Fast and Accurate Atomistic Calculation of Free Energies for Biomolecular Modeling and Simulation

1. Introduction

Atomistic modeling and simulation methods enable a modern molecular approach to biomedical research. The ability of these methods to address biologically relevant problems is largely determined by their accurate treatment of electrostatic interactions in the target biomolecular structure surrounded by water. An accurate yet efficient determination of the electrostatics in protein-ligand interactions is of profound importance in molecular design and drug discovery. Several cellular processes, such as signal transduction, gene expression, and protein synthesis, are controlled by the binding of biomolecules. In pharmaceutical applications and structure-based drug discovery, it is necessary to have accurate and fast prediction of the binding free energy of drugs to biomolecular targets, e.g. proteins and nucleic acids (Shirts, 2010; Lau, 2011). The computational prediction of binding free energies is however complex and challenging, and its outcomes can depend strongly on the molecular modeling technique, particularly the underlying solvent model (Mobley D. L., 2009; Mobley D. L., 2008; Shivakumar, 2009).

The implicit solvent model -- which treats solvent as a continuum with the dielectric and non-polar properties of water -- offers a good balance between accuracy and speed. Within the implicit solvation framework, the Generalized Born (GB) and the Poisson Equation (PE) are widely used (Gilson, 1995; Scarsi, 1997; Onufriev, 2000). Due to its relative simplicity and efficiency, the GB model is almost exclusively used in molecular dynamics simulations where it has shown impressive success in a variety of areas, from protein folding to molecular docking. However, persistent inaccuracies in conformational ensembles generated by the method in molecular dynamics simulations can ultimately lead to erroneous conclusions. Some of the accuracy problems are unique to the GB model, but some are already present at the more fundamental continuum electrostatics level. Specifically, the PE and even more sophisticated continuum models still fail to reach chemical accuracy (better than 1 kcal/mol) simultaneously for small drug-like molecules and amino-acids – a prerequisite for chemical accuracy of atomistic modeling in general. Moreover, despite the huge strides in this direction, implicit solvent models are not still fast enough when it comes to practical Computer Aided Drug Design (CADD).

Our goal here is to enable some of the practical atomistic computation at a higher level of accuracy than practical implicit solvent models can currently deliver, while keeping, or even exceeding, their computational efficiency. Most promising models will be implemented as open source code, including AMBER molecular dynamics tools (freely available at http://ambermd.org/). We introduce three major Aims in this paper specified in Section 2, followed by the precise description of Aim 1. It is worth mentioning that Aim 1 is almost achieved (Forouzesh, 2017), while Aim 2 and Aim 3 are still in progress.

2. Specific Aims

The purpose of the specific aims is to describe concisely and realistically the goals of the proposed research and summarize the existing results and expected outcomes, including the impact of the proposed research will exert on the research fields involved.

Aim 1: Analysis and Optimization of Grid-Based Surface Generalized Born (GB) Model for Calculating Electrostatic Binding Free Energies. Accurate and efficient computation of solvation free energies (ΔG_{pol}) is central to calculating binding free energies ($\Delta\Delta G_{pol}$) and the associated applications. In this study (Forouzesh, 2017), we described in detail an implementation of grid-based molecular surface GB model for computing the (ΔG_{pol}), named GBNSR6. Within this model, the effective Born radii are estimated using a physics-inspired method called "field-view", re-purposed here for computing r^{-6} integrals over the molecular surface. The parameters that enter the numerical integration formulae were originally optimized for $\frac{1}{r^2}$, and may not be optimal for $\frac{1}{R^6}$. Thus, we will optimize these parameters particularly for r^{-6} integrals over the molecular surface, and speed-up the effective Born radii calculation.

Aim 2: Optimal Dielectric Boundary for Practical Continuum Solvent Calculations. We will develop a novel approach to constructing optimal dielectric boundary, central to many continuum solvation models, including the GB model. The new boundary will be optimized for calculating protein-ligand binding free energies. We will use new optimization techniques for a virtually exhaustive search for best agreement with the explicit solvent. The approach differs from past efforts in several critical aspects. (1) Physical realism will be enforced, and transferability enhanced, by constraining the DB by experimentally observable atom-(water oxygen) Radial

Distribution Functions (RDF) (see Figure 1). (2) The optimal parameters of the DB will be well-suited for protein-ligand binding calculations through inclusion of protein-ligand energies and the corresponding molecular structures into the training set. Multidimensional parameter (atomic radii, probe radius) at this scale was all but impossible in the past, but will come within reach through the proposed global optimization approach.

