THEORETICAL AND IN SILICO MODELING OF BIOLOGICAL SYSTEMS: FROM PROTEIN STRUCTURE PREDICTION TO CELLULAR DYNAMICS

A Dissertation
presented to
the Faculty of the Graduate School
at the University of Missouri-Columbia

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by
BOGDAN BARZ
Dr. Ioan Kosztin, Dissertation Supervisor
DECEMBER 2009
The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

THEORETICAL AND IN SILICO MODELING OF BIOLOGICAL SYSTEMS: FROM PROTEIN STRUCTURE PREDICTION TO CELLULAR DYNAMICS

presented by Bogdan Barz,

a candidate for the degree of doctor of philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

________________________________________
Professor Ioan Kosztin

________________________________________
Professor Dong Xu

________________________________________
Professor Gabor Forgacs

________________________________________
Professor Paul Miceli

________________________________________
Professor Shi-Jie Chen
Acknowledgments

I would like to thank my advisor Dr. Ioan Kosztin for his guidance, patience and unwavering support. I would also like to thank my collaborators Dr. Dong Xu, Dr. Gabor Forgacs and Dr. Tuck C. Wong; Dr. Paul Miceli and Dr. Shi-Jie Chen, members of my committee; and my colleagues Dr. Lorant Janosi, Jhuma Das, Dr. Adrian Neagu and Dr. Elijah Flenner for their various contributions to this work. I would like to express my gratitude to Dr. Nicolae Aldea for guiding my first steps in scientific research with enthusiasm and dedication. Financial support was provided through grants from the National Science Foundation [FIBR-0526854] and the National Institutes of Health [R21/33-GM078601]. Computer time was generously provided by the University of Missouri Bioinformatics Consortium.
# Table of Contents

List of Tables ................................................................. v

List of Figures ................................................................. vi

List of Abbreviations ......................................................... x

1 Introduction ................................................................. 1

2 Quality assessment of *in silico* protein structure prediction .......... 5

   2.1 Introduction .......................................................... 6

   2.2 MDR Method ........................................................... 7

   2.3 Molecular Dynamics protocol ........................................ 10

   2.4 Test set A ............................................................... 10

   2.5 Test set B ............................................................... 11

   2.6 CASP8 competition .................................................... 14

   2.7 Conclusions ........................................................... 20

3 Conformation and oligomerization of gp41 fusion peptide .............. 23

   3.1 Introduction ........................................................... 24

   3.2 Transition from \(\alpha\)-helical monomers to \(\beta\)-sheet oligomer ..... 26

       3.2.1 FTIR experiments .............................................. 26

       3.2.2 Molecular dynamics simulations ............................. 28

       3.2.3 Constant and variable surface tension simulations ...... 31

       3.2.4 SMD simulations ............................................... 33

       3.2.5 Conclusions: \(\alpha\) to \(\beta\) transition ...................... 35

   3.3 Stability of gp41 oligomers .......................................... 36

       3.3.1 Parallel \(\alpha\)-helical dimer .................................. 37

       3.3.2 Parallel and anti-parallel \(\beta\)-sheet dimers ................ 40

       3.3.3 \(\beta\)-sheet trimer ............................................. 43

       3.3.4 Conclusions: Oligomer stability .............................. 45

4 Calculating free energy profiles and diffusion coefficients from nonequlilibrium molecular dynamics simulations ......................... 46

   4.1 Introduction ........................................................... 47

   4.2 Reaction coordinate and PMF ...................................... 49

   4.3 Harmonic guiding potential, Umbrella Sampling method ............ 51

   4.4 Stiff spring approximation ......................................... 52
## List of Tables

2.1 Ranking of 5 predicted models for 16 proteins using 5 ranking parameters. Columns 2 and 3 contain the cRMSD of the best and poorest models among the 5. Columns 4 through 8 contain the cRMSD of the models selected with each of the five parameters T100, T200, T300, T400 and T500. The last two rows contain the percentage of cases for which the five ranking parameters selected the best model or a model within 0.5 Å from the best one. ................................................................. 9

2.2 MDR ranking for the 22 proteins in subset A. For each structure is presented the value of the cRMSD with the native (columns 2-12) at the position of the MDR rank (column 1). The bolded value corresponds to the best structure among the three. ................................................................. 11

2.3 Ranking results for test set B. The first three columns of the table display the name of the protein, the length of the amino acid sequence and the cRMSD of the best model in the set of 100 structures. The fourth column contains the number of residues that have helical or extended conformation. The next three columns display the cRMSD of the structure with the lowest MDR, OPUS-PSP and DFIRE-A score. The values in the last three columns are the cRMSDs of the TOP5 models with the lowest MDR, OPUS-PSP and DFIRE-A scores. ................................................................. 13

2.4 Top ten servers in the CASP8 Free Modeling (FM) category. The ranking is based on the total GDT-TS (last column) obtained by summing up the GDT-TS scores of the best of the submitted 5 models for each of the 13 FM models. ................................................................. 15

2.5 GDT-TS scores of the best of top 5 structures, corresponding to 41 target proteins from the Human and Server section of CASP8, generated with Rosetta and ranked with the Rosetta scoring function (RSR) and the MDR methods, respectively. ................................................................. 19
List of Figures

2.1 Change in ΔRMSD during gradual heating from 40 K to 600 K. Each color corresponds to a model of different quality (expressed in cRMSD) as follows: : black - native, red - 2.5 Å, blue - 4.5 Å, orange - 6.0 Å; b) Similar plot as in a) for a different protein: black - native, red - 3.5 Å, blue - 4.5 Å, orange - 6.0 Å. c) Example of change in cRMSD during the heating of a predicted model from 40 K to 600 K with superimposed gray bars that indicate 5 time intervals considered for the calculation of 5 ranking parameters (T100, T200, T300, T400 and T500) defined by the average change in cRMSD during each time interval. 8

2.2 Cartoon representation of crystallographic structures for the 7 proteins in test set B. The coloring is a continuous RGB scheme distributed over the total number of residues. The images were rendered with VMD. 12

2.3 Scatter plots of MDR score as a function of cRMSD for each of the 7 proteins in test set B. Each point corresponds to one of the 100 models for each of the 7 proteins, except the one found at 0 Å cRMSD which corresponds to the native structure. 13

2.4 Structure of MUFOLD-MD server. First module includes the structure generation and filtering and the second model includes the MDR ranking of the filtered structures. 14

2.5 Comparison of the GDT-TS scores for structures, corresponding to 13 target proteins from the Free Modeling section of CASP8, obtained as follows: Best 10k = best of the 104 structures generated by Rosetta used in the MUFOLD-MD server; Best 64 = best of the top 64 structures selected by the RSR method; Top5 RSR = best of the top 5 structures selected with RSR; Top5 MDR = best of the top 5 structures selected with MDR; and Top5 Robetta = best of the 5 structures produced by the BAKER-ROBETTA server. 16

2.6 Cartoon representation of the domain 3 (residues 236 to 279) of target T0510 for both the MUFOLD-MD prediction model 4 (a) and the crystallographic structure (b). 17

2.7 Cartoon representation of the domain 2 (residues 124 to 180) of target T0416 for both the MUFOLD-MD prediction model 2 (a) and the crystallographic structure (b). 17
2.8 Top: Comparison of the GDT-TS scores of the best of top 5 structures, corresponding to 41 target proteins from the Human and Server section of CASP8, selected by the MUFOLD-MD server using the MDR (Top5 MDR) and RSR (Top5 RSR) methods, respectively. Bottom: Difference between the GDT-TS scores obtained with the MDR and RSR methods for the same structures as in the Top panel. The green (red) bars indicate that the GDT-TS score of the structure picked by the MDR (RSR) method was at least 5 points bigger than the other one. For the blue bars the difference between the GDT-TS scores (in absolute value) was less than 5 points.

3.1 Top panel: Area per lipid (APL) as a function of the water/lipid ratio ($W_o$). Lower panels A-D: The amide-I region (1600 - 1700 cm$^{-1}$) of the FTIR spectrum (solid curve) of gp41-FP in the AOT reversed micelles corresponding to $W_o$ values indicated by arrows in the top panel. The dashed curves represent the deconvoluted FTIR bands.

3.2 (a) Side and (b) top views of the $\alpha$-helical gp41-FP dimer inserted perpendicular into the upper leaflet of a POPE lipid bilayer (system A1). The peptides are shown in cartoon representation, the phosphorous atoms of the lipid head groups are shown in van der Waals representation while the rest of the lipid molecules and the solvent water molecules are shown in line representation. Similarly, side and top views of the gp41-FP dimers in $\alpha$-helical (system A2) and $\beta$-sheet (system B1) conformations inserted into the POPE lipid bilayer parallel to its surface are shown in (c), (e) and (d), (f), respectively. The images were rendered with VMD [1].

3.3 Time evolution of the color coded secondary structure elements of the two gp41-FP monomers during the 30 ns long equilibrium MD trajectories S1 and S2 of system A1 described in the text. The residue numbers are on the vertical axis. In each panel the top (bottom) graph refers to the first (second) monomer. The images were rendered with VMD [1].

3.4 Time evolution of the color coded secondary structure elements of the two gp41-FP monomers during the 5 ns long MD trajectories S3, S4 of system A2, and S5 of system B described in the text. The residue numbers are on the vertical axis. In each panel the top (bottom) graph refers to the first (second) monomer. The images were rendered with VMD [1].

3.5 Distance between N termini of the two peptides in SA during SIM1. Insets represent snapshots of SA right after energy minimization (left; side view), and after 5ns of equilibration (right; top view).

