Thesis submitted for transfer from M.Phil to PhD

The Use of Free-Energy Simulations as Scoring Functions

Julien MICHEL
Department of Chemistry
University of Southampton
September 2003
Supervisor : Dr. J. W. Essex
Advisor : Dr. G. Attard

University of Southampton
FACULTY OF SCIENCE
CHEMISTRY

THE USE OF FREE-ENERGY SIMULATIONS AS SCORING FUNCTIONS

by Julien Michel

ABSTRACT

Solvation models based on the Generalised Born/Surface Area (GB/SA) theory [1] have been parameterised to reliably reproduce solvation energy of small organics to within $1\,kcal.mol^{-1}$. These models have then been used to compute Potentials of Mean Force (PMF) for the association of various molecules. Results indicates that one of the models should be sufficiently accurate to be used in protein simulations. Improvements have been sought by testing the Surface Area Surface Integral (SASI) theory [2] . It is shown that the SASI method is an improvement over pure Surface Area based methods for the modelling of non polar solvation.

STATEMENT OF AIMS

The aim of this work is to develop a set of algorithms that are fast enough and yet rigorous enough to be successfully applied to the prediction of the binding energy of putative drugs to a specified protein. A successful method will bridge the gap between expensive free-energy simulations and empirical scoring functions that are currently used to predict binding energy.

ABBREVIATIONS

1K	one thousand
1M	one million
AA	All Atom
CM	Contact Minimum
FEP	Free Energy Perturbation
GA	Genetic Algorithm
GB	Generalised Born
LJ	Lennard-Jones
MC	Monte Carlo
MD	Molecular Dynamics
PDA	Pairwise Descreening Approximation
PB	Poisson-Boltzmann
PMF	Potential of Mean Force
SA	Surface Area
SASA	Solvent Accessible Surface Area
SASI	Surface Area Surface Integral
SSM	Solvent Separated Minimum
UA	United Atom
vdW	van der Waals

Contents

Chapter 1
Introduction

1.1 Computer aided Drug design

Most drugs act by perturbing or modifying the function of a biomolecule such as a protein or a membrane. Over the last two decades, thanks to a regular increase in computer power, computer simulations have allowed the study of these effects at the atomic level. Methods that can predict the affinity of a small organic for a given biomolecule have drawn considerable interest from academics and pharmaceutical companies for their potential usefulness in designing new drugs [3,4] .

The process by which the drug (ligand) interacts with its target (receptor) is based on known principles of thermodynamics and quantum mechanics. Unfortunately, it is also well understood that a rigorous application of these principles to computer models is too time consuming. It is therefore necessary to use simplified models, that do not model exactly what happens in reality, but can nonetheless provide useful insights.

Evaluating the affinity of a drug/target couple can be thought as a two steps process. First, the *binding mode*, that is, the structural arrangement of the drug/target complex has to be determined. This is often difficult as typical biomolecules have huge numbers of degrees of freedom and is referred as the *docking* problem [5] .

Once the binding mode is known, it is necessary to predict how tightly bound the

drug/target complex is. This is referred as the *scoring* problem. The strength of the interaction is directly related to the *free energy* of the complex. Thus, computer simulations that predict free energies can be used to search for promising drug candidates.

1.2 Predicting free energies of binding

It is unfortunately, very difficult to use free energy calculations routinely. This is due to two distinct problems. First, to derive free energies, it is necessary to *sample* the ensemble of possible drug/complex arrangements. The number of possible conformations is usually very large and long simulations are required to obtain reliable results. Second, the systems modelled must be realistic enough to yield meaningful results. This means that a significant portion of the often large receptor has to be modelled, along with solvent molecules surrounding the complex. This makes evaluation of the energy of the system very expensive.

These problems prevents free energy calculations from being routinely applied to large libraries of compounds. Instead, *scoring functions* are often used. Traditional scoring functions consist in a set of easily computed molecular properties that are related to free energies of binding by empirical coefficients [6] . Scoring functions usually do not work well when applied to systems that differs from those that were used to derive their parameters. They are also seldom very accurate and are mainly used to pick a few promising candidates from a large list of compounds.

In this project, protocols based on free energy simulations will be developed to predict free energies of binding.

Because traditional free energy simulations are too time consuming, a number of approximations will have to be introduced. For example, atomistic solvent models will be replaced by faster, implicit solvent models [1,7] based on laws of classical electrostatics. Implicit solvent models reduce the complexity of the system studied, and also yield increased sampling as only solute atoms have to be considered.

It is also expected that a fraction of the receptor will be modelled as a grid potential. This will further reduce the complexity of the system and allow for a fast evaluation of the total energy of the system.

Methods developed to increase sampling such as Parallel Tempering [8–10] are also likely to be considered.

It is also expected that the developed protocol will benefit a previously devised docking algorithm [11] .

Chapter 2
Methods

2.1 Statistical Mechanics

The validity of any computational model results from the successful reproduction and prediction of experimental observables, typically thermodynamic and structural properties. Statistical mechanics is a branch of physics which relates data generated at the microscopic level (atomic positions,velocities..) to macroscopic properties (temperature, pressure).

The instantaneous mechanical state of an N particle chemical system, characterised by their positions in space $\mathbf{r}$ and their momenta $\mathbf{p}$, belongs to a $6N$ dimensional space called *phase space*. It is possible in theory to determine a macroscopic property A_{obs} by taking the time average of its instantaneous value A_{inst}, as determined by the state of the system in phase space, over a time period.

$$A_{obs} = \langle A_{inst} \rangle_{time} = \lim_{t \to \infty} \frac{1}{t} \int_{t=0}^{t_{obs}} A_{inst}(\mathbf{p}^N(t), \mathbf{r}^N(t)) dt \tag{2.1}$$

Observable are collected from systems containing a large number of atoms (order of 10^{23}) and simulating such systems is not feasible. Instead, a single system evolving in time is replaced by a large number of replicas of the system that are considered simultaneously.

The collection of states of a system is denoted an *ensemble* and the time average of A is replaced by an ensemble average :

$$\langle A \rangle_{ensemble} = \int \int d\boldsymbol{p}^N d\boldsymbol{r}^N A(\boldsymbol{p}^N, \boldsymbol{r}^N) \rho(\boldsymbol{p}^N, \boldsymbol{r}^N) \tag{2.2}$$

where $\rho(\boldsymbol{p}^N, \boldsymbol{r}^N)$ is the probability density of the ensemble.

The probability density of a state i of energy E_i, is the ratio between the number of particles in that state, n_i, and the total number of particles N in the system. In the canonical ensemble (NVT), that is an ensemble where the number of particles, the volume and the temperature of the system remain constant, this ratio is equal to:

$$\rho_{NVT}\left(\boldsymbol{p}^N, \boldsymbol{r}^N\right) = \frac{n_i}{N} = \frac{\exp(-E_i/k_B T)}{Q_{NVT}} \tag{2.3}$$

where E_i is the energy of state i, k_B the Boltzmann constant, T the temperature, and Q_{NVT} the so called canonical partition function. The classical partition function in the canonical ensemble is defined as:

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \int \int d\boldsymbol{p}^N d\boldsymbol{r}^N \exp(-E(\boldsymbol{p}^N, \boldsymbol{r}^N)/k_B T) \tag{2.4}$$

where $E(\boldsymbol{p}^N, \boldsymbol{r}^N)$ is E_i. The factor $N!$ arises from the indistinguishability of the particles and $1/h^{3N}$ is required to ensure that the partition function is equal to the quantum mechanical result for a particle in a box.

2.2 Metropolis Monte Carlo

Ideally we would want to be able to compute the partition function, but this is unfeasible for any system of interest because of the number of states to consider. It is still possible, however, to generate meaningful results without visiting all phase space.

Monte Carlo methods generate new positions in phase space for a system by randomly modifying parts of that system and accepting these changes with a certain probability. For a molecular system, a huge fraction of the phase space corresponds to states with an extremely small Boltzmann factor and thus do not contribute to the partition function. A mean to generate states that lie in the regions of phase space that make important contributions to the partition function is required. Metropolis [12] proposed such a method.

Simply put, new states are generated and weighted with a probability proportional to their Boltzmann factor.

$$prob(A \rightarrow B) = \exp(\frac{-(V(\mathbf{r}_B^N) - V(\mathbf{r}_A^N))}{k_\beta T}) \tag{2.5}$$

When Metropolis MC is used, the probability that state B is accepted is determined by equation 2.5 . The move is then accepted if a randomly generated number lying between [0,1] is lower than $prob(A \rightarrow B)$. If state B has a lower energy than state A then the exponential will be greater than one and the move will always be accepted.

2.3 Free Energy calculations

It can be shown that Gibbs and Helmholtz free energies are related to the partition function. In the micro-canonical ensemble, the expression for Gibbs is :

$$G = -k_\beta T \ln Q \tag{2.6}$$

Equation 2.6 shows that the calculation of the absolute free energy of a system requires the calculation of its partition function. However, the partition function is a 3N dimensional integral where N is the number of particles in the system and in practice, this integral is too difficult to evaluate for systems of interest in biomolecular simulations [13] .

Fortunately, if one is interested in the free energy *difference* between two states A and B, eq 2.7 can be used.

$$G_B - G_A = \Delta G = -RT \ln < \exp \frac{-\Delta H}{RT} >_A \tag{2.7}$$

where $\Delta H = H_B - H_A$ and $<>_A$ refers to an ensemble average over a system represented by Hamiltonian H_A [14] .

If state A and state B do not overlap well in phase space, then the value of the free energy difference using 2.7 will not be accurate, because the phase space of B is not correctly sampled while simulation A is being done.

One method to avoid this problem is to use the *Free Energy Perturbation method*.

It is possible to write $H(\lambda) = \lambda H_B + (1 - \lambda) H_A$ where λ varies from 0 ($H(0) = H_A$) to 1 ($H(1) = H_B$).

One can then connect state A and B by a set of more similar states and eq 2.7 can be rewritten as a sum of energy differences.

$$G_B - G_A = \Delta G = \sum_{\lambda=0}^{1} -RT \ln < \exp \frac{-\Delta H'}{RT} >_\lambda \tag{2.8}$$

where $\Delta H' = H_{\lambda+d\lambda} - H_\lambda$.

Another commonly used method is *thermodynamic integration*, which requires ensemble averages of the derivative of the Hamiltonian with respect to λ to be evaluated at different values of λ.

$$\Delta G = \int_{\lambda=0}^{\lambda=1} [\frac{\partial H}{\partial \lambda}]_\lambda d\lambda \tag{2.9}$$

2.4 Potential Energy function for Atomic Models

Statistical mechanics permits us to compute thermodynamic properties such as free energies. When we use a Monte Carlo method to generate statistics for the appropriate ensemble we need to know the potential energy U of the system.

When using a molecular mechanics force field, the potential energy of a system is defined by its coordinates. With molecular dynamics it is also necessary to compute forces. Force fields only typically consider the nuclei positions; electron motion is often ignored, and usually bond lengths, valence angles, torsions, and non-bonded interactions are included. Several force fields exist (AMBER,[15] OPLS,[16] and CHARMM[17] for example), using the same components and differing through parameterisation or definition of these components. Force field can be united atom (UA) or all atom (AA). In united atom force fields, the hydrogen atoms bonded to carbon atoms are not modelled explicitly, rather the mass of the hydrogen atom is added to the carbon atom to form a "united atom". The non-bonded parameters associated with the carbon atom are also modified to take this change into account. The functional form of the total potential energy, U_{total}, in the AMBER force field is as follows:

$$U_{total} = U_{bond} + U_{angle} + U_{dihedral} + U_{non-bonded}. \tag{2.10}$$

The bond and angle contributions are described by harmonic potentials :

$$U_{bond} = \sum_{bonds} K_b(r - r_{eq})^2 \tag{2.11}$$

$$U_{angle} = \sum_{angles} K_\theta(\theta - \theta_{eq})^2 \tag{2.12}$$

where r corresponds to the bond length, θ to the valence angle, and r_{eq} and θ_{eq} to the associated equilibrium values. K_b and K_θ are force constants. The torsional term $U_{dihedral}$ is computed as:

$$U_{dihedral} = \sum_{dihedrals} A_n(1 + \cos(n\phi - \delta)), \tag{2.13}$$

where ϕ is the dihedral angle, n is the multiplicity (which gives the number of minimum points in the function as the torsion angle changes from 0 to 2π), δ is the phase angle and A_n is the force constant. Finally, the non-bonded energy is composed of an electrostatic and a Lennard-Jones term:

$$U_{non-bonded} = \sum_i \sum_{j>i} \left\{ \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\}, \tag{2.14}$$

where the sum is over all atom pairs i, j. The q_i are the partial atomic charges, ϵ_{ij} and σ_{ij} are the Lennard-Jones well-depth energy and collision-diameter parameters, and r_{ij} is the inter-atomic distance.

2.5 Implicit modelling of solvation

Drug/receptor interactions occurs in an aqueous phase. Therefore, realistic computer simulations have to consider solvent to model this process properly. Traditional simulations must include thousands of water molecules to solvate properly a protein. Complicated methods such as periodic boundary conditions or Ewald summation are also needed. Thus, it is not uncommon that during the simulation, most of the time is spent computing non bonded interactions for solvent molecules, which are usually not the prime interest of the simulation.

A serious alternative to these explicit solvent simulation is to consider the solvent as a high-dielectric continuum interacting with charges that are embedded in solute molecules

of lower dielectric. The solute response to the reaction field of the solvent dielectric can then be modelled by applying laws of classical electrostatics.

Advantages of an implicit over an explicit solvent are two-fold.

1. Thousands of solvent molecules do not have to be modelled explicitly, reducing the complexity of the system and the CPU cost.

2. By definition, every solute move is always at equilibrium with the solvent.

The Poisson Boltzmann (PB) equation is one of the most accurate ways to model these electrostatic interactions [7] . Analytical solutions of the PB equation for solutes of arbitrary shape are not available and are usually obtained by finite-difference or boundary-element numerical methods. Solving the PB equation is quite expensive for large molecules and other more efficient and approximate methods have been proposed.

One of these methods is the generalised Born (GB) approach which we have adopted in our work and is presented in the next section.

2.5.1 Electrostatics : The Generalised Born method

The Born Model

Born has shown [18] that an analytical equation for the electrostatic energy of an isolated ion can be derived from classical electrostatic theory.

Classical electrostatic theory states [19] that the total electrostatic energy in a dielectric medium is defined as :

$$G = \frac{1}{8\pi} \int_V \vec{E}.\vec{D}.dV \tag{2.15}$$

$$\vec{D} = \epsilon.\vec{E} \tag{2.16}$$

$\vec{E}$ and $\vec{D}$ are the electric field and electric displacement, ϵ is the dielectric constant of the medium and dV a volume element. $\vec{E}$ can be obtained from Gauss Law :

$$\int_S \vec{E}.d\vec{S} = \frac{Q}{\epsilon} \tag{2.17}$$

The surface integral on the left is the area integral over any closed surface. Q is the total charge that lies within the space delimited by $\vec{S}$.

