Załącznik Nr 1
do Zarządzenia nr 16 Rektora UW
z dnia 18 maja 2006 r.

University of Warsaw
Faculty of Physics

Jan Malinowski
Student's book no.: 372202

Biophysical driving forces for peptide folding: insights from computer simulations of Chignolin

first cycle degree thesis
field of study Applications of Physics in Biology and Medicine
speciality Molecular Modelling and Bioinformatics

The thesis written under the supervision of
Dr Piotr Setny
Centre of New Technologies
University of Warsaw

Warsaw, september 2019

Oświadczenie kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Statement of the Supervisor on Submission of the Thesis

I hereby certify that the thesis submitted has been prepared under my supervision and I declare that it satisfies the requirements of submission in the proceedings for the award of a degree.

Date Signature of the Supervisor:

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Statement of the Author(s) on Submission of the Thesis

Aware of legal liability I certify that the thesis submitted has been prepared by myself and does not include information gathered contrary to the law.

I also declare that the thesis submitted has not been the subject of proceedings resulting in the award of a university degree.

Furthermore I certify that the submitted version of the thesis is identical with its attached electronic version.

Date Signature of the Author(s) of the thesis

Summary

W pracy tej badano przebieg zmian wartości poszczególnych potencjałów termodynamicznych w procesie zwijania się małego peptydu – chignoliny (tłum. własne). W tym celu, przy pomocy metod dynamiki molekularnej, wygenerowano jej 4 rozwinięte modele, które zostały użyte w symulacji zwijania. Wyniki tych symulacji posłużyły następnie do uzyskania wartości poszczególnych wkładów energetycznych, które zostały kolejno poddane analizie. Ostatecznie udało się pokazać, że energia potencjalna peptydu U jest siłą napędową procesu zwijania, któremu przeciwdziała wkład polarny energii swobodnej hydratacji G_p.

Key words

zwijanie białek, energia swobodna hydratacji, potencjały termodynamiczne, model *implicit solvent*, dynamika molekularna, chignolina

Area of study (codes according to Erasmus Subject Area Codes List)

13.2 Physics

The title of the thesis in Polish

Biofizyczne siły napędowe procesu zwijania: spostrzeżenia z komputerowych symulacji peptydu chignoliny

Contents

1. Introduction . . . 6
1.1. The importance of water in life processes . . . 6
1.2. Hydration free energy . . . 6
1.3. The work purpose . . . 8

2. Methodology . . . 9
2.1. Molecular dynamics . . . 9
2.2. Implicit Solvent Model . . . 11
2.3. Analysis methods . . . 13

3. Experimental system . . . 14
3.1. Chignolin peptide . . . 14
3.2. Experimental procedures . . . 15

4. Results and discussion . . . 17
4.1. Results of molecular dynamic simulations . . . 17
4.2. Variability of different energy contributions during folding . . . 19
4.3. Clusterisation of structures based on RMSD . . . 23
4.4. Energy distribution within clusters . . . 25

5. Summary . . . 27

Supplementary information . . . 28
. . .

Chapter 1

Introduction

1.1. The importance of water in life processes

Water is deeply rooted in human culture. Many philosophers ascribed almost a mystical "life giving" role to it, e.g. Thales of Miletus saw in it the ἀρχή (first principle) of the universe [1]. Today we know that human body consists of up to 60% of water [2]. This notion, together with the development of modern chemistry and understanding of water's properties as a solvent, led many scientists to the perception of aqueous environment as only a passive environment for chemical reactions.

Recent studies, especially those dealing with water in biomolecular systems, have shown the incompleteness of this former view [3]. In this new look water gained an active role in various processes from protein folding to molecular recognition [4]. According to Kauzmann's theory, water particles are the source of the hydrophobic effect, which is the driving force of the hydrophobic collapse – an event that leads proteins to their native fold [5]. It is also the main reason of biological membranes assembly. In turn, in molecular recognition, the presence of individual water particles is responsible for specific hydrogen bonding interactions, necessary to achieve specific and stable binding.

Despite years of active research, water still remains relatively poorly understood. It is difficult to pinpoint water-related effects in the laboratory experiments because there are no techniques capable of examination of the interaction details between it and protein. This happens because experimental techniques can usually determine only the properties of the system as a whole without distinguishing the solute and solvent. [6]. Thanks to computer simulations we can investigate some aspects of hydration effects that can't be determined experimentally. One such property are hydration contributions to thermodynamic free energy G, which is the thermodynamic potential whose changes are the driving force for biophysical processes.

1.2. Hydration free energy

As it was mentioned above, proteins exist in aqueous environment which actively helps them getting and stabilizing the right fold. But how exactly this happens? According to the thermodynamics, Gibbs free energy G, the thermodynamic potential that characterises systems in NpT conditions (fixed pressure p, temperature T and number of particles N), is widely used in the biomolecular field

$$G = U + pV - TS,$$

where U is the internal energy, p and V are subsequently pressure and volume, S is the entropy and T is the temperature of the system. Considering folding of a protein in water solution this can be seen that the protein seeks its native state at the expense of the pressure, volume and entropy changes of the whole system. Solvent contributions to this process can be determined by considering the thermodynamic cycle presented in the Fig. 1.1.

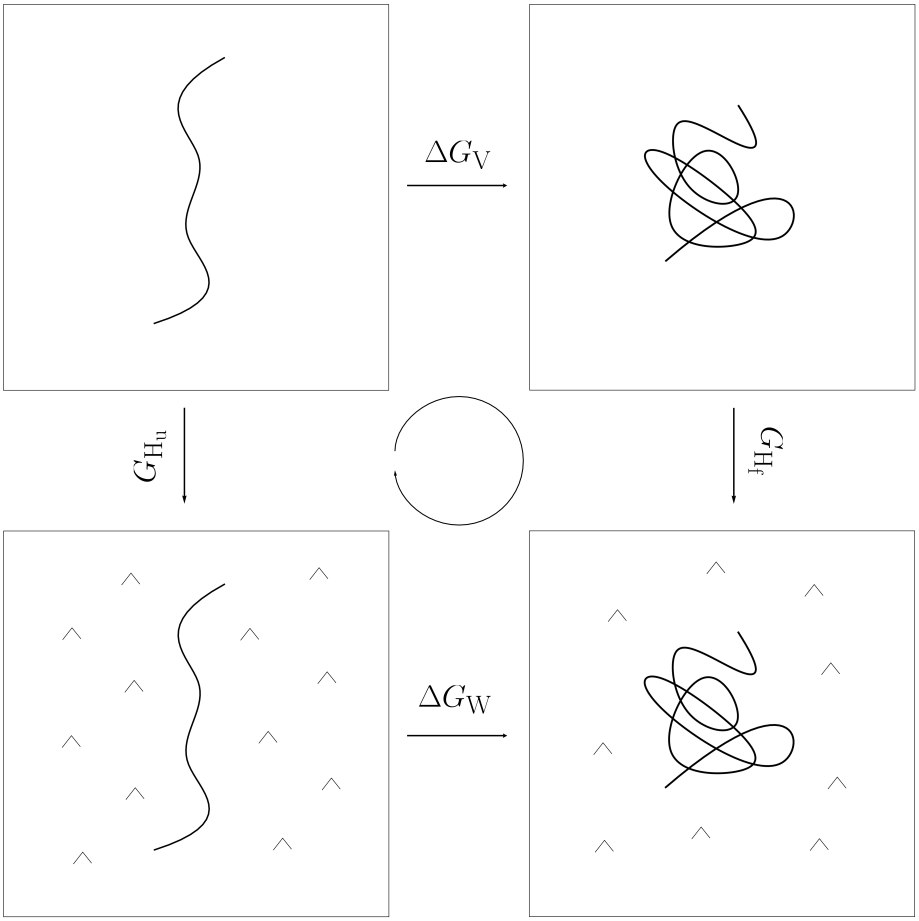

Figure 1.1: Protein folding thermodynamic cycle in the presence and absence of the solvent. Left boxes contain the unfolded proteins, whereas right folded ones. Solvent is absent in the upper boxes and present in lower ones.

Due to differences in protein conformational distribution in vacuum and in solution, this cycle is strictly appropriate for individual microstates, that is for singular conformations representing folded and unfolded states. From the cycle above, a change in free energy during folding in water, ΔG_W can be extracted as

$$\Delta G_{\mathrm{W}} = \Delta G_{\mathrm{V}} + G_{\mathrm{H_f}} - G_{\mathrm{H_u}}, \tag{1.1}$$

where G_{V} is free energy change during folding in vacuum, and $G_{H_{\mathrm{f}}}$ and $G_{H_{\mathrm{u}}}$ are hydration free energies of folded and unfolded microstates, respectively.

Determination of the change of the free energy during folding, ΔG_{fold} can be done by averaging the the results from the Eq. 1.1. Using fixed protein models taken from trajectory snapshots, leads to the formula for the total free energy change during folding

$$\Delta G_{fold} \cong \Delta G_{\mathrm{V}} + \Delta G_H =< U^{\mathrm{f}} > - < U^{\mathrm{u}} > + < G_{\mathrm{H}}^{\mathrm{f}} > + < G_{\mathrm{H}}^{\mathrm{u}} > -T\Delta S. \tag{1.2}$$

Here the entropic term accounts for the difference in phase-space volume sampled by folded and unfolded states. This equation allows us to perceive the behaviour of the peptide and it's surroundings as the result of the competition between different energetic contributions. But which of them actually stabilise the native structure? How are these contributions behaviours related to each other?