3.6 Time evolution of the non-bonded (sum of van der Waals and electrostatic) interaction energy between the two monomers of SA (dash-dotted line), SB1 (solid line), and SB2 (dashed line) from 2 to 10 ns of the corresponding MD simulations (SIM1, SIM2 and SIM3, respectively). The gray lines correspond to the non-bonded interaction energies between peptide 1 and 2 and peptide 2 and 3, respectively of system ST during SIM4.
3.7 Time evolution of the color coded secondary structure elements in SB1 (top panel) and SB2 (bottom panel) during simulations SIM2 and SIM3, respectively. The residue numbers are on the vertical axis. The figure was generated with the Timeline plugin of VMD. .................................................. 41

3.8 Snapshots of SB1 (left) and SB2 (right), shown in cartoon representation. The mean locations of hydrogen bonds between the monomers are shown as green segments. The thickness of each segment is proportional to the cumulative lifetime of the corresponding hydrogen bond. .............................. 42

3.9 a) Time evolution of the color coded secondary structure elements in ST during simulation SIM4. The top, middle, and bottom graphs refer to the first, second, and third monomers, respectively. The residue numbers are on the vertical axis. b) Snapshot of the equilibrated ST trimer with the mean location of hydrogen bonds between the monomers shown as green segments. The thickness of each segment is proportional to the cumulative lifetime of the corresponding hydrogen bond. .............................. 44

4.1 Gaussian forward (long-dashed line), reverse (dotted line) and dissipative (solid line) work distribution functions within the stiff-spring approximation. The shaded region in $P_{F/R}(W)$ is the one sampled in F/R SMD pullings. The tail region of $P_d(W)$ corresponding to negative dissipative work is also highlighted. ................................................................. 49

4.2 Lateral and top view of the unit cell ($a = 20 \text{ Å}, b = 23 \text{ Å}$ and $c = 52.5 \text{ Å}$) of the simulated water and SWNT system using Van der Waals and surface representations. Water molecules cross the SWNTs in single files. Figure rendered with VMD [1]. ................................................................. 58

4.3 (a) PMF $U_0(z)$ of a water molecule along the $z$-axis of one of the SWNTs obtained through equilibrium MD simulations. The included snapshot illustrates a completely filled SWNT with five water molecules located about the corresponding PMF minima. (b) Comparison between $U_0(z)$ (thin line) and the same PMF $U_{US}(z)$ (thick line) obtained from umbrella sampling. Graphics rendered with the program VMD [1]. ................................................................. 60

4.4 (Color online) Work along (a) forward and (b) reverse SMD pullings (dashed lines). The mean work $\overline{W}_{F/R}$ is shown as thick solid line. (c) Comparison between $U_0(z)$ (dashed line) and $U_{FR}(z) = (\overline{W}_F - \overline{W}_R) / 2$ (solid line), obtained from fast forward and reverse SMD pullings. Vertical dashed lines indicate the extremities of SWNTs. ................................................................. 63

4.5 (a) Mean dissipative work determined from the FT method (thick-solid) and the cumulant approximation applied separately to forward (thin-solid) and reverse (thin-dashed) pullings. (b) Validity test of JE along forward (solid) and reverse (dashed) processes. The PMF $U$ in the corresponding dissipative work $W_{dF/R} = W_{F/R} \mp U$ is determined from the FR (thick) and the cumulant approximation (thin) methods, respectively. ................................. 65
5.1 Left: Interaction potential energies (top) and forces (bottom) between CPs from same (blue) and different (red) cells. Right: Aggregate made of 500 model cells, each with 10 CPs. Two of the neighboring cells are presented in close caption and colored in red and green. CPs are represented as solid spheres. The (transparent) cellular surface is determined by the envelope of the intracellular CPs. The images were rendered with VMD.

5.2 Fusion of highly viscous polymer droplets [2]. The red circles represent schematics of the spherical caps that model the shape of the fusing droplets.

5.3 Fusion of two multicellular spherical aggregates made of CHO cells (a) and CPD cells (b). The white circles are used for measuring the instantaneous radius of the fusing aggregate and the radius of the interface between the two aggregates. Original images were produced with the confocal microscope (a) and VMD (b). The overlapping with the circles was done in CanvasX.

5.4 Fusion of spherical aggregates made of CHO cells and CPD cells. The volume conservation is verified through angle dependence of $\rho(\theta) = R/R_0$ during the fusion process in experiment (a) and simulation (c). The time dependence of $\sin^2(\theta) = (r/R)^2$ during fusion from experiment and simulation are displayed in c) and d), respectively. By fitting the experimental data with the theoretical curve we calculated the characteristic fusion time $\tau$ and the time corresponding to the initial state $\tau_0$.

5.5 Top - Snapshots of (a) uncompressed and (b) compressed cushion tissue droplets in culture medium, between the parallel plates of a tissue surface tensiometer [3]. Plot in (c) shows the relaxation of the compression force (blue - measured data; red - fit with Eq. (5.9)). Bottom - Corresponding plots for the CPD simulation.

5.6 Fusion of 10 cellular aggregates. Initial and final configurations of CHO [4] (CPD) systems are shown in a) and b) (c and d). Different colors were used for neighboring aggregates.
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOT</td>
<td>aerosol OT</td>
</tr>
<tr>
<td>CP</td>
<td>cellular particle</td>
</tr>
<tr>
<td>CPD</td>
<td>cellular particle dynamics</td>
</tr>
<tr>
<td>CT</td>
<td>C-terminus</td>
</tr>
<tr>
<td>F</td>
<td>forward</td>
</tr>
<tr>
<td>FM</td>
<td>free modeling</td>
</tr>
<tr>
<td>FP</td>
<td>fusion peptide</td>
</tr>
<tr>
<td>GlpF</td>
<td>glycerol uptake facilitator</td>
</tr>
<tr>
<td>HGP</td>
<td>harmonic guiding potential</td>
</tr>
<tr>
<td>JE</td>
<td>Jarzynski equality</td>
</tr>
<tr>
<td>MD</td>
<td>molecular dynamics</td>
</tr>
<tr>
<td>MDR</td>
<td>molecular dynamics ranking</td>
</tr>
<tr>
<td>MDS</td>
<td>multi-dimensional scaling</td>
</tr>
<tr>
<td>MFPT</td>
<td>mean first passage time</td>
</tr>
<tr>
<td>PBC</td>
<td>periodic boundary conditions</td>
</tr>
<tr>
<td>PDB</td>
<td>protein data bank</td>
</tr>
<tr>
<td>PME</td>
<td>particle mesh Ewald</td>
</tr>
<tr>
<td>PMF</td>
<td>potential of mean force</td>
</tr>
<tr>
<td>POPE</td>
<td>1-Palmitoyl-2-Oleoyl-sn-glycero-3-PhosphoEthanolamine</td>
</tr>
<tr>
<td>R</td>
<td>reverse</td>
</tr>
<tr>
<td>RC</td>
<td>reaction coordinate</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------</td>
</tr>
<tr>
<td>RMSD</td>
<td>root mean square deviation</td>
</tr>
<tr>
<td>SMD</td>
<td>steered molecular dynamics</td>
</tr>
<tr>
<td>SSA</td>
<td>stiff-spring approximation</td>
</tr>
<tr>
<td>SWNT</td>
<td>single wall nanotube</td>
</tr>
<tr>
<td>US</td>
<td>umbrella sampling</td>
</tr>
<tr>
<td>VMD</td>
<td>visual molecular dynamics</td>
</tr>
<tr>
<td>WHAM</td>
<td>weighted histogram analysis method</td>
</tr>
</tbody>
</table>
Abstract

This thesis describes four distinct studies (involving a variety of concepts and methods from the physical sciences) of complex biological systems and processes that cover a wide range of time and length scales (extending from the molecular to the multi-cellular level), namely: 1) identification of near-native structures from a large set of in silico predicted structures for a given target protein by probing the stability of the structures against simulated heating using Molecular Dynamics (MD) simulations; 2) all atom MD simulation study of protein-cell membrane interactions designed to elucidate the relation between the conformational structure of the GP41 fusion peptide of HIV-1 virus and the density of the lipid membrane; 3) calculation of the potential of mean force for water transport through single wall carbon nanotubes from nonequilibrium MD simulations; and 4) development and application of a novel computational method, referred to as the cellular particle dynamics (CPD) method, for the simulation of the dynamics of multicellular systems.
1 Introduction
In this thesis I provide a description of the four major projects that I contributed to during my Ph.D. studies. In each of these projects a variety of concepts and methods from the physical sciences are employed to study complex biological systems and processes on a wide range of time and length scales.

The first project (Chp. 2) deals with protein folding, which is one of the most fundamental and difficult problems in biomolecular modeling. The goal of this project is to develop (i) \textit{ab initio} and homology methods for predicting and refining the 3D (tertiary) structure of a protein from the sole knowledge of its amino-acid sequence (primary structure), and (ii) efficient methods for assessing the quality of the predicted structure for target proteins. The knowledge of the high (almost atomic) resolution structure of a protein is crucial for understanding and controlling its functions. The experimental determination of the 3D structure of a protein through X-ray crystallography is tedious and expensive. Thus, in our post genomic era, replacing or complementing the experimental determination of protein structures with reliable and much faster and efficient \textit{in silico} prediction methods is a top priority in the field of molecular biology. Here I describe a new method, based on MD simulations, that enables the selection of near native structures from a large set of \textit{in silico} predicted structures for a given target protein.

The second project (Chp. 3) is an extensive all atom MD simulation study in the field of protein-cell membrane interactions. The interactions between proteins and lipid membranes are complex and highly specific. They play a crucial role in understanding the molecular level mechanism of a wide variety of processes, such as viral infection of a cell or molecular and ion transport across the cell membrane. In the case of HIV-1 viral infection the interaction between the transmembrane subunit (fusion peptide) gp41 of the envelope glycoprotein gp160 and the cellular membrane is an important initial step in the fusion of the viral envelope membrane with the cell membrane. The gp41 N-terminal fusion domain contains 16, mostly hydrophobic, residues and, depending on the environmental conditions, it can adopt either a monomeric alpha-helical or an oligomeric beta-sheet
conformation. Based on a series of MD simulations and results from FTIR experiments, here I give a detailed analysis of the conformational stability and possible transformations of the fusion peptide gp41 in the cell membrane.

The scope of the third project (Chp. 4) is to develop and test a novel approach for calculating the molecular flux across channel proteins and artificial, hydrophobic nanotubes. Channel proteins are transmembrane proteins that facilitate the passive transport (driven by concentration and/or chemical potential gradient) across the plasma membrane of certain ions and molecules in a highly selective manner. Channel proteins behave like hydrophobic nanopores embedded in the cell membrane, while the transported ions/molecules can be modeled as overdamped Brownian particles that diffuse along the axis of the channel guided by an effective potential of mean force (PMF) due to their interaction with the rest of the atoms in the system. Here I present a novel method, referred to as the FR method, for calculating simultaneously both the PMF and the position dependent diffusion coefficient of a molecule that moves through a channel protein or artificial nanopore. The FR method is tested on single file water transport through single walled carbon nanotubes (SWNTs) that connect two water reservoirs. A SWNT is a hydrophobic nanopore that can be regarded as a toy model for a complex channel protein.

Finally, the fourth project (Chp. 5) deals with the theoretical and computer modeling of multicellular systems. There is ample experimental evidence that in many cases (e.g., embryogenesis, tumor growth, angiogenesis, cell sorting and self-assembly of artificial cell aggregates used in bioprinting) living tissues and multicellular aggregates behave as highly visco-elastic liquids. The property of tissue liquidity is the result of cellular adhesion and motility. Here I describe a novel approach, referred to as the cellular particle dynamics (CPD) method, for the computer simulation of the time dependent dynamics of multicellular systems. In CPD cells are modeled as an ensemble of cellular particles that interact via short range contact interactions, characterized by an attractive (adhesive interaction) and a repulsive (excluded volume interaction) component. The time evolution of the spatial con-
formation of the multicellular system is determined directly by recording the trajectories of all cells through integration of their equations of motion. The viability of the CPD method is demonstrated in the case of fusion of two spherical multicellular aggregates. Comparison between the CPD simulation and experimental results provides a direct way to relate cellular level model parameters to experimentally measurable tissue level biophysical quantities (e.g., surface tension, viscosity and shear modulus). By design, the CPD method is rather flexible and most suitable for multiscale modeling of multicellular system. The spatial level of detail of the system can be easily tuned by changing the number of particles in a cell. Thus, CPD can be used equally well to describe both cell level processes (e.g., the adhesion of two cells) and tissue level processes (e.g., the formation of 3D constructs of millions of cells through bioprinting).
2 Quality assessment of *in silico* protein structure prediction
2.1 Introduction

It is well known that the secondary and tertiary structures of a protein are determined by its aminoacid (AA) sequence. Over the last years both homology and ab initio prediction groups have tried to perfect a method for predicting the 3D configuration of a protein given its sequence. In general, homology methods are very successful in predicting the structure of target proteins (e.g., within 1 Å of their unknown native structures) with high (>85%) sequence homology with other proteins of known structure [5]. However, homology methods are impractical in the case of protein targets with poor (<25%) sequence homology, and for predicting such structures one needs to rely on other folding methods, e.g., ab initio, which are less accurate [6]. Once a large set of models is predicted, by either of these methods, a selection criterion is needed for identifying the structure which is the closest to the native one.

