Chapter 3

Implicit solvent models

3.1 Introduction

In this chapter a number of approaches to approximate solvent effects in molecular dynamics are outlined. The solvent degrees of freedom are not considered explicitly in these treatments, hence the name *implicit solvent models*. The interaction between solvent and solute is instead described as function of the solute coordinates solely. All effects due to the solvent, such as polar screening of charges, or even van der Waals interactions and entropy costs for cavity formations, are accounted for by suitable functions. Therefore, the effective energy function of a macromolecule, containing both intramolecular interactions and solvent-molecule coupling terms, has the form:

$$W(\vec{r^M}) = H_{mm}(\vec{r^M}) + \Delta G_{\text{solv}}(\vec{r^M}) \tag{3.1}$$

if the molecule consists of M atoms with Cartesian coordinates $\vec{r_i} = (x_i, y_i, z_i)$, with $i = 1, \ldots, M$. In section (3.6) this expression is derived based on equilibrium statistical mechanics. The solvation energy term ΔG_{solv} is usually divided into three contributions. The first is the cavity formation, that is the rearrangement of solvent molecules due to the solute. The second term is the hydrophobicity or tendency of polar solvent to avoid the contact with the non polar regions of the protein. The third contribution is the electrostatic solvation, that is the shielding effect of polarized solvent on the electrostatic interactions. The first two terms have both an entropic and enthalpic character, while the electrostatic component is purely enthalpic and is mainly considered in the models discussed in this chapter.

An implicit solvent approach in molecular dynamics drastically reduces the computational time, since the number of degrees of freedom is decreased by roughly one order of magnitude. Furthermore the equilibration of solvent induced by large conformational changes in the solute, like folding or unfolding, occurs instantaneously. In fact, during a molecular dynamics simulation the solvation free energy is updated at each step according to the new solute coordinates. This implies that the solvent reaches an instantaneous thermodynamic equilibrium with the solute. The rearrangement of explicit solvent molecules is instead slow and the solvent-solute interactions need to be averaged over relatively long times in order to provide meaningful results [63]. Hence, in case of conformational transitions studied by molecular dynamics, the use of implicit solvent has some advantage, since it provides a more efficient sampling by reducing the solute-solvent interactions to their mean field characteristics. In chapters 4 and 5 two applications featuring implicit solvent models are presented.

In many cases a detailed description of solvent-solute interactions is however most important [64], for instance in the simulation of hydrogen bonds. Thereby explicit representation of solvent molecules is required, since implicit solvent models cannot account for specific solvent-solute hydrogen bonding. In chapter 6 the pattern of intra-helical hydrogen bonds in a bacterial cytochrome c is simulated. The use of explicit water molecules allows for a correct simulation of surface intra-protein hydrogen bonds.

In the subsequent sections the continuum Poisson-Boltzmann formalism and the generalized Born approximation are presented. The former constitutes the exact theoretical treatment of electrostatic interactions in continuous media. The latter is based on the same framework and yields an approximate calculation of the solvent-induced reaction field energy.

Another approach to implicit solvent models focuses on the screening of of electrostatic interactions by solvent charges. The EEF1 energy function model [65], developed along this line, is also presented in this chapter.

The reduction from explicit to implicit solvent, in the formalism of equilibrium statistical mechanics, is then discussed in the last section.

3.2 Poisson-Boltzmann equation

The exact description of electrostatic interactions in a continuous dielectric medium is provided by the Poisson equation. Given a spatial charge distribution $\rho(\vec{r})$ in an environment with uniform dielectric constant ε, the electrostatic potential $\phi(\vec{r})$ satisfies the equation:

$$\Delta\phi(\vec{r}) = -4\pi\frac{\rho(\vec{r})}{\varepsilon} \tag{3.2}$$

If the dielectric medium is not homogeneous, that is the dielectric constant depends on the position, $\varepsilon = \varepsilon(\vec{r})$, the equation is generalized as follows:

$$\nabla \cdot [\varepsilon(\vec{r})\nabla\phi(\vec{r})] = -4\pi\rho(\vec{r}) \tag{3.3}$$

In the simple case of constant ε, the solution to Poisson equation is the well known Coulomb potential:

$$\phi(\vec{r}) = \int_V \frac{\rho(\vec{r'})d\vec{r'}}{\varepsilon|\vec{r}-\vec{r'}|} \tag{3.4}$$

A protein can be represented as a set of atomic point charges distributed in space and immersed in a low dielectricum. The surrounding solvent is given by a high dielectric medium, containing ions.