1.3. The work purpose

The purpose of this work is to catch a glimpse of understanding of these biophysical driving forces by examining the behaviour of the Chignolin peptide during it's folding process. To achieve this goal few actions were planned. Firstly, molecular dynamics simulations were carried out in order to get several folding trajectories. The above mentioned energy contributions were determined by post-processing of those results. Lastly, the values of G_{H} and U were analysed during the peptide folding, with the aim to analyse their changes between the unfolded and folded states.

Chapter 2

Methodology

2.1. Molecular dynamics

Simulations of large biomolecules in *explicit solvent*, long enough to give insight into large-scale conformational changes, were hardly possible to carry out until XXI century due to the high computational cost. Dynamic development of computer hardware, especially central and graphical processing units, is gradually enabling more extensive theoretical research on many biological processes e.g. protein folding and binding.

Classical molecular dynamics is one of the methods which allows sampling the phase space of systems consisting of many atoms using Newton's equation of motion

$$\vec{\mathbf{F}}_i = m_i \vec{\mathbf{a}}_i,$$

where $\vec{\mathbf{F}}_i$ is a cumulative force acting on atom i, m_i and $\vec{\mathbf{a}}_i$ are its mass and acceleration, respectively. Solving this equation allows finding the accelerations, velocities and positions of all elements in the system in the subsequent moment. Forces acting on each particle are determined by the calculation of the potential gradient.

The exact formula of this potential can be determined by means of quantum mechanics simulation, but it is also too computationally demanding for the protein-sized systems. Owing to the Born-Oppenheimer approximation it is possible to assume that the electron density depends on the momentary positions of nuclei by integration of the electronic degrees of freedom in calculations, and thus focusing on the nuclei translations [7]. Moreover the usage of Hellmann-Feynmann theorem allows the calculation of forces acting on the atomic nuclei, using classical Coulomb's electrostatic law, from electron density and the rest atomic nuclei [8–11]. Subsequent assumption that molecules remain close to their energy minima and that the chemical bonds are stable and thus can not be broken during the simulation, forms the basis for the introduction of simple potential energy models as a function of nuclei positions:

$$U_{total}(\vec{\mathbf{r}}) = U_{bond}(\vec{\mathbf{r}}) + U_{angle}(\vec{\mathbf{r}}) + U_{dihedral}(\vec{\mathbf{r}}) + U_{elec}(\vec{\mathbf{r}}) + U_{vdW}(\vec{\mathbf{r}}).$$

Here, subsequent potential contributions correspond to U_{bond} - chemical bond stretching, U_{angle} and $U_{dihedral}$ - deformations of the planar angles between atoms and dihedral angles, respectively, U_{elec} - electrostatic energy, and U_{vdW} - steric repulsion and dispersion between atoms [12]. For the first two terms, the harmonic potential is used, whereas the third one is defined as a simple periodic function. Examples of these potentials are listed below

$$U_{bonded} = \frac{1}{2}k_b(r - r_0)^2,$$
$$U_{angle} = \frac{1}{2}k_a(\theta - \theta_0)^2,$$
$$U_{dihedral} = \frac{1}{2}\sum_{i=1}^{n} V_i[1 + \cos(i\omega - \gamma)],$$

where k_b, k_a are the force constants, $r - r_0$ and $\theta - \theta_0$ are corresponding distances from the potential minima, i depicts one of the n minima with the V_i rotational barrier, ω the torsional angle, and γ the phase factor. Electrostatic contribution U_{elec} is calculated using the classical Coulomb's formula $\frac{-1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}}$ for the interaction between atoms i and j, having partial charges q_i, q_j, r_{ij} the distance between them, and ε_0 is the absolute permittivity of the vacuum. For the modelling of short range repulsive and dispersion interactions, Lennard-Jones potential [13] is used

$$U_{\mathrm{LJ}} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right],$$

where ε_{ij} is the depth of the potential well and σ_{ij} represents the distance at which the potential has a zero place, i and j are atom indices. Mathematical formulas of all the written above potentials with the set of parameters is called a force field. There are a few force fields e.g. AMBER, CHARMM and GROMOS which differ in exact formulation of these potentials. Necessary parameters are determined experimentally or by means of quantum mechanics *ab inito* calculations.

After the determination of resultant force acting on each atom, accelerations are applied and new coordinates of molecule are found using one of the motion equation integration methods. The widely used example of such is the velocity Verlet algorithm [14], which allows finding new position and velocity of an atom i with the usage of equation below

$$\vec{\mathbf{x}}_i(t + \Delta t) = \vec{\mathbf{x}}_i(t) + \vec{\mathbf{v}}_i(t)\Delta t + \frac{1}{2}\vec{\mathbf{a}}_i(t)\Delta t^2,$$
$$\vec{\mathbf{v}}_i(t + \Delta t) = \vec{\mathbf{v}}_i(t) + \frac{\vec{\mathbf{a}}_i(t) + \vec{\mathbf{a}}_i(t + \Delta t)}{2}\Delta t,$$

where $\vec{\mathbf{x}}_i(t)$, $\vec{\mathbf{v}}_i(t)$, $\vec{\mathbf{a}}_i(t)$ denote subsequently position, velocity and acceleration of the atom i at time t, and Δt depicts the length of the simulation timestep.
In order to represent physical conditions some thermodynamic properties have to have specified values. Sets of the parameters which are chosen to be fixed determine statistical ensembles. Three of these typically considered ensembles for MD simulations are:

- microcanonical (NVE) with fixed number of particles, volume and energy of the system,
- canonical (NVT) with specified temperature, volume and number of particles,
- isothermal-isobaric (NpT) with fixed temperature, pressure and number of particles.

This is accomplished using additional algorithms representing thermostats and barostats applied to the system. A sample thermostat used in this work is the velocity rescaling algorithm which rescales the velocities of atoms to keep the constant temperature. The role of the barostat is to scale the system volume in order to keep pressure of the system close to the

desired value. According to the ergodic theorem, mean values of the properties of the stationary process for the sufficiently long sampling converge to their respective statistical ensemble averages, and thus are connected via statistical physics with well defined thermodynamic properties [15]. This validates the whole molecular simulation based approach.

2.2. Implicit Solvent Model

Despite the great technological improvements of the last few years, simulations of complex molecules in their natural, aqueous environment are still computationally challenging. One of the factors contributing to this is the need to include into calculations each water molecule independently. Since there are usually orders of magnitude more H_2O particles than solute ones, this so-called *explicit solvent* approach, although guarantees high scientific reliability, does so at the expense of high computational cost. One of the ideas of overcoming this limit was the creation of *implicit solvent* models. In this approach solvent is represented as continuous medium instead of explicit atomic particles, which is illustrated in the figure below.

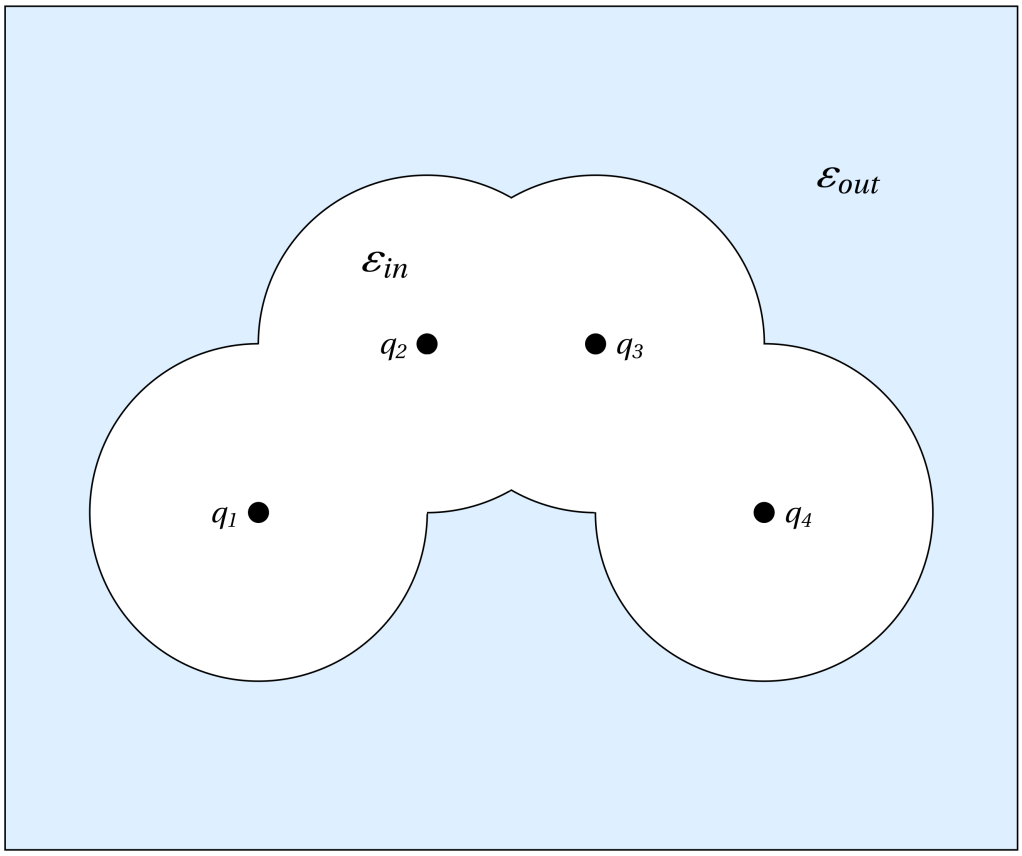

Figure 2.1: Illustration of implicit model solvent in simulation box with molecule. q_i, where $i = 1, 2, 3, 4$ are point atomic charges and ε_{in} and ε_{out} are dielectric constants inside and outside the molecule, respectively. The line between them is the boundary which separates low and high dielectric regions.