The most common way of discriminating between predicted structures of a given protein is by using either knowledge based or physics based energy functions [7, 8]. Knowledge based potentials that can be applied to reduced representations of proteins with either one center [9], two centers [10] or more centers (heavy atoms) [11] of interaction per amino acid are widely used for identifying and ranking near native structures from a pool of generated decoys. The main concern when using any of these energy functions is recognizing both secondary and tertiary structure features that resemble to the native structure [12]. Most of the physics based energy functions (CHARMM [13, 14], GROMOS [15, 16], AMBER [17]) require all atom models of the protein. The improved physics based energy functions [8] that use a generalized Born implicit solvent provide fairly accurate discrimination between folded and miss folded proteins. Here we present an alternative method for discriminating between predicted structures of a protein by testing their stability during simple heating using a physics based all atom potential (i.e., CHARMM27). In general (but not always), the smaller the RMSD of the structure of a protein is (with respect to its
native one) the more stable this structure is. Thus, one can rank the quality of these structures by comparing the relative stability of the predicted structures against gradual heating, simulated by all-atom MD. We refer to this approach as the MD-Ranking (MDR) ranking method. We describe in detail the MDR method in the next section of this chapter. The third and fourth sections contain applications of the MDR method to two sets of proteins. A comparison with two knowledge based scoring functions is included in the fourth section. In the fifth section we present results from the Critical Assessment of Techniques for Protein Structure Prediction 8 (CASP8) competition where we participated with the MUFOLD-MD server for the 3D structure prediction and performed very well in the Free Modeling (FM) category. The final section contains the concluding remarks of this chapter.

2.2 MDR Method

The MDR method is a new assessment method for near-native structures generated by various protein structure prediction methods and uses dynamical properties of proteins observed through MD simulations. The core of the MDR method is the stability of the structures, tested by monitoring the change of their cRMSD (RMSD of the C\(\alpha\) atoms) during the MD simulation of their scheduled heating in vacuum at a rate of 1 K/ps. The heating simulation is preceded by the energy minimization needed for removing bad contacts between atoms. The change in cRMSD (\(\Delta\)RMSD) is calculated every picoseconds during heating from 40 K to 600 K between the heated structure and the structure at the beginning of the energy minimization. Fig. 2.1 illustrates \(\Delta\)RMSD as a function of temperature for two proteins, each with three predicted models of known quality. The crystallographic structure is also included for comparison. Fig. 2.1a is a case where the difference in the \(\Delta\)RMSD curves for the three models is favorable for MDR selection. The black curve corresponding to the native structure is the lowest, indicating the native as the most stable one. The best model (red curve) which has the cRMSD of 2.5 Å has a similar behavior to the native up
to 150 K when it starts unfolding at a higher rate. The medium and low quality models (blue and orange curves) are the most unstable and have a higher ∆RMSD already at the beginning of the heating process. According to our hypothesis the most stable structure is the most similar to the crystallographic one making the model with 2.5 Å cRMSD the one selected by MDR. For some proteins, i.e., Fig. 2.1b, the low quality model (orange) is more stable than the high quality one (red). In these cases the MDR method will have a very low success rate.

Figure 2.1: Change in ∆RMSD during gradual heating from 40 K to 600 K. Each color corresponds to a model of different quality (expressed in cRMSD) as follows: : black - native, red - 2.5 Å, blue - 4.5 Å, orange - 6.0 Å; b) Similar plot as in a) for a different protein: black - native, red - 3.5 Å, blue - 4.5 Å, orange - 6.0 Å. c) Example of change in cRMSD during the heating of a predicted model from 40 K to 600 K with superimposed gray bars that indicate 5 time intervals considered for the calculation of 5 ranking parameters (T100, T200, T300, T400 and T500) defined by the average change in cRMSD during each time interval.

To find an optimal parameter that will rank predicted structures according to their similarity with the native structure, we calculated the average ∆RMSD for five time intervals: the first 100 ps, 200 ps, 300 ps, 400 ps and 500 ps of the heating process. The time intervals are displayed in Fig. 2.1c as grey bars. To test the efficiency of these 5 parameters, labeled as T100, T200, T300, T400 and T500, we selected randomly 16 proteins from the July 2005 release of Astral database [18], generated 15,000 low resolution structures (positions of Cα atoms) by using the Multi-dimensional Scaling (MDS) method [19, 20] and selected 5 models for each protein using a filter consisting of a clustering method [21]. The coordinates of the missing side-chain heavy atoms of the selected models were determined by using the program PULCHRA [22] and the hydrogen atoms were added by using PSFGEN,
Table 2.1: Ranking of 5 predicted models for 16 proteins using 5 ranking parameters. Columns 2 and 3 contain the cRMSD of the best and poorest models among the 5. Columns 4 through 8 contain the cRMSD of the models selected with each of the five parameters T100, T200, T300, T400 and T500. The last two rows contain the percentage of cases for which the five ranking parameters selected the best model or a model within 0.5 Å from the best one.

which is part of the visual molecular dynamics (VMD) software [1]. The all-atom models were subject to MD heating and the ΔRMSD was recorded each ps. Next, we calculated the average ΔRMSD for each of the 5 time intervals and ranked the 5 models considering the ones with smallest average ΔRMSD the most similar to the native structure. The final results given by the 5 ranking parameters are displayed in Table 2.1. When the selected structure is within 0.5 Å from the best among the five, the ranking parameter T100 has the highest accuracy, of 81.3 %, which is similar to the accuracy of the ranking parameter T400. When looking at the performance of each ranking parameter of selecting the best model among the 5 structures, the ranking parameter T100 gives the best selection, with 68.8% success rate. Based on these results we decided that the best ranking parameter

<table>
<thead>
<tr>
<th>No.</th>
<th>Smallest cRMSD</th>
<th>Largest cRMSD</th>
<th>Best T100</th>
<th>Best T200</th>
<th>Best T300</th>
<th>Best T400</th>
<th>Best T500</th>
<th>Correct guess (%)</th>
<th>Guess within 0.5 Å (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.2</td>
<td>6.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>2.2</td>
<td>68.8</td>
<td>81.3</td>
</tr>
<tr>
<td>2</td>
<td>5.5</td>
<td>9.5</td>
<td>5.5</td>
<td>5.5</td>
<td>5.5</td>
<td>5.5</td>
<td>5.5</td>
<td>43.8</td>
<td>68.8</td>
</tr>
<tr>
<td>3</td>
<td>4.3</td>
<td>12.0</td>
<td>4.3</td>
<td>4.8</td>
<td>4.3</td>
<td>4.3</td>
<td>4.3</td>
<td>56.3</td>
<td>75.0</td>
</tr>
<tr>
<td>4</td>
<td>2.4</td>
<td>14.8</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>62.5</td>
<td>81.3</td>
</tr>
<tr>
<td>5</td>
<td>3.3</td>
<td>6.7</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>5.4</td>
<td>56.3</td>
<td>75.0</td>
</tr>
<tr>
<td>6</td>
<td>4.4</td>
<td>10.2</td>
<td>4.4</td>
<td>4.8</td>
<td>4.8</td>
<td>4.8</td>
<td>4.8</td>
<td>56.3</td>
<td>75.0</td>
</tr>
<tr>
<td>7</td>
<td>3.0</td>
<td>4.9</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>62.5</td>
<td>81.3</td>
</tr>
<tr>
<td>8</td>
<td>3.5</td>
<td>10.3</td>
<td>3.5</td>
<td>4.2</td>
<td>3.5</td>
<td>4.2</td>
<td>3.5</td>
<td>56.3</td>
<td>75.0</td>
</tr>
<tr>
<td>9</td>
<td>2.7</td>
<td>12.9</td>
<td>2.7</td>
<td>2.7</td>
<td>2.7</td>
<td>2.7</td>
<td>2.7</td>
<td>68.8</td>
<td>81.3</td>
</tr>
<tr>
<td>10</td>
<td>2.1</td>
<td>10.8</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td>43.8</td>
<td>68.8</td>
</tr>
<tr>
<td>11</td>
<td>3.2</td>
<td>9.9</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
<td>56.3</td>
<td>75.0</td>
</tr>
<tr>
<td>12</td>
<td>3.4</td>
<td>8.3</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td>56.3</td>
<td>75.0</td>
</tr>
<tr>
<td>13</td>
<td>5.7</td>
<td>13.2</td>
<td>12.3</td>
<td>12.3</td>
<td>10.6</td>
<td>10.6</td>
<td>10.6</td>
<td>62.5</td>
<td>81.3</td>
</tr>
<tr>
<td>14</td>
<td>2.9</td>
<td>12.3</td>
<td>2.9</td>
<td>9.4</td>
<td>9.4</td>
<td>2.9</td>
<td>2.9</td>
<td>56.3</td>
<td>75.0</td>
</tr>
<tr>
<td>15</td>
<td>5.0</td>
<td>9.2</td>
<td>5.3</td>
<td>5.3</td>
<td>5.3</td>
<td>5.3</td>
<td>5.3</td>
<td>62.5</td>
<td>81.3</td>
</tr>
<tr>
<td>16</td>
<td>3.5</td>
<td>5.6</td>
<td>3.7</td>
<td>3.7</td>
<td>3.7</td>
<td>3.7</td>
<td>3.7</td>
<td>56.3</td>
<td>75.0</td>
</tr>
</tbody>
</table>
(MDR score) of the predicted structures, from both accuracy and required computational
time points of view, is their average $\Delta$RMSD during the first 100 K of the heating process,
i.e., the heating interval between 40 and 140 K. Additional details about the MD simulation
protocol are presented in the next section, followed (in Sections 2.4 and 2.5) by applica-
tions of the MDR method (using the ranking parameter T100) to two sets of proteins with
known crystallographic structures.

2.3 Molecular Dynamics protocol

The MD simulations were performed in vacuum with the parallel code NAMD2.6 [23].
As force field parameters we used CHARMM27 [13, 14]. Van der Waals interactions were
cut at 12 Å and smoothly switched between 10 and 12 Å. For full long-range electrostatic
interactions the particle mesh Ewald (PME) method [24] was used. The SHAKE algorithm
[25] was used for rigid bonds involving all hydrogen atoms. Newton’s equation of motion
was integrated for each atom every 2 fs using the velocity Verlet method. During the
gradual heating the temperature was increased by 1 K each 100 ps and was controlled
by Langevin dynamics with a damping coefficient of 5 ps$^{-1}$. Before running the MD
simulations all the structures were optimized by removing the bad contacts through energy
minimization. The energy minimization was done with the congruent gradient method
built-in the NAMD code and was divided in three stages: with the heavy atoms fixed - 500
steps, with the C$_\alpha$ atoms fixed - 2000 steps, and with no atoms fixed - 2000 steps.