For a uniformly charged spherical shell of radius α and interior dielectric ϵ_{vac} inside and outside, one can obtain :

$$\vec{E}_{int} = 0 \qquad \vec{D}_{int} = 0 \qquad r < \alpha \tag{2.18}$$

$$\vec{E}_{out} = \frac{q}{\epsilon_{vac} r^3}.\vec{r} \qquad \vec{D}_{out} = \frac{q}{r^3}.\vec{r} \qquad r > \alpha \tag{2.19}$$

With q the total charge of the sphere.

The total electrostatic energy of the system is :

$$G_{vac} = \frac{1}{8\pi}\int_V \vec{E}.\vec{D}.dV \tag{2.20}$$

$$= \frac{1}{8\pi}[\int_{in} \vec{E}_{in}.\vec{D}_{in}dV + \int_{out} \vec{E}_{out}.\vec{D}_{out}dV] \tag{2.21}$$

$$= \frac{1}{8\pi\epsilon_{vac}}\int_{out} \frac{q^2}{r^4}dV \tag{2.22}$$

By integrating 2.22 from α to ∞ we find :

$$G_{vac} = \frac{q^2}{2\epsilon_{vac}\alpha} \tag{2.23}$$

If the same spherical system is now considered in a dielectric medium with an interior dielectric of ϵ_i and and exterior dielectric constant of ϵ_{solv}, the total electrostatic energy can similarly shown to be :

$$G_{solv} = \frac{q^2}{2\epsilon_{solv}\alpha} \tag{2.24}$$

The electrostatic energy to transfer a spherical charged ion of radius α from a medium of dielectric ϵ_{vac} to another of dielectric ϵ_{solv} is the difference between 2.24 and 2.23. This is the Born Equation [18]

$$\Delta G_{born} = (\frac{1}{\epsilon_{solv}} - \frac{1}{\epsilon_{vac}})\frac{q^2}{\alpha} \tag{2.25}$$

Generalised Born

The Born model of solvation can be generalised to a molecule of arbitrary shape by treating each atom as a sphere of radius α_i , a charge q_i and interior dielectric ϵ_i.

If we assume initially that each sphere is separated by a distance large enough so that they appear as point charges to other spheres (i.e single dielectric medium), then the total electrostatic energy of the system is the sum of the Coulombic interaction and the Born solvation energy.

$$G_{tot} = \frac{1}{2}\sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon_{solv} r_{ij}} - \frac{1}{2}(\frac{1}{\epsilon_{vac}} - \frac{1}{\epsilon_{solv}}) \sum_i \frac{q_i^2}{\alpha_i} \tag{2.26}$$

Unfortunately, equation 2.26 is not valid for molecular systems where the radius α_i and the distance r_{iJ} are usually too close for the former to be negligible.

Still [1] has shown that by splitting the Coulombic interaction in two terms, one can write equation 2.27

$$G_{tot} = \frac{1}{2}\sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon_{vac} r_{ij}} - \frac{1}{2}(\frac{1}{\epsilon_{vac}} - \frac{1}{\epsilon_{solv}}) \sum_i \sum_{j \neq i} \frac{q_i q_j}{r_{ij}} - \frac{1}{2}(\frac{1}{\epsilon_{vac}} - \frac{1}{\epsilon_{solv}}) \sum_i \frac{q_i^2}{\alpha_i} \tag{2.27}$$

Terms 2 and 3 can then be recombined in a single formula :

$$\Delta G_{tot} = \frac{1}{2}\sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon_{vac} r_{ij}} + \Delta G_{Genborn} \tag{2.28}$$

Where $\Delta G_{Genborn}$ is :

$$\Delta G_{Genborn} = -\frac{1}{2}(\frac{1}{\epsilon_{vac}} - \frac{1}{\epsilon_{solv}}) \sum_i \sum_j \frac{q_i q_j}{\sqrt{r_{ij}^2 + B_i B_j e^{\frac{-r_{ij}^2}{4 B_i B_j}}}} \tag{2.29}$$

Where B_i and B_j are Born radii.

This expression reduce to the Born equation for the case of a single spherical ion and gives the Coulomb energy as $r_{ij} \to \infty$. It has been shown to be particularly adequate to model solvation.

It is important to realise that equation 2.29 has no physical basis. It results effectively from an interpolation between different theoretical results : the Born equation, the Onsager dipole energy equation and the Coulomb equation for widely separated charges.

Much of the difficulty with 2.29 consist in computing the Born radii B_i. The Born radius of one atom is affected by neighbouring atoms and is no longer equal to α_i. In the Generalised Born formalism, B_i is defined as, the radius that would give the actual electrostatic energy of the molecule-dielectric system by the Born equation if all other atoms of the system were uncharged (only displace the dielectric). This corresponds to defining a spherically averaged dielectric boundary for atom i (the angular dependence is not taken into account). The evaluation of the integral itself is not straightforward as it depends on the position of all other atoms of the solute with respect to the solvent/solute boundary. B_i can be derived by the Poisson equation but this nullifies the advantage of the model.

In the original paper from Still [1] , the Born radii are computed using a numerical method which can be summarised as :

1. Consider a shell of thickness T_k surrounding the van der Waals surface of atom k.
2. Weight the interior radius of this shell using the ratio of solvent accessible surface area A_k to the actual surface area.
3. Repeat the weight for the exterior radius and calculate the difference between weighted interior and exterior radii.
4. Sum the difference between weighted interior and exterior radii for a series of concentric shells up to shell M which encompasses the whole of the van der Waals surface of the molecule.
5. For shell M no weight is applied and the radius is simply added to the previous summation term, to obtain an effective Born radius, which is then used in equation 2.29 .

The method is illustrated by figure 2.1. A formal description of this algorithm is given by equation 2.30 .

$$\frac{1}{B_i} = \sum_{k=1}^{M} \frac{A_k}{4\pi r_k^2} \left[\left(\frac{1}{r_k - 0.5T_k} \right) - \left(\frac{1}{r_k + 0.5T_k} \right) \right] + \frac{1}{r_{M+1} - 0.5T_{M+1}} \tag{2.30}$$

Because this method is time consuming, analytical approaches that are inexact have also been developed [20,21] . It is hoped that most of the errors arising from these inexact

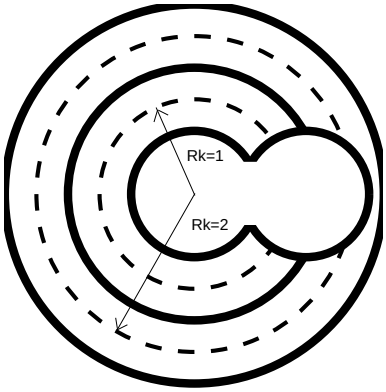

Figure 2.1: Generation of consecutive shells by equation 2.30

solutions are systematic and can be corrected by empirical terms. In our work we have used the Pairwise Descreening Approximation developed by Hawkins et al [20] to compute Born radii (details not shown).

$$B_i^{-1} = \alpha_i^{-1} - \frac{1}{2}\sum_{j\neq i}[\frac{1}{L_{ij}} - \frac{1}{U_{ij}} + \frac{R_{ij}}{4}(\frac{1}{U_{ij}^2} - \frac{1}{L_{ij}^2}) + \frac{1}{2R_{ij}}ln\frac{L_{ij}}{U_{ij}} + \frac{S_{ij}^2\alpha_j^2}{4R_{ij}}(\frac{1}{L_{ij}^2} - \frac{1}{U_{ij}^2})] \quad (2.31)$$

$$L_{ij} = 1 \text{ if } R_{ij} + S_{ij}\alpha_j \leq \alpha_i$$

$$L_{ij} = \alpha_i \text{ if } R_{ij} \text{ - } S_{ij}\alpha_j \leq \alpha_i < R_{ij} + S_{ij}\alpha_j$$

$$L_{ij} = R_{ij} - \alpha_j \text{ if } \alpha_i \leq R_{ij} \text{ - } S_{ij}\alpha_j$$

$$U_{ij} = 1 \text{ if } R_{ij} + S_{ij}\alpha_j \leq \alpha_j$$

$$U_{ij} = R_{ij} + S_{ij}\alpha_j \text{ if } \alpha i < R_{ij} \text{ - } S_{ij}\alpha_j$$

Where R_{ij} is the distance between two spheres centred on atoms, α_i the intrinsic Born radius of atom i (or otherwise, the Born radius that would give its solvation energy if it was alone) and S_{ij} a screening factor that scale the Born radius of atom j. This factor was introduced to correct systematic errors in the PDA approximation, because PDA over-estimates the Born radius by not taking into account the fact that two atomic spheres, j, j' can overlap. In this case S_{ij} accounts for the overlapping region of dielectric being displaced twice. This means that scaling factors should have a value between 0 and 1.

The GB equations, or PB for that matter do not tell the whole story of solvation. It is also necessary to take into account other effects that are described in the next section.

2.5.2 The apolar component of solvation

Solvation is not entirely determined by the distribution of charges of the solute inside a cavity.

To insert a solute inside solvent, a cavity the size of the solute has to be created. In the case of water, the hydrogen bonding network is disrupted and solvent molecules have to reorganise and reorient around the solute.

van der Waals forces also play a role as solute atoms are able to establish interactions with solvent atoms. Solvent atoms lie usually far enough away from solute atoms for the van der Waals forces to be predominantly attractive.

In most continuum solvation models, both effects are taken into account with a single dependence on the solvent accessible surface area (SASA) of the solute.

$$G_{nonpol} = G_{cav} + G_{vdW} = \sum_{k=1}^{N} \sigma_k . SASA_k \tag{2.32}$$

The SASA is the surface that defines the region of space that solvent is excluded from upon insertion of the solute. For that purpose it is assumed that water can be represented as a sphere of radius 1.4 Å , and the SASA is defined by rolling that sphere over the van der Waals surface of the solute. The set of coefficients σ_k are usually empirically derived.

Using the SASA of a solute to model solvation has the drawback that buried atoms do not interact at all with the solvent, which is not the case in explicit solvent simulations.

Recently, Zacharias [2] has proposed a method that takes into account cavity penalty and van der Waals forces separately. This method has drawn our attention for it does not appear to suffer from deficiencies of a traditional SASA method.

The SASI method can be formulated as follows.

Let us assume that the water distribution around the N atoms i of a solute is approximated by a continuous density function ρ_{iw}, depending only on the distance r_{iw} between water and solvent atoms, the average solvent density ρ_w , and a solute atom-water correlation function $g_{iw}(r_{iw})$. Then

$$\rho_{iw}(r_{iw}) = \rho_w . g(r_{iw}) \tag{2.33}$$

We now consider that the dispersive interaction between the solute and the solvent falls off with the sixth power of the distance (LJ potential) and is characterised by a solute

atom-water oxygen dispersion parameter C_{iw}. By integrating from the solute excluded volume over space we can write that the solute dispersive energy is :

$$U_{disp} = \sum_{i}^{N} \rho_w . C_{iw} \int_{V_w} dr_{iw} g(r_{iw}) r_{iw}^{-6} \tag{2.34}$$

Floris and Tomasi have shown [22] , from work of Huron and Claverie [23] that this volume integral can be transformed into an integral over the surface of the solute excluded volume.

$$U_{disp} = \sum_{i}^{N} \rho_w . C_{iw} \int_{S_s} d\sigma_s A(\vec{r}_{iw}) . \vec{n}_s \tag{2.35}$$

where $\vec{n}_s$ is a vector normal to the surface s, $d\sigma_s$ a surface element and $A(\vec{r}_{iw})$ is defined as :

$$A(\vec{r}_{iw}) = [\frac{1}{r_{iw}^3} \int_{r_{iw}}^{\infty} \frac{g_{iw}(x)}{x^4} dx] \vec{r}_{iw} \tag{2.36}$$

Where x is a point in solvent occupied space.

If we assume that the solvent density is uniform outside the solute cavity, then $g_{iw}(x) = 1$ and 2.36 can be solved.

$$A(\vec{r}_{iw}) = \frac{\vec{r}_{iw}}{3r_{iw}^6} \tag{2.37}$$

The dispersion interaction can now be determined by a discrete summation over k surface elements of surface ΔS_k distributed on the surface of the solute excluded volume.

$$U_{disp} = \sum_{i}^{N} \rho_w . C_{iw} \sum_{k} \Delta S_k A(\vec{r}_{ik}) . \vec{n}_{sk} \tag{2.38}$$

Where $\vec{n}_{sk}$ is a normal vector to S for element k and $\vec{r}_{ik}$ the distance between atom i and surface element k. See figure 2.2 for clarity.

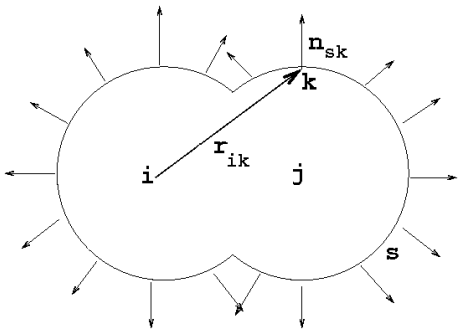

Figure 2.2: Illustration of equation 2.38 for a diatomic molecule

Chapter 3
A review of Generalised Born Models

3.1 Modelling electrostatic effects

Following the paper of Still et al [1] considerable effort has been devoted to the development and application of Generalised Born methods. The original method was semi-analytical, the computation of the Born radii requiring a CPU expensive numerical method. Since then analytical approximations to the computation of Born radii have been proposed [20,21] . Note that these really are approximations and do not give exact results. Some other faster recent numerical methods have also been discovered [24,25] . While these methods are not as fast as the analytical ones they are, in principle, more accurate.

Modifications to the original functional form proposed by Still have been suggested by a number of workers. Ghosh et al [26] proposed an alternative approach where the GB theory is recast in the form of surface integrals which has some practical computational advantages over the volume integral formalism. Jayaram has shown that another functional form yielded better agreement for the computation of pkA shifts in dicarboxylic acids [27] . The new functional form, however, was less effective at predicting the solvation energy of organics. Srinivasan et al. modified the equation so that salt effects could be incorporated in the model [28] . The modified model gave a response to salt effects which is roughly similar to results obtained with Poisson-Boltzmann calculations, with a very

good agreement at low salt concentrations.

In standard GB models, the dielectric boundary is defined by the van der Waals surface, which can leave numerous small "solvent" cavities within large solutes. These cavities are too small to be filled with solvent but yet will be considered so by the GB equation. This can explain why GB models tend to underestimate the value of the Born radius of buried atoms in proteins. To compensate for this "missed volume",Onufriev et al[29] introduced a correction factor which is conceived as the ratio of the correct solvent inaccessible volume to the van der Waals volume.