Although the treatment of water in this manner leads to the reduction of reliability, it allows efficient examination of few phenomena e.g. screening of multiple ligands binding to a receptor [16].

Hydration free energy can be divided to sum of two parts: nonpolar, connected with work that has to be done in order to create a cavity in a solvent to make place for the new particle, and polar, related to the work needed to transfer atomic charges into it. It is typically postulated that the former one depends on the area of solvent accessible surface, determined using a pseudoatomic sphere rolled on the solute surface, and can be calculated from the formula

$$G_{np} = \gamma A,$$

where γ is the surface tension and A is the surface area of the molecule.

One possible solution for the polar term is given by the Poisson-Boltzmann model. According to the electrostatic theory, electrostatic potential at point $\mathbf{r}$ can be found by solving Poisson equation

$$\nabla \cdot \varepsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) = -\rho(\mathbf{r}), \tag{2.1}$$

where ε is the dielectric constant, ϕ is the electrostatic potential and ρ is the charge density. Adding the dissolved electrolytes, whose local density is given by Boltzmann distribution, to the system, leads to the equation of Poisson-Boltzmann model

$$\nabla \cdot \varepsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) = -\rho(\mathbf{r}) - \sum_i n_i^0 q_i e^{-\beta q_i \phi(\mathbf{r})}, \tag{2.2}$$

where n_i^0 is the average number density of the ions of type i, q_i is their charge and $\beta = 1/k_B T$, where k_B is the Boltzmann constant and T is the temperature.
Finally for the weak electrostatic field it is possible to expand the function inside exponent in Eq. 2.2 into Taylor series and cut off the $\mathcal{O}(n^2)$ components, obtaining the formula below

$$\nabla \cdot \varepsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) = -\rho(\mathbf{r}) + \sum \beta n_i^0 q_i^2 \phi(\mathbf{r}), \tag{2.3}$$

which is the linearized version of Poisson-Boltzmann equation [17]. Example of the program, which solves it, is the **Adaptive Poisson Boltzmann Solver** [18, 19]. It uses information about the atomic structure, charges and van der Waals radii of atoms, to integrate the equation above in iterative manner. The structure under study is placed on a grid and with each step the electrostatic potential is calculated more accurately. The adaptive term in the method name stems from the iterative character of determining of the grid resolution and boundary conditions for electrostatic potential in the presence of the solute.

In order to determine the polar contribution the Eq. 2.3 and Eq. 2.1 are solved for the molecule in solvent and vacuum, respectively. The knowledge of electrostatic potentials allows the determination of the polar contribution G_p as a difference of respective electrostatic energies

$$G_p = \frac{1}{2} \sum_i q_i (\phi_W(\mathbf{r}_i) - \phi_V(\mathbf{r}_i)),$$

where ϕ_W and ϕ_V are potentials at point i in solvent and vacuum, respectively, and q_i is the atomic partial charge at this point. G_p summed with G_{np} gives the hydration free energy G_h. [20].

2.3. Analysis methods

For the determination of the similarity between different models of one structure, two measures were chosen: root-mean-square deviation of atomic positions (RMSD) and, so called, Q parameter. The first one corresponds to standard formula

$$\mathrm{RMSD}(a,b) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(a_{i_x}-b_{i_x})^2+(a_{i_y}-b_{i_y})^2+(a_{i_z}-b_{i_z})^2\right]}.$$

Here a and b denote two different models of one structure having n atoms, a_{i_k} and b_{i_k} are proper coordinates of atom i, where $k = x, y, z$. The parameter Q represents the fraction of defined native contacts between atoms in the structure. In this work the *Soft Cut* version of it was used from the MDAnalysis package [21, 22], which calculates Q value relative to the reference structure using the formula below

$$Q(B) = \frac{1}{n}\sum_{(i,j)}\frac{1}{1+e^{\beta(r_{ij}(B)-\lambda r_{ij}(A))}}, \quad (2.4)$$

where $r_{ij}(A)$, $r_{ij}(B)$ are the distances between the residues i and j (minimal distance between their atoms) in models A and B, and β, λ are the parameters [23].

Chapter 3

Experimental system

3.1. Chignolin peptide

Chignolin is an oligopeptide designed based on the scaffold of the G-peptide, which consists of 10 amino acid residues (GYDPETGTWG). It is the smallest known peptide with unique native tertiary fold (Fig. 3.1). It is stabilised by the hydrophobic interaction between Tyr2, Pro4 and Trp9 and 4 hydrogen bonds between Asp3, Glu5, Gly7 and Thr8 [24]. Chignolin is often used as a model protein in theoretical studies due to its high folding rate and its small size that contribute to relatively low computational cost of folding simulations. In this work 1UAO Protein Data Bank (PDB) record [25], with peptide model created by means of NMR was used as a source of the folded state geometry.

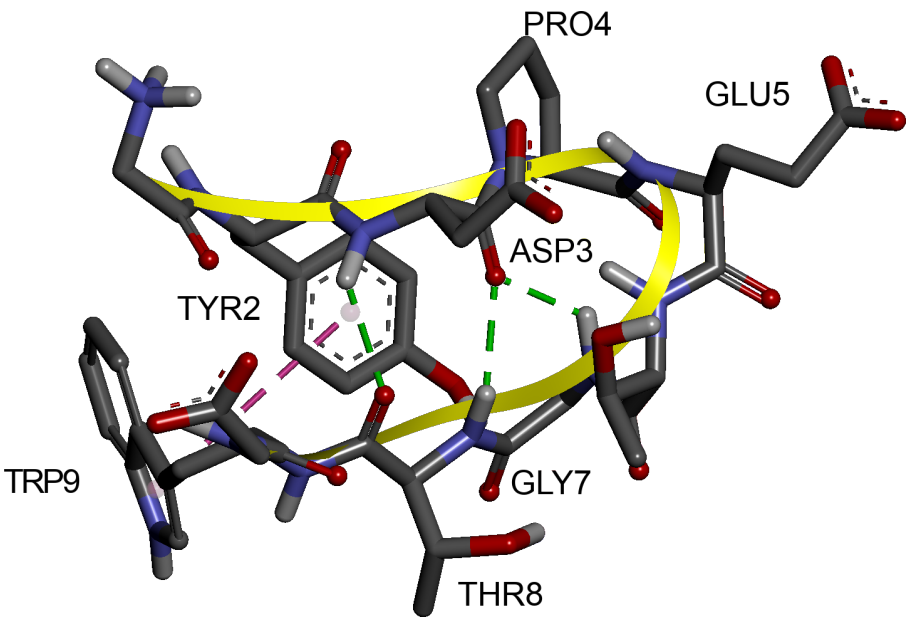

Figure 3.1: One of the models of Chignolin from 1UAO pdb record. Hydrogen bonds and hydrophobic interactions are represented by green and cyan dashed lines, respectively. For clarity, there are only 3 of 4 hydrogen bonds and 1 of 2 hydrophobic interactions denoted in this figure [26].

3.2. Experimental procedures

The NMR structure was selected from the 1UAO PDB record for further work. The model was inspected for possible abnormalities using the MolProbity server [27,28]. Cubic simulation box length was set in such a manner that the distance between any of the peptide's atom from the wall was at least 15Å . Rigid TIP3P water model [29, 30] was added to the simulation box using the gmx solvate tool. Ions were inserted to first neutralise the system and then to reach the concentration of $0.15\frac{\text{mol}}{\text{l}}$ using the gmx genion program. The cutoff for van der Waals interactions was set to 10 Å and electrostatic potentials were calculated explicitly up to 10 Å . Beyond this distance the Particle Mesh Ewald method was used to determine the long-range electrostatic interactions [31, 32]. All simulations were carried out using the GROMACS package [33–40] with AMBER03 force field [41], periodic boundary conditions and 2 fs timestep. Before molecular dynamics simulations, energy of the system was minimised using the steepest descent algorithm for 1000 steps.

The purpose of the first molecular dynamics simulation was to produce a set of unfolded models, whose subsequent folding would be examined. For its first stage Berendsen barostat [42] and velocity rescaling thermostat [43] were used to keep the NPT values constant, and for the second part just the latter one because of volume fixing in NVT ensemble. Parameters of this simulation are listed below in Table 3.1.