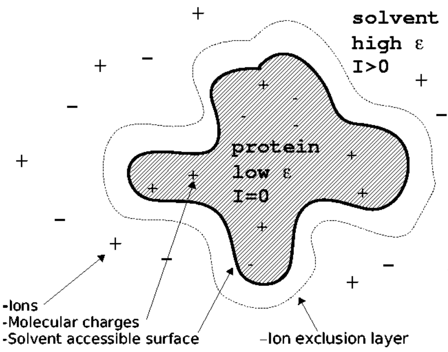

Figure 3.1: Sketch of a molecule embedded in a ionic solution, according to the continuum representation [66].

One can therefore write a Poisson equation for a protein, defining the charge distribution $\rho(\vec{r})$ as sum of delta functions representing point charges. The dielectric function assumes two different values in two space regions separated by the protein surface, namely a low value within the protein interior, like 2 or 4, and a value close to 80 in aqueous solution. The correct value of ε within the solute is matter of debate [67, 68, 69].

In order to obtain a realistic description of a solvated protein, ions in solution surrounding the molecule are also needed. Their effect is included using the ion distribution resulting from a Boltzmann statistics in a mean field approximation. This distribution is given by:

$$\rho_{ion}(\vec{r}) = \sum_s c_s(\vec{r}) q_s \exp(-\beta q_s \phi(\vec{r})) \tag{3.5}$$

if the index s runs over the number of present ion species, c_s is the local concentration of species s and q_s represents its charge. The Boltzmann constant is also present, since $\beta = \frac{1}{k_B T}$ In this way the Poisson-Boltzmann equation is obtained:

$$\begin{aligned} \nabla \cdot [\varepsilon(\vec{r}) \nabla \phi(\vec{r})] &= -4\pi[\rho(\vec{r}) + \rho_{ion}] \qquad (3.6) \\ &= -4\pi\rho(\vec{r}) - 4\pi \sum_s c_s(\vec{r}) q_s \exp(-\beta q_s \phi(\vec{r})) \qquad (3.7) \end{aligned}$$

A linearized version of the Poisson-Boltzmann equation can be written by expanding the exponential as power series in in the electrostatic potential ϕ up to first order:

$$\sum_s c_s(\vec{r}) q_s \exp(-\beta q_s \phi(\vec{r})) = \sum_s c_s(\vec{r}) q_s - \beta \sum_s c_s(\vec{r}) q_s^2 \phi(\vec{r}) + \dots \tag{3.8}$$

The term of order zero vanishes under the hypothesis of electroneutrality of the ionic solution:

$$\sum_s c_s(\vec{r})q_s = 0 \tag{3.9}$$

so that the linearized Poisson Boltzmann equation (LPBE) becomes:

$$\nabla \cdot [\varepsilon(\vec{r})\nabla\phi(\vec{r})] - 8\pi I(\vec{r})\phi(\vec{r}) = -4\pi\rho(\vec{r}) \tag{3.10}$$

using the definition of the ionic strength $I(\vec{r})$:

$$I(\vec{r}) = \frac{1}{2}\sum_s c_s(\vec{r})q_s^2 \tag{3.11}$$

3.3 Numerical solution of the LPBE and applications

The LPBE can be solved analytically only if the system geometry is simple enough. This is the case for few systems [70]. The geometry dictated by a protein shape is far too complex to be treated analytically and requires numerical algorithms for the electrostatic potential to be determined. A common solving procedure bases on a finite difference method on a grid [71, 72]. The system is mapped onto a cubic grid, with a step size l, which is usually in the order of 0.5 Å or less.

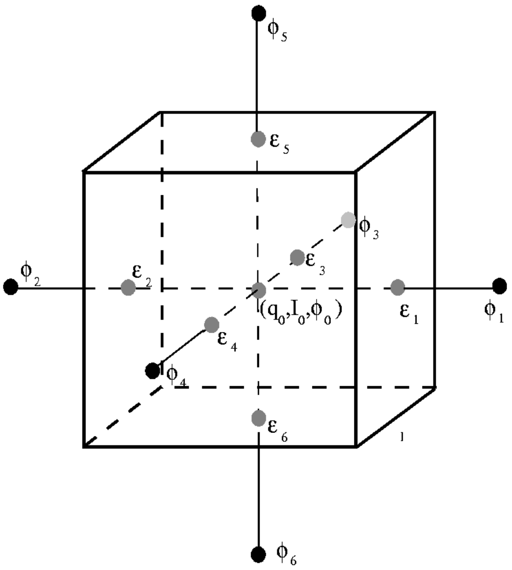

Figure 3.2: Elementary grid element in the solution of the LPBE.