Generalised Born solvation models have been shown to be adequate to reproduce solvation energies of small solutes. This usually requires some form of parameterisation. Some workers have derived parameters for the full solvation model against experimental energies of solvation. Hawkins et al [30] parameterised solvation models for semi-empirical quantum chemistry methods . They achieved a mean error close to 0.50 kcal.mol^{-1} with models using around 50 parameters on a dataset of 261 compounds. For the MMFF [31] forcefield, Cheng et al [32] obtained an average mean error of 0.64 kcal.mol^{-1} on a dataset of 82 neutral compounds by using 21 parameters. For the AMBER [15] forcefield, Jayaram et al [33] used 14 parameters to reproduce the solvation energy of 32 compounds to within 1.1 kcal.mol^{-1} . For the OPLS [16] forcefield, Qiu et al [21] reproduced solvation energy of 35 compounds to within 0.9 $kcal.mol^{-1}$ using 6 parameters. In a very recent study, Zhang et al reported a new solvation model for AMBER that uses 43 parameters to reproduce solvation energy of 328 neutral molecules to within 0.85 kcal.mol^{-1} and 30 charged molecules to within 4.36 kcal.mol^{-1} [34] .

Other workers first calibrate the GB equation only against solution of the PB equation [35–37] or results of free energy perturbations calculations in an explicit solvent [38] .

Overall it seems that achieving an accuracy better than 1 kcal.mol^{-1} on average requires many parameters to be introduced. Given the limits of Poisson-Boltzmann based models, it might be unrealistic to expect a better agreement with experiment for GB models.

Generalised Born models have also been used to examine intermolecular interactions moderated by solvent. Osapay et al have shown that a GB model predicted a strength of interaction between different hydrogen bonds found in proteins which was in good agreement with explicit solvent simulations and PB calculations [39] . The same study

was done on protein-DNA complexes by Dixit et al [40] . Scarsi et al have computed the ramachandran plot of alanine dipeptide and shown that it was qualitatively similar to explicit solvent results, although the latter showed more variations [41] .

Srinivasan et al [28,42,43] looked at the ability of PB and GB models of solvation to provide a useful understanding of the relative energetics of A and B form of helices for DNA and RNA. While GB models tends to overestimate the effect of solvent screening for a pair of buried atoms, individual errors tends to cancel out and total energies are in good agreement with PB results and explain the relative differences between the A form and B form.

Edinger et al [44] compared the ability of different solvation models to rank conformations of peptides versus results obtained with the Poisson-Boltzmann equation. Three solvation models were considered : a GB model, a distance dependent dielectric model and a model based on surface area only. They concluded that the GB model was the most appropriate to rank conformations of peptides similarly to Poisson-Boltzmann results.

A number of studies have been concerned by the ability of GB models to reproduce reliably the dynamics of proteins [45,46] . Results appears to be mixed. While some workers reported GB simulations in excellent agreement with explicit solvent simulations, others observed qualitatively different behaviour.

Masunov et al [47] have computed potentials of mean force for the association of various pairs of protein side chains, using the explicit/implicit solvent model SSBP (Spherical Solvent Boundary Potential) [48] . The SSBP model uses a small number of explicit solvent molecules to form the primary solvation shell, the remaining solvent being represented by an implicit model. The PMFs thus obtained where compared to a GB model [37] used in the CHARMM program [17] , the EEF1 (Effective Energy Function 1) solvation model [49] and a primitive electrolyte model (The Coulombic interactions are screened with a dielectric of ϵ 80). They concluded that the GB model was the closest to the results obtained with the more accurate SSBP model.

3.2 Modelling apolar effects

Apolar effects are thought to arise from two main causes, the work needed to create a cavity large enough to insert the solute into the solvent and the solute/solvent van der Waals interactions.

To date, most of the solvation models that have been used in classical simulation model apolar effects with a term proportional to the total solvent accessible surface area (SASA) of the system. This is based on the on the fact that SASA is a good descriptor of the experimental solvation energy of linear alkanes [50] . This does not hold true, however, for cyclic alkanes [51] .

While it is reasonable to expect the cavity penalty to be purely dependent on the SASA, there are evidences that dispersive interactions play a significant role. Gallichio et al[52] studied linear, branched and cyclic alkane solvation, using explicit solvent and free-energy perturbation methods, to obtain the different components of the solvation energy. They concluded that the creation of a cavity in solvent is associated with a large penalty, which is balanced by the favourable dispersive interactions between the solute and the solvent. An important finding was that although the cavity term was strongly dependent on the total SA of the solute, the dispersive term was only weakly correlated.

Graziano et al [53] have studied the hydration of aromatic hydrocarbons and found that the more negative solvation energy of aromatics (-0.9 $kcal.mol^{-1}$ for benzene vs 2.2 $kcal.mol^{-1}$ for butane, which has approximately the same SASA) was due to stronger solute-solvent dispersive interactions, with the ability of benzene to form weak hydrogen bonds playing little part.

Recently, Gallichio et al [38] have proposed a solvation model based on the Surface-Generalised Born formalism [26] with a novel non-polar term. In their work, the non polar energy is expressed as the sum of surface-area dependent term (for the cavity formation) and a constant, surface-area independent term. Both terms are based on atom-types and they introduced 38 different terms, derived by fitting to the difference between the experimental solvation energy and the polar contribution calculated by the SGB method, for a dataset of approximately 200 compounds . Gallichio recognised that these dispersive terms should vanish as the atoms becomes more and more buried. He introduced thus a switching function based on the Born radius. The switching function was chosen so that the dispersive terms would be quickly reduced to zero once the Born radius of one atom is greater than 3 angstroms.

An interesting study was reported recently by Zacharias [2] . By focusing on united atoms alkanes only, Zacharias introduced a dispersive solute-solvent equation based on the carbon-oxygen dispersive term from OPLS. The dispersive solute-solvent contribution

is computed by integrating over the solute-excluded volume using surface area-elements. This is in fact the method used in the PCM solvation model of Floris and Tomasi, widely used in quantum chemistry [22] . By using this approach Zacharias is able to predict correctly the energy difference between hexane and cyclohexane and to obtain a much better fit against the experimental solvation energy of alkanes.

Chapter 4
Results

4.1 Parameterisation of a GB/SA model for water

4.1.1 Dataset

Very recently, a new method to derive partial charges has been proposed [54] . The AM1-BCC charges emulates charges obtained by a RESP/HF6-31G** molecular electrostatic potential. They could therefore be used along with the the forcefield AMBER [15] . It is easy to assign torsions with AMBER because generic parameters are available if needed. The AMBER7 package also provides a set of tools which should greatly facilitate automatic parameter assignment and the very recent development of the GAFF [55] force field, compatible with AMBER, could provide most of the parameters needed for a wide range of ligands.

Since AMBER promises a quick setup of ligands when free energies of binding will be computed with our model, this forcefield was adopted for the rest of this work.

Experimental vacuum to water transfer energies of a wide range of simple organic compounds [56] were gathered. A detailed list is given in the appendix. A set of compounds for AMBER was then built using the AM1/BCC semi-empirical method to calculate partial charges. This set contains a balanced mixture of organic functions encountered within

drugs. It was also required that these chemicals would not undergo any conformational change upon solvation. This hypothesis was not rigorously tested but compounds showing possible problems were not included (for instance : 1,2 dichloroethane, triethylamine). It is necessary that solutes in the set keep the same conformation in vacuum and in aqueous phase, otherwise the experimental solvation energy would contains other terms than the GBSA term.

The set was split in two groups, a training set with approximately 75% of the compounds and a validation set. Compounds in the validation set were not used to derive coefficients for our solvation model. This allows us to check the transferability of the parameters.

Table 4.1: Composition of the dataset

Family	Training	Validation
Linear Alkanes	6	2
Branched Alkanes	3	1
Cycloalkanes	4	1
Alkenes	6	2
Alkynes	3	1
Arenes	6	2
Alcohols	7	3
Aldehydes	4	2
Ketones	6	2
Carboxylic acids	4	1
Esthers	6	2
Ethers	6	2
Aliphatic amines	10	3
Aromatic amines	6	2
Nitriles	3	1
Amides	3	1
Thiols	6	2
Multi-functionals	8	3
Halides	18	7
Nitro	4	2
O charged	2	1
N charged	9	3
S charged	3	1
Total	133	47

4.1.2 Fittable parameters

Once the dataset was ready, parameters that will be used in the solvation models had to be considered. After analysis of the different equations of the model a number of choices

appeared.

1. Intrinsic Born radius
   (a) Taken as van der Waals radius
   (b) van der Waals radius with offset
   (c) Dependent on charge of atom
   (d) Independent of van der Waals radius
2. Scaling Factors
   (a) Based on atomic mass
   (b) Based on atom-types/connectivity
3. Surface tensions
   (a) Single value for all atoms
   (b) Different values, positive, null or negative

The intrinsic Born radius is the Born radius of a completely isolated atom, it corresponds to α_i in equation 2.31 . Since this value cannot be determined experimentally for most atoms, it is derived empirically. It was decided that the intrinsic Born radius of one atom would be derived by multiplying its van der Waals radius by an empirically determined offset.

The scaling factors are the empirical terms that corrects for the systematic error caused by the Pairwise Descreening Approximation. They correspond to S_{ij} in equation 2.31. Because scaling factors based on atom-types/connectivity would require a lot of parameters it was decided to use a small set, based only on atomic mass.

Surface tensions term are the terms that relate the SASA of a solute to its non polar energy of solvation. They correspond to σ_k in equation 2.32 . A single parameter for all atoms was initially considered.

It is worthwhile to make a comment about possible approaches to the generalised Born solvation model. Two different approaches to the parameterisation of this model in the literature are usually observed. One considers the generalised Born model essentially as "Poisson-Boltzmann with further approximations" and optimises the parameters to

reproduce the electrostatic energies computed by solving numerically the Poisson equation. This is probably a good way to obtain accurate Born radii, *within* the context of a continuum electrostatic theory. This approach does not guarantee however results that closely match experimental values, because Poisson Boltzmann theory also makes a number of assumptions.

A more pragmatical approach is to consider the Generalised Born model as a set of equations that can produce numbers in good agreement with experiment if carefully chosen parameters are used. Therefore the parameters can also be derived by matching the computed values directly to the experimental free energies of hydration. This approach can yield parameters that seem unphysical with respect to their definition, but this allows errors introduced by a number of approximations to be corrected (Coulomb field, instant change of dielectric at the boundary, effective Born radius, functional form, quality of the partial charges etc ...).

However it might not be wise to use a set of unphysical parameters. Therefore, a second approach has also been considered.

In the second approach, two different solvation models are effectively derived. The first solvation model is obtained by optimising intrinsic Born radii and surface tension only, using a finite-difference method to compute Born radii. The algorithm originally employed by Still [1] is used for that purpose. While this algorithm is quite slow, its main advantage is that it yields accurate Born radii.

Once this solvation model is ready, the numerical integration method is replaced by the Pairwise Descreening Approximation of Hawkins et al [20] . Because of the errors caused by the PDA approximation, a set of scaling factors is introduced. The main difference with the first approach is that these scaling factors are now optimised to reproduce the polarisation energy (GB) obtained from the more accurate method, instead of the experimental free energy of hydration. This means that the scaling factors compensate *only* for errors in the PDA approximation and no others, unknown, errors caused by approximations of the Generalised Born theory.

4.1.3 Resulting models

Optimisation of the chosen parameters was done with a Genetic Algorithm (GA) taken from the GAlib library [57] .

Genetic Algorithms are very effective methods to optimise complex function with a large number of variables. Because of this, they are much better suited than simplex based methods to derive optimal parameters for our solvation models.

The author has written a program in C to collect input files and perform the required computations for the full solvation model.

Many models based on possible combinations of parameters have been tested. Three different models that have been further tested in the next section are presented here.

The mean error is given for the training set and the validation set respectively.

1. **GBSAv1** - 8 parameters

   Scaling factors:

   sH 0.24 sC 0.36 sNsp3 1.04 sNsp2 0.75 sO 0.85 sS 1.04 sX 1.18

   Offset for Intrinsic Born radius:

   None

   Surface Tension:

   0.00765 H,C,N,O,F only

   Mean Error : 1.00/1.34 $kcal.mol^{-1}$

2. **GBSAv2** - 11 parameters

   Scaling factors:

   sH 0.50 sC 0.52 sNsp3 1.24 sNsp2 0.73 sOsp2 1.31 sOsp3 0.52 sS 0.99 sX 1.14

   Offset for Intrinsic Born radius:

   O2 0.91 N3 0.66

   Surface Tension:

   0.00675 H,C,F only

   Mean Error : 0.80/0.96 $kcal.mol^{-1}$

3. **GBSAv3** - 13 parameters

   Scaling factors:

   sH 0.81 sC 0.77 sN 0.70 sO 0.88 sS 0.83 sX 0.93

   Offset for Intrinsic Born radius:

   Osp2 1.0 Osp3 0.66 O2 0.88 N 0.73 N3 0.87 others 0.86

Surface Tension:
0.0070 H,C,F only
Mean Error : 1.15/1.23 $kcal.mol^{-1}$

GBSAv1 has been derived by optimising all the parameters at once against the experimental solvation energies. sNsp2 and sNsp3 designate scaling factors sp2 Nitrogen and sp3 Nitrogen atoms. sX is a single scaling factor applied to all halides. The surface tension term is 0 for elements that are not listed on the surface tension list. The surface area of some heavy atoms (S,Cl,Br,I) and sometimes Nitrogen and Oxygen in some models was ignored as it was found that the solvation free energy of compounds containing these elements tended to be systematically underestimated. Because the surface area term is always a penalty to the free energy of hydration, this made solvation free energies of these compounds more negative and reduced the error against experiment. It is not clear however, if this is a problem due to the GB or SA equation.

GBSAv2 introduces sOsp2 and sOsp3 which are scaling factors for sp2 and sp3 oxygen atoms. GBSAv2 also uses offsets to the van der Waals radii to determine the intrinsic Born radius of charged hetero atoms (O2 and N3 AMBER atom types). This was done to correct some errors for ionic compounds.

Unlike the two previous models, GBSAv3 was derived in a two step fashion mentioned in the last section. The surface tension term was first optimised along with the group of offsets against the experimental solvation free energies using an accurate finite differences method to determine Born radii [1] . This yielded a model with a mean error of 1.02 $kcal.mol^{-1}$ on the training set. The Pairwise Descreening Approximation of Hawkins et al [20] was then introduced. With the PDA method it was possible to reproduce the previously obtained polarisation energies to within 0.30 $kcal.mol^{-1}$ on average. When the finite difference method was substituted by the PDA approximation, the mean error increased from 1.02 to 1.15 $kcal.mol^{-1}$ for the training set.

GBSAv3 was latter slightly modified to correct some problematic potentials of mean force. More precisely the offset for O2 atoms was set at 0.85,and N3 at 0.95 . This new version is called GBSAv3-2.

At first glance it seems that GBSAv3, that uses more parameters than GBSAv1 or GBSAv2, should have a lower mean unsigned error. This is not the case because the scaling factors have been parameterised to compute accurate Born radii and not to

reproduce experimental solvation free energy. In fact it is possible to argue that the scaling factors are not empirical parameters in this situation and that GBSAv3 uses only seven empirical parameters.