Table 3.1. Parameters of the unfolding simulation

stage	**ensemble**	**temperature**	**duration**
1	NPT	300 K	1 ns
2	NVT	1000 K	1000 ns

Subsequently, 4 random unfolded structures from the simulation stage 2 were selected for the folding simulations. During the cooling procedure only velocity rescaling was used and finally, the production phase at $T = 310$ K was carried out using the Nose-Hoover [44, 45] thermostat and Parinello-Rahman barostat [46, 47]. Parameters used in the second set of molecular dynamic simulations are shown in Table 3.2.

Table 3.2. Parameters of the second simulation

stage	**ensemble**	**temperature**	**duration**
1	NVT	310 K	0.5 ns
2	NPT	310 K	1000 ns

RMSDs of heavy atoms (relative to the model 0 from NMR structure) and potential energies of the peptide models were calculated using the rms and energy tools from the GROMACS package.

For the calculation of hydration contributions, peptide structures extracted from simulation frames, saved every 10 ps, were prepared for the Adaptive Poisson-Boltzmann solver (APBS) calculations in pqr format by adding the information about atomic radii and partial charges [48, 49]. The dimensions of the fine grid were chosen to be equal to the maximal distance of two atoms in examined structure plus 12 Å margins on each side, and the coarse mesh domain lengths were set to be 18 Å above the fine ones. The polar contribution, G_p, was calculated using the 161 grid points for grid-based discretisation, 1.1 Å radius sphere was used for the determination of molecular surface, dielectric constants of the solute and

the solvent were set to 1 and 78.5, respectively. The nonpolar contribution to hydration free energy was calculated assuming its linear dependence on peptide solvent accessible surface. The regression coefficient was calculated using the information about surfaces and free energy changes between four different states of Chignolin: unfolded, prefolded, misfolded and folded, whose relative nonpolar free energies were determined earlier in full-atom simulations [50]. A plot presenting this procedure is placed in Supplementary section. For those calculations the 0.65 Å radius of solvent molecules was chosen to determine the surface size.

Chapter 4

Results and discussion

4.1. Results of molecular dynamic simulations

Trajectories obtained during 4 simulations were used to calculate RMSDs of heavy atoms atoms, against the chosen reference model from 1UAO PDB record, which are plotted in Fig. 4.1.

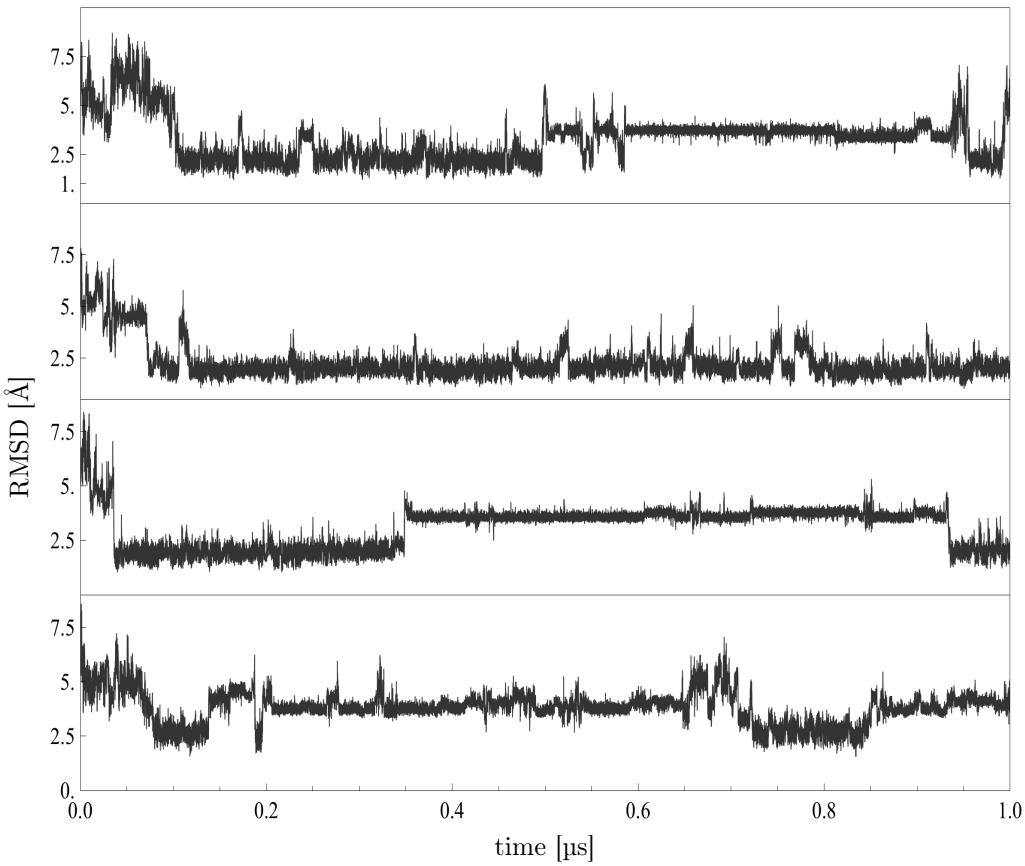

Figure 4.1: RMSD against the reference model during four simulations.

Through the analysis of RMSDs shown in Fig. 4.1 we can observe several folding and unfolding events of Chignolin peptide. Visual analysis of the RMSD plot allows us to distinguish 3

states: stable folded (RMSD about 2.5 Å), stable misfolded (RMSD between 2.5 Å and 5 Å) and unfolded (RMSD above 5 Å). During the simulation we can observe repetitive probing of all of the above states which suggests that sufficient sampling of configuration around the folded state was achieved. In the first, third and fourth simulation the peptide folded to native structure twice, whereas during the second simulation only once. The times needed to find the native fold and its lifetimes are presented in Table 4.1.

Table 4.1. Average times to fold and native state lifetimes

#	**time to fold** [ns]	**native state duration** [ns]
1	94	393/54
2	63	+
3	26	310/+
4	70	57/123

Here # denotes the simulation number, and + means that the peptide was folded until the end of the simulation.

An average time to fold the peptide to the native state from elongated structure was 63 ± 28 ns, and an average lifetime of the native state was 187 ± 155 ns. Successful folding of the peptide in these simulations allows to examine the changes of free energy during this process.

4.2. Variability of different energy contributions during folding

In order to examine the behaviour of the driving forces for peptide folding, the results obtained by APBS (hydration free energy contribution) and gmx energy (protein potential energy in vacuum) are presented in Fig. 4.2.

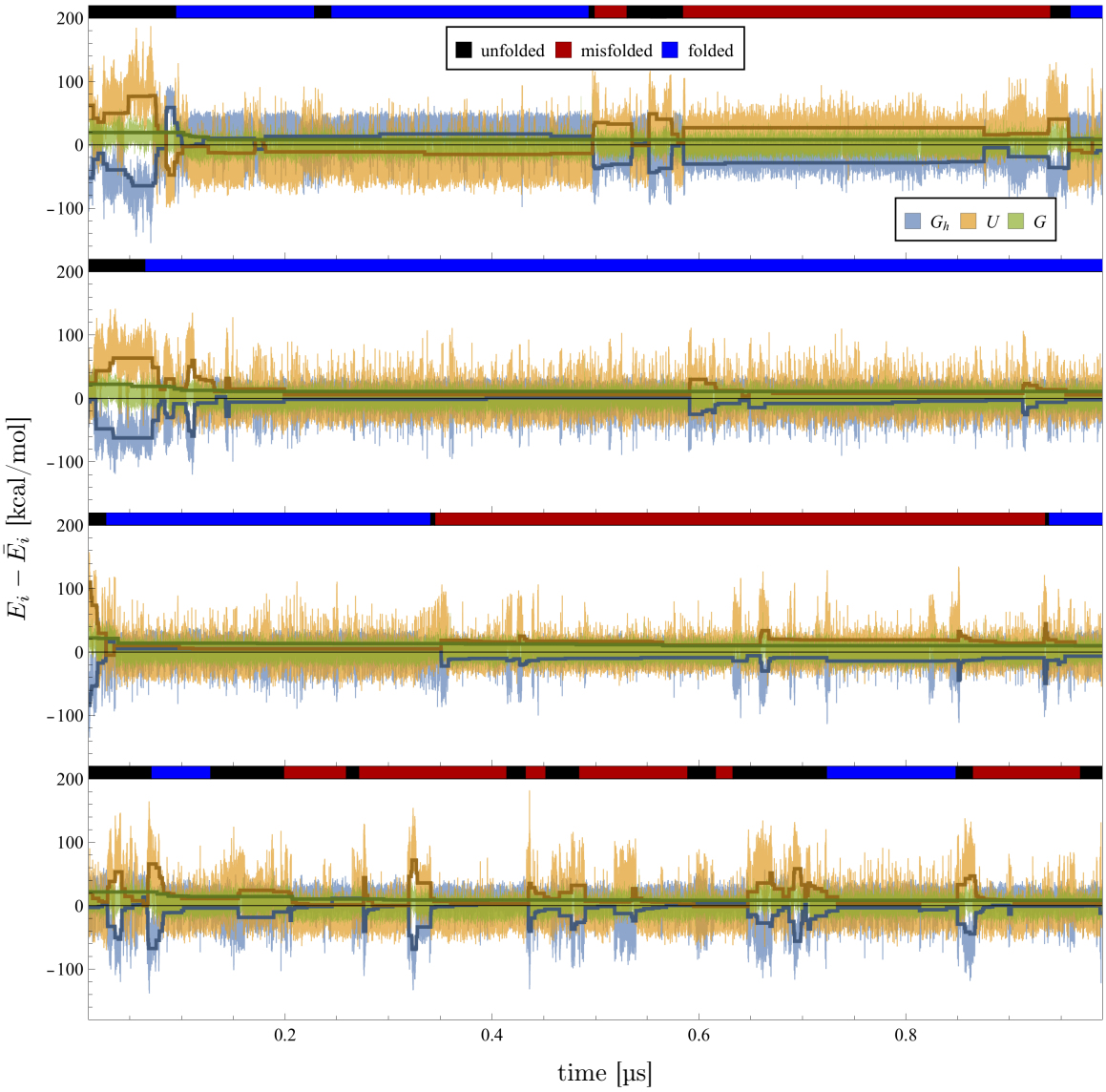

Figure 4.2: Various energy contributions changes during the folding simulations.