In the center of the elementary cube of the grid (see fig. (3.2)) a point charge q_0 is placed, together with the ionic strength I_0 and the potential ϕ_0. On the six faces of the cube the dielectric function $\varepsilon(i)$, with $i = 1, \ldots, 6$ is defined. The electrostatic potential ϕ_i, charge q_i and ionic strength I_i in the six points corresponding to the centers of the adjacent cube, are related to ϕ_0, q_0 and I_0 by means of the linearized Poisson-Boltzmann equation, written in integral form for the elementary cube:

$$\int_{V_{\text{cube}}} \nabla \cdot [\varepsilon(\vec{r})\nabla\phi(\vec{r})] - \int_{V_{\text{cube}}} 8\pi\beta I(\vec{r})\phi(\vec{r}) = -4\pi \int_{V_{\text{cube}}} \rho(\vec{r}) \tag{3.12}$$

This is equivalent to:

$$\int_{V_{\text{cube}}} \nabla \cdot [\varepsilon(\vec{r})\nabla\phi(\vec{r})] = 8\pi\beta I_0 \phi_0 l^3 + 4\pi q_0 l^3 \tag{3.13}$$

The left side is transformed via Gauss' theorem into a surface integral:

$$\int_{V_{\text{cube}}} \nabla \cdot [\varepsilon(\vec{r})\nabla\phi(\vec{r})] = \int_{S_{\text{cube}}} \varepsilon(\vec{r})\nabla\phi(\vec{r}) \cdot \vec{n} dS \tag{3.14}$$

The discretized form of $\nabla\phi$ is the incremental ratio of ϕ from the center of one cube to the next one, so that the flux appearing in eq. (3.14) is explicitly calculated and the equation becomes:

$$\sum_{i=1}^{6} 2\frac{\varepsilon_i(\phi_i - \phi_0)l^2}{l} - 8\pi\beta I_0 \phi_0 l^3 = -4\pi q_0 \tag{3.15}$$

The resulting expression for ϕ_0 is then:

$$\phi_0 = \frac{\left(\sum_{i=1}^{6} \varepsilon_i \phi_i\right) + \frac{4\pi q_0}{l}}{\left(\sum_{i=1}^{6} \varepsilon_i \phi_i\right) + 8\pi\beta I_0 l^2} \tag{3.16}$$

Starting from arbitrary values, each point is calculated iteratively, until a required convergence is reached.

The grid resolution might be critical for a good result. Step sizes in the order of 0.3 Å would be required, what is often not feasible for a large molecular system like a protein, due to computational limits. One can instead apply a procedure called *focusing* in which, after calculating the whole system on a coarse grid (typically with step size 1 or 2 Å), smaller grids with higher resolution are generated and subsequently positioned on portions of the system. On the boundary of the small grids the potential is defined by interpolating values resulting from the coarser grid, and the calculation is then repeated inside the small grid.

The determination of the electrostatic potential on a grid suffers from an artifact, due to the self energy of the system. This quantity, which diverges for a point charge in continuum limit, is finite in a discretized system, yet it depends on the grid resolution and geometry, like number and position of grid points. Therefore, the numerically evaluated absolute electrostatic energy of a molecular system is an ill-defined quantity. The significance of Poisson-Boltzmann calculations relies on the comparison between values obtained for the same molecular system under different spatial conformations, charge states, or solvent environments. Very good results are for instance obtained in the calculation of internal pKa's and protonation as well as redox equilibria in proteins [73, 74, 75].

The description of molecular electrostatic energy by Poisson-Boltzmann equation allows for a very accurate determination of the electrostatic potential of a solvated molecule, which in turn provides the polar part of the solvation free energy, as explained next.

When a charge, distributed on a molecular system with dielectric constant $\varepsilon_{int} = \varepsilon_p$, is transferred from the uniform phase $\varepsilon_{ext} = \varepsilon_{int} = \varepsilon_p$ to a solvent with $\varepsilon_{ext} = \varepsilon_w$, it experiences a reaction field, given by the difference between the original electrostatic potential, when everywhere $\varepsilon = \varepsilon_p$, and the potential obtained as solution of the Poisson-Boltzmann equation in presence of the solvent:

$$\phi_{reac} = \phi_{sol} - \phi_{vac} \tag{3.17}$$

Thus, the electrostatic component of the solvation energy for a molecular system of atomic partial charges q_i is the electrostatic energy due to the reaction field:

$$\Delta G_{\mathrm{pol}} = \frac{1}{2}\sum_i q_i \phi_{reac}(\vec{r}_i) \tag{3.18}$$

while for a continuous charge distribution $\rho(\vec{r})$ one has:

$$\Delta G_{\mathrm{pol}} = \frac{1}{2}\int \rho(\vec{r})\phi_{reac}(\vec{r})d^3r \tag{3.19}$$

Nonpolar contributions to solvent-solute interactions are not included in the Poisson-Boltzmann formalism. Also solvent entropy, which plays an important role in protein stability, as discussed in section (1.2), is not accounted for. These limitations can be overcome by adding suitable non polar and entropy terms [63].