The mean error increase between the training set and validation set is 0.3 kcal.mol^{-1} for GBSAv1, 0.2 kcal.mol^{-1} for GBSAv2, less than 0.1 kcal.mol^{-1} for GBSAv3 and with GBSAv3-2, compounds in the validation set are slightly better predicted on average (the mean error decreases by 0.05 kcal.mol^{-1}). This suggests that, at least for GBSAv3, the models have not been biased toward compounds in the training set.

To further assess the predictive power of the method, parameters for GBSAv1 have been derived by systematically excluding one family of compounds from table 4.1. The resulting model was then used to predict solvation energies of that missing family. It was also possible to check if the final parameters were similar to the parameterisation against a full dataset.

This has shown that the parameters were not too sensitive to the composition of the dataset. Models that did not include one family typically saw their mean error on that family increase by 0.0-0.2 kcal.mol^{-1} for neutrals and 0.5-3.0 kcal.mol^{-1} for ionics against the full model. This suggest that the model has some transferability to unknown classes of compounds.

Finally, errors on different families of compounds for the different solvation models are summarised on table 4.2.

Table 4.2: Mean error for each family with different solvation models [a]

Family	GBSAv1	GBSAv2	GBSAv3-2
n-alkanes	0.21	0.42	0.28
branched alkanes	0.51	0.69	0.48
cyclic alkanes	0.41	0.30	0.60
alkenes	0.34	0.35	0.33
alkynes	0.46	0.41	0.33
arenes	1.09	0.43	0.51
alcohols	1.28	1.15	0.79
aldehydes	1.94	1.21	0.67
ketones	0.98	0.42	1.30
carboxylic acids	0.70	0.53	0.56
esters	2.24	0.41	0.82
ethers	0.66	1.38	0.59
aliphatic amines	1.21	0.60	2.15
aromatic amines	0.78	0.95	1.32
nitriles	0.73	0.45	1.31
amides	0.59	0.80	2.09
thiols	0.32	0.45	0.56
multi-functionals	1.20	1.72	2.44
halides	0.40	0.37	0.86
nitro	2.10	1.30	1.14
ionics -O charged	3.91	0.85	1.53
ionics -N charged	0.80	1.75	2.11
ionics -S charged	1.95	1.33	2.26

[a] Error on training set in kcal.mol^{-1}

As can be seen, each model outperforms the others on some particular families, but getting consistent improvement over the whole dataset is quite challenging.

These models can be compared with existing ones in the literature and mentioned in chapter 3. Other workers are able to achieve a better agreement with experiment (around 0.5 kcal.mol^{-1}), but they usually requires a much larger number of parameters . Because experimental data is limited, it is unwise to use a large number of parameters as this can lead to models with poor transferability.

An important issue is the adequateness of these models for our purposes. They are parameterised on small organics but will be used to study systems where intermolecular interactions between different species plays a significant role.

A more stringent test for an implicit solvent model is to check if it can reliably reproduce Potentials of Mean Force that have been derived using explicit solvent models. Such a study has been conducted and is presented in the next section.

4.2 Application: Potentials of mean force

A number of potentials of mean force (PMFs) for the association of various species in solution have been computed using the GB/SA solvation models derived in the last section. These potentials of mean force show how the free energy of association of a pair of molecules varies as they approach each other.

PMFs for the systems studied here have been previously reported in the literature. It is therefore possible to compare our results with previous work. Rigorous comparison can not be done because even for explicit solvent simulation, issues such as adequate treatment of long range electrostatics, solvent model, potential energy function, often lead to quite different results for the same system. Thus, such comparisons can and should only serve us to identify qualitative differences and general features.

4.2.1 Methods

Each system has been set-up with AM1-BCC generated charges and is constructed with equilibrium bond lengths, bond angles for AMBER94/99.

Free Energy were computed with the free-energy perturbation method. The coordinate is defined by the distance between two atoms, usually the centre of mass of the solute. Windows were placed approximately every 0.2 angstroms along the coordinate.

A custom version of MCPRO1.5 [58] was modified so that FEP could be carried with a GBSA model.

Because the solvent is always at equilibrium with the solute when using a continuum solvation model, sampling is much faster. Therefore we need a smaller amount of moves to converge the free energy differences. It also seems that the number of moves can be further reduced when the solutes present a high-level of symmetry (such as methane or benzene).

Understandably, continuum solvent models cannot predict solvent separated contact minimum which have sometimes been shown to be significant or even more stable than the first contact minimum.

The results are presented in two parts. In the first we study *true PMFs*, where the systems under study are not constrained to remain in a specific orientation. This requires adequate sampling of all configurations for a given distance and can reliably be accomplished only for simple systems with few or no internal degrees of freedom.

In the second part we study more complex systems constituted by amino-acids side chains. True PMFs are generally unavailable because of the large number of conformations that have to be averaged. Instead, PMFs along given orientations have been computed and can be compared with previous work from others.

4.2.2 True Potentials of mean force

Methane pair

This PMFs is a test case for hydrophobic interactions. Because the polarisation term is negligible for this compound, the solvent effects are only due to the surface area dependent term. The molecules are kept rigid. A very small amount of sampling over each window is required to converge the free energy differences. We report here PMFs generated with 1K moves for equilibration and 10K moves for collection. By contrast, Jorgensen et al[59] have carried out a simulation on this system (United Atom) with explicit solvent and used 500K moves for equilibration, followed by 2M moves for collection. The PMF computed in vacuum and with a GB/SA term are reported below on figure 4.1 .

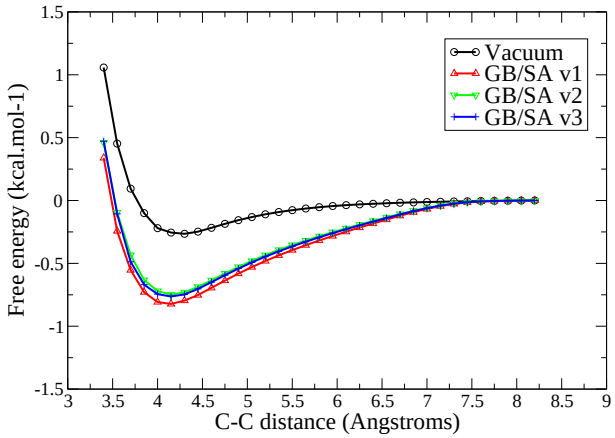

Figure 4.1: Methane-Methane PMF

A CM (Contact-Minimum) is present at a separation of 4.1 Angstroms. In vacuum, the well-depth is about -0.3 $kcal.mol^{-1}$ and arises from attractive Lennard-Jones interaction. In solution, our various GB/SA models perform similarly with a net attraction of around -0.80 $kcal.mol^{-1}$ this is because the surface tension term does not vary much. Since the GB term is negligible for methane, the energy change in solution is solely related to the change in SASA and the association of the two species is favoured to reduce

the total SASA of the system. Jorgensen has computed a binding free energy of -0.42 +/-0.34 $kcal.mol^{-1}$ for that system. Thus our model is slightly more attractive. We note however, that the hydration free energy of a single methane molecule is under-estimated with our solvation models (between 1.1-1.2 $kcal.mol^{-1}$ instead of 2.0 $kcal.mol^{-1}$). If our models predicted a more accurate solvation energy, the dimer would become more stabilised.

Benzene pair

This PMF has been computed along the distance of the centre of mass of each benzene molecule. Our parameters for benzene are quite similar to the ones employed in a study by Jorgensen et al[60] . A Monte Carlo optimisation in vacuum reveals that the global minimum is a roughly T-shaped dimer with a separation of 4.8 Angstroms. The interaction energy is -2.40 $kcal.mol^{-1}$. Jorgensen has found a slightly more perpendicular configuration at a separation of 5.0 Angstroms, with an interaction energy of -2.31 $kcal.mol^{-1}$.

The PMF in vacuum and with our GBSA models for the association of two rigid benzene molecules are reported below. For each window 11K moves were performed and averaging was usually done on the last 10K.

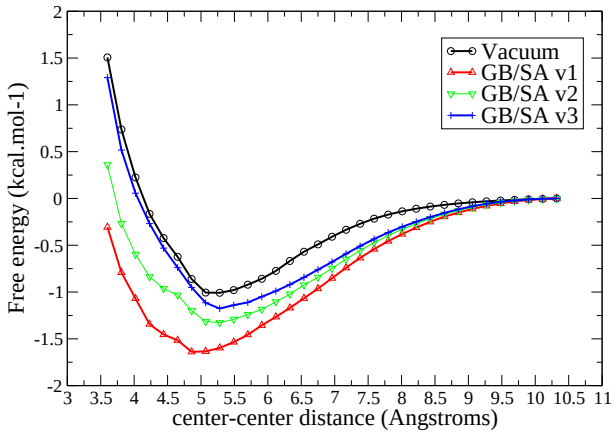

Figure 4.2: Benzene-Benzene PMF

According to these results, the strength of the interactions between two benzenes is stronger in water than in vacuum. This is expected because of the hydrophobic effect. However figure 4.2 shows that our solvation models exhibit more marked differences than with methane. A CM is observed around 5.1-5.3 angstroms with a well depth

ranging from 1.2 to 1.6 $kcal.mol^{-1}$.In the study of Jorgensen a CM at 5.5 angstroms with a well-depth of -1.5 $kcal.mol^{-1}$ was found. Jorgensen also noted that integration of his computed PMF overestimated the association of benzene in water according to experimental measurements[61] .

If the association was purely driven by hydrophobic forces, we would expect the PMF in solution to be more stabilised by about 0.7 $kcal.mol^{-1}$ than in vacuum, because of the difference of SASA for the two separated species, and the dimer at the CM (Difference of 100 $\mathring{A}^{-2}$ with a surface tension around 0.007 $kcal.mol^{-1}.\mathring{A}^{-2}$). This is almost the case in version GBSAv1, but in GBSAv2 and GBSAv3, as the two species get closer, the GB term becomes more positive, reducing the amount of stabilisation due to the reduction of SASA. This is expected as the dimer exhibits an attractive Coulombic attraction and part of the GB equation is anti-correlated with the Coulombic term. Benzene is expected to be able to form weak-hydrogen bonds with water and desolvation would reduce its ability to form these hydrogen bonds[53] . Another interesting fact is that the solvation energy of benzene with GBSAv1 is -2.2 $kcal.mol^{-1}$, and -1.3 $kcal.mol^{-1}$ in GBSAv3 (versus an experimental value of -0.9 $kcal.mol^{-1}$). A naive reasoning would have suggested that desolvation would be more disfavoured by GBSAv1 than GBSAv3, but it is actually the contrary. This is caused by the large difference in scaling factors for H and C between these two models. In GBSAv1, as the two molecules get closer, the Born radii do not increase much, because the descreening influence of one atom is strongly reduced by the scaling factors. If there are little or no variations in Born radii, then the GB term does not vary much, and thus does not oppose the Coulombic attraction. With GBSAv3, larger variations are seen because the scaling factors reduce less the descreening, causing ultimately the PMF to be weaker.

Jorgensen predominantly observed around the CM a range of distorted T-shaped pairs, including some roughly parallel stacked and displaced pairs. In the vicinity of the CM with our GBSA models we tend to see more perpendicular T-shaped pairs and rarely parallel stacked and displaced pairs. As noted before, at a short distance of around 4 angstroms, where only face to face staking is possible, the net interaction is repulsive, even though the gas phase interaction energy is almost -2.0 $kcal.mol^{-1}$. This is because this conformation, configurationaly restricted, is entropically unfavourable.

Formic acid pair

In gas phase the double hydrogen bonded conformation is mainly observed. There is evidence however that very little of this interaction remains in water, where hydration is preferred. The two solute were kept rigid and PMF computed along the carbon-carbon distance. 110K moves were performed for each window, averaging was usually done on the last 100K moves.

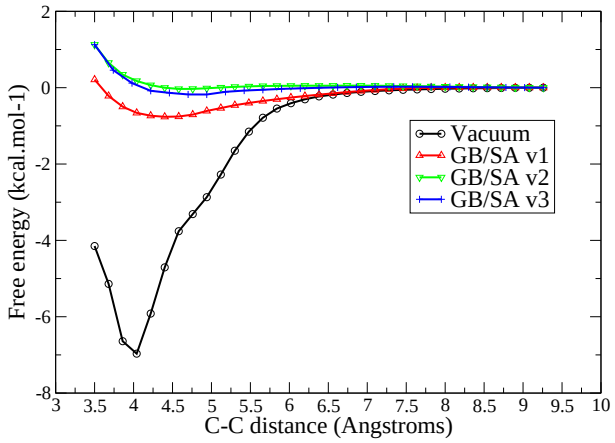

Figure 4.3: Formic Acid -Formic Acid PMF

As can be seen on figure 4.3, the effects of solvation are very marked on this system. In vacuum the PMF exhibit a CM at 4.0 angstroms and a well-depth of -7.0 $kcal.mol^{-1}$. A simulation of the same system with GBSAv1 yield a more broad CM around 4.5 angstroms with a well-depth of -0.7 $kcal.mol^{-1}$. Experimental studies suggest that the dimerisation energy of formic acid in vacuum is around -3.2 $kcal.mol^{-1}$ +/- 0.8 $kcal.mol^{-1}$ and that little or no dimerisation seem to occur in water (see[62] and references therein). GBSAv1 still exhibit a small interaction, but it vanishes with models 2 and 3.

N-methylacetamide pair

The association of two N-methyl-acetamide molecules in solution is representative of a polar-polar interaction mainly driven by the formation of an hydrogen bond at short separation. It is also a prototype model for the interaction of protein backbones.

Bond lengths and angles were kept fixed. Torsional motion was sampled about the central C-N bond and for the methyl hydrogens. The PMF were computed along the distance between the middle of the C-N bond of each molecule. 110K moves were performed for each window, and averaging was usually done on the last 100K moves.

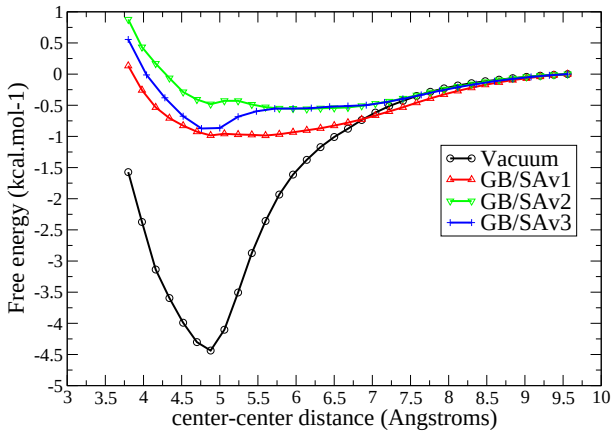

Figure 4.4: NMA-NMA PMF

Figure 4.4 shows that in vacuum the PMF exhibit a CM at 4.90 angstroms with a well-depth of -4.4 $kcal.mol^{-1}$. This correspond to an optimal hydrogen bonded configuration with a nitrogen-oxygen distance of about 3.0 angstroms. In the GB/SA simulation this attraction is considerably reduced, because of solvent competition, the contact minimum is wider and the net attraction varies between -0.5 and -1.0 $kcal.mol^{-1}$. However in a study by Jorgensen[63] no attraction was observed. This suggest that our GBSA models do not sufficiently shield the Coulombic term.