In this plot we observe concomitant changes of hydration free energy G_h and the potential energy of the peptide U. In order to improve the illustration of the folding process, rectangular bar was added to the plot, whose colours correspond to the three previously distinguished states. The clusterisation of peptide conformations during the simulations and the assignment of frames to respective clusters are explained in Chapter 4.3. One can notice that during folding and unfolding events in all simulations there is a visible growth of G_h with a simultaneous decrease of U, whose changes largely cancel out, leaving relatively little variation in the total

free energy G.

Analysis of this plot suggests that G_h and U energy contributions are anticorrelated. The dependence between G_h and U from all simulations is plotted in Fig. 4.3.

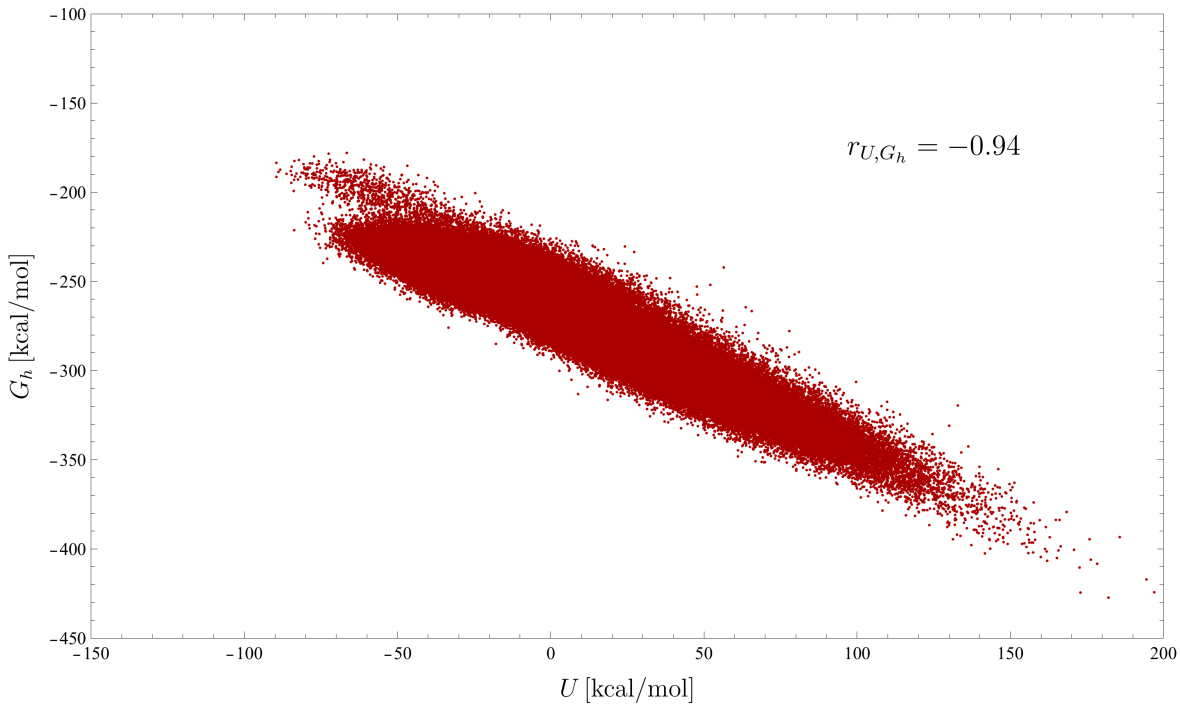

Figure 4.3: $G_h(U)$ from all simulations.

Indeed, as illustrated in Fig. 4.3, the highly negatively correlated dependence of these two energy contributions is observed with the Pearson's correlation coefficient $r_{U,G_h} = -0.94$. Along with the decrease of the potential energy of peptide U, G_h is increasing, which suggests that the folding takes place at the expense of the potential energy and gain in hydration free hydration energy. For better understanding of the behaviour of these driving forces, it was decided to examine their distributions along a reaction coordinate chosen to represent the folding process. RMSD is the most obvious choice in biomolecular research. In order to check its physicality in illustrating the folding process, it was compared with another measure called **Q** parameter, which reflects the fraction of defined native contacts between residues in the reference structure (Chapter 2.3). One of the structures obtained by NMR was used as the reference model for those calculations. Parameters λ and β in Eq. 2.4 were set to 1.25 and default (5.0), respectively. More detailed information about this choice is given in Supplementary section.

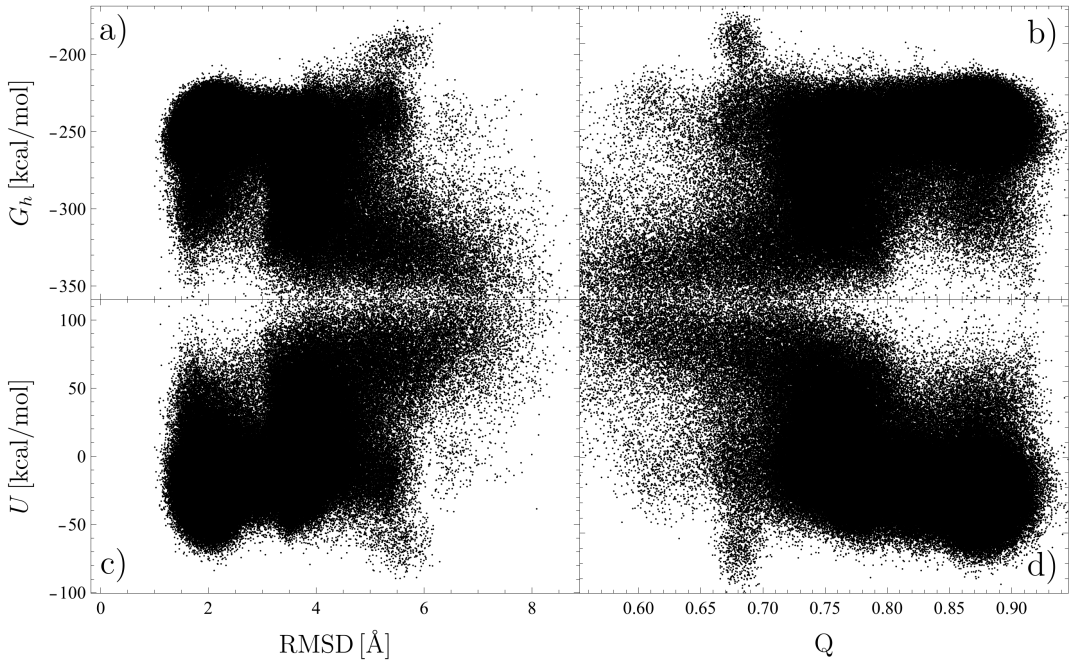

Figure 4.4: Dependence between: a) G_h and RMSD, b) G_h and Q, c) U and RMSD, d) U and Q, from all simulations.

In this figure one can notice that the dependencies of G_h and U from the RMSD almost reflect the ones from Q parameter. The fraction of native contacts is correlated with G_h and negatively correlated with U, and, reversely RMSD is correlated positively with U and negatively with G_h. Moreover, we can find similar shapes on these two kinds of plots which demonstrates that the two measures similarly represent the folding process, at least in the case of chignolin. From this picture we can see that the U is the driving force for folding and G_h opposes it.

4.3. Clusterisation of structures based on RMSD

All trajectories obtained in molecular dynamics were concatenated into one super-trajectory. The RMSD of its heavy atoms against itself was then calculated and presented as the matrix, whose brightness corresponds to the RMSD (Fig. 4.5).