In molecular dynamics simulations the application of the Poisson-Boltzmann equation is in principle possible (see for instance [76]) but the cost involved in solving it directly limits its use. However, progress is being made with simulation schemes that overcome the problem of a complete calculation of the electrostatic potential at every time step, for instance by updating the forces due to solvent less frequently [77, 78] or by optimizing the repeated solution for similar conformations [79].

3.4 Generalized Born approximation

Since the solution of the Poisson-Boltzmann equation provides an accurate description of the electrostatic part of solvation, but is numerically quite expensive, there is a clear interest in exploring efficient approximations, compatible with molecular dynamics.

The *generalized Born approximation* is based on Poisson-Boltzmann theory, but the iterative self-consistent solution to the electrostatic potential is replaced by an approximate calculation of the solvent-induced reaction field energy. It is presented here following a review of Bashford and Case [80], which is suggested for further reading.

The basic assumption is that the electrostatic contribution to the solvation free energy is provided by a pairwise sum over interacting partial charges q_i in the solute alone, as given by the following expression:

$$\Delta G_{\mathrm{pol}} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right)\sum_{i,j}\frac{q_i q_j}{f_{GB}(r_{ij})} \tag{3.20}$$

where $f_{GB}(r_{ij})$ is a function that interpolates between the distance r_{ij} of the pair (i,j) valid at large distances and an “effective Born radius” R_i at short distances, defined in analogy with the Born formula for a single ion. The above expression is a formal generalization of the reaction field energy , eq. (3.18), experienced by a single spherical charge of radius a and internal dielectric constant ε_p in an implicit solvent with dielectric constant ε_w [81]:

$$\Delta G_{\mathrm{Born}} = -\frac{q^2}{2a}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right) \tag{3.21}$$

The function $f_{GB}(r_{ij})$ is defined as follows:

$$f_{GB}(r_{ij}) = \left[r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}^2}{4R_i R_j}\right)\right]^{\frac{1}{2}} \tag{3.22}$$

When $r_{ij}^2 >> R_i R_j$ the reaction field energy neglects the size of the atoms, whereas at short distance the Born radii are dominant and expression (3.22) approaches eq. (3.21).

The exact values of the Born radius R_i of charge i can in principle be determined by means of the Poisson-Boltzmann equation. Considering charge q_i alone in the solute interior, if the reaction potential is known, the corresponding reaction field energy as obtained from eq. (3.18) can be set equal to the Born energy of eq. (3.21), setting also $R_i = a$:

$$\Delta G^i_{\mathrm{pol}}\frac{q_i \phi_{reac}(r_i)}{2} = -\frac{q^2}{2R_i}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right) \tag{3.23}$$

The above equation is however not applicable for practical purposes if one wants to reduce the computational effort of solving the Poisson-Boltzmann equation. Another expression for R_i can be instead derived, which finally leads to a result, by means of diverse numerical or analytical procedures and further approximations. For the derivation of the Born radius of charge i it is useful to consider the alternative formulation for electrostatic energy in terms of electric field and displacement vector:

$$W_i = \frac{1}{2}\int \rho_i(\vec{r})\phi_i(\vec{r})d^3r = \frac{1}{8\pi}\int \vec{E}_i \cdot \vec{D}_i d^3r \tag{3.24}$$

The vectors are indexed by i since the charge q_i generates both vector fields.

At this point the *Coulomb field approximation* is invoked, which states that the displacement vector $\vec{D}_i$ of a point charge q_i maintains a Coulombic shape also in case of non-spherical dielectric volume:

$$\vec{D}_i \simeq \frac{q_i \vec{r}}{r^3} \tag{3.25}$$

so that the electrostatic energy for charge i in presence of an external dielectric constant ε_{ext} and of an internal dielectric constant ε_p becomes:

$$\begin{aligned} W_i^{\text{ext}} &= \frac{1}{8\pi}\int \frac{\vec{D}_i}{\varepsilon} \cdot \vec{D}_i d^3 r \\ &\simeq \int \frac{|D_i|^2}{\varepsilon} d^3 r \\ &= \int_{\text{int}} \frac{q_i^2}{r^4 \varepsilon_p} d^3 r + \int_{\text{ext}} \frac{q_i^2}{r^4 \varepsilon_{ext}} d^3 r \end{aligned} \tag{3.26}$$

if the integration is splitted into the internal (solute) and the external (solvent) volume part. The reaction field is given at each point by the difference between the potential in a homogeneous dielectric environment and the potential in presence of a solute, as stated in eq. (3.18). Thus,the reaction field energy for charge i is given by the difference in electrostatic energy upon transferring the charge from the uniform ε_p to the solvent ε_w, as stated in eq. Therefore:

$$\Delta G^i_{\text{pol}} = W_i^{\text{w}} - W_i^{\text{p}} = -\frac{1}{8\pi}\left(\frac{1}{\varepsilon_p} - \frac{1}{\varepsilon_w}\right)\int_{\text{ext}} \frac{q_i^2}{r^4} d^3 r \tag{3.27}$$