Interestingly, GBSAv2 yields a smaller attraction than GBSAv1. However the predicted solvation energy of NMA with GBSAv2 is -8.4 $kcal.mol^{-1}$, while it is -9.0 $kcal.mol^{-1}$ with GBSAv1. Once again, there is no correlation between the solvation energy of an isolated molecule, and the strength of interaction of a dimer. In GBSAv2, a smaller interaction is achieved because the GBSA term around the CM is about 2.0 $kcal.mol^{-1}$ more positive than at long separation. In GBSAv1, the solvation contribution varies by only 1.2 $kcal.mol^{-1}$, thus the coulombic term is less shielded in GBSAv1 than in GBSAv2. With GBSAv3, the predicted solvation energy of a single NMA molecule is only -6.9 $kcal.mol^{-1}$, yet the PMF is as attractive as in GBSAv1.

Once again we see that a PMF can not be deduced from the value of the solvation energy of the isolated species. Furthermore it appears here that agreement with experimental values of free energy of solvation *doesn't* ensure best results. The relevant factor is the Born radius, because it is responsible for the *variations* in GB energy as the dimer is formed. The functional form plays a role also, as one should recall that the GB equation is an interpolation built to reproduce theoretical results at short and large separation. It

is quite possible that alternative functional forms would represent better the variation of the GB energy at intermediate distances (between 3-6 angstroms).

In the study by Jorgensen a PMF for the dimerisation of N-methylacetamide in chloroform and water was reported. In chloroform the dimer exhibited mostly a hydrogen bonded configuration, while in water this interaction was uncommon and replaced by a stacked/displaced geometry that exposes the amides edge to minimise loss of hydrogen bonding to water. Analysis of snapshots in our simulation shows a wide range of conformations, some of them hydrogen bonded, but the stacked/displaced conformation does not occur often. As this conformation yields a favourable dipole alignment, it could occur in a GBSA simulation, but it is quite possible that this conformation occurs because of hydrogen-bonding with water. Such non-linear behaviour is probably out of reach of a continuum solvent model.

We note that the free energy of association computed here applies for a pair of unconstrained molecules. In a protein backbone, the environment usually restrains pairs of amides to a smaller range of configurations, which implies that the free energy of interaction would be more favourable for secondary structures such as alpha-helices or beta-sheets, and would increase as the amides becomes more buried. This is because the Coulombic interaction between buried atoms will not be shielded as much as if they were fully exposed to the solvent [39] .

4.2.3 Constrained Potentials of mean force

Accurate potentials of mean force between charged systems are challenging, because the high Coulombic term must be correctly balanced by a high solvation energy term to yield free energy that are usually orders of magnitude smaller than these terms, thus a slight error in either terms can lead to a large error in the free energy.

The system studied here are models of amino-acids side chains and have been studied by Masunov et al[47] with the CHARMM force field. In this study, constrained PMF were computed with the Spherical Solvent Boundary Potential (SSBP), a hybrid solvation model[48] . The results were compared to PMFs generated with the primitive electrolyte model (where ϵ is set to 80), the EEF1 solvation model[49] and a GBSA implementation in CHARMM.[37]

In the systems studied, bond lengths were kept rigid, appropriate angle bending and

torsion from AMBER94 were applied. 55K moves were performed for each window, and averaging usually done on the last 50K.

For some Potentials of Mean Force, a slightly modified version of GBSAv3, called GBSAv3-2 was used. Because the only difference between GBSAv3 and GBSAv3-2 are offsets for AMBER atom types O2 and N3, only PMFs containing these atom types were made with GBSAv3-2.

In the following part these notations have been used :

. Glu0 : Glutamic acid side chain, neutral.

. Glu- : Glutamic acid side chain, charged.

. Arg+ : Arginine side chain, charged

. Lys+ : Lysine side chain, charged

. HisP : Protonated Histidine side chain

. HisD : Neutral Histidine side chain, hydrogen on N-delta

. HisE : Neutral Histidine side chain, hydrogen on N-epsilon

Glu0-Glu0 pair

A PMF for the interaction of two neutral Glutamic acid side chains has been derived for a coplanar approach. The C-C distance corresponds to the distance between the carbon atoms from each carboxy group.

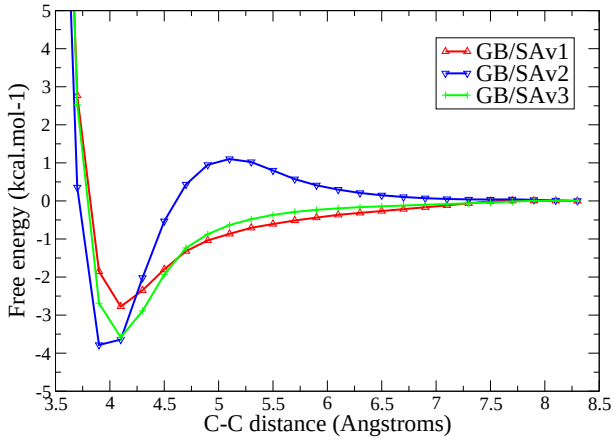

Figure 4.5: PMF Glu0 Glu0

Figure 4.5 shows that all models exhibits a CM around 4.0 angstroms. The well depth is -2.7 $kcal.mol^{-1}$ for GBSAv1 and -3.5 $kcal.mol^{-1}$ for GBSAv2 and GBSAv3. In the study by Masunov, a well-depth of approximately -1.5 $kcal.mol^{-1}$ has been found. This suggest that our model is too attractive. Quite interestingly, the PMF computed with the GBSA model from CHARMM was slightly repulsive.

However note, the peculiar behaviour of GBSAv2. We cannot see a barrier for association with a continuum model, as it is usually caused, in explicit solvent simulations, by a transition state between the contact minimum (CM) and a solvent-separated minimum (SSM). In fact, analysis of the simulation shows that this strange behaviour is caused by an error in the computation of the Born radius for atom HO.

It is important to realise that the value of a Born radius depends on the coordinates of all the atoms in the system studied. Because over the course of the PMF, the two glutamic acid side chains are getting closer, it is normal that the Born radii of atoms in the carboxylic group increases.

The Born radius is computed with the equation 2.31 seen in chapter 2. Because this equation is inexact, scaling factors have to be used to correct for systematic errors.

At short distance, the Born radius of HO is over-estimated because of the large value of the scaling factor of O in GBSAv2 (set at 1.31). Nearby oxygen atoms appear so 'big' that the HO atom 'feels' more buried and thus Coulombic interactions are less shielded, and the GB term suddenly drops, ultimately resulting in a stronger attraction. This is more clear on fig 4.6 below where XX is the non bonded energy between the two molecules. Obviously something is wrong as the GBSA term is no longer anti-correlated with the Coulombic terms at short distance.

This happened because in GBSAv2 the scaling factors are not required to yield correct Born radii, rather they have been optimised to diminish the error against experimental solvation free energies. These experimental solvation free energies contains contributions from physical processes that are neglected by the GBSA theory and thus cannot be exactly predicted by a GBSA model. If unconstrained, in an effort to improve agreement with experiment, scaling factors can adopt values greater than one which is contradicting their initial definition. This had been observed previously by Hawkins et al [30] before, but, to our knowledge, this is the first time that it is shown that this can affect free energies.

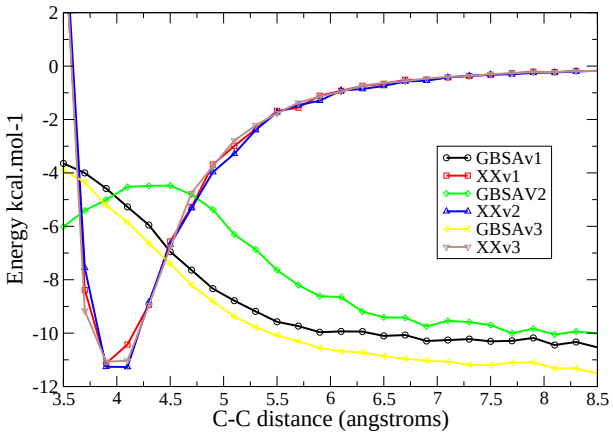

Figure 4.6: Energy averages Glu0 Glu0

Glu- Glu- pair

The same coplanar approach is applied.

All the PMFs on figure 4.7 are repulsive, GBSAv2 more than the rest. This PMF is similar to the one obtained by Masunov for the CHARMM GBSA model, although it is even more repulsive. The SSBP simulation by Masunov exhibits a broad CM with a well-depth of +1 $kcal.mol^{-1}$ approximately. This suggest that the PMFs obtained by a GBSA model are too repulsive for this kind of system.

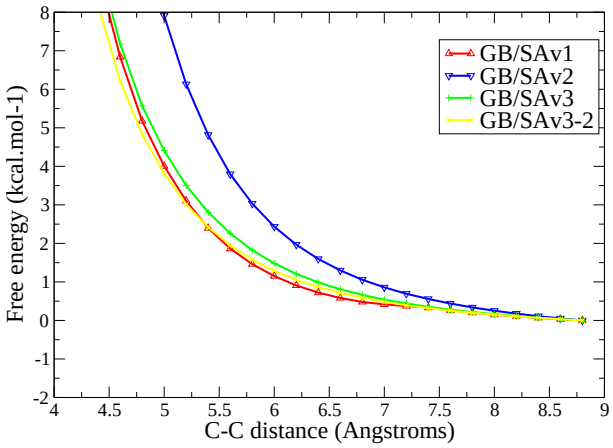

Figure 4.7: PMF Glu- Glu-

Glu0 Glu- pair

The same kind of behaviour observed with GBSAv2 for Glu0-Glu0 appears here on figure 4.8. It is caused by the same reasons. Only GBSAv1 exhibit a weak attraction of

about -0.5 $kcal.mol^{-1}$, GBSAv3 shows no attraction. The SSBP simulation of Masunov estimates the well depth to be about -2.0 $kcal.mol^{-1}$. The GBSA model from CHARMM is repulsive and exhibits no association.

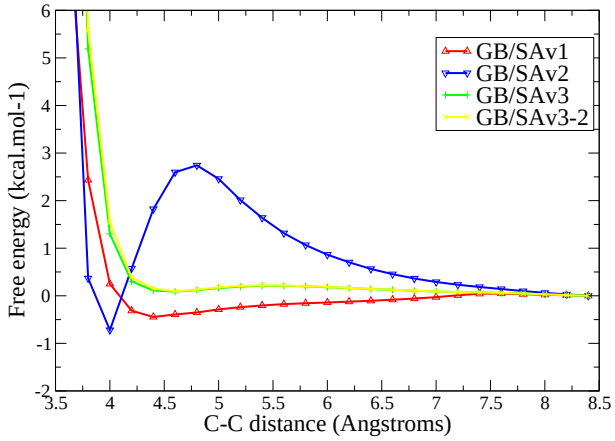

Figure 4.8: PMF Glu0 Glu-

Glu- Arg+ pair

A coplanar approach is maintained and allows formation of double H-bond between the two groups. A simpler version of this PMF where the hydrocarbon side chains have not been taken into account has also been computed for AMBER by Rozanska et al[64] . The C-C distance is the distance between the carbon atom from the carboxylate moiety and the carbon atom from the guanadiminium group.

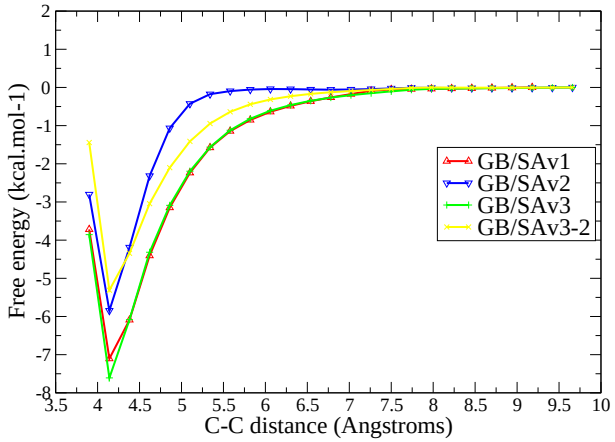

Figure 4.9: PMF Glu- Arg+

The resulting PMFs are quite sensitive to the solvation model used. The well-depths range from -7.6 to -5.3 $kcal.mol^{-1}$. It is interesting to note that only slight modifications

to GBSAv3 were required to reduce the interaction between the two species by about 2.3 $kcal.mol^{-1}$. Masunov found a well-depth of -4.5 $kcal.mol^{-1}$ with a SSBP simulation, and close to -4.0 $kcal.mol^{-1}$ with the GB model from CHARMM. Rozanska found a well-depth of -2.7 $kcal.mol^{-1}$ using Ewald lattice summation.

Arg+ Arg+ pair

Two possible orientations have been considered. In the first one, the two guanadiminium moieties are constrained to stay in the same plane.

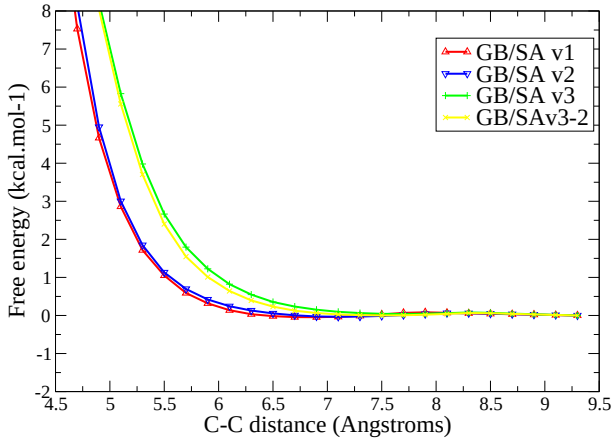

Figure 4.10: PMF Arg+ Arg+ coplanar approach

All the PMFs on figure 4.10 are repulsive until about 6.5 angstroms. No CM is really observed. The SSBP simulation of Masunov finds a CM at 4.6 angstroms with a well-depth of 0.0 $kcal.mol^{-1}$. The GBSA model from CHARMM is similar to ours but falls off with distance more quickly . It is repulsive by 2 $kcal.mol^{-1}$ at 4.6 angstroms. The difference most likely arise from the different parameters for the non-bonded interaction of the polar hydrogens of the guanadiminium moiety in AMBER (r* = 0.60 Angstroms) and CHARMM (r* = 0.22).

The two side chains can also approach each other in a stacked fashion which has also been considered.

By analysing the PDB database, Soetens et al have observed a stacked conformation as the preferred mode of interaction between two hydrated arginine side-chains when the C-C distance of the guanimidium groups is small (below 4 angstroms)[65] . The amount of stabilisation due to this interaction is controversial and Soetens reports values ranging between -10.0 $kcal.mol^{-1}$ and -2.7 $kcal.mol^{-1}$, depending on the water model used .

Masunov reports a well depth of approximately -1.0 $kcal.mol^{-1}$ for the SSBP simulation, but the solvent separated minimum (SSM) that lies at about 6.5 angstroms is slightly deeper. Soetens noticed only a very shallow SSM but argued that it could have been due to insufficient sampling. The GBSA model of CHARMM yields a well-depth of -1.0 $kcal.mol^{-1}$ at about 3.6 angstroms.