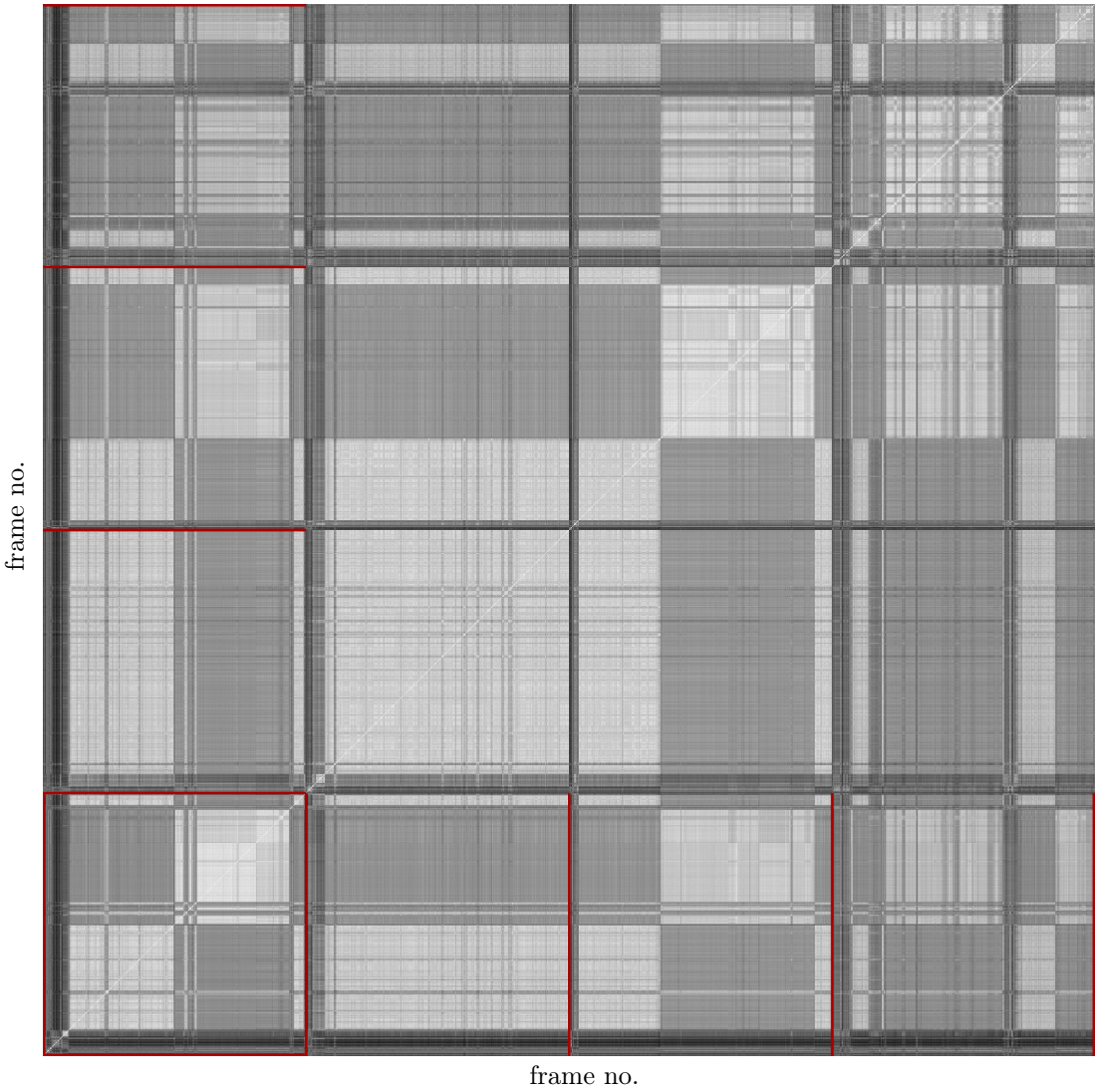

Figure 4.5: RMSD matrix of super-trajectory against itself. Brighter colours correspond to lower RMSD between two frames. Red lines separate different simulations.

Analysis of this matrix allows to distinguish 3 groups of conformations which were earlier seen in the RMSDs plots (Fig. 4.1). In order to assign the individual states to these classes, the gromos clustering method, implemented in the gmx cluster program, was used with the 2.5 Å cutoff. The RMSDs of centroids of two most numerous clusters with respect to the native state, represented by an NMR model, are 0.74 Å and 3.52 Å, which allows their classification as folded and misfolded states. The centroid of the cluster representing the *folded* state was structurally superposed with the one of the NMR models, which is shown in the Fig. 4.6.

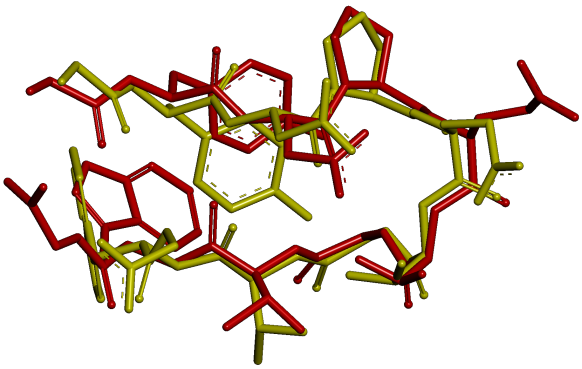

Figure 4.6: Superposition of centroid of the *folded* cluster (red) and one of the NMR models of Chignolin (yellow) [26].

In order to assign simulation frames to particular clusters for subsequent processing, the median filter was applied to the data using the window size of 1000 frames (0.01 µs), what allowed the overcoming of erroneous assignments caused by the fluctuations of the RMSD during the simulations. Such data were then used to assign the conformations to the clusters with following condition:

- if the distance between the centroid of the first/second cluster and examined conformation was <= 2.5 Å, it was assigned to that cluster,
- elsewhere, it was assigned to the third cluster e.g. the unfolded one.

It was checked that no frame was simultaneously at the distance <= 2.5 Å to both the first and the second cluster, thus this method created the division into disjoint classes. The result of this procedure is illustrated in Figure 4.7.

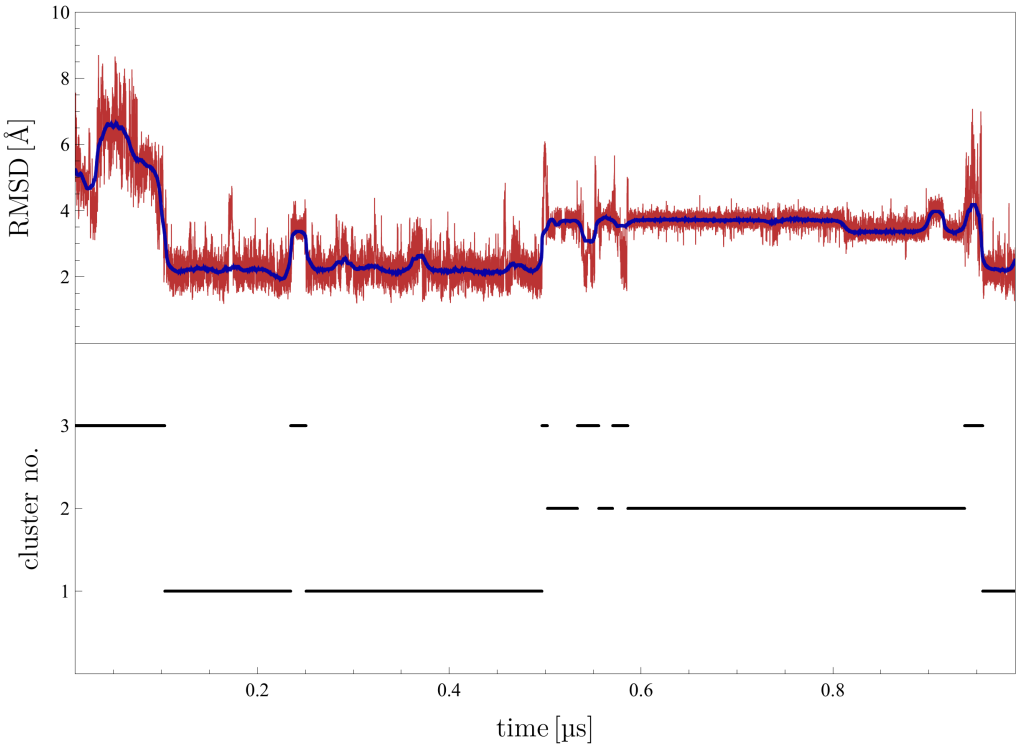

Figure 4.7: The plot of RMSD and cluster number from time in first simulation.

4.4. Energy distribution within clusters

Having this division, the distributions of total free energy G, polar G_p and nonpolar G_{np} contributions to hydration free energy, and potential energy of the peptide U are presented in Fig. 4.8.