2.4 Test set A

Two sets of proteins were used for testing the MDR method. For set A 22 proteins were
randomly selected from the July 2005 release of Astral database. 15,000 low resolution
structures (positions of C$_\alpha$ atoms) were generated for each of the 22 proteins by using the
Table 2.2: MDR ranking for the 22 proteins in subset A. For each structure is presented the value of the cRMSD with the native (columns 2-12) at the position of the MDR rank (column 1). The bolded value corresponds to the best structure among the three.

MDS method. Further, 3 models of known quality were selected for each protein, i.e., high (cRMSD ~ 3 Å), medium (cRMSD ~ 4.5 Å) and low quality (cRMSD ~ 6 Å). The coordinates of the missing side-chain heavy atoms were determined by using the program PULCHRA and the hydrogen atoms were added by using PSFGEN.

For each protein in set A we calculated the MDR score of the three models and ranked them accordingly. The MDR method identified the best structure with cRMSD < 3 Å with 80% success rate (8 cases out of 10) and the best structure with 3 Å < cRMSD < 4 Å with 77% success rate. The results are presented in Table 2.2.

2.5 Test set B

As test set B we used 7 small proteins with sequence lengths varying from 54 to 75 amino acids from the 4state reduced set of decoys of Parker & Levitt [26]. Cartoon representations of the native configurations are displayed in Fig. 2.2. The method for generating the low resolution decoys is based on a 4-state off-lattice model. From a large pool of generated structures the authors selected only conformations that had a low score using a variety of scoring functions and built the all-atom decoys from the C_{α} atoms with the program segmod [27]. For each of the 7 proteins we selected randomly 100 structures from ~600-700 all-atom decoys. The cRMSD of the selected decoys varied between 0.88 and 9 Å. The MDR method was used for ranking each of the 100 decoys of the 7 proteins in set B.
Figure 2.2: Cartoon representation of crystallographic structures for the 7 proteins in test set B. The coloring is a continuous RGB scheme distributed over the total number of residues. The images were rendered with VMD.

For proteins 1ctf, 1r69, 2cro and 3icb the models with lowest MDR score (TOP1) have less than 2 Å cRMSD, as shown in Table 2.3, column 5. For proteins 1sn3, 4pti and 4rxn the MDR method performed less accurately and the TOP1 models have a cRMSD greater than 4.5 Å. When looking at the best of the five models with smallest MDR score (TOP5), there is a significant improvement for protein 4pti and the selected model has a cRMSD of 1.84 Å.

The good correlation between the MDR score and the cRMSD of the 100 decoys for some of the proteins, displayed as scatter plot in Fig. 2.3, could be related to the secondary structure of their crystallographic structure. The secondary structure of proteins 1ctf, 1r69, 2cro and 3icb, displayed in Fig. 2.2, for which MDR method gives very good ranking contains at least 4 alpha helices, except for 1ctf which has 3 alpha helices and 3 beta strands. Proteins 1sn3 and 4rxn, for which the MDR method cannot select a structure with cRMSD < 4 Å have both 3 beta strands and 1 alpha helix in their tertiary conformations while 4pti has 2 alpha helices and 2 beta strands. A more quantitative characterization of the secondary structure elements is given in column 4 of Table 2.2 which contains the number
Table 2.3: Ranking results for test set B. The first three columns of the table display the name of the protein, the length of the amino acid sequence and the cRMSD of the best model in the set of 100 structures. The fourth column contains the number of residues that have helical or extended conformation. The next three columns display the cRMSD of the structure with the lowest MDR, OPUS-PSP and DFIRE-A score. The values in the last three columns are the cRMSDs of the TOP5 models with the lowest MDR, OPUS-PSP and DFIRE-A scores.

<table>
<thead>
<tr>
<th>Set B</th>
<th>Seq. Length</th>
<th>Best in 100</th>
<th>α/β</th>
<th>TOP1 MDR</th>
<th>TOP1 OPUS</th>
<th>TOP1 DFIRE</th>
<th>TOP5 MDR</th>
<th>TOP5 OPUS</th>
<th>TOP5 DFIRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ctf</td>
<td>68</td>
<td>1.67</td>
<td>36/18</td>
<td>1.68</td>
<td>2.15</td>
<td>2.93</td>
<td>1.67</td>
<td>1.68</td>
<td>1.68</td>
</tr>
<tr>
<td>1r69</td>
<td>63</td>
<td>1.37</td>
<td>44/0</td>
<td>1.53</td>
<td>1.67</td>
<td>2.59</td>
<td>1.53</td>
<td>1.37</td>
<td>1.37</td>
</tr>
<tr>
<td>1sn3</td>
<td>65</td>
<td>1.67</td>
<td>8/17</td>
<td>5.94</td>
<td>7.07</td>
<td>6.36</td>
<td>3.34</td>
<td>1.97</td>
<td>2.43</td>
</tr>
<tr>
<td>2cro</td>
<td>65</td>
<td>0.88</td>
<td>43/0</td>
<td>0.88</td>
<td>0.88</td>
<td>1.95</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>3icb</td>
<td>75</td>
<td>1.00</td>
<td>44/2</td>
<td>1.51</td>
<td>1.75</td>
<td>1.61</td>
<td>1.49</td>
<td>1.51</td>
<td>1.00</td>
</tr>
<tr>
<td>4pti</td>
<td>58</td>
<td>1.69</td>
<td>13/15</td>
<td>5.07</td>
<td>2.06</td>
<td>2.23</td>
<td>1.84</td>
<td>1.78</td>
<td>2.23</td>
</tr>
<tr>
<td>4rxn</td>
<td>54</td>
<td>1.46</td>
<td>9/11</td>
<td>4.70</td>
<td>2.77</td>
<td>4.53</td>
<td>4.70</td>
<td>2.12</td>
<td>2.17</td>
</tr>
</tbody>
</table>

Figure 2.3: Scatter plots of MDR score as a function of cRMSD for each of the 7 proteins in test set B. Each point corresponds to one of the 100 models for each of the 7 proteins, except the one found at 0 Å cRMSD which corresponds to the native structure.

As comparative ranking methods we used two knowledge-based scoring functions, OPUS-PSP and DFIRE-A. OPUS-PSP is an empirical scoring function that uses a new method of packing the side chain groups using information from all-atom structure [28].
DFIRE-A stands for distance-scaled, finite ideal-gas reference state (DFIRE-A) method [9] and uses only information about the C\textsubscript{\textalpha} atoms of the protein. The MDR method performed better or similar to the TOP1 ranking of OPUS-PSP and DFIRE-A scoring functions in five out of seven cases, i.e., proteins 1ctf, 1r69, 1sn3, 2cro and 3icb. The results are presented in Table 2.3. When the native structure was included among the 100 decoys, the MDR method identified it as TOP1 in 6 of 7 cases, DFIRE-A in 4 of 7 cases and OPUS-PSP in 7 of 7 cases.

### 2.6 CASP8 competition

During the Critical Assessment of Techniques for Protein Structure Prediction competition from 2008 for predicting 3D models of proteins with unknown crystallographic structure based on their AA sequence, we participated with the MUFOLD-MD server which was part of the MUFOLD team. The server consists of two modules as presented in Fig. 2.4. In the first module the low-resolution 3D structure of the protein is determined from its AA sequence by using the \textit{ab initio} method implemented in the ROSETTA 2.2.0 software [29–32]. For a 100 aminoacids protein, it takes about 4 hours to generate 10,000 ROSETTA structures on a computer cluster with 48 Intel Xeon EM64T-2.8GHz CPUs. Within the same module the generated structures are ranked by using Rosetta’s score based method and 64 structures with the lowest Rosetta score are selected. Finally, in the second module, the MDR score is used to establish the final ranking of the previously determined top
Table 2.4: Top ten servers in the CASP8 Free Modeling (FM) category. The ranking is based on the total GDT-TS (last column) obtained by summing up the GDT-TS scores of the best of the submitted 5 models for each of the 13 FM models.

64 structures. Thus, the fully automated MUFOLD-MD server is capable to predict the structure of a protein from its aminoacid sequence in just a few hours. Due to the ab initio structure generator provided by the ROSETTA 2.2.0 software package our server had better performance in the Free Modeling (FM) category of the competition. The FM category contained 13 domains from different proteins as presented in the first row of Table 2.4. For each FM target the MUFOLD-MD server provided 5 models selected with the MDR method. Once the crystallographic structures were released the quality of the submitted models was evaluated by three independent assessors and was made available through the CASP8 website (http://predictioncenter.org/casp8). The official criterion for assessing the quality of the models was the Global Distance Test Total Score (GDT-TS) which is part of the Local - Global Alignment (LGA) software [33], a method for finding 3D similarities in protein structures. By definition, GDT-TS = 100(GDT-P1 + GDT-P2 + GDT-P4 + GDT-P8)/4, where GDT-Pn denotes the percentage of residues in the model structure falling within a defined distance cutoff n Å of their position in the experimental structure.

Based on the official CASP8 results, in the FM category MUFOLD-MD is ranked number one, before the BAKER-ROBETTA server, as shown in Table 2.4. The ranking of the servers was done according to the total GDT-TS score (last column) obtained by summing up the GDT-TS scores of the best of the submitted five models for each of the 13 FM targets. We note that the success of MUFOLD-MD in the FM category is attributed not only to Rosetta but also to our MDR method that overall outperformed (though only marginally)
Figure 2.5: Comparison of the GDT-TS scores for structures, corresponding to 13 target proteins from the Free Modeling section of CASP8, obtained as follows: Best 10k = best of the 104 structures generated by Rosetta used in the MUFOLD-MD server; Best 64 = best of the top 64 structures selected by the RSR method; Top5 RSR = best of the top 5 structures selected with RSR; Top5 MDR = best of the top 5 structures selected with MDR; and Top5 Robetta = best of the 5 structures produced by the BAKER-ROBETTA server.

the structure ranking based on the Rosetta scoring function, referred to as the RSR method. This point is illustrated in Fig. 2.5, where we compare GDT-TS scores for structures of the 13 FM targets obtained as follows: (i) best of the 104 structures generated by Rosetta (dark blue); (ii) best of the top 64 structures obtained using the RSR method (light blue); (iii) best of the top 5 structures selected by RSR (red); (iv) best of the top 5 structures selected with our MDR method (green); and (v) best of the 5 models submitted by the BAKER-ROBETTA server. In most of the cases the MDR method outperformed both RSR and ROBETTA as well. Most likely, the MDR method would have performed even better if the quality of the top 64 structures determined through RSR would have been better.

Several models submitted by our server were described by the assessors as exceptional and we present in detail two of them in the following paragraphs.

One of the best models submitted by the MUFOLD-MD server was the fourth model of target T0510 which was evaluated as the best by all three assessors. Target T0510 has three domains which cover the amino acid sequence as follows: D1: 1-163, D2: 168-235, and D3: 236-279. Even though the domain used by our server extends from residue 162 to 288, the C-terminus region (residues 255-279) of the fourth predicted model matches very
Figure 2.6: Cartoon representation of the domain 3 (residues 236 to 279) of target T0510 for both the MUFOOLD-MD prediction model 4 (a) and the crystallographic structure (b).