Setting this expression equal to the Born formula eq. (3.21) with $a = R_i$, one gets:

$$\frac{1}{R_i} = \frac{1}{4\pi}\int_{\text{ext}} \frac{1}{r^4} d^3 r = \frac{1}{4\pi}\left[\int_V \frac{1}{r^4} d^3 r - \int_{\text{int}} \frac{1}{r^4} d^3 r\right] \tag{3.28}$$

where the integral over the solvent is replaced by an integral over the whole space minus a integral over the solute. Moreover, if the point charge i is considered to be spread over a small spherical surface of radius α_i, to avoid singularities in the integration, the integral over the whole space yields $4\pi\alpha_i^{-1}$, such that the Born radius results:

$$\frac{1}{R_i} = \frac{1}{\alpha_i} - \frac{1}{4\pi}\int_{\text{int},r>\alpha} \frac{1}{r^4} d^3 r \tag{3.29}$$

where α_i is usually defined as the van der Waals radius of atom i.

The playground for different generalized Born approaches is the calculation of the integral in eq. (3.29). Many different methods have been applied, for instance the transformation of the volume integral into a surface integral [82, 83], or the evaluation of the volume integral on a grid [84]. Also analytical techniques have been developed, pairwise summations to mimic the volume integration [85] or pairwise integration using gaussian atomic functions [86].

Salt effects can also be incorporated at the level of Debye-Hückel theory. This is achieved by performing the following substitution in eq. (3.20):

$$\left(\frac{1}{\varepsilon_p}-\frac{1}{\varepsilon_w}\right)\rightarrow\left(\frac{1}{\varepsilon_p}-\frac{\exp(-\kappa f_{GB}(r_{ij}))}{\varepsilon_w}\right) \tag{3.30}$$

where κ is the Debye-Hückel screening parameter [87]. A qualitative argument to explain this substitution is that in both limits of large and small distances this expression leads to the correct equations [63].

Recent developments of the generalized Born approximation address the question of defining the solvent boundary at the molecular surface. This is a critical issue, since it may result in microscopic solvent-inaccessible voids of high dielectric in the interior of large biomolecules.

A switching function [88] was introduced to modulate the solvent-solute boundary and correspondingly the dielectric function in the generalized Born approach implemented in CHARMM and applied in chapter 5. This approach is based on an analytical definition of the molecular volume $v(\vec{r})$ as a superposition of atomic functions. The switching function $\mathcal{H}(\vec{r})$ is used to define this molecular volume. This is a smooth volume exclusion function going from zero in the interior of the solute to one in the solvent region. It is function of all atomic positions.

The switching function can be expressed as product of polynomial atomic volume exclusion functions $H_i(r)$, with $\vec{r}_i$ indicating the atomic position:

$$\mathcal{H}(\vec{r},\{\vec{r}_i\})=\prod_i H_i(|\vec{r}-\vec{r}_i|) \tag{3.31}$$

Each polynomial function H_i is defined as follows:

$$H_i(|\vec{r}-\vec{r}_i|) \;=\; \begin{cases} 0, & r\leq R^i_{PB}-w \\ \frac{1}{2}-\frac{3}{4w}\left(r-R^i_{PB}\right)-\frac{1}{4w}\left(r-R^i_{PB}\right)^3, & R^i_{PB}-w<r<R^i_{PB}+w \\ 1, & r\geq R^i_{PB}+w \end{cases} \tag{3.32}$$

Thus it is zero below the van der Waals radius R^i_{PB} of atom i, is equal to one at distances larger than this radius, and a smooth connection between the two values is done within a window of length $2w$ which is in the order of tenths of an Ångstrom. If w is set to zero, the van der Waals surface is regained. The molecular volume function used in the integration is linked to the switching function via the equation:

$$v(\vec{r})=1-\mathcal{H}(\vec{r},\{\vec{r}_i\}) \tag{3.33}$$

The solvent dielectric function, involved in the calculation of the integral in eq. (3.29), goes smoothly from zero in the protein interior to one in the bulk solvent and has the form:

$$\varepsilon(\vec{r})=1+(\varepsilon_w-1)\mathcal{H}(\vec{r},\{\vec{r}_i\}) \tag{3.34}$$

The implementation of generalized Born formalism in CHARMM31 discussed above and applied in chapter 5 also contains an empirical correction to the Coulomb field approximation, in order to reproduce deviations from the spherical symmetry of the molecular system [89].