Aim 3: Next Level Accuracy in a Practical Generalized Born (GB) Model for Molecular Dynamics (MD) Simulations. We will develop a practical "accuracy limit" Generalized Born model -- more accurate than the popular GB models currently implemented in major modeling packages -- yet as computationally efficient as the fastest analytical GB models most often used today. In contrast to previous GB models used in MD, the new model will be based on an "R6" approach (Izadi, 2015) that closely mimics Poisson solutions with optimal dielectric boundary. By providing a close agreement with the Poisson reference, the new model will address several well-known deficiencies of the latest generation of fast and practical GB models (Anandakrishnan, 2015). We will construct and investigate in the context of MD simulations, a *surface- based* GB model. All GB models available for practical MD simulations today are *volume-based*. The difference is important: for biological macromolecules, surface to volume ratio is low, the fact that we will exploit to make the model faster than leading volume-based ones, without compromising accuracy (see Figure 2). Also, to the best of our knowledge, the theoretically well-grounded "R6" GB has never been adapted for MD.

3. Aim 1: Analysis of Grid-Based Surface Generalized Born (GB) Model for Calculating Electrostatic Binding Free Energies

Atomistic simulation is one of the most widely used theoretical tools in biomedical research. High accuracy of solvent representation in these simulations is paramount for biological applications, ranging from structure-based drug design to protein structure prediction and refinement. Arguably the most accurate among classical models of solvation is the one in which individual water molecules are treated explicitly on the same footing with the target biomolecule. Yet, accuracy of this explicit solvent representation comes at extremely high price, computationally. For example, a recent study (Lindorff-Larsen, 2011) of folding-unfolding transitions of 12 of the fastest folding proteins required extremely long simulations on one-of-a-kind specialized supercomputer. Efficient estimates of binding free energies, particularly important in many areas including structure-based drug design, can take hours or even days per structure, or may not even be possible due to convergence issues.

In this section, we briefly describe a Cartesian grid-based molecular surface numerical algorithm GBNSR6 for calculating the effective Born radii (and the solvation free energy), and present the results of different tests in the context of protein−ligand binding. The algorithm adapts to the R6 integration a "field-view" surface integration method (Cai, 2011) previously implemented in the free and open source AmberTools package. More details are discussed in (Forouzesh, 2017).

Model. The polar component of the solvation free energy ΔG_{pol} is calculated using GBNSR6 by the following formula:

$$\Delta G_{pol} \approx \frac{-1}{2}\left(\frac{1}{\varepsilon_{in}}-\frac{1}{\varepsilon_{out}}\right)\frac{1}{1+\beta\alpha}\sum_{ij} q_i q_j\left(\frac{1}{f_{ij}^{GB}}+\frac{\alpha\beta}{A}\right)$$

where ε_{in} and ε_{out} are the dielectric constants of the solute and solvent, respectively, $\beta = \frac{\varepsilon_{in}}{\varepsilon_{out}}$ and $\alpha = 0.571412$. A is the electrostatic size of the molecule, which can be computed analytically, q_i is the partial charge of atom i. The most common functional form of $f_{ij}^{GB} = [r_{ij}^2 + R_i R_j \exp(-\frac{r_{ij}^2}{R_i R_j})]^{1/2}$ is employed where R_i is the so-called effective Born radius of atom i, and r_{ij} is the distance between atoms i and j. The dielectric constants are set to $\varepsilon_{in} = 1$ and $\varepsilon_{out} = 80$. In this work, the effective Born radii R are calculated by the following "R^6" equation:

$$R_i^{-3} = (\frac{-1}{4\pi} \oint_{\partial V} \frac{\boldsymbol{r} - \boldsymbol{r_i}}{|\boldsymbol{r} - \boldsymbol{r_i}|^6} . d\boldsymbol{s})$$

where ∂V represents the molecular surface, $d\boldsymbol{s}$ is the infinitesimal surface element vector, $\boldsymbol{r_i}$ is the position of atom i, and $\boldsymbol{r}$ indicates the position of the infinitesimal surface element. In this work, the spherical surface of a given molecule is approximated by the corresponding orthogonal grid surface (see Figure 3).

Datasets. The proposed model is tested on two data sets: first, a set of 15 small protein−ligand complexes, which are used to test the accuracy (dataset1). Second, a set of molecules of widely different sizes, which are used to test the speed (dataset 2).

Accuracy. The method's performance is investigated in terms of its accuracy against a commonly used reference, as well as the sensitivity to the grid parameters (see Figure 4).

Speed. The elapsed time for computing ΔG_{pol} on dataset2 is measured, shown in Figure 5. This test was done on a commodity PC.