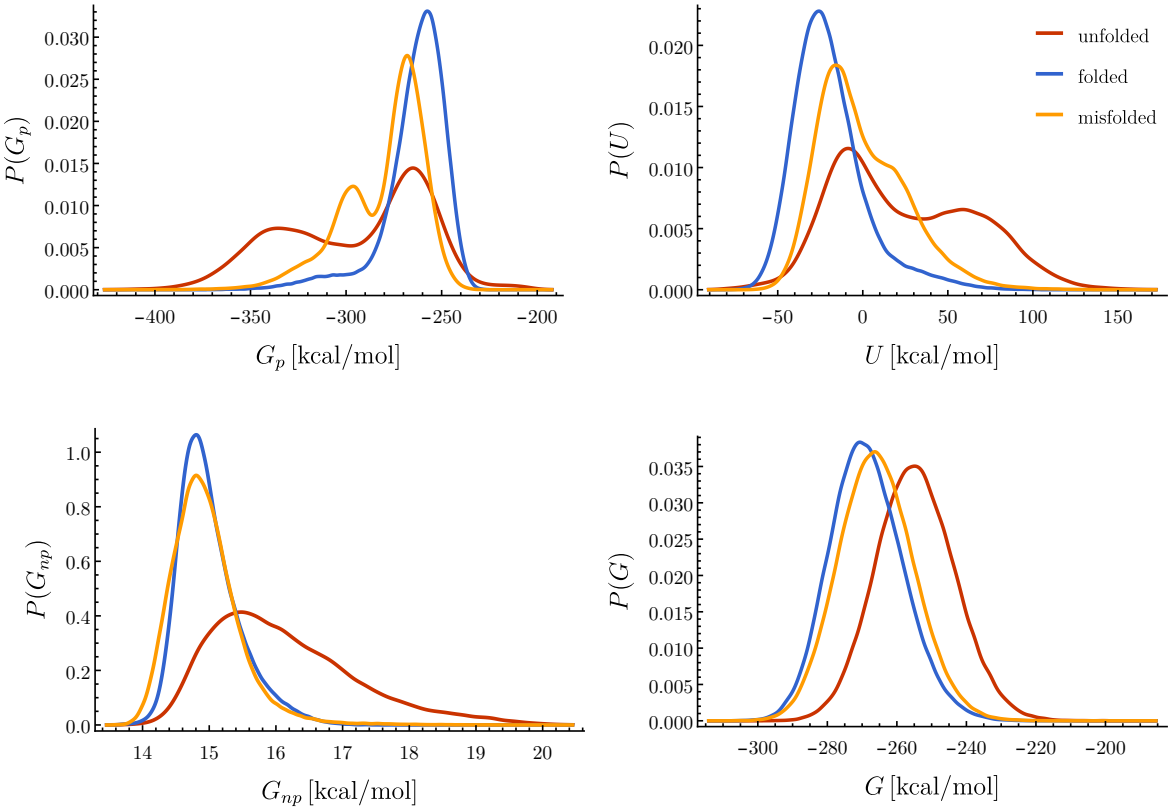

Figure 4.8: Distributions of G_p, U ,G_{np}, G in distinguished clusters.

The first thing to notice is the fact that G_p and U distributions reflect themselves almost exactly. We observe here one unimodal distribution from the *folded* cluster and two bimodal distributions from the *misfolded* and *unfolded* ones. As expected most states from the *folded* distribution are characterised by the highest values of G_p and the lowest of U, and in the *unfolded* one reversely. The appearance of the two modes in these two distributions is somehow surprising and their existence could be linked with the occurrence of the interaction between the N-terminus and C-terminus, whose electrostatic contribution dominates energy terms. In conclusion the main driving force of the folding is the potential energy of peptide U, which competes with the G_p. In the G_{np} plot we observe nonpolar contribution which acts as expected - the accessible surface area correlated with it is smaller for the partially and fully folded conformations. The last plot depicts the total free energy of the clusters and again we notice the lowest values for the most folded states and the highest for the most unfolded, which leads to the conclusion that the formation of peptide intramolecular bonds during folding overcomes the loss of peptide-solvent interactions. It is worth noting that all the histograms are Gaussian-like with the narrowest distribution for the *folded* state and the widest for the *unfolded* one. It allows the linkage between the distribution width and the total entropy of the respective macrostates. According to the statistical physics, in the

first order approximation, entropy is dependent on the variance of the energy distribution. Accordingly, the estimation of variance of $P(G)$ distribution shown in Fig. 4.8, would allow the determination of the $-T\Delta S$ term in Eq. 1.2, and thus, the total free energy of folding [51]. In agreement with intuition, the folded state with relatively lowest variance of G, should have lower entropy than the unfolded one. These notions allow us to believe that we received correct results for our computational experiment.

Chapter 5

Summary

In this work we examined different energy contributions to free energy changes during the folding process of a short peptide Chignolin. Several folding trajectories were generated by means of the classical molecular dynamics simulations. Their frames were then post-processed in order to find the energy terms. They were also clustered into three groups: *unfolded*, *misfolded* and *folded* by the respective RMSD of their heavy atoms against the reference structure. It was shown that the peptide's potential energy U is the main driving force for folding. This force competes mainly with the polar term of the hydration free energy G_p, which means that the favoured creation of the peptide intramolecular interactions comes at the cost of breaking the interactions between water and protein. This result demonstrates that, at least in the case of the considered peptide, solvent effects tend to destabilise the native fold.

Supplementary information

Determination of γ coefficient for the calculation of G_{np}

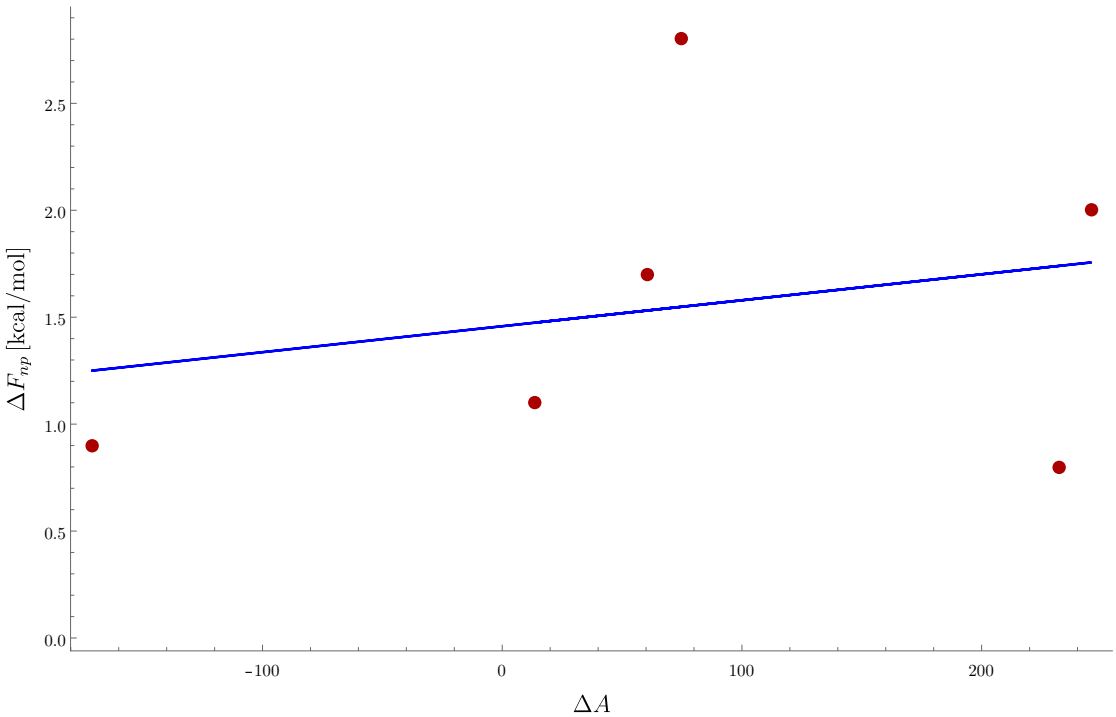

Figure 1*:Fitting of the γ value. Red dots depict the values which were used for the fitting procedure. Blue line determines γ.

Red dots represents the changes of polar contribution to hydration free energy between the unfolded, prefolded, misfolded and folded states and respective changes of surface area [50].

Explanation of the choice of β and λ parameters

We based the selection of parameters on the idea that the distribution of Q values, obtained based on our simulations, should best separate the folded state from other states. One such distribution is presented in the Fig. 2*.

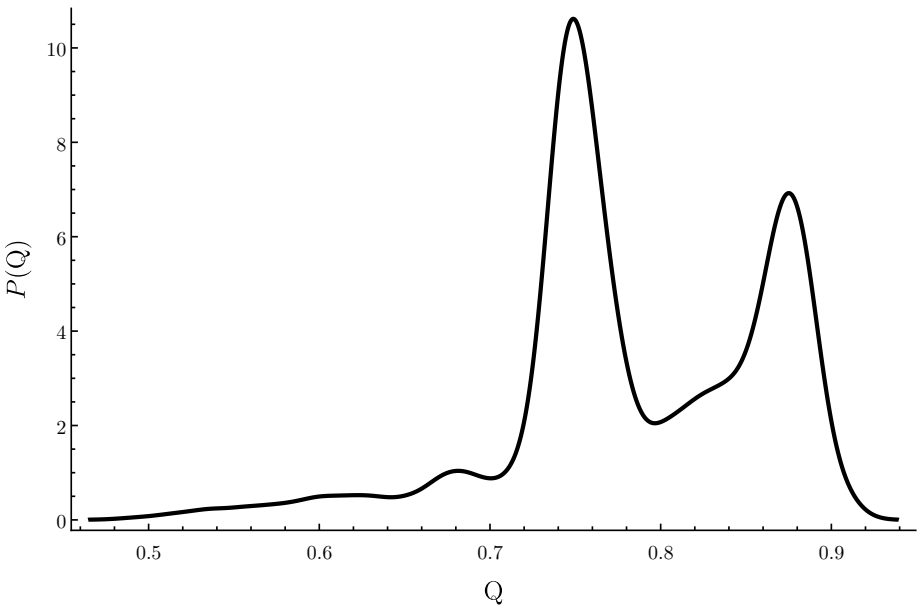

Figure 2*:Distributions of Q parameter for one simulation.