3.5 EEF1 energy function

This is a so called *solvent exclusion model*, which provides an alternative formulation for the solvation free energy of a protein. The effect of water on a polypeptide is estimated from the solvation of each single atom, modified by the presence of other solute groups that exclude solvent. The parametrization of the function is derived from experimental data on solvation, and requires further empirical corrections. Despite the strong approximation and its empirical character, the model works for molecular dynamics, namely it succeeds in describing protein thermodynamics under native conditions, discriminating between native and unfolded conformations, and giving unfolding pathways in agreement with explicit water simulations [90]. Moreover, the speed of a molecular dynamics simulation with EEF1 water model is only 50% slower than a simulation *in vacuo*.

The solvation free energy of a given conformation $\vec{r^M}$, as in eq. 3.1, can be written as a volume integral of a density $f(\vec{r})$:

$$\Delta G_{\mathrm{solv}} = \int_V f(\vec{r})d\vec{r} \tag{3.35}$$

The density $f(\vec{r})$ contains contributions from solvent-solute and solvent-solvent interactions, both of enthalpic and of entropic origin. When the conformation of the macromolecule changes, the solvation free energy of each group is also changing because of two reasons: first, the solvent is excluded from a volume now occupied by the new conformation of the macromolecule; second, density and orientation of solvent molecules are modified in the remaining space. The latter property, which causes charge screening, is supposed to have an important role for polar residues only.

The basic assumption of the EEF1 model is that for a polyatomic solute the solvation free energy can be written as a sum over atomic contributions:

$$\Delta G_{\mathrm{solv}} = \sum_i \Delta G^i_{\mathrm{solv}} \tag{3.36}$$

Such an expression can be formally derived by considering the solute solvent interactions as an additive function of the different groups [65].

Taking into account only the contribution to solvation energy due to the solvent exclusion effect, one can write:

$$\Delta G^i_{\mathrm{solv}} = \Delta G^i_{\mathrm{ref}} - \sum_j \int_{V_j} f_i(\vec{r})d\vec{r} \tag{3.37}$$

where $\Delta G^i_{\mathrm{ref}}$ is the reference solvation free energy, that is the solvation free energy of group i in a appropriate small molecule where the group is largely exposed to solvent. The volume V_j is occupied by group j and the sum runs over all groups j surrounding i. To simplify the calculation, the integral over V_j is approximated by the product $f_i(r_{ij})V_j$:

$$\Delta G^i_{\mathrm{solv}} = \Delta G^i_{\mathrm{ref}} - \sum_{j\neq i} f_i(r_{ij})V_j \tag{3.38}$$

Therefore, the free energy of solvation of group i in the macromolecule is given by the reference value in a nearly completely solvated state minus the reduction in solvation due to the presence of the surrounding groups. The model does not take into account the finite size of water molecules, thus all cavities are instantaneously filled with solvent. However, this does not lead to a significant error if cavities are surrounded by nonpolar groups, for which the solvation energy is small [65].

The function $f_i(r_{ij})$ is assumed to be Gaussian:

$$f_i(r) = \frac{\alpha_i}{4\pi r^2}\exp\left[-\left(\frac{r-R_i}{\lambda_i}\right)^2\right] \tag{3.39}$$

and contains a number of parameters. R_i is the van der Waals radius of atom i (which is the corresponding parameter in CHARMM19 set), λ_i is a correlation length, corresponding to the length of the first solvation shell. The choice of a gaussian function with this correlation length ensures that about 80% of the solvation energy is provided by the first solvation shell, what was shown to be the case in computer simulations of water and also of Lennard-Jones fluids [65].

The α_i coefficient is related to the free energy of solvation of group i in isolation, $\Delta G^i_{\mathrm{free}}$, that is given by the integral of $f_i(r)$ over the whole space. This free energy differs from $\Delta G^i_{\mathrm{ref}}$, which is affected by the presence of a small compound linked to the atom. The parameter α_i is defined such that the solvation free energy of atoms deeply buried inside a protein is close to zero, with deviations depending on the number of atoms surrounding the given group. For the ionic groups, which are mostly exposed to the solvent, $\Delta G^i_{\mathrm{free}} = \Delta G^i_{\mathrm{ref}}$ is set.

The free energy values in the reference state, $\Delta G^i_{\mathrm{ref}}$, are derived from experimental measurements [91].

The excluded volume effect, which defines the free energy of solvation, accounts for the self energy of a charge transferred from a high to a low dielectric medium. The charge screening effect is not included in expression (3.37), but can be taken into account easily by defining a distance dependent dielectric constant, namely $\varepsilon \propto r^{-1}$, which shortens the range of electrostatic interactions. With this choice, the short-range hydrogen bonding interactions are left almost unaltered, while at large distances interactions are shielded. However, ionic side chains interact strongly with each other and with polar groups also in presence of the distance dependent dielectric constant, which is in contrast with experimental data. Therefore, a further empirical approximation is required, namely the originally charged side chains are neutralized, while their polarity is increased and a penalty is added to their solvation energy in order to prevent them to be buried in the interior of the protein.