Appendices:

Appendix 1: Figures

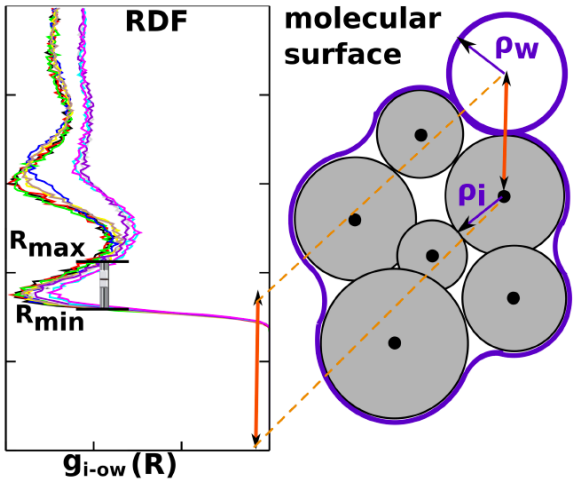

Figure 1. Optimal dielectric boundary (DB) based on radial distribution function (RDF) as physical constraints. The probe radius (ρ_w) and intrinsic atomic radii (ρ_i) are optimized simoltaneoously, under the physically justified constraint that $\rho_w + \rho_i$ is bound within the first peal of the RDF; that is between R_{min}n and R_{max}. The RDFs are computed in explicit solvent for a diverse set of biomolecular structures.

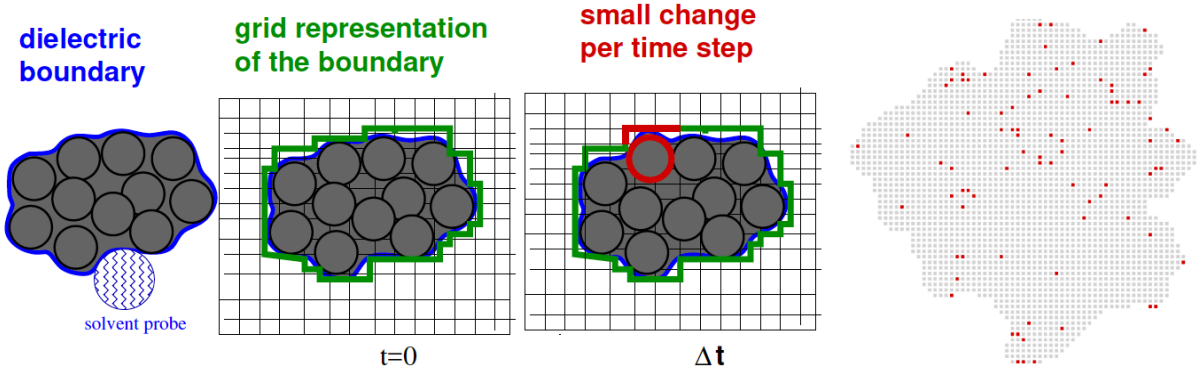

Figure 2. GBNSR6-MD: Adaption of the accurate grid-based GBNSR6 model for MD simulations. Little change in the dielectric boundary (DB) per time-step drastically reduces the updates needed. Rightmost panel: The molecular surface generated for a 726 atom protein using the field-view method. Black squares: grid elements at $t = 0$. Red squares: grid elements ($\sim 5\%$) that have changed after $\Delta t = 2$ fs.

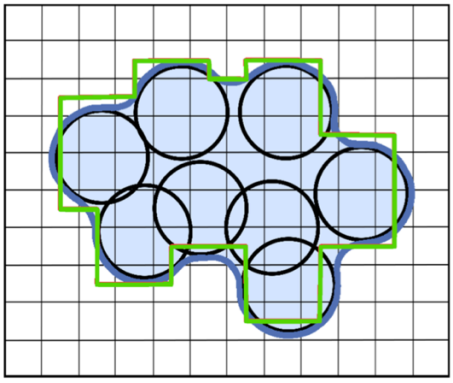

Figure 3. Finite-difference discretization of an abstract molecule. The blue spherical surface shows the ideal Solvent Exclude Surface (SES) that represents the dielectric boundary, and green lines are the approximating square surface elements. The background black mesh is the uniform Cartesian grid.