It was decided that the β and λ, which will produce distribution, representing the folded (Q 0.9) and other (Q<0.8) states, whose peaks would be the most distant, will be chosen for the calculations. After few steps it was noticed that for the wide range of β parameters, the corresponding distributions are not significantly different, thus this one was taken as the default ($\beta = 5.0$). In order to find the best value of λ parameter, its different values, ranging from 1.1 to 1.7 with 0.01 step, were systematically tested and the results were presented in Fig. 3*.

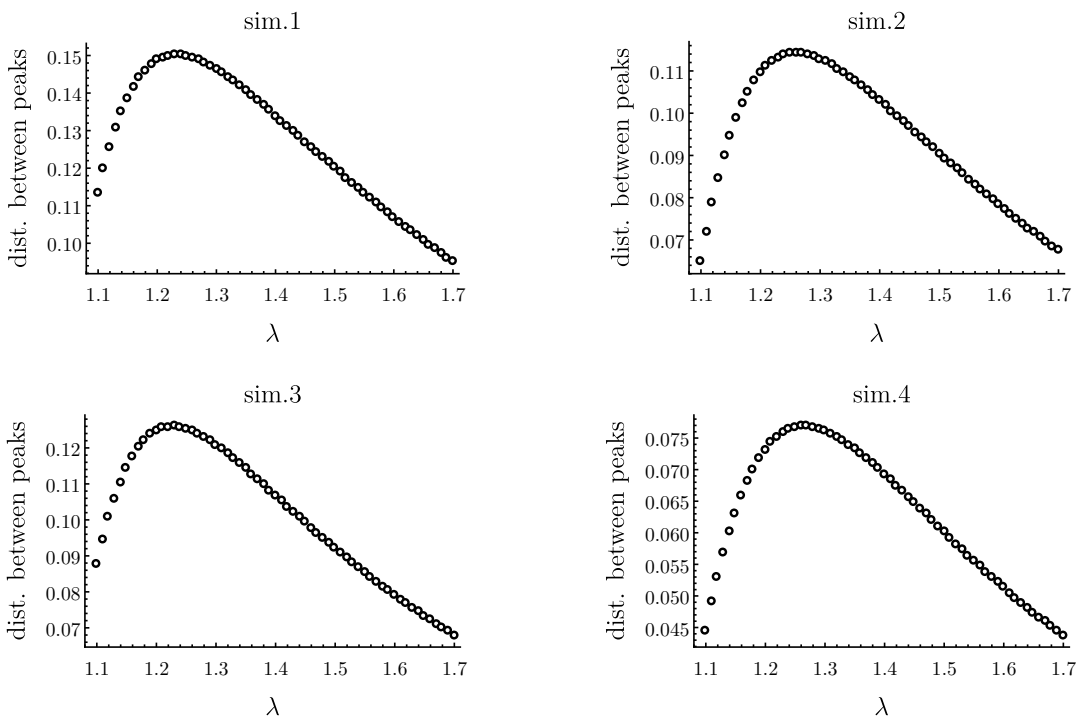

Figure 3*:The distance between most prominent peaks depending on λ for all simulations.

λ parameter values, which lead to Q distribution with the most distant peaks, were then averaged leading to the final λ=1.25 for the calculation of Q parameter in Chapter 4.3.

Additional information

All plots were created using the Mathematica software [52]. Fig. 2.1 was made in GIMP program [53]. Post-processing was done using the Python3 NumPy libraries [54, 55].

Bibliography

[1] G. Reale. *Filosofia antica.* Jaca Book, Milan, 1992.

[2] P. E. Watson, I. D. Watson, and R. D. Batt. Total body water volumes for adult males and females estimated from simple anthropometric measurements. *Am. J. Clin. Nutr.*, 33(1):27–39, 1980.

[3] P. Ball. Water as an active constituent in cell biology. *Chem. Rev.*, 108(1):74–108, 2008.

[4] Y. Levy and J. N. Onuchic. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, 35:389–415, 2006.

[5] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.

[6] P. Ball. Water is an active matrix of life for cell and molecular biology. *Proc. Natl. Acad. Sci. U.S.A.*, 114(51):13327–13335, 2017.

[7] M. Born and J. R. Oppenheimer. Zur quantentheorie der molekeln. *Ann. Phys. (Leipzig)*, 389(20):457–484, 1927.

[8] P. Güttinger. Das verhalten von atomen im magnetischen drehfeld. *Z. Physik*, 73(3):169–184, 1932.

[9] W. Pauli. Principles of wave mechanics. In *Handbuch der Physik*, volume 24, page 162. Springer, Berlin, 1933.

[10] R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56(4):340–343, 1939.

[11] H. Hellmann. *Einführung in die Quantenchemie.* Franz Deuticke, Leipzig, 1937.

[12] A. R. Leach. *Molecular modelling: principles and applications.* Prentice Hall, 2001. OCLC: 45008511.

[13] J. E. Jones. On the determination of molecular fields. —II. from the equation of state of a gas. *Proc. Royal Soc. Lond.*, 106(738):463–477, 1924.

[14] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159(1):98–103, 1967.

[15] R. D. Skeel. What makes molecular dynamics work? *SIAM J. Sci. Comput.*, 31(2):1363–1378, 2009.

[16] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78(1):1–20, 1999.

[17] J. Wang et al. Poisson-boltzmann solvents in molecular dynamics simulations. *Comput. Phys. Commun.*, 3:1010–1031, 2008.

[18] N. A. Baker et al. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, 98(18):10037–10041, 2001.

[19] E. Jurrus et al. Improvements to the APBS biomolecular solvation software suite. *Protein Sci.*, 27(1):112–128, 2018.

[20] M. K. Gilson. Introduction to continuum electrostatics, with molecular applications. In D. A. Beard, editor, *Biophysics textbook online*. Biophysical Society, Bethesda, 2000.

[21] N. Michaud-Agrawal et al. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, 32(10):2319–2327, 2011.

[22] R. J. Gowers et al. MDAnalysis: A python package for the rapid analysis of molecular dynamics simulations. In *Proceedings of the 15th Python in Science Conference*, pages 98–105, 2016.

[23] R. Best, G. Hummer, and W. A. Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.*, 110, 2013.

[24] S. Honda et al. 10 residue folded peptide designed by segment statistics. *Structure*, 12(8):1507–1518, 2004.

[25] S. Honda and K. Yamasaki. Nmr structure of designed protein, chignolin, consisting of only ten amino acids (ensembles). Protein Data Bank, Rutgers University, 2004.

[26] Image of 1UAO (S. Honda, K. Yamasaki, Y. Sawada and H. Morii (2004) structure *Structure* 8:1507–1518) created with Discovery Studio (Dassault Systèmes BIOVIA, Discovery Studio Visualizer, v.19.1.0.18287, San Diego: Dassault Systèmes, 2018).

[27] I. W. Davis et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, 35:W375–W383, 2007.

[28] V. B. Chen et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D*, 66(1):12–21, 2010.

[29] D. van der Spoel, P. J. van Maaren, and H. Berendsen. A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field. *J. Chem. Phys.*, 108:10220–10230, 1998.

[30] A. Wallqvist and R. D. Mountain. Molecular Models of Water. I. Derivation and Description. *Rev. Comp. Ch.*, 13(Chap. 4), 1999.

[31] C. L. Wennberg et al. Lennard-jones lattice summation in bilayer simulations has critical effects on surface tension and lipid properties. *J. Chem. Theory Comput.*, 9(8):3527–3537, 2013.

[32] C. L. Wennberg et al. Direct-space corrections enable fast and accurate lorentz–berthelot combination rule lennard-jones lattice summation. *J. Chem. Theory Comput.*, 11(12):5737–5746, 2015.

[33] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91(1):43–56, 1995.

[34] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7(8):306–317, 2001.

[35] D. van der Spoel et al. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

[36] B. Hess et al. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[37] S. Pronk et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[38] S. Páll et al. Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In Stefano Markidis and Erwin Laure, editors, *Solving Software Challenges for Exascale*, Lecture Notes in Computer Science, pages 3–27. Springer International Publishing, 2015.

[39] M. J. Abraham et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[40] S. Páll and B. Hess. A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.*, 184(12):2641–2650, 2013.

[41] Y. Duan et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24(16):1999–2012, 2003.

[42] H. J. C. Berendsen et al. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[43] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[44] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81:511–519, 1984.

[45] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[46] M. Parrinello and A. Rahman. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.*, 45(14):1196–1199, 1980.

[47] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[48] T. J. Dolinsky et al. PDB2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic Acids Res.*, 32:W665–W667, 2004.

[49] T. J. Dolinsky et al. PDB2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, 35:W522–W525, 2007.

[50] P. Setny and A. Dudek. Explicit solvent hydration benchmark for proteins with application to the PBSA method. *J. Chem. Theory Comput.*, 13(6):2762–2776, 2017.

[51] S. Chong and S. Ham. Configurational entropy of protein: A combined approach based on molecular simulation and integral-equation theory of liquids. *Chem. Phys. Lett.*, 504(4):225–229, 2011.

[52] Wolfram Research, Inc. Mathematica, Version 11.3. Champaign, IL, 2018.

[53] The GIMP Development Team. Gimp, 2019.

[54] T. E. Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[55] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy array: A structure for efficient numerical computation. *Comput Sci Eng*, 13(2):22–30, 2011.