Figure 2.7: Cartoon representation of the domain 2 (residues 124 to 180) of target T0416 for both the MUFOOLD-MD prediction model 2 (a) and the crystallographic structure (b).
well the crystallographic structure with a RMSD of \( \sim 2\,\text{Å} \). In the original ranking of the 64 structures given by the Rosetta energy function, model 4 was ranked 9th. Thus, the MDR method brought a considerable improvement by re-ranking the 64 structures selected by Rosetta energy function and including the model 4 among the 5 models submitted. The cartoon representations of model 4 and the crystallographic structure of domain D3 are shown in Fig. 2.6.

Another very good prediction submitted by MUFOLD-MD was the second model of target T0416 - domain 2 which was ranked as the best model by two assessors. Target T0416 has two domains, the first extending over residues 6-115,194-315 and the second extending over residues 124-180. In our predictions we used only one domain for the entire protein. During the selection of the 64 structures with the lowest value of Rosetta energy function, model 2 was ranked the 25th confirming again the improved selection of the MDR method. A cartoon representation of both model 2 and the crystallographic structure for domain 2 of target T0416 is shown in Fig. 2.7.

To further demonstrate the superiority of the MDR method with respect to RSR, we have compared the GDT-TS scores of the best of top 5 structures, corresponding to 41 CASP8 targets from the Human and Server category, generated by MUFOLD-MD and selected by employing the MDR and RSR methods, respectively. The results are shown in Table 2.5 and Fig. 2.8. While in most of the cases the difference between the GDT-TS scores was smaller than 5\% (blue bars in Fig. 2.8), the MDR result (green bars) was substantially better than the RSR one (red bars) in almost twice as many cases. The ranking was performed using the top 64 Rosetta-generated structures determined through RSR. In this analysis, we used only 41 target proteins out of the listed 57 proteins from the Human and Server section of CASP8. We dropped 16 targets for various reasons, i.e., 3 were canceled by the organizers (T0387, T0403, T0467), 2 were not included in the official CASP8 results (T0484, T0500), 7 for which Rosetta generated five or less structures (T0389, T0413, T0440, T0407, T0421, T0449, T0457), and 4 for which the quality of the
<table>
<thead>
<tr>
<th>Target</th>
<th>GDT-TS</th>
<th>Target</th>
<th>GDT-TS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top5 RSR</td>
<td>Top5 MDR</td>
<td></td>
</tr>
<tr>
<td>T0480</td>
<td>69.12</td>
<td>63.24</td>
<td>T0423</td>
</tr>
<tr>
<td>T0474</td>
<td>65.76</td>
<td>72.28</td>
<td>T0454</td>
</tr>
<tr>
<td>T0499</td>
<td>63.84</td>
<td>70.54</td>
<td>T0462</td>
</tr>
<tr>
<td>T0469</td>
<td>55.00</td>
<td>39.62</td>
<td>T0466</td>
</tr>
<tr>
<td>T0415</td>
<td>53.67</td>
<td>50.69</td>
<td>T0393</td>
</tr>
<tr>
<td>T0437</td>
<td>47.50</td>
<td>47.50</td>
<td>T0465</td>
</tr>
<tr>
<td>T0482</td>
<td>44.74</td>
<td>44.74</td>
<td>T0496</td>
</tr>
<tr>
<td>T0417</td>
<td>44.18</td>
<td>45.28</td>
<td>T0495</td>
</tr>
<tr>
<td>T0473</td>
<td>44.12</td>
<td>46.69</td>
<td>T0493</td>
</tr>
<tr>
<td>T0396</td>
<td>43.14</td>
<td>60.29</td>
<td>T0451</td>
</tr>
<tr>
<td>T0476</td>
<td>40.06</td>
<td>46.31</td>
<td>T0401</td>
</tr>
<tr>
<td>T0498</td>
<td>37.50</td>
<td>36.98</td>
<td>T0405</td>
</tr>
<tr>
<td>T0446</td>
<td>37.39</td>
<td>35.68</td>
<td>T0425</td>
</tr>
<tr>
<td>T0409</td>
<td>37.14</td>
<td>31.07</td>
<td>T0489</td>
</tr>
<tr>
<td>T0460</td>
<td>36.52</td>
<td>34.55</td>
<td>T0443</td>
</tr>
<tr>
<td>T0464</td>
<td>36.30</td>
<td>35.96</td>
<td>T0397</td>
</tr>
<tr>
<td>T0468</td>
<td>35.34</td>
<td>32.47</td>
<td>T0399</td>
</tr>
<tr>
<td>T0411</td>
<td>33.33</td>
<td>33.96</td>
<td>T0429</td>
</tr>
<tr>
<td>T0391</td>
<td>32.72</td>
<td>31.99</td>
<td>T0434</td>
</tr>
<tr>
<td>T0471</td>
<td>32.25</td>
<td>39.69</td>
<td>T0395</td>
</tr>
<tr>
<td>T0492</td>
<td>32.19</td>
<td>60.62</td>
<td></td>
</tr>
</tbody>
</table>

| Average GDT-TS | 33.15 | 34.34 |

Table 2.5: GDT-TS scores of the best of top 5 structures, corresponding to 41 target proteins from the Human and Server section of CASP8, generated with Rosetta and ranked with the Rosetta scoring function (RSR) and the MDR methods, respectively.
Figure 2.8: Top: Comparison of the GDT-TS scores of the best of top 5 structures, corresponding to 41 target proteins from the Human and Server section of CASP8, selected by the MUFOOLD-MD server using the MDR (Top5 MDR) and RSR (Top5 RSR) methods, respectively. Bottom: Difference between the GDT-TS scores obtained with the MDR and RSR methods for the same structures as in the Top panel. The green (red) bars indicate that the GDT-TS score of the structure picked by the MDR (RSR) method was at least 5 points bigger than the other one. For the blue bars the difference between the GDT-TS scores (in absolute value) was less than 5 points.

generated structures was extremely low, with GDT-TS score less than 10 (T0419, T0427, T0431, T0487).

2.7 Conclusions

The MDR method is a novel method for assessing the quality of structures predicted with homology or ab initio methods. We have shown that our method, based on MD simulations, is able to discriminate between high quality, i.e., cRMSD < 3 Å and low quality structures, i.e., cRMSD > 4 Å, with ~80% success rate. The MDR method performs very well for proteins that have high number of helical elements in their crystallographic structure, as seen in Section 2.5. The affinity of CHARMM force field for helical conformations [34] could provide an explanation for the good correlation between the MDR score and the cRMSD in the case of proteins with high ratio of helical to extended secondary structure elements in test set B. For these cases the MDR method performs better than the widely
used knowledge based scoring function methods such as OPUS-PSP or DFIRE-A. The lack of solvent (explicit or implicit) in our MD simulations most certainly has a negative impact on the quality of our MDR method as is well known that the polar aqueous solvent plays an important role in the stability of globular proteins. However, using any type of solvent during the MD heating process would considerably increase the computational time required by MDR ranking, rendering the method less attractive.

While the success of the MUFOLO-MD server at CASP8 competition in the FM category is greatly attributed to the Rosetta software, we also demonstrated that the MDR method outperformed (though by not too much) both the Rosetta scoring function based ranking (RSR) method, and the one used by the BAKER-ROBETTA server. Due to our completely new approach used in the MDR method, based on dynamic rather than static information of proteins, there is a high probability that the MDR method would select a model which is structurally very different from the ones selected with knowledge based methods. This was proved in CASP8 competition as model 2 submitted by the MUFOLO-MD server was found to be in a different structural cluster by the assessors when compared with models submitted by other servers. One of the weak points of the MUFOLO-MD server was the lack of accurate domain prediction. While the Rosetta software has very good performance in predicting ab initio structures for fairly short proteins (100-150 AA), when faced with sequences longer than 200 AA its predictions can be far from the real structure. In most of the cases from CASP8 our domain prediction was either very poor or missing. Because of this, Rosetta’s performance for targets with more than 200 aminoacids was minimal, thus decreasing the performance of the MUFOLO-MD server.

It should be noted that the success of the MDR method for ranking the predicted low-resolution structures in test sets A and in the MUFOLO-MD server relies heavily on how well the high-resolution (all-atom) structures are constructed by PULCHRA [35] and PSF-GEN. As an alternative one could use the Rosetta software to generate all atom models for a target protein, however, in most cases this would require prohibitively long computational
times, thus making this approach undesirable.
3 Conformation and oligomerization of gp41 fusion peptide
3.1 Introduction

Enveloped viruses such as human immunodeficiency virus (HIV) and influenza virus infect their target cells by a process involving cell-specific binding to the cell membrane followed by fusion of the viral enveloped membrane with cellular membranes [37]. The enveloped viral protein (fusion protein) is responsible for the actual membrane fusion step, leading to the release of viral contents into the target cell and subsequent infection. For many viruses, a small segment of the fusion protein usually located at the N-terminus is responsible for the early stage in the membrane fusion process [38]. This domain is usually referred to as the fusion domain or fusion peptide (FP). In the case of HIV-1 the glycoprotein gp41 is responsible for the membrane fusion process [39] and usually they are assembled as coiled-coil homotrimers [38]. The N-terminal gp41-FP is highly homologous with corresponding domains of other enveloped viruses [40], and its first 16 residues (AVGIGALFLGFLGAAG) are mostly hydrophobic.

A variety of experimental methods (NMR, FTIR, CD, ESR) as well as MD simulations have been used to study the structure of the FPs and the mode of interaction with the membrane. Studies to date have shown that both gp41 FP and the influenza HA2 FP display considerable structural plasticity. It can adopt either the monomeric $\alpha$-helical or the oligomeric $\beta$-sheet conformations. For shorter FPs (16 or 23 residues), a consensus has been reached that at higher peptide loading, i.e., lower [Lipid]/[Peptide] ratio ([L]/[P]), the oligomeric $\beta$-sheet is favored, while at higher [L]/[P], monomeric $\alpha$-helical form dominates. However, even at comparable [L]/[P] in different membrane mimics, a variety of results were obtained. $\alpha$-helical structures are present in micelles [41–43] while $\beta$-sheet structures are dominant in lipid bilayers with smaller curvature of the membrane than in micelles [44]. At the same time, previous investigations have demonstrated that the conformation and oligomeric state of the gp41 FP can be altered by changing the APL while,

This chapter is based on published article [36].
apparently, keeping the membrane curvature intact [45, 46].

These results suggest that a decrease in the APL leads to a shift from monomeric α-helical structure to oligomeric β-sheet structure. Thus we proposed that the conformation and the oligomeric state of the gp41 FP are determined by the APL and the effect of the curvature of the membrane surface on the conformation of the gp41-FP is mainly manifested through the induced change in the APL. Section 3.2 describes an all-atom MD simulation study that, combined with results from Fourier-transformed infrared spectroscopy (FTIR) measurements, seems to validate the above hypothesis.