The implementation of this model in CHARMM is related to the extended model for topology and parameters CHARMM19 (see section 2.3.1), in which nonpolar hydrogen atoms are included in the definition of the corresponding carbon groups.

3.6 Statistical mechanics of a solvated protein

Anfinsen [2] formulated the hypothesis that the native state of a protein is essentially a unique conformation corresponding to the thermodynamically most stable state. This means that the molecule during folding is able to explore the conformational space within the experimental time scales, overcoming barriers, until it reaches the equilibrium. At the equilibrium the conformations are populated according to the Boltzmann distribution.

The validity of the thermodynamic hypothesis is suggested by many experiments on protein folding. Nevertheless larger and more complex proteins might fold under kinetic control, thus reaching the kinetically most accessible state instead of the most stable one.

The statistical description in terms of energy landscape (see section 1.7) is able to represent both situations, namely thermodynamic and kinetic control of folding into the native state. In all cases one can suppose that although the macromolecular degrees of freedom might not equilibrate during a conformational change, the solvent molecules actually reach the thermodynamic equilibrium within a couple of picoseconds, or longer in the case of internal cavities, but always within the experimental time scales. This means that in the study of protein conformational stability the solvent can be considered at equilibrium and it can be averaged over the corresponding degrees of freedom.

One can describe the statistical mechanics of a solvated macromolecule using the formalism of the canonical ensemble. The following treatment is based on a recent review by Lazaridis and Karplus [92].

The macromolecule consists of M atoms with Cartesian coordinates $\vec{r_i} = (x_i, y_i, z_i)$, with $i = 1, \dots, M$ and internal coordinates $\vec{q_j}$, with $j = 1, \dots, 3M-6$. The solvent is made of N rigid molecules with coordinates $\vec{s_k}$ and $k = 1, \dots, N$. Each $\vec{s_k}$ contains the Cartesian coordinates of the center of mass and three Euler angles specifying the orientation. We can define the hamiltonian H of the interacting system of macromolecule and solvent and write the canonical partition function:

$$Q = \frac{Z}{N! \Lambda^{3M} \Lambda^{3N}} \tag{3.40}$$

where Z is the classical configuration integral, given by:

$$Z = \int \exp(-\beta H) d\vec{r}^M d\vec{s}^N \tag{3.41}$$

with $\beta = \frac{1}{k_B T}$. The solvent degrees of freedom can be easily integrated formally if the Hamiltonian is additive, namely if one can separate the solvent-solvent(ss) interactions from the molecule-molecule (pp) and from the molecule-solvent (ps) interactions:

$$H = H_{ss} + H_{ps} + H_{pp} \tag{3.42}$$

The configurational integral can be written as:

$$\begin{aligned} Z &= \int d\vec{r^M} \exp(-\beta H_{pp}) \int d\vec{s^N} \exp(-\beta H_{ss} - \beta H_{ps}) \\ &= \int d\vec{r^M} \exp(-\beta H_{pp}) \langle \exp(-\beta H_{ps}) \rangle_{ss} Z_{ss} \end{aligned} \tag{3.43}$$

where:

$$Z_{ss} = \int \exp(-\beta H_{ss}) d\vec{s^N} \tag{3.44}$$

and the ensemble average on the solvent degrees of freedom is:

$$\langle \exp(-\beta H_{ps}) \rangle_{ss} = \frac{\int d\vec{s^N} \exp(-\beta H_{ss} - \beta H_{ps})}{Z_{ss}} \tag{3.45}$$

By eliminating the solvent degrees of freedom one can define an effective energy function W given by the sum of two terms: the macromolecular interaction energy, H_{mm}, and the solvation free energy ΔG_{solv}:

$$\Delta G_{solv} = -k_B T \ln \langle \exp(-\beta H_{ps}) \rangle_{ss} \tag{3.46}$$

so that the configurational integral in eq. (3.41) becomes:

$$Z = Z_{ss} \int \exp[-\beta(H_{pp} + \Delta G_{solv})] d\vec{r^M} = \int \exp[-\beta W] d\vec{r^M} \tag{3.47}$$

The effective energy only depends on the macromolecular degrees of freedom. Energy and entropy of solvation are included if the solvent is at equilibrium. The effective energy W defines a hypersurface in the conformational space of the macromolecule which is nothing but the energy landscape.