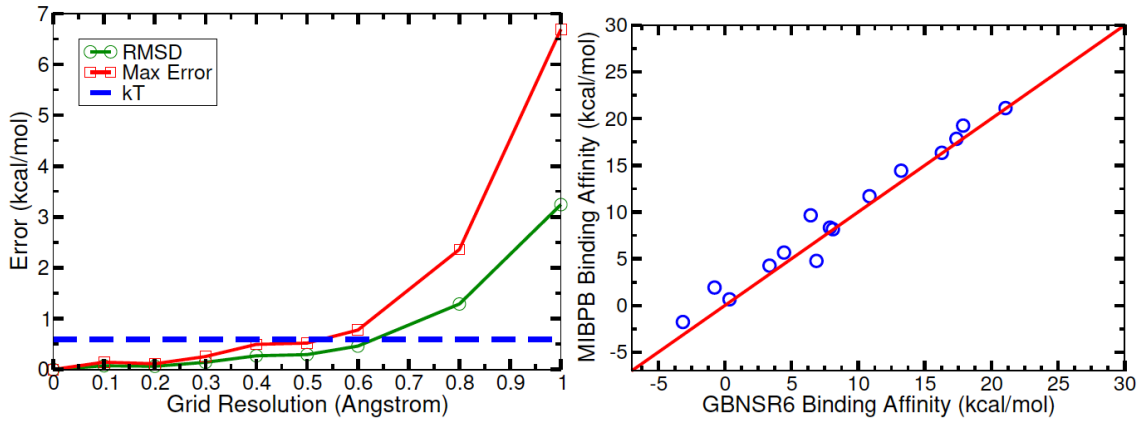

Figure 4. Left: Error in computing $\Delta\Delta G_{pol}$ due to finite grid resolution h relative to $h = 0$ Å extrapolation on dataset1. Regarding the maximum acceptable error (k_bT= 0.59 kcal/mol) in many applications, $h = 0.5$ Å turned out to be the coarsest grid resolution which does not violate the threshold. Right: Correlation between $\Delta\Delta G_{pol}$ computed by GBNSR6 and the numerical PB reference MIBPB ($h = 0.5$ Å). Red line $x = y$ indicates a perfect match. $r^2 = 0.97$.

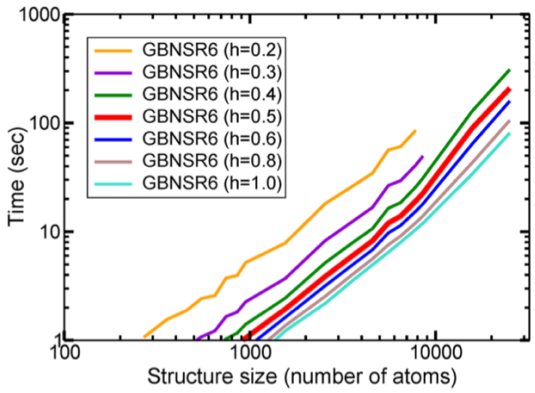

Figure 5. GBNSR6 CPU time for computing ΔG_{pol} on dataset2. One can observe that by setting the grid spacing to $h = 0.5$ Å, the computational cost is relatively low, while the error in calculating $\Delta\Delta G_{pol}$ is is still less than 0.59 kcal/mol (see Figure 4).

Appendix 2: Bibliography

Anandakrishnan, R. D. (2015). Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophysical journal, 108*(5), 1153-1164.

Cai, Q. Y. (2011). On-the-fly numerical surface integration for finite-difference Poisson–Boltzmann methods. *Journal of chemical theory and computation, 7*(11), 3608-3619.

Forouzesh, N. I. (2017). Grid-Based Surface Generalized Born Model for Calculation of Electrostatic Binding Free Energies. *Journal of chemical information and modeling, 57*(10), 2505-2513.

Gilson, M. K. (1995). Theory of electrostatic interactions in macromolecules. *Current opinion in structural biology, 5*(2), 216-223.

Izadi, S. A. (2015). Protein–Ligand Electrostatic Binding Free Energies from Explicit and Implicit Solvation. *Journal of chemical theory and computation, 11*(9), 4450-4459.

Lau, A. Y. (2011). The hidden energetics of ligand binding and activation in a glutamate receptor. *Nature structural & molecular biology, 18*(3), 283-287.

Lindorff-Larsen, K. P. (2011). How fast-folding proteins fold. *Science, 334*(6055), 517-520.

Mobley, D. L. (2008). Treating entropy and conformational changes in implicit solvent simulations of small molecules. *The Journal of Physical Chemistry B, 112*(3), 938-946.

Mobley, D. L. (2009). Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. *Journal of chemical theory and computation, 5*(2), 350-358.

Onufriev, A. B. (2000). Modification of the generalized Born model suitable for macromolecules. *The Journal of Physical Chemistry B, 104*(15), 3712-3720.

Scarsi, M. A. (1997). Continuum electrostatic energies of macromolecules in aqueous solutions. *The Journal of Physical Chemistry A, 101*(43), 8098-8106.

Shirts, M. R. (2010). *Free energy calculations in structure-based drug design.* Cambridge: Cambridge University Press.

Shivakumar, D. D. (2009). Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. *Journal of chemical theory and computation, 5*(4), 919-930.