While some studies have shown that both the α-helical and β-sheet structures of the FP are fusogenic [47–49], results of other work suggested that specific conformation was responsible for membrane fusion [46, 49]. However, the results of numerous such studies are filled with controversies. Structural characterizations of the oligomers formed by the FP also revealed rather contradictory results. SS NMR study of the 13C and 15N backbone - labeled FP23 in membrane by Weliky et al. suggested a mixture of parallel and anti-parallel β-sheet structures [50]. Gordon et al. using 13C-enhanced FTIR found anti-parallel structure for FP23 in solution, and in membranes at high peptide loading [51]. The same conclusion was reached by Castano et al. for FP23 in membrane FTIR and Brewster angle microscopy [44]. The anti-parallel structure was found to dominate at high peptide loading. However, a recent work of Sackett and Shai [52], using a much longer peptide segment, FP70, that includes the N-terminal haptad repeat (NHR) domain, found that the FP was assembled in a parallel β-sheet structure from 13C-enhanced FTIR. They argued that interactions in the NHR domain stabilize the parallel β-sheet structure which is the structure most favorable for inducing membrane fusion by penetrating and traversing the membrane.

From the above mentioned studies it is not clear whether the helical or the extended conformation is the one responsible for the fusion with membrane. Moreover it is not clear which conformation is more stable in the lipid membrane. In Section 3.3 we use all-atom
MD simulations to investigate the relative stability of four different gp41 FP23 oligomers (i.e., \(\alpha\)-helical dimer, \(\beta\)-sheet parallel and anti-parallel dimers, and \(\beta\)-sheet anti-parallel trimer) inserted into a POPE phospholipid bilayer under physiological conditions. Our results suggest that the parallel \(\beta\)-sheet dimer is the most stable structure among the four oligomers studied, thus supporting those experimental findings according to which this structure is the most favorable for initiating the HIV-1 membrane fusion.

3.2 Transition from \(\alpha\)-helical monomers to \(\beta\)-sheet oligomer

In this section we investigate the APL dependence of the conformation of gp41 by employing MD simulations. The goal is to test the hypothesis according to which a decrease in the APL results in a shift from monomeric \(\alpha\)-helical to oligomeric \(\beta\)-sheet conformation of gp41.

3.2.1 FTIR experiments

First, a brief description of the FTIR experiment, performed by Dr. Wong [36], that motivated the MD simulation study is provided. A reversed micellar system formed by Aerosol-OT (AOT, sodium dioctyl sulfosuccinate) was chosen as the membrane mimicking system for the experiment. The AOT reversed micelles provide a water/lipid interface with negative curvature, and with APL ranging from as low as 30 to about 50 \(\text{Å}^2\) [53], a range substantially lower than the membrane models used by previous studies, typically of 60–80 \(\text{Å}^2\) for planar bilayer surfaces and for micellar surfaces with positive curvature. The use of the reverse micelles adds membrane surface with negative curvature as mimics, which has not been used in the past in the studies of the conformation and membrane interaction for FPs, and it drastically extends the range of APL in this study to adequately test our hypoth-
Energy minimization and equilibration of the systems were carried out in several stages. First, the systems were subjected to 3×10^4 minimization steps by harmonically restraining the C_σ increased (or equivalently APL decreased) in time through the application of a corresponding to the built (equilibrated) systems A1, A2 and B were 61×61×31 layers of water, where the area of the membrane was kept fixed, corresponding to APL \approx 44 \text{ Å}^2 when the APL has decreased below 50 Å² \cite{42}. The FTIR spectra of the FP23/AOT reversed micelle sample show that as the solubil-

![Graph](image_url)

Figure 3.1: Top panel: Area per lipid (APL) as a function of the water/lipid ratio \(W_o\). Lower panels A-D: The amide-I region (1600 - 1700 cm\(^{-1}\)) of the FTIR spectrum (solid curve) of gp41-FP in the AOT reversed micelles corresponding to \(W_o\) values indicated by arrows in the top panel. The dashed curves represent the deconvoluted FTIR bands.

The reversed micellar system also offers a system where the membrane curvature, and thus the APL, can be fine-tuned by varying the water/lipid ratio \(W_o = \frac{\text{[water]}}{[\text{AOT}]}\) \cite{53}. The FTIR spectra of the FP23/AOT reversed micelle sample show that as the solubil-

ity of the micelle solution decreases from a value of 18 to a value of 5 (which corresponds to a decrease in APL from 51 Å² to 39 Å²) a FTIR peak appears at position 1627 cm\(^{-1}\) and its intensity increases relatively to the residual TFA peak at 1674 cm\(^{-1}\). This peak is specific to the β-sheet conformation of gp41. At the same time, at high values of APL a
broad signal at 1650 cm$^{-1}$ (obtained by deconvolution of spectral lines) is observed and is commonly assign to $\alpha$ helical or coiled structures. Unfortunately it is not easy to distinguish these two structures. Due to the low intensity and the broadness of this peak, it is also conceivable that it could have arisen from an artifact due to incomplete subtraction of the background signal. We believe that when the APL has decreased below 50 Å$^2$ ($W_o < 18$ in the AOT system) the lateral compression causes the $\alpha$-helical monomers to unfold into turn structures, and then the monomers in the transient turn structures aggregate and form $\beta$-sheets. The oligomeric $\beta$-sheet structure ultimately becomes the dominant form at the lowest end of the APL range used. Thus the results support our hypothesis that the APL determines the oligomeric state and the conformation of the gp41 FP. The monomeric $\alpha$-helical conformers of the FP turn increasingly into oligomeric $\beta$-sheet structure as APL decreases. However, the APL values sampled by the FTIR experiments extend only up to 50 Å$^2$ and in this experiment the presence of the $\alpha$-helical structure at this value is not clear. Therefore, to explore the conformational space of the gp41 oligomers at higher APL values one needs to resort to MD simulations.

### 3.2.2 Molecular dynamics simulations

To investigate the APL dependence of the conformational state and stability of the gp41-FP by means of MD simulations we have built three model systems (Fig. 3.2), referred to as A1, A2 and B1, starting from its all atom structure obtained from FTIR spectroscopy (PDB entry code 1ERF [51]) and employing the molecular modeling and visualization program VMD [1]. The $\alpha$-helical dimer was assembled by placing two protonated helical gp41 monomers parallel to each other and separated by a distance of $\sim 5$ Å. By rotating the monomers as rigid bodies about their axes, the energetically most favorable dimer conformation was obtained by: (i) minimizing the non-bonding interaction energies (van der Waals and electrostatic), and (ii) maximizing the number of hydrogen bonds between the monomers. This optimization process was carried out using VMD and its MDEnergy
Figure 3.2: (a) Side and (b) top views of the α-helical gp41-FP dimer inserted perpendicular into the upper leaflet of a POPE lipid bilayer (system A1). The peptides are shown in cartoon representation, the phosphorous atoms of the lipid head groups are shown in van der Waals representation while the rest of the lipid molecules and the solvent water molecules are shown in line representation. Similarly, side and top views of the gp41-FP dimers in α-helical (system A2) and β-sheet (system B1) conformations inserted into the POPE lipid bilayer parallel to its surface are shown in (c), (e) and (d), (f), respectively. The images were rendered with VMD [1].

The parallel β-sheet dimer was built starting from the helical dimer A1. By using IMD [55], the distance between residues 1 and 6 of both A1 monomers was increased from ∼ 23 Å to ∼ 47 Å, the peptide length corresponding to B1. Furthermore, the distance between the parallel monomers was reduced to a value (∼ 5 Å) that allowed for hydrogen bond formation.

Next, the gp41 oligomers were inserted in a pre-equilibrated patch of POPE lipid bilayer. The helices in A1 were oriented perpendicular to the plane of the bilayer. The penetration depth of the A1 helices into the lipid membrane was about 60% of their length. Thus, the hydrophilic tails (residues 17-23) of the monomers stayed outside the lipid bilayer, leading to the attachment of the protonated N-terminus to the hydrophilic lipid headgroups as observed in experiments. The first 16 residues of the oligomers A2 and B1 were completely inserted into the POPE bilayer, while the hydrophilic tails of the peptides were kept at the surface of the upper leaflet. All lipid molecules within a distance of 0.6 Å of
the protein atoms were removed. Finally, each system was hydrated by adding two ∼ 15 Å thick pre-equilibrated water layers to both sides of the lipid membrane. The electroneutrality of the systems was restored by adding 4 Cl- ions in A1, A2 and B1 using the Autoionize plugin of VMD. The final systems had a total of 26608 (A1), 35068 (A2) and 34573 (B1) atoms, respectively.

All the MD simulations in this chapter were performed with the parallel code NAMD2.5 [23]. We have used the force field parameters CHARMM22 [13, 14] for proteins and CHARMM27 [14] for lipids. Water molecules were described with the TIP3P model [56]. van der Waals interactions were cut at 12 Å and smoothly switched off between 10 and 12 Å. Full long-range electrostatic interactions were used by employing the particle mesh Ewald (PME) method [24]. The energy minimization was done with the congruent gradient method within the NAMD2.5 code. The SHAKE algorithm [25] was used for rigid bonds involving all hydrogen atoms. Newton’s equation of motion was integrated with a time step of 2 fs using the velocity Verlet method with periodic boundary conditions (PBC) to avoid finite size effects. All systems were treated as NpT ensembles [57]. The normal pressure (105 Pa) was kept constant by using a hybrid Nose-Hoover piston method with a period and decay time of 100 fs. The area in the xy plane of the lipid membrane was kept constant throughout the simulations allowing the periodic cell to fluctuate in the normal z direction. The temperature was kept constant at 310 K by employing a Langevin thermostat with a damping coefficient of 5 ps−1.

A set of five MD simulations in the NpT ensemble, referred to as S1,...,S5, were performed using the equilibrated systems A1, A2 and B1. The 30 ns long simulations S1 and S2 were designed to test the stability of the α-helical conformation of the gp41-FP monomers in system A1 against helix to coil transition triggered by the variation of APL. While in S1 the membrane area was kept fixed, corresponding to APL of 70 Å², in S2 the membrane area was decreased in time through the application of a negative surface tension σ = -50 mN/m. The purpose of simulations S3 and S4 of system A2 was to investigate
the APL dependence of the transformation of the $\alpha$-helical gp41-FP monomers into turns followed by the formation of $\beta$-bridges as precursors of the dimer in the $\beta$-sheet conformation. Since the time scales on which lipid molecules undergo lateral diffusion in the membrane and monomeric $\alpha$-helical gp41-FPs unfold within the membrane are orders of magnitude larger than the longest equilibrium MD simulations (\sim 100 \text{ ns}), we have resorted to nonequilibrium steered molecular dynamics [58] (SMD) simulations. Using SMD the forced unfolding of the $\alpha$-helical gp41-FP monomers (system A2) was brought about by pulling apart the $C_\alpha$ atoms of residues 1 and 16 through a spring while the reduction of the APL was mimicked by applying an attractive force $F_0 = 109 \text{ pN}/C_\alpha$ atom (turned on gradually during the first 1.5 ns of the SMD simulation) between the gp41-FP monomers. Finally, the purpose of the 15 ns long equilibrium MD simulation S5, corresponding to APL = 74 Å$^2$, was to investigate the conformational stability of the $\beta$-sheet gp41-FP dimer (system B), and to compare this with the unfolded structure of A2 at the end of the SMD simulations S3 and S4.