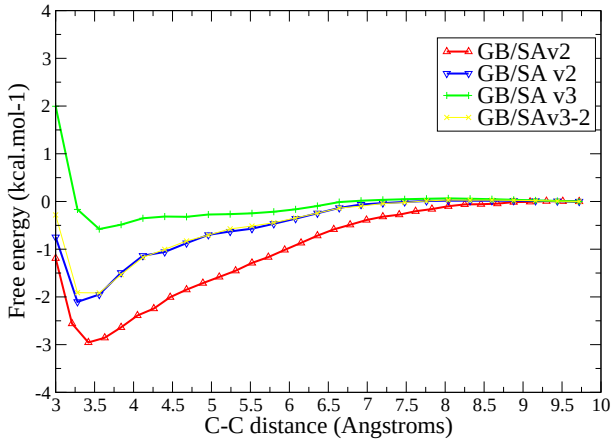

Figure 4.11: PMF Arg+ Arg+ stacked approach

The PMFs from figure 4.11 are quite sensitive to the solvation model. Interaction energy ranges between 3.0 and 0.5 $kcal.mol^{-1}$ approximately. We note that, although the side chains were kept parallel, they were allowed to rotate and the two hydrocarbons moieties could assume a parallel or anti-parallel conformation, while Masunov kept the orientation fixed (anti-parallel).

As noted above quantitative comparison is difficult. We see however, that our results agree with the analysis of Soetens : the stacked approach is clearly preferred over the coplanar at short distance. Trends between stacked , planar and a third T-shaped geometry are more complex at longer distance and could not presumably be reproduced by a continuum model.

HisP Glu- pair

Masunov reports a well-depth of about -1.0 $kcal.mol^{-1}$ at a CM of 2.8 angstroms with the SSBP potential. The GBSA model from CHARMM is attractive by about -2.5 $kcal.mol^{-1}$. PMFs on figure 4.12 are more attractive, with well-depths between 5.0 and 7.5 $kcal.mol^{-1}$. Since other PMFs including Glu- appears correct, the stronger attraction is probably caused by the protonated Histidine.

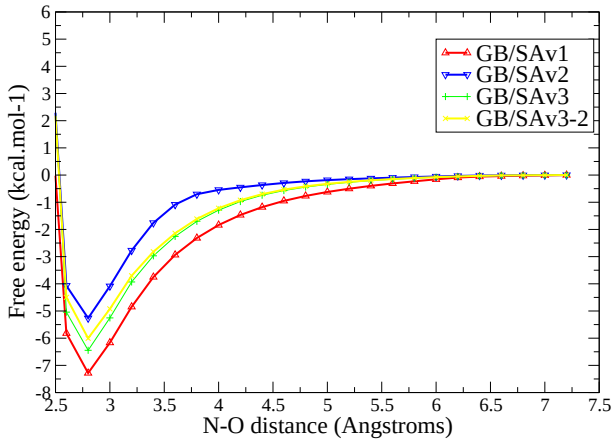

Figure 4.12: PMF HisP Glu- coplanar approach

HisD Glu- pair

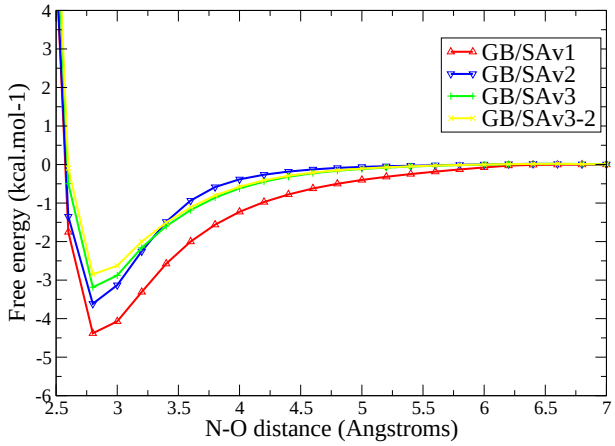

Figure 4.13: PMF HisD Glu- coplanar approach

Masunov reports a well-depth of about -2.5 $kcal.mol^{-1}$ at a CM of 3.0 angstroms with the SSBP potential. The GBSA model from CHARMM is attractive by about -1.0 $kcal.mol^{-1}$. Figure 4.13 shows that GBSAv3 and GBSAv3-2 comes closest to the SSBP results with well depths of -3.2 and -2.8 kcal.mol^{-1} respectively.

HisD HisE pair

Masunov reports a well-depth of about -1.75 $kcal.mol^{-1}$ at a CM of 3.0 angstroms with the SSBP potential. The GBSA model from CHARMM is attractive by about -2.0 $kcal.mol^{-1}$. With our models, GBSAv3 comes closest to the SSBP results with a well-depth of -3.0 $kcal.mol^{-1}$.

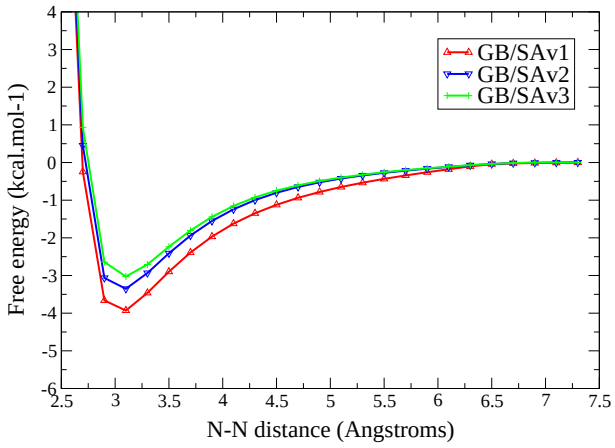

Figure 4.14: PMF HisD HisE coplanar approach

HisP HisE pair

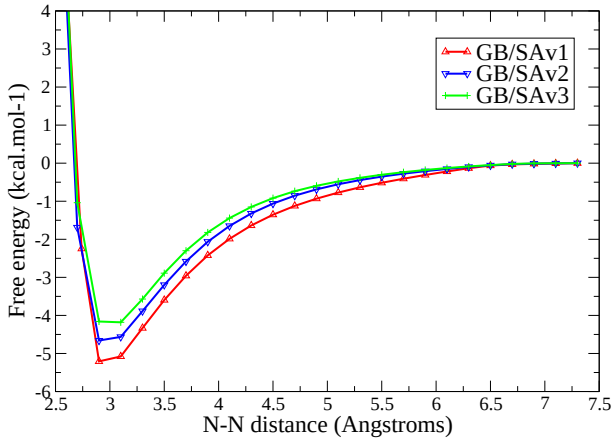

Figure 4.15: PMF HisP HisE coplanar approach

Masunov reports a well-depth of about -2.0 $kcal.mol^{-1}$ at a CM of 3.0 angstroms with the SSBP potential. The GBSA model from CHARMM is attractive by about -1.0 $kcal.mol^{-1}$. Figure 4.15 shows that GBSAv3 comes closest to the SSBP results with a well-depth of -4.0 $kcal.mol^{-1}$.

HisP HisP pair

Here two possible approaches have been considered. In the first one, the two imidazoles rings are constrained to stay in the same plane.

As can be seen on figure 4.16 all the GBSA models exhibits a litlle or no attraction. They are quite similar to the result Masunov reports for the GBSA model from

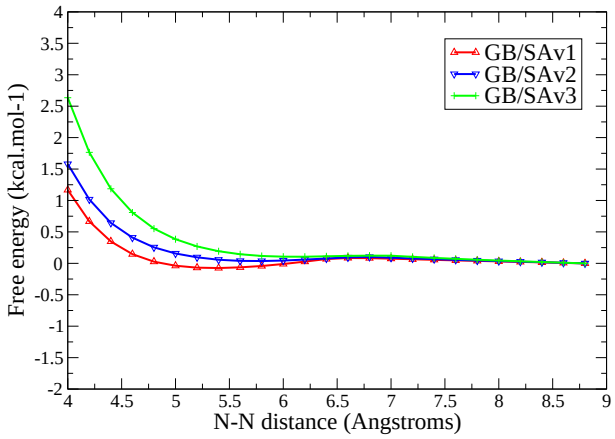

Figure 4.16: PMF HisP HisP coplanar approach

CHARMM. However the SSBP simulation exhibits a well-depth of -2.0 $kcal.mol^{-1}$ around 4.7 angstroms.

The second considered approach is a stacking between the two rings.

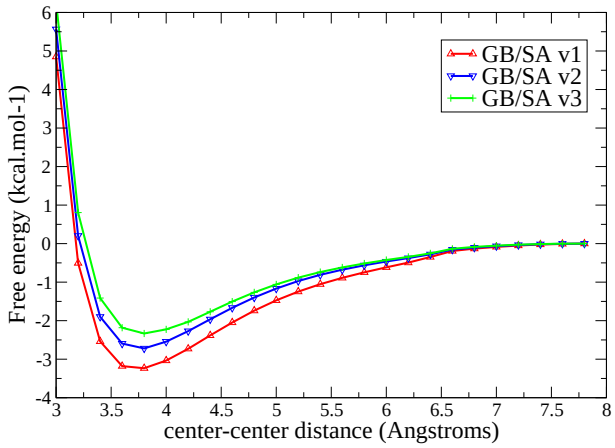

Figure 4.17: PMF HisP HisP stacked approach

Masunov reports a SSM deeper (+0.0 $kcal.mol^{-1}$) than the CM (+1.0 $kcal.mol^{-1}$ at 3.8 angstroms).The GBSA model from CHARMM is entirely repulsive. Present models behave quite differently, and yield well-depths between -2.0 and -3.0 $kcal.mol^{-1}$ at about 3.8 angstroms (see figure 4.17).

Lys+ Glu- pair-coplanar

Masunov reports a CM at 3.2 angstroms with a well-depth of about -2.2 $kcal.mol^{-1}$. The GBSA model from CHARMM exhibits a CM at 3.5 angstroms with a well depth

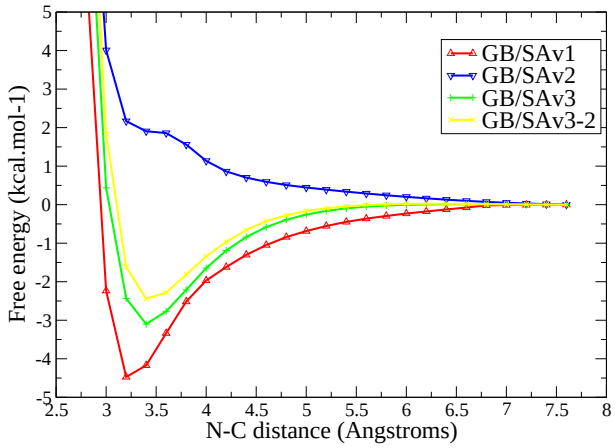

Figure 4.18: PMF Lys+ Glu- coplanar approach

of -1.75 $kcal.mol^{-1}$. Our model GBSAv1 is strongly attractive with a well-depth of -4.5 $kcal.mol^{-1}$ and a CM at 3.2 angstroms. GBSAv3 and GBSAv3-2 are in better agreement for the well-depth (-3.1 and -2.4 kcal.mol^{-1} respectively) and the CM is around 3.4 angstroms. GBSAv2 exhibits no attraction, presumably because of errors in the computation of Born radii.

4.2.4 Discussion

We have tested the ability of our solvation models to reproduce correctly the intermolecular interactions between various molecules in solution. The systems have been chosen to represent a wide range of interactions (hydrophobic, aromatic, polar, ionic). A number of PMFs appears to be too attractive with our GBSAv1. GBSAv2 was thought to perform better because it yields more negative solvation energies on average. However this new model fall short of expectations when applied to PMFs because the Born radii were not accurately computed anymore. The errors in the computation of the Born radii are caused by the use of large , greater than one, scaling factors. We also realised that an accurate prediction of the solvation energy *does not* ensure a proper balance of the intermolecular terms. Rather, the *variations* of the GB energy as the two species approach controls the strength of the final interaction. The GB energy term is controlled by the Born radii and the functional form of the GB equation. Therefore we should arguably focus more on deriving adequate Born radii, than minimising the difference between the experimental and computed solvation free energy.

With this alternative approach we are able to obtain a model, GBSAv3-2, that appears

to perform better than the others. In general our model appears more attractive than the GBSA model from CHARMM. This sometimes leads to better agreement with the SSBP simulations, sometimes worse. The CHARMM model was designed to reproduce Born radii obtained by solving the Poisson equation, and their dataset was a collection of peptides and proteins. GBSAv3 is first optimised against solvation free energies of small organics with an accurate calculation of Born radii, and then scaling factors are introduced to allow the use of the faster, PDA approximation.

We do not want to give the impression that the ability to predict the solvation energy of an isolated molecule is not important. A good solvation model should predict the correct solvation energy of an isolated molecule, and the correct variations of the Born radii as two interacting molecules are getting closer. However, because our solvation models are only an approximation to the actual physical process, an accurate treatment of the former does not imply necessarily that the latter will be correct.

While GBSAv3 should predict Born radii more accurately than GBSAv1 and GBSAv2, it sill uses a simple dependence on the SASA to model the cost associated by creating a cavity in the solvent (Gcav) and to establish dispersive-repulsive interactions with the solvent. Because of the possible problems highlighted in chapter 3 for a pure surface area based method, we have considered an alternative way to model non polar solvation.

4.3 SASI, a better modelling of solvation ?

4.3.1 Implementation and Testing

C code was added to a the program previously written by the author and used to compute various energy terms of our solvation models. The surface elements were created using the Shrake and Rupley method [66] .

Dr Zacharias was kind enough to provide his code in Fortran and a subset of his dataset (25 UA alkanes, linear, branched and cyclic). This allowed us to check that his reported results could be reproduced within two decimals of accuracy with our code.

All atom models were then created or taken from our dataset for the 25 alkanes sent by Dr Zacharias. The original SASI method works only for UA carbons and uses a probe radius of 1.8 Å . This method was modified so that it can handle any atom present in the

AMBER parm99 forcefield. It was found that setting the probe radius to 1.8 Å for the definition of the solute excluded volume did not give better results than with a probe of 1.4 Å .Therefore it was decided to use a probe of 1.4 Å which allows the computation of the SASA at the same time. SASA was computed with a Shrake and Rupley algorithm [66] with a density of 2048 points/sphere, this high value was chosen so that results from Zacharias study could accurately be reproduced. As the polarisation energy is negligible for these molecules, the GB term was not computed.

Pitera et al [67] have shown that well buried atoms that do not contribute at all to the SASA can amount for a non negligible amount of the total dispersive energy. This was done by performing thermodynamic integration on spheres of various radii filled with UA methylene atoms where buried atoms were mutated into dummy atoms. The systems studied were solvated by a box of SPC water. Having shown that the free energy difference between the 'full' and 'hollow' species was not null, they pointed that ignoring interior atoms would affect dimerisation energies.