Introducing the internal coordinates $\vec{q}_j$ instead of the Cartesian ones we can rewrite the integral in 3.47 after integrating over the six external degrees of freedom describing the translation of center of mass and the rigid rotation:

$$Z = Z_{ss} V 8\pi^2 \int \exp[-\beta \tilde{W}(\vec{q})] d\vec{q} \tag{3.48}$$

including the constant Jacobian of the coordinate transformation into $d\vec{q}$.

For a molecule at thermal equilibrium one can show that the probability of finding the system at a given configuration $\vec{q}$ is:

$$p(\vec{q}) = \frac{\exp[-\beta \tilde{W}(\vec{q})]}{\int \exp[-\beta \tilde{W}(\vec{q})] d\vec{q}} \tag{3.49}$$

One can show that:

$$\int p(\vec{q})\ln p(\vec{q})d\vec{q} = -\ln Z + \ln Z_{ss} + \ln V8\pi^2 - \beta\int p(\vec{q})\tilde{W}(\vec{q})d\vec{q} \tag{3.50}$$

Since the Helmholtz free energy of the system is obtained from the canonical partition function Q as follows:

$$A = -KT\ln Q = -k_BT\ln Z + k_BT\ln(N!\Lambda^{3M}\Lambda^{3N}) \tag{3.51}$$

one can derive the logarithm of the configurational integral Z from the expression given in eq. (3.50) and write the free energy:

$$\begin{aligned} A &= k_BT\ln(N!\Lambda^{3M}\Lambda^{3N}) - k_BT\ln Z_{ww} - KT\ln V8\pi^2 \\ &\quad -k_BT\ln\int \exp[-\beta\tilde{W}(\vec{q})]d\vec{q} \\ &= A_0 + k_BT\ln\frac{\Lambda^{3M}}{V8\pi^2} + \int p(\vec{q})\{H_{pp}(\vec{q}) - \Delta G_{solv}(\vec{q})\}d\vec{q} \\ &= +k_BT\int p(\vec{q})\ln p(\vec{q})d\vec{q} \\ &= A_0 + k_BT\ln\frac{\Lambda^{3M}}{V8\pi^2} + \langle\tilde{W}\rangle - TS^{conf} \end{aligned} \tag{3.52}$$

A_0 is the free energy of the pure solvent, the second term is the ideal contribution coming from translation and rigid rotation of the molecule. The third term is the effective internal energy, comprising interaction energy and solvation energy and the fourth is the conformational entropy. The Gibbs free energy differs from the Helmholtz free energy in that it contains an additive term pV where p indicates the pressure. Since under typical native conditions this term is constant, it can be neglected and thus the expression in eq. (3.52) is equivalent to the Gibbs free energy.

3.6.1 Native state and thermodynamic stability

If the native conformation of a protein is reached under thermodynamic control, the system is at thermal equilibrium. Therefore, under physiological conditions the phase space is defined by an ensemble of conformations obeying the probability distribution in eq. (3.49) and with a free energy defined in eq. (3.52). The phase space can be divided into two subsets N and D consisting of two configurations of the molecule, namely the folded (native) and the unfolded (denatured) conformations. Then the configurational integral Z consists of the sum of two terms:

$$Z = Z_N + Z_D \tag{3.53}$$

with

$$Z_N = Z_{ss}V8\pi^2\int_N \exp(-\beta\tilde{W})d\vec{q} \tag{3.54}$$

and the same expression is valid for subset D, where the integration is restricted to the subset D of phase space. Also the probability distributions $p_N(\vec{q})$ and $p_D(\vec{q})$ can be defined by restricting the integration in eq. (3.49) respectively to subset N or D. Using eq. (3.52) the free energy for subset N is

$$A_N = A_0 + k_B T \ln \frac{\Lambda^{3M}}{V 8\pi^2} + \langle \tilde{W} \rangle_N - T S_N^{conf} \tag{3.55}$$

and the same for subset D:

$$A_D = A_0 + k_B T \ln \frac{\Lambda^{3M}}{V 8\pi^2} + \langle \tilde{W} \rangle_D - T S_D^{conf} \tag{3.56}$$

Therefore, the free energy difference between sets N and D is:

$$A_N - A_D = \langle \tilde{W} \rangle_N - \langle \tilde{W} \rangle_D - T(S_D^{conf} - S_N^{conf}) = \Delta \langle H_{pp} \rangle + \Delta \langle \Delta G^{solv} \rangle - T \Delta S^{conf} \tag{3.57}$$

The last equation defines the Helmholtz free energy of folding under native conditions. The Gibbs free energy of folding is then obtained from the Helmholtz free energy of eq. (3.57) by adding the $p\Delta V$ term:

$$\Delta G = \Delta \langle H_{pp} \rangle + \Delta \langle \Delta G^{solv} \rangle - T \Delta S^{conf} + p \Delta V \tag{3.58}$$