### 3.2.3 Constant and variable surface tension simulations

According to our hypothesis that the conformational and oligomeric state of the gp41-FP is mainly determined by the APL, the dramatic reduction of the APL during S2 should lead to the destruction of the $\alpha$-helical structure of the monomers A1, and their eventual transition into the more stable $\beta$-conformation at APL < 50 Å$^2$. To test this the secondary structure elements of the gp41-FP were calculated along both S1 and S2 MD trajectories by employing the STRIDE [59] algorithm as implemented in the program VMD [1] with the results shown in Fig. 3.3. At the beginning of the simulations S1, S2 the $\alpha$-helical regions of the A1 monomers extend only from residues 4 to 12 and 4 to 9, respectively. While during both S1 and S2 there are noticeable dynamic fluctuations in the lengths of the $\alpha$-helices (colored red in Fig. 3.3) the trend in the structural changes of the monomers is rather different in the two cases. In S1 the $\alpha$-helices are quite robust against turn- (gray regions
Figure 3.3: Time evolution of the color coded secondary structure elements of the two gp41-FP monomers during the 30 ns long equilibrium MD trajectories S1 and S2 of system A1 described in the text. The residue numbers are on the vertical axis. In each panel the top (bottom) graph refers to the first (second) monomer. The images were rendered with VMD [1].
in Fig. 3.3) and looser \( \pi \)-helix-structure (orange regions in Fig. 3.3) formation. During S2 the transition of the gp41-FP monomers into turn- and \( \pi \)-helical-structures appears to be gradual and lasting, in agreement with the hypothesis that a decrease in APL does not favor the \( \alpha \)-helical structure of the gp41-FP. Furthermore, careful analysis of the S1 and S2 MD trajectories indicate that the unwinding of the \( \alpha \)-helices are triggered by the occasional strong stochastic pulling exerted by the negative POPE phosphate group on the protonated N-terminus of the gp41-FP. Because of geometrical constraints, by decreasing the APL the rewinding of the helix becomes less favorable compared to its unwinding, thus at the end leading to the melting of the helical structure.

### 3.2.4 SMD simulations

As already mentioned, the transition of gp41-FP from \( \alpha \)- to \( \beta \)-structure in a lipid membrane occurs on a time scale that is beyond the reach of conventional MD simulation, even if a large negative surface tension is applied to mimic the reduction of the APL. In the non-equilibrium SMD simulations S3 and S4 the unwinding of the \( \alpha \)-helices in the gp41-FP monomers (system A2) was accelerated by applying an elastic stretching force between \( C_{\alpha 1} \) and \( C_{\alpha 16} \). In addition, to mimic reduction of the APL, in case of S4 the gp41-FP monomers were pushed one towards the other by applying a constant force between the corresponding centers of mass of residues 3-9 and 10-16, respectively.

The time evolution of the secondary structure elements during the 5 ns long SMD simulations S3 and S4 was monitored similarly to S1 and S2, and the results are shown in Fig. 3.4. Due to the applied SMD forces in both S3 and S4 it took only a few ns for the initial \( \alpha \)-helices (red) to melt into turns (gray) followed by the emergence of \( \beta \)-structures (green and blue). It should be noted that a quantitative analysis of S3 and S4, as well as, their direct comparison with S1 and S2 is rather difficult because of the large irreversible perturbation exerted on the system by the SMD forces. However, a direct comparison between S3 and S4 is meaningful and can be used to provide further insight into how the decrease in APL
Figure 3.4: Time evolution of the color coded secondary structure elements of the two gp41-FP monomers during the 5 ns long MD trajectories S3, S4 of system A2, and S5 of system B described in the text. The residue numbers are on the vertical axis. In each panel the top (bottom) graph refers to the first (second) monomer. The images were rendered with VMD [1].
accelerates and, therefore, favors the \( \alpha \) - to \( \beta \) -structure transition in gp41-FP dimers. As shown in Fig. 3.4, compared to S3 in S4 the melting of the \( \alpha \)-helices and the formation of the \( \beta \)-bridges and precursors of \( \beta \)-sheets occur faster, the latter also being more extended. Indeed, in Fig. 3.4 the locations of the \( \beta \)-structures in A2 at the end of S4 correspond to the ones in system B1 (gp41-FP dimer in initial \( \beta \)-sheet configuration) during the last 5ns of simulation S5. Furthermore, the stability of the gp41-FP dimer in \( \beta \)-sheet structure depends crucially on the position and number of hydrogen bonds between the monomers which only in case of S4 converge to the optimal ones from system B.

### 3.2.5 Conclusions: \( \alpha \) to \( \beta \) transition

Previous structural studies of the HIV gp41 fusion peptides in membrane environments led to a variety of structures, resulting in suggestions of conformational plasticity of the FP [60, 61], and its relevancy to the fusogenicity of the fusion domain. In our study, a major factor affecting the structure (and oligomeric state) adopted by the HIV gp41-FP, the surface area of the membrane or APL, has been revealed. Our hypothesis, that the structure and oligomeric state of the FP is determined by the APL of the membrane surface, and that the effects of changes in the curvature of the membrane surface on the structure of the FP are manifested by the changes in the APL, has been verified in the above FTIR experiments and further corroborated by MD simulations. The FTIR studies of Dr. Wong allowed us to extend the APL range to a low value region (35.50 Å\(^2\)) heretofore unexplored. The extension of the APL range, from normal micelles to planar bilayers/monolayers to reversed micelles, enables us to confirm convincingly that a decrease in the APL, due to changes in curvature or due to other factors such as the addition of Ca++ ions [45, 46] or an applied lateral pressure [44], leads to the oligomerization of the gp41-FP and the formation of \( \beta \)-sheet structure, while larger APL, such as in the case of normal micelles, favors monomeric helical structure.

The implication of the results of this work on the fusion process induced by the HIV
fusion peptide is a more complicated issue. First the APL used in this study is substantially smaller than usually found in lipid bilayers. However, it is generally believed that negative curvature (and smaller APL) in cell membranes is developed during the early stage of the fusion process. In such a situation, small APL comparable to those in the current study may be achieved.

So far we were mainly concerned with the transition from $\alpha$ helical structure of the FP into the parallel $\beta$ sheet configuration. However, there are experimental results [44] that suggest the anti-parallel $\beta$ sheet form of the FP as the more stable structure and the one responsible for the fusion of the virus with the cellular membrane. In the next section we compare the stability of several gp41 oligomers, i.e., parallel helical dimer, parallel and the anti-parallel $\beta$-sheet dimers and the anti-parallel $\beta$-sheet trimer.

### 3.3 Stability of gp41 oligomers

To investigate the relative stability of different gp41 FP23 oligomers we have considered four systems: (1) a $\alpha$-helical dimer (referred to as system SA); (2) a parallel $\beta$-sheet dimer (SB1); (3) an anti-parallel $\beta$-sheet dimer (SB2); and (4) an anti-parallel $\beta$-sheet trimer (ST).

Systems SA and SB1 are identical with systems A1, respectively B1, from the first section. The structure of the anti-parallel $\beta$-sheet dimer (SB2) was derived from SB1 by inverting the orientation of one of the monomers. The $\beta$-sheet trimer (ST) was built by adding a third stretched helical monomer to SB1, oriented anti-parallel to the other two monomers, with equidistant separation between all pairs of monomers. The procedure for inserting the SB2 and ST in the POPE patch was identical with the one followed in Subsection 3.2.2 for systems A2 and B1. The electro-neutrality of the systems was restored by adding 4 and 6 Cl- ions in SB2 and ST, respectively. The final systems had a total of 26608 (SA), 34573 (SB1), 34912 (SB2) and 37107 (ST) atoms, respectively. After $\sim$
105 steps of energy minimization (using the congruent gradient method), each system was subjected to equilibrium MD simulations. The lengths of the MD runs were as follows: 30 ns for SA (simulation SIM1), 15 ns for SB1 (simulation SIM2), 10 ns for SB2 (simulation SIM3), and 12 ns for ST (simulation SIM4).

Analysis of the MD trajectories for each FP oligomer allows us to: (i) determine the energetically most favorable conformation and (ii) identify the important interactions contributing to their structural stability. For each MD run, we monitored the secondary structure, the interaction energy and the $\pi$-stacking interaction between the FP monomers.

### 3.3.1 Parallel $\alpha$-helical dimer

The main goal of our MD simulation SIM1 of the helical dimer SA was to investigate the effect of the dimeric interaction on the behavior of the two gp41 helical monomers inserted in a lipid bilayer. Already during energy minimization, the N-terminus of one of the SA monomers attached to a phosphorus atom due to the strong electrostatic interaction between the protonated N-terminus and the negatively charged phosphate groups (left inset to Fig. 3.5). Not surprisingly the N-terminus of the second SA monomer attached to another phosphorus atom at the beginning of simulation SIM1. The fusogenic importance of the protonated N-terminus in connection with other positively charged aminoacids (e.g., ARG22) of the 23 residues long gp41 FP has been observed in previous FTIR experiments [62]. The time evolution during SIM1 of the distance between the two N-termini is shown in Fig. 3.5.

As soon as SIM1 started, both SA monomers moved rather quickly towards the surface of the upper leaflet of the lipid bilayer. During this motion the initially parallel orientation of the monomers was completely lost, with the projections of the axes of the SA monomers on the plane of the membrane being approximately perpendicular to each other (right inset to Fig. 3.5). Based on previous studies involving single gp41 helical monomers, it is widely accepted that an oblique insertion of the monomer is required to destabilize the
Figure 3.5: Distance between N termini of the two peptides in SA during SIM1. Insets represent snapshots of SA right after energy minimization (left; side view), and after 5ns of equilibration (right; top view).

membrane. The tilted orientation of the helical monomer is associated with its asymmetric hydrophobicity gradient [63–66]. The angle made by the axis of the peptide with the plane of the membrane is reported to be $\sim 50^\circ$ [65]. Throughout SIM1, the tilt angle of the axes of the helical SA monomers (with respect to the plane of the lipid membrane) fluctuated about an average value of $30^\circ \pm 5.5^\circ$ and $2.4^\circ \pm 5.5^\circ$, respectively. The difference from the angle reported in literature could be due to the interaction between the two monomers.

The time evolution of the secondary structure elements of SA during SIM1 was monitored. Since SIM1 is identical with simulation S1 from Section 3.2 we will use Fig. 3.3 when we refer to the time evolution of the secondary structure elements of system SA during SIM1. As shown in Fig.3.3, top panel, initially in both SA monomers the helical structure comprised residues 3 to 16. During SIM1 the helices remained relatively stable with apparent fluctuations of their residue lengths. For the first monomer, the size of the helix (practically unchanged during the first 16 ns of SIM1) reduced from ten residues (3
to 12) to only six residues (3 to 8). It is interesting to note that after complete loss of the helical structure during the time interval from 21 to 24 ns the helix refolded to its previous residue length. The helical structure of the second monomer appeared to be more robust against fluctuations. During SIM1 (30 ns), the length of the second helix fluctuated between eight (4 to 11) and six residues (4 to 9). These results are in agreement with other studies that show a slight break in the helix at residue GLY10 in DPC micelles [67, 68]. Also, previous MD simulations show that the 12 N-terminus residues of gp41 in helical form are necessary and sufficient to destabilize the membrane without requiring the full 23 residues long gp41 FP [64].