In order to study systems similar to Pitera et al, a program has been written in C to generate a Body Centred Cubic (bcc) lattice filled with carbon atoms. From this lattice spheres of various radii can be generated . They were used to study the effect of buried atoms on the total dispersive energy of the molecule. These spheres can be thought of as very rough models of proteins.

Finally the X-Ray structure of ubiquitin was downloaded from the PDB databank (PDB code : 1ubq). An all atom model for this protein was then build and SASI calculation were performed on this more realistic example.

4.3.2 Prediction of solvation energy

Agreement with MD simulations

Zacharias found that by using a probe of 1.8 Å and a carbon-oxygen water dispersive term A_{co} of 1246.0 $kcal.mol^{-1}\mathring{A}^6$ taken from the OPLS force field, the dispersive energy for UA alkanes agreed quite well with results from Molecular Dynamics simulation of AA alkanes in explicit water with the OPLS force field [52] . A_{co} is obtained from ϵ, the Lennard-Jones well-depth energy of carbon and oxygen, and σ, the collision-diameter parameter.

$$A_{co} = 4\epsilon_{co}\sigma_{co}^{12} \tag{4.1}$$

Results including some of our AA models are shown in the table below.

Table 4.3: Alkane Water Interaction Energies with Explicit MD,UA SASI and AA SASI [a]

alkanes	U_{disp} MD [b]	U_{disp} UA SASI [c]	U_{disp} AA SASI
methane	-3.3	-3.2	-4.0
ethane	-5.4	-5.4	-5.8
propane	-7.2	-7.2	-7.4
butane	-9.0	-9.0	-9.0
pentane	-10.8	-10.8	-10.5
hexane	-12.4	-12.5	-11.9
isobutane	-8.9	-8.8	-8.8
neopentane	-10.4	-10.1	-10.0
cyclopentane	-10.0	-10.6	-9.8
cyclohexane	-11.6	-12.1	-10.9
Mean Error	0.00	0.17	0.35

[a] Energy in kcal.mol^{-1}
[b] From Gallicchio [52]
[c] From Zacharias ,[2] r_{probe} is 1.8 Å

Our results do not agree as well as with the UA models. While the dispersive energy starts more negative for methane, it ends up less negative for larger molecules. However the largest difference is no more than 0.7 kcal.mol^{-1} . It is quite possible that even better agreement could be obtained by adjusting the dispersive interaction terms of carbon or hydrogen. Although our aim is certainly not to derive a method that reproduce results of another (done with OPLS), it is interesting to see that the SASI method gives results in close agreement with those derived from an explicit solvent simulation.

Predictive power of the SASI method

We then checked if U_{disp} is a good predictor for the free energy of hydration.

Table 4.4: Predictive power of SASI[a]

Method	Coefficients	Mean Unsigned Error [b]	Correlation r
SASA - UA	1.24+0.0029*SASA	0.51	0.22
SASI - UA	-2.06+0.044*SASA+0.618*Udisp	0.12	0.96
SASA - AA	1.40+0.0020*SASA	0.51	0.17
SASI - AA	-1.57+0.043*SASA+0.815*Udisp	0.31	0.70

[a] List of alkanes used: methane, ethane, propane, butane, pentane, hexane, heptane, octane, 2,2-dimethylpropane, 2-methylbutane, 2-methylpentane, 2-methylpropane,3-methylpentane, cyclopentane,methylcyclopentane, cyclohexane, methylcyclohexane, 1,2-dimethylcyclohexane, cycloheptane, cyclooctane.
[b] in kcal.mol^{-1}

Agreement for all-atom models is not as good as with a united atom model (see fig 4.19 and fig 4.20). It is due to a less favourable prediction for branched alkanes. At the present moment it is not know why AA models perform less well, particularly 3-methylpentane which is very badly predicted (outlier on fig 4.20). Given the good agreement of Udisp between the various methods, the surface area might be the cause. However note that in both cases, the SASI method helped improve the correlation coefficient by a factor of over 4 compared to the traditional pure SASA method. Another possible cause is the fact that all the alkane have been modelled in a trans conformation which might not necessarily be the one they adopt in solution.

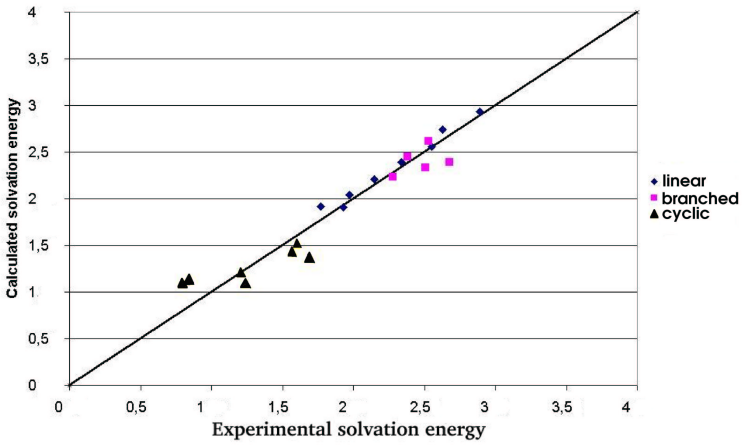

Figure 4.19: Predicted vs Experimental solvation free energy. UA-SASI model

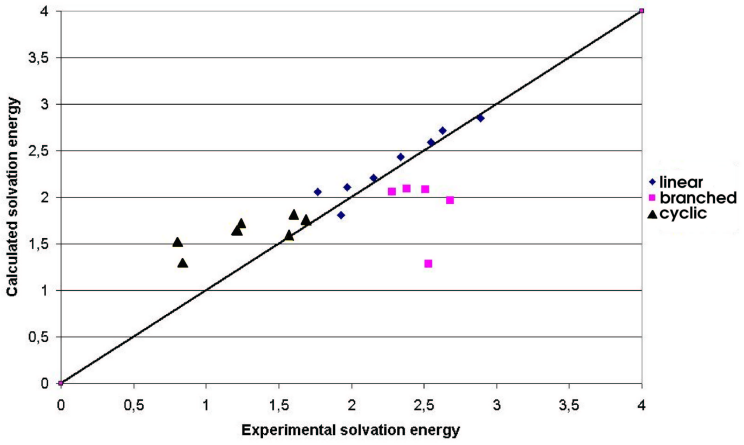

Figure 4.20: Predicted vs Experimental solvation free energy. AA-SASI model

Conformational energy differences

A model purely based on SASA exhibits a too weak conformational dependence on the free energy of hydration. In table 4.5 below the solvation free energy of different rotamers of hexane is compared with work from Gallicchio, using the formulas derived in table 4.4 .

Table 4.5: Solvation energy of rotamers of hexane [a]

Rotamer	MD [52]	SASA UA	SASI UA	SASA AA	SASI AA
ttt	0.000	0.000	0.000	0.000	0.000
tgt	-0.395	-0.012	-0.065	-0.011	-0.113
tgg	-1.181	-0.048	-0.439	-0.024	-0.251

[a] Energy difference with the all trans (ttt) state in kcal.mol^{-1}, t is trans , g is gauche

Once again, the AA model does not perform as well as a UA model, but results with a SASI method are much better than a pure SASA method. This is because changes in dispersive energy term are only weakly correlated with changes in surface area, and while both terms are reduced as hexane is 'folding', SASA is reduced more quickly. Because the empirical coefficient associated to the solvent accessible surface area in the SASI model is larger than the one in the pure SASA model, the solvation energy decreases more rapidly. As Zacharias pointed out, results from the MD simulation should not be entirely trusted, as the computed absolute free energies do not agree very well with experiment.

Finally the agreement of the different methods for the prediction of the difference in free energy of hydration between linear alkanes and their cyclic counterpart is reported in table 4.6 .

Table 4.6: Difference of free energy of hydration between linear and cyclic alkanes[a]

Alkanes	Experiment [b]	SASA UA	SASI UA	SASA AA	SASI AA
pentane-cyclopentane	1.13	0.08	1.18	0.06	0.79
hexane-cyclohexane	1.31	0.11	1.47	0.08	0.87
heptane-cycloheptane	1.83	0.14	1.65	0.10	1.20
octane-cyclooctane	2.05	0.16	1.80	0.12	1.56
Mean Error	0.00	1.46	0.16	1.49	0.47

[a] in kcal.mol^{-1}. linear alkanes are all trans.
[b] from Ben-Naim and Marcus[51]

Improvement over the traditional SASA method is striking.

4.3.3 Interaction of buried atoms with solvent

Another interesting feature of the SASI method is that atoms that do not contact the molecular surface (i.e, their SASA is null) can still establish dispersive interactions with the solvent.

In this section, the author tried to reproduce computations done by Pitera and van Gunsteren [67] using the SASI model. Unfortunately, despite our best efforts, our systems differs and this makes comparison difficult. Although a protocol similar to the one used by Pitera was adopted, our spheres have more atoms and are more dense.

Nevertheless, interesting insight can be obtained from this study.

Energy of interior atoms in spheres

Table 4.7: Dispersive energy of model systems[a]

Solute	No Atoms	No Buried	U_{disp} tot	U_{disp} buried	Ratio	Pitera [b]
sphere 6 Å	113	59	-63.5	-18.2	28.7%	18.7%
sphere 9 Å	387	211	-139.8	-35.0	25.0%	20.7%
sphere 12 Å	941	641	-254.9	-79.7	31.2%	20.4%

[a] in kcal.mol^{-1}
[b] See ,[67] supplementary information

For each of the sphere in table 4.7, the solute-solvent dispersive energy was computed for the whole molecule (U_{disp} tot) and for the buried atoms (U_{disp} buried) using equation 2.38 from the SASI method. The ratio between these two energies is also reported.

Pitera's study was done with the GROMOS96 forcefield and the number of atoms contained in his spheres diverges notably from ours. Our models are more dense by

about 30 %. Given these differences we do not expect quantitative agreement between the methods. It is interesting however, to see that the ratio of dispersive energy between buried atoms and the total number of atoms is in the range of Pitera's results. The SASI method indicates that in large, dense, spherical molecules (not unlike proteins) buried atoms could account for about 30% of the dispersive energy.

Interaction energy of dimers

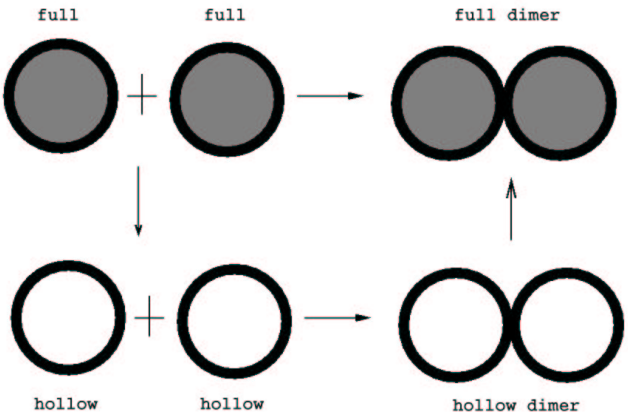

Figure 4.21: Thermodynamic cycle for the dimerisation of full and hollow spheres

A hypothetical binding process where two single spheres 'dimerise' to form an unbound complex (see figure 4.21) can be imagined . In that case Pitera has shown that there was a difference in the free energy of binding between a 'full' dimer and a 'hollow' dimer. With only a solute-solvent dispersive energy term, dimerisation is disfavoured because part of each sphere cannot establish interactions with solvent as efficiently as before. There is a difference between a 'full' and a 'hollow' dimer because interior atoms are also affected by the binding process. Therefore, a 'full' model is more penalised when it dimerises because part of its interior atoms cannot interact as efficiently as before with the solvent.

A solvation model that represents these effects only with SASA would not see any differences between the 'full' and 'hollow' dimers. This is because the dimers have the same SASA. Therefore it would in fact overestimate the stability of the 'hollow' dimer.

Table 4.8: Interaction Energy of dimers[a]

Solute	No Atoms	No Buried	$\Delta\Delta G_{bind(full->hollow)}$	Pitera [b]
spheres 9/9 Å	774	430	4.7	3.2
spheres 12/12 Å	1882	1290	10.4	5.2

[a] in kcal.mol^{-1} . Spheres are separated by 1 Å
[b] See [67]

Table 4.8 shows that there is a difference in the dimerisation energy of a 'full' and an 'hollow' dimer, and therefore that the influence of buried atoms on the solute-solvent dispersive energy should not be neglected.

The present results suggests that the SASI method favours more the 'full' solute than Pitera MD simulation does. As the systems studied here are more dense than in Pitera study, this was expected.

Note that in a complete solvation model, a term based on the reduction of SASA would favour binding, the actual sign of the binding energy would depend on the balance between the surface area term and the dispersive energy.

Application to ubiquitin

Finally, we have applied the SASI method to a protein already studied by Pitera. Unfortunately, Pitera used an united atom model with GROMOS96 parameters, while we use an all atom model with AMBER99 parameters, so comparison should be made cautiously.

Table 4.9: Dispersive energy of ubiquin[a]

Solute	No Atoms	No Buried	U_{disp} tot	U_{disp} buried	Ratio
AMBER-AA SASI	1231	680	-285.1	-74.1	26.0%
GROMOS96-UA MD[b]	761	266	-320.0	-70.7	22.4%

[a] in kcal.mol^{-1}
[b] See ,[67] supplementary information

Here again, in the SASI method buried atoms contribute more than they do in the Pitera study. However, note that our model has a higher percentage of buried atoms than in the UA model of Pitera. Given the differences in method, model and forcefield the agreement between the continuum approximation of SASI and the explicit solvent MD simulation is remarkable.

4.3.4 Discussion

It appears that the explicit inclusion of a dispersive energy term in a solvation model considerably increases the agreement between predicted and experimental solvation energy of alkanes. The SASI method also increase the sensitivity of the free energy of hydration to conformational changes by a factor of 10 over a traditional method based purely on solvent accessible surface area, bringing results in closer agreement with the explicit solvent simulations of Gallicchio [52] and the experimental solvation free energy differences between linear alkanes and cyclic alkanes. Furthermore, with the SASI method, completely buried atoms have a non zero contribution to the solvation energy and therefore the SASI method is not subject to the deficiencies of traditional continuum solvation models as pointed out by Pitera and van Gunsteren [67] , although quantitative analysis of the results should be performed with caution.

A possible concern with this method is the increased CPU cost. While there is little overhead on small molecules, it can takes longer to compute the dispersive term than the surface area for large molecules (around 2-3 times on ubiquitin with 1231 atoms). Another drawback is that at the present moment, it is not possible to use a method other than Shrake and Rupley to compute the dispersive energy term. As it makes sense to use the same method to compute the surface area, it would not be easy to use other, faster methods. It should be stressed, however, that no effort to increase the speed of the calculation has been made. Standard methods such as cutoffs and neighbour list could be used to speed up the calculation.

Finally it is important to note that to derive our model we have assumed a uniform solvent density around the solute. While the method works well for alkanes, it could not reproduce fluctuations of solvent density along a molecule containing atoms attracting more solvent than others. Unfortunately, it is not possible to replicate the study we have done on hydrocarbons for polar and charged compounds, because the polarisation energy term becomes predominant and it is not possible to separate polarisation energy and non-polarisation energy from experimental measurements.

