62 had a R^2_{max}>0,5). However, this also means that the protocol that presented the best results in this set of complexes might not be the most suitable for a complex with a different receptor or a complex composed by the same receptor, but with a ligand very different from those tested. Nevertheless some of the observations and conclusions reached during this work, may be likely applicable to other systems involving Protein-Protein Interactions.

8. Future prospects

Advances in molecular biology, the interest of many laboratories to do research on Protein-Protein Interactions (PPI’s) as well as many genome-scale analysis of PPIs in several organisms led to an enormous amount of data that can be used to clarify many biochemical pathways, protein functions and diseases associated to deregulation of these pathways.[1]

Furthermore, the use of computer-based approaches to get insight on biological/biochemical interactions increases exponentially every year. Computational techniques are expected to be employed extensively in many areas of science and technology.[14] Despite the advances discussed before, some challenges still remain. Drug design still demands a good understanding of the intermolecular forces involved, protein structure and protein function. Moreover, after a scaffold is developed or discovered, many combinations of functional groups need to be evaluated to obtain a lead compound. To make this optimization possible through a computational approach, both the conformational space of the ligand and the 3D structure of the protein receptor must be known.[1]

Several other problems have to be faced when designing drug-like inhibitors, such as selectivity toward a specific target, oral bioavailability, metabolic stability, immunogenicity and costs associated with production and storage.[15] Although it cannot be the ultimate solution, molecular modeling can aid in facing most of these problems and the work discussed in this thesis aimed to evaluate a method that might be useful in drug development and optimization.

Even if the investigation only regarded a few Bcl-Xl ligands, the discussed protocol might be extended to different cases, yet more tests should be made. However, it should be noted that, due to the non-deterministic nature of MD simulations, MM-PBSA and MM-GBSA calculations should be conducted on replicated trajectories to achieve statistical significance. Other variables, such as a simulation time longer than 12 ns, a different explicit solvent model (such as TIP4Pew) or a different setup in the MD protocol might also be evaluated. By disposing of good predictions, it will be possible to design better ligands for Bcl-Xl (receptor) that will be tested in laboratory and possibly allow better results in treatment of cancer.

9. References

[1] - Rushikesh Sable, Seetharama Jois; Surfing the Protein-Protein Interaction Surface Using Docking Methods: Application to the Design of PPI Inhibitors; *Molecules*, **20**, 11569-11603 (2015); DOI:10.3390/molecules200611569

[2] - Genheden S., Ryde U.; The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities.; *Expert Opin. Drug Discov.*., **10(5)**:449-61 (2015); DOI: 10.1517/17460441.2015.1032936

[3] - Mobley D. L., Klimovich P. V.; Perspective: Alchemical free energy calculations for drug discovery.; *J. Chem. Phys.*, **137(23)**:230901; DOI: 10.1063/1.4769292

[4] - http://ambermd.org (seen in March 2015)

[5] - Irene Maffucci and Alessandro Contini; Tuning the Solvation Term in the MM-PBSA/GBSA Binding Affinity Predictions; *Frontiers in Computational Chemistry,* **1***,* 82-120 (2015)

[6] – Onufriev A. V., Aguilar B.; Accuracy of continuum electrostatic calculations based on three common dielectric boundary definitions.; *J. Theor. Comput. Chem***., 13(3)**. pii: 1440006 (2014)

[7] - Yamagishi J., Okimoto N., Morimoto G., Taiji M.; A new set of atomic radii for accurate estimation of solvation free energy by Poisson-Boltzmann solvent model.; *J. Comput. Chem.*, **35(29)**:2132-9 (2014); DOI: 10.1002/jcc.23728

[8] - Irene Maffucci and Alessandro Contini; Explicit Ligand Hydration Shells Improve the Correlation between MM-PB/GBSA Binding Energies and Experimental Activities; *Journal of Chemical Theory and Computation*, **9 (6)**, 2706–2717 (2013); DOI: 10.1021/ct400045d

[9] - Li S., Bradley P.; Probing the role of interfacial waters in protein-DNA recognition using a hybrid implicit/explicit solvation model.; *Proteins*, **81(8)**:1318-29 (2013); DOI: 10.1002/prot.24272

[10] - Lessene G., Czabotar P. E., Sleebs B. E., Zobel K., Lowes K. N., Adams J. M., Baell J. B., Colman P. M., Deshayes K., Fairbrother W. J., Flygare J. A., Gibbons P., Kersten W. J. ,Kulasegaram S., Moss R. M., Parisot J. P., Smith B. J., Street I. P., Yang H., Huang D. C., Watson K. G.; Structure-guided design of a selective BCL-X(L) inhibitor.; *Nat. Chem. Biol.*, **9(6)**:390-7 (2013); DOI: 10.1038/nchembio.1246

[11] - https://www.chemcomp.com/MOE-Molecular_Operating_Environment.html (seen in June 2015)

[12] - Humphrey W., Dalke A., Schulten K.; VMD: visual molecular dynamics.; *Journal of Molecular Graphics*, **14(1)**:33-8 (1996); DOI: 10.1016/0263-7855(96)00018-5

[13] - Hou T., Wang J., Li Y., Wang W.; Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations.; *J. Chem. Inf. Model.*, **51(1)**:69-82 (2010); DOI: 10.1021/ci100275a

[14] - Yamashita T., Ueda A., Mitsui T., Tomonaga A., Matsumoto S., Kodama T., Fujitani H.; The feasibility of an efficient drug design method with high-performance computers.; *Chem. Pharm. Bull (Tokyo)*, **63(3)**:147-55 (2015); DOI: 10.1248/cpb.c14-00596

[15] - Villoutreix B. O., Kuenemann M. A., Poyet J. L., Bruzzoni-Giovanelli H., Labbé C., Lagorce D., Sperandio O., Miteva M. A..; Drug-Like Protein-Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology; *Mol Inform*.; **33(6-7)**, 414-437 (2014); DOI: 10.1002/minf.201400040