Chapter 5
Conclusion

Most of the work done so far has been focused on deriving a solvation model that is fast and accurate enough for our purposes. One of current solvation models, GBSAv3-2 appears to meet this goal. It is interesting to note that while this model is less accurate than GBSAv2 for the prediction of free energy of solvation, it behaves much better when used to compute free energies (Potentials of Mean Force). This suggests that a good agreement with experiment does not always translate in a good solvation model and great care should be taken to ensure that the right physics goes into the right part of the model.

Our models are not as good as those developed by others to predict free energies of solvation [30,38] . It is true however, that these models use more than three times the number of parameters to yield average mean error in the range of 0.5 $kcal.mol^{-1}$. Given the current level of accuracy obtained on potentials of mean force, we do not think that it would be useful to increase the number of parameters. It might even be risky as too many parameters could overfit the model. Furthermore, in order to check that it behaves sensibly, a model with more parameters would need extensive testing using potentials of mean force.

Our tests of the SASI method appear encouraging. It is not clear, however, given that the solvation energy of biomolecules is dominated by the polarisation term, if a more expensive treatment of non polar solvation is worth it. We suggest at least that

the SASI method should be combined with a GB model to test its performance against our set of potentials of mean force.

Once this is done, it is likely that the best solvation model will be coded in a Python code currently in development in the laboratory. A dataset of drugs/protein system will have to be gathered. Free energies of binding for drug/receptor complexes will then be studied using this solvation model. If it can be shown that reliable results can be obtained, the method will be optimised to allow fast computation of free energies. Part of the receptor could be modelled as a grid potential. Parallel tempering methods could increase sampling[8–10] . The inclusion of a small number of explicit solvent molecules that are structurally important for a drug to interacts with its receptor, will also be considered.

Bibliography

[1] W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, *J Am Chem Soc*, **112**, 6127, (1990).

[2] M. Zacharias, *J Phys Chem A*, **107**, 3000, (2003).

[3] J. Drews, *Science*, **287**, 1960, (2000).

[4] J. Bajorath, *Drug Discov Today*, **6**, 989, (2001).

[5] R. D. Taylor, P. J. Jewsbury and J. W. Essex, *J Comput Aid Mol Des*, **16**, 151, (2002).

[6] W. P. Walters, M. T. Stahl and M. A. Murcko, *Drug Discov Today*, **3**, 160, (1998).

[7] J. Tomasi and M. Persico, *Chem Rev*, **94**, 2027, (1994).

[8] U. H. E. Hansmann, F. Eisenmenger and Y. Okamoto, *Chem Phys Lett*, **297**, 374, (1998).

[9] U. H. E. Hansmann, *Chem Phys Lett*, **281**, 140, (1997).

[10] U. H. E. Hansmann and Y. Okamoto, *Curr Opin Struc Biol*, **9**, 177, (1999).

[11] R. D. Taylor, P. J. Jewsbury and J. W. Essex, *J. Comp. Chem.*, **24**, 1637, (2003).

[12] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. Teller and E. Teller, *J Chem Phys*, **21**, 1087, (1953).

[13] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, **Oxford University Press**, (2001).

[14] P. Kollman, *Chem Rev*, **93**, 2395, (1993).

[15] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, **117**, 5179, (1995).

[16] W. L. Jorgensen and J. Tirado-rives, *J. Am. Chem. Soc.*, **110**, 1666, (1988).

[17] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comp. Chem.*, **4**, 187, (1983).

[18] M. Born, *Z. Phys.*, **1**, 45, (1920).

[19] J. Jackson, *Classical Electrodynamics*, **Wiley ed**.

[20] G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *Chem Phys Lett*, **246**, 122, (1995).

[21] D. Qiu, P. S. Shenkin, F. P. Hollinger and W. C. Still, *J Phys Chem A*, **101**, 3005, (1997).

[22] F. Floris and J. Tomasi, *J Comput Chem*, **10**, 616, (1989).

[23] M. J. Huron and P. Claverie, *J Phys Chem*, **15**, 2123, (1972).

[24] M. Scarsi, J. Apostolakis and A. Caflisch, *J Phys Chem A*, **101**, 8098, (1997).

[25] M. Lee, F. R. Salabury and C. L. Brooks, *J. Chem. Phys*, **116**, 10606, (2002).

[26] A. Ghosh, C. S. Rapp and R. A. Friesner, *J Phys Chem B*, **102**, 10983, (1998).

[27] B. Jayaram, Y. Liu and D. L. Beveridge, *J Chem Phys*, **109**, 1465, (1998).

[28] J. Srinivasan, M. W. Trevathan, P. Beroza and D. A. Case, *Theor Chem Acc*, **101**, 426, (1999).

[29] A. Onufriev, D. A. Case and D. Bashford, *J Comput Chem*, **23**, 1297, (2002).

[30] G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *J Phys Chem-us*, **100**, 19824, (1996).

[31] T. A. Halgren, *J Comput Chem*, **17**, 490, (1996).

[32] A. Cheng, S. A. Best, K. M. Merz and C. H. Reynolds, *J Mol Graph Model*, **18**, 273, (2000).

[33] B. Jayaram, D. Sprous and D. L. Beveridge, *J Phys Chem B*, **102**, 9571, (1998).

[34] W. Zhang, T. Hou, X. Qiao and X. Xu, *J. Phys. Chem B*, **107**, 9071, (2003).

[35] M. S. Lee, F. R. Salsbury and C. L. Brooks, *J Chem Phys*, **116**, 10606, (2002).

[36] B. N. Dominy, *Mol Simulat*, **24**, 259, (2000).

[37] B. N. Dominy and C. L. Brooks, *J Phys Chem B*, **103**, 3765, (1999).

[38] E. Gallicchio, L. Y. Zhang and R. M. Levy, *J Comput Chem*, **23**, 517, (2002).

[39] K. Osapay, W. S. Young, D. Bashford, C. L. Brooks and D. A. Case, *J Phys Chem-us*, **100**, 2698, (1996).

[40] S. B. Dixit and B. Jayaram, *J Biomol Struct Dyn*, **16**, 237, (1998).

[41] M. Scarsi, J. Apostolakis and A. Caflisch, *J Phys Chem B*, **102**, 3637, (1998).

[42] J. Srinivasan, T. E. I. Cheatham, K. P. and D. A. Case, *J.Am.Chem.Soc*, **120**, 9401, (1998).

[43] J. Srinivasan, J. Miller, K. P. and D. A. Case, *J.Biomol.Struct.Dyn.*, **16**, 671, (1998).

[44] S. R. Edinger, C. Cortis, P. S. Shenkin and R. A. Friesner, *J Phys Chem B*, **101**, 1190, (1997).

[45] B. Egwolf and P. Tavan, *J Chem Phys*, **118**, 2039, (2003).

[46] P. F. B. Goncalves and H. Stassen, *J Comput Chem*, **23**, 706, (2002).

[47] A. Masunov and T. Lazaridis, *J Am Chem Soc*, **125**, 1722, (2003).

[48] D. Beglov and B. Roux, *J Chem Phys*, **100**, 9050, (1994).

[49] T. Lazaridis and M. Karplus, *Proteins*, **35**, 133, (1999).

[50] W. Hasel, T. Hendrickson and W. Still, *Tetrahedron Comput. Methodol.*, **1**, 103, (1988).

[51] A. Ben-Naim and Y. Marcus, *J Chem Phys*, **81**, 2016, (1984).

[52] E. Gallicchio, M. M. Kubo and R. M. Levy, *J Phys Chem B*, **104**, 6271, (2000).

[53] G. Graziano and B. Lee, *J Phys Chem B*, **105**, 10367, (2001).

[54] A. Jakalian, D. Jack and C. I. Bayly, *Abstr Pap Am Chem S*, **220**, 1, (2000).

[55] J. Wang, R. M. Wolf, D. A. Case and P. A. Kollman, *In preparation*.

[56] T. H. Zhu, J. B. Li, G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *J Chem Phys*, **109**, 9117, (1998).

[57] M. Wall, *http://lancet.mit.edu/ga/*.

[58] W. L. Jorgensen, *MCPRO 1.5 Yale University, New Haven, CT*, (1996).

[59] W. L. Jorgensen, J. K. Buckner, S. Boudon and J. Tiradorives, *J Chem Phys*, **89**, 3742, (1988).

[60] W. L. Jorgensen and D. L. Severance, *J Am Chem Soc*, **112**, 4768, (1990).

[61] E. E. Tucker and S. D. Christian, *J Phys Chem*, **83**, 426, (1979).

[62] C. Colominas, J. Teixido, J. Cerneli, F. Luque and M. Orozco, *J Phys Chem B*, **102**, 2269, (1998).

[63] W. L. Jorgensen, *J Am Chem Soc*, **111**, 3770, (1989).

[64] X. Rozanska and C. Chipot, *J Chem Phys*, **112**, 9691, (2000).

[65] J. C. Soetens, C. Millot, C. Chipot, G. Jansen, J. G. Angyan and B. Maigret, *J Phys Chem B*, **101**, 10910, (1997).

[66] A. Shrake and J. Rupley, *J Mol Biol*, **79**, 351, (1973).

[67] J. W. Pitera and W. F. Gunsteren, *J Am Chem Soc*, **123**, 3163, (2001).

Appendix

Energies in kcal.mol^{-1}.

Table 5.1: Dataset - Training I

Compound	Free energy of hydration
ethane	1.80
propane	2.00
butane	2.20
hexane	2.50
heptane	2.60
octane	2.90
2-methylpentane	2.50
2,4-dimethylpentane	2.90
neopentane	2.50
cyclopropane	0.80
cyclohexane	1.20
methyl-cyclohexane	1.70
cis-1,2-dimethylcyclohexane	1.60
2-methylpropene	1.20
1-butene	1.40
cyclopentene	0.60
(E)2-pentene	1.30
1-pentene	1.70
trans-1,3-butadiene	0.60
propyne	-0.30
1-butyne	-0.20
1-hexyne	0.30
anthracene	-4.20
benzene	-0.90
ethylbenzene	-0.80
o-xylene	-0.90
m-xylene	-0.80
toluene	-0.90
2-butanol	-4.50
3-pentanol	-4.30
4-methylphenol	-6.10
ethanol	-5.00
1-octanol	-4.10
phenol	-6.60
t-butanol	-4.50
benzaldehyde	-4.00
butanal	-3.20
ethanal	-3.50
propanal	-3.40
2-butanone	-3.60
2-hexanone	-3.30
3,3-dimethylbutanone	-2.90

Table 5.2: Dataset - Training II

Compound	Free Energy of Hydration
3-pentanone	-3.40
acetone	-3.90
cyclopentanone	-4.50
acetic acid	-6.70
propanoic acid	-6.50
butanoic acid	-6.40
pentanoic acid	-7.00
butylacetate	-2.55
ethylacetate	-3.10
methylacetate	-3.30
methylbenzoate	-2.20
methylpentanoate	-2.60
propylacetate	-2.90
1,2-dimethoxyethane	-4.80
1,4-dioxane	-5.05
anisole	-2.45
diethylether	-1.60
methylisopropylether	-2.00
Tetrahydrofuran	-3.50
ammonia	-4.30
azetidine	-5.60
diethylamine	-4.10
dimethylamine	-4.30
dipropylamine	-3.70
nbutylamine	-4.30
npropylamine	-4.40
piperazine	-7.40
pyrrolidine	-5.50
trimethylamine	-3.20
2,4-dimethylpyridine	-4.90
2,6-dimethylpyridine	-4.60
2-ethylpyrazine	-5.50
2-methylpyrazine	-5.60
4-methylpyridine	-4.90
pyridine	-4.70
acetonitrile	-3.90
benzonitrile	-4.10
propionitrile	-3.90
acetamide	-9.70
N-methylacetamide	-10.10
propionamide	-9.40
benzenethiol	-2.55
diethylsulfide	-1.30
dimethyldisulfide	-1.80
ethanethiol	-1.30
methanethiol	-1.20

Table 5.3: Dataset - Training III

Compound	Free Energy of Hydration
thioanisole	-2.70
1-methylthymine	-10.40
2-methoxyethanol	-6.80
4-amino-3,5,6-trichloro-2-carboxy-pyridine	-12.00
9-methyladenine	-13.60
morpholine	-7.20
N-methylmorpholine	-6.30
p-hydroxybenzaldehyde	-10.50
trifluoroethanol	-4.30
1,1-difluoroethane	-0.10
1-bromobutane	-0.40
1-chlorobutane	-0.10
1-iodopentane	-0.10
2-bromopropane	-0.50
2-chloropropane	-0.25
bromoethane	-0.70
chlorobenzne	-1.10
fluorobenzene	-0.80
fluoroform	-0.80
hexafluoropropane	3.90
iodobenzene	-1.70
iodoethane	-0.70
methyliodide	-0.90
o-chlorotoluene	-1.15
p-bromotoluene	-1.40
p-dichlorobenzene	-1.00
trichloromethane	-1.10
1-nitrobutane	-3.10
1-nitropropane	-3.30
2-methyl-1-nitrobenzene	-3.60
nitroethane	-3.70
acetate	-80.0
propionate	-79.0
ammonium	-86.0
anilinium	-66.0
pyridinium	-59.0
methylammonium	-71.0
nbutylammonium	-66.0
N,N-dimethylanilinium	-52.0
piperidinium	-60.0
tbutylammonium	-63.0
trimethylammonium	-57.0
thiophenolate	-67.0
propanethiolate	-76.0
$(CH3)_2SH^+$	-61.0

Table 5.4: Dataset - Validation

Compound	Free Energy of Hydration
methane	2.00
pentane	2.30
isobutane	2.30
cyclopentane	1.20
ethylene	1.30
1-propene	1.30
1-pentyne	0.00
naphtalene	-2.40
phenanthrene	-3.95
methanol	-5.10
nbutanol	-4.70
npropanol	-4.80
octanal	-2.30
pentanal	-3.00
4-heptanone	-2.90
acetophenone	-4.60
hexanoic acid	-6.20
methylbutanoate	-2.80
methylpropanoate	-2.90
dimethylether	-1.90
methylpropylether	-1.70
ethylamine	-4.50
methylamine	-4.60
piperidine	-5.10
2-methylpyridine	-4.80
aniline	-4.90
butanonitrile	-3.60
N,N-dimethylacetamide	-8.50
hydrogensulfide	-0.70
methylethylsulfide	-1.40
butenyne	0.00
m-hydroxybenzaldehyde	-9.50
p-bromophenol	-7.10
2-iodopropane	-0.50
bromobenzene	-1.50
chloroethane	-0.60
chlorofluoromethane	-0.80
dibromomethane	-2.10
fluoromethane	-0.20
tetrafluoromethane	3.20
2-nitropropane	-3.10
benzoate	-76.0
dimethylammonium	-64.0
npropylammonium	-67.0
tetramethylammonium	-52.0
$CH3(SH_2)^+$	-61.0