Figure 3.6: Time evolution of the non-bonded (sum of van der Waals and electrostatic) interaction energy between the two monomers of SA (dash-dotted line), SB1 (solid line), and SB2 (dashed line) from 2 to 10 ns of the corresponding MD simulations (SIM1, SIM2 and SIM3, respectively). The gray lines correspond to the non-bonded interaction energies between peptide 1 and 2 and peptide 2 and 3, respectively of system ST during SIM4.

The time evolution (during the first 10 ns of SIM1) of the non-bonding (sum of van der Waals and Coulomb electrostatic) interaction energy between the two SA monomers was also calculated by employing the CHARMM force field. The result is shown in Fig. 3.6
(dash-dotted line). The small value of this energy, together with the lack of hydrogen bonds between the SA monomers suggest that there is very little (if any) concerted attractive interaction between the SA monomers, which tend to act independently from each other. By contrast, the non-bonding interaction energy between the SB1 and SB2 monomers, also shown in Fig. 3.6, are much larger, suggesting that the attraction between the monomers of the β dimers may have an important role in their functions.

### 3.3.2 Parallel and anti-parallel β-sheet dimers

The β dimers SB1 and SB2 reached the surface of the lipid bilayer at the very beginning of the simulations SIM2 and SIM3, respectively. While during SIM2 the N termini of the parallel SB1 dimer were buried most of the time within the membrane, during SIM3 the N-termini of the anti-parallel dimer SB2 were located outside the membrane. The distance between the N-termini of the SB1 monomers was fluctuating around 10 Å.

Similarly to SA, we have followed the time evolution of the secondary structure elements of SB1 and SB2 during SIM2 and SIM3, respectively. The results are plotted in Fig. 3.7. For SB1, the β-sheet structures were formed at the very beginning of SIM2 and were prevalent throughout the MD simulation involving mostly three groups of residues: 3 to 4, 7 to 8 and 12 to 15 for the first monomer and 3 to 4, 7 to 8 and 12 to 16 for the second monomer. Occasionally, for short times, the rest of the residues exhibited an extended structure but predominantly they had a random coil or turn structure. During almost the entire SIM3, both SB2 monomers had a β-sheet structure extending from residue 7 to 13. Representative snapshots of the dimers SB1 and SB2 are shown in Fig. 3.8 using cartoon representation. The green segments indicate the mean location of the hydrogen bonds. The thickness of the green segments is proportional to the cumulative lifetime of the corresponding hydrogen bonds. The presence of hydrogen bonds was determined every 10 ps and their occurrence was averaged over a running window of 1 ns. On average, SB1 had the most hydrogen bonds (4.5), followed by SB2 (3.7) and SA (0.07). Based on this
result, one may conclude that the SB1 dimer was somewhat more stable than SB2 and both β dimers were much more stable than the α dimer SA.

![Figure 3.7](image)

Figure 3.7: Time evolution of the color coded secondary structure elements in SB1 (top panel) and SB2 (bottom panel) during simulations SIM2 and SIM3, respectively. The residue numbers are on the vertical axis. The figure was generated with the Timeline plugin of VMD.

We investigated the tilt angles of both SB1 and SB2 dimers with respect to the plane of the membrane. In the case of SB1 the β-sheet formed by the two peptides was parallel with the plane of the membrane throughout the entire simulation SIM2. For SB2, the position of the N-termini, which were in direct contact with the bulk water, appeared to be related to the overall tilt orientation. We observed that within a few picoseconds into SIM3 the plane of the β-sheet structure reoriented perpendicular the membrane plane. Then, during the 10 ns of SIM3, the angle between the β-sheet of SB2 and the plane of the membrane decreases gradually from \( \sim 90^\circ \) to \( \sim 50^\circ \). We note that Castano et al.
observed tilt angles of the anti-parallel β-sheet gp41 of $\sim 36^\circ$ in multibilayers of DOPC/Cholesterol/DOPE/DOPG, and $\sim 30^\circ$ in single bilayer of DMPC. While our tilt angles are larger than these experimental values, based on their time evolution, it is likely that a continuation of SIM3 would further reduce the gap between the experimental and simulated tilt angles.

Next, we focused on the interactions (stacking) between aromatic residues that may provide additional stability for the parallel β dimer SB1. The existence of $\pi - \pi$ interactions was suggested in the case of gp41 parallel β trimer between the benzene rings of
residues PHE8 and PHE11 [52]. In order to quantitatively characterize the $\pi-\pi$ interactions between residue pairs PHE8-PHE8 and PHE11-PHE11 in SB1 and residue pairs PHE8-PHE11 and PHE11-PHE8 in SB2, we have followed the time evolution of the: (1) distance $d$ between the centers of mass of the benzene rings in the corresponding residues, and (2) the angle $\theta$ (between the normal vectors to the planes of the benzene rings). A strong $\pi-\pi$ interaction requires $d \sim 5-6$ Å and $\theta \sim 0^\circ$. The analysis of the data indicates a fairly strong $\pi-\pi$ interaction only between the PHE8 residues of the SB1 monomers. Thus, from this viewpoint the parallel SB1 dimer should possess extra stability compared to the anti-parallel SB2 dimer. Also, the same conclusion can be reached from the plot in Fig. 3.6, according to which the non-bonding attractive interaction energy between the SB1 monomers is larger than the one between the SB2 monomers.

### 3.3.3 $\beta$-sheet trimer

We have also studied the behavior of a gp41 $\beta$-sheet trimer (ST), inserted in a hydrated POPE lipid bilayer, by a 12ns long MD simulation (SIM4). During SIM4, ST was located most of the time at the surface of the bilayer. The first 16 residues of the three ST monomers formed a plane parallel to the plane of the membrane. Similarly to the $\beta$-dimers SB1 and SB2, to investigate the stability of ST during SIM4 we have followed the time evolution of the secondary structure elements, the dynamics of the hydrogen bonding between the ST monomers, and the existence of $\pi$-stacking interactions involving residues PHE8 and PHE11.

The time evolution of the secondary structure elements in the ST monomers are shown in Fig. 3.9a, while a representative snapshot of ST (using cartoon representation) is shown in Fig. 3.9b. During SIM4, all three monomers exhibit transient extended structures with fluctuating lengths. Clearly there is a correlation between the proximity of the monomers and the location of $\beta$-strands on them. The 2nd (middle in Fig. 3.9) monomer has the most $\beta$-strands, involving residues 5 to 8 and 12 to 18, which form hydrogen bonds with
Figure 3.9: a) Time evolution of the color coded secondary structure elements in ST during simulation SIM4. The top, middle, and bottom graphs refer to the first, second, and third monomers, respectively. The residue numbers are on the vertical axis. b) Snapshot of the equilibrated ST trimer with the mean location of hydrogen bonds between the monomers shown as green segments. The thickness of each segment is proportional to the cumulative lifetime of the corresponding hydrogen bond.
the corresponding extended structures on the other two monomers, i.e., residues 10 to 14 of monomer 1 (top in Fig. 3.9) and residues 2 to 7 of monomer 3 (bottom in Fig. 3.9). As shown in Fig. 3.9b, the hydrogen bonds that bind the β-strands in the monomers lead to a relatively stable β-sheet like structure for the ST trimer. Similarly to the anti-parallel SB2 dimer, the anti-parallel ST trimer also lacks the stability enhancing π–π interaction between the benzene rings of the PHE11 and PHE8 residues of adjacent monomers.

3.3.4 Conclusions: Oligomer stability

We found that the interaction between the two α-helical monomers in the lipid membrane is very weak, thus rendering the helical dimer unstable. The almost perpendicular orientation in the plane of the membrane, small number of hydrogen bonds and low interaction energy between the two monomers suggests that the two peptides behave as independent monomers. By contrast, the β-sheet dimers are strongly bound due to the significant number of hydrogen bonds between the two peptides. Furthermore, the parallel β-sheet dimer appears to be more stable than the anti-parallel one, as indicated by the larger interaction energy, number of hydrogen bonds and the stabilizing π-stacking interactions between the phenylalanine residues. The anti-parallel β-sheet trimer behaved similar to the anti-parallel β-sheet dimer the main difference being the higher occurrence of β-sheet elements for the peptide in the middle, most probably due to its favorable position between the other two peptides.

In conclusion, these results indicate the parallel β-sheet dimer as the most stable gp41 oligomer in agreement with our results from the previous section that support a transition from an initially parallel α helical dimer to a β-sheet oligomer.
Calculating free energy profiles and diffusion coefficients from nonequilibrium molecular dynamics simulations
4.1 Introduction

The study of the structure-function relationship of large biomolecules often requires to follow their dynamics on a meso- or even macroscopic time scale while retaining its atomic scale spatial resolution. A typical example is molecular and ion transport through channel proteins [70]. While structural details are needed at atomic resolution in order to determine the forces that guide the diffusion of the particles across the channel, the duration of the permeation process may exceed by several orders of magnitude the time scale of all atom MD simulations (tens of nanoseconds). In such cases, a simplified alternative approach is to model the transported molecule in the channel as an overdamped Brownian particle that diffuses along the axis of the channel in the presence of an effective potential of mean force (PMF) that describes its interaction with the rest of the atoms in the system [71]. A PMF is the Landau free energy profile along a reaction coordinate (RC), or order parameter [72], and it can be determined from the equilibrium statistical distribution function of the system by systematically integrating out all degrees of freedom except the RC [73, 74]. In principle, both the effective diffusion coefficient and the PMF, quantities that enter the Langevin equation of motion, which determines the dynamics of the transported molecule, can be determined from MD simulations. In practice, however, the calculation of free energy differences and PMFs are rather difficult and computationally expensive [73, 75]. Since even the longest equilibrium MD trajectories can sample only a small region of the RC domain of interest, the one situated in the vicinity of the corresponding PMF minimum, simple equilibrium MD simulations are not suitable for PMF calculations. The traditional method for calculating PMFs by means of biased equilibrium MD simulations is umbrella sampling (US) [73, 76, 77]. However, US may become inefficient and computationally unaffordable when too many sampling windows are required. This may happen when the amplitude of the equilibrium fluctuations of the RC is very small compared to the size of the RC interval.

This chapter is based on published article [69].
in which the PMF is sought. In such cases the RC can be sampled efficiently by employing SMD in which the system is guided, according to a predefined protocol, along the RC by using a harmonic guiding potential (HGP). By choosing a sufficiently large value for the elastic constant of the HGP, i.e., within the stiff-spring approximation [78, 79], the distance between the target and actual value of the RC at a given time can be kept below a desired value.

In general, for a large system ($\gtrsim 10^5$ atoms) computationally one can afford only a limited number ($\sim 10$) of such non equilibrium SMD pullings, and the real challenge is to find a way to reconstruct the PMF (at least semiquantitatively) along the RC using this limited amount of data. In principle, the equilibrium PMF can be reconstructed from the Jarzynski equality (JE) that relates the equilibrium free energy difference $\Delta F$ between two states to the exponential average of the external work $W$ done along all non equilibrium paths that connect those states and are subject to the preestablished variation protocol of the RC [80, 81]. The validity of JE depends crucially on a small fraction of trajectories with $W_d < 0$, which transiently violate the second law of thermodynamics. Since such trajectories (whose number decreases exponentially with the mean dissipative work $\overline{W}_d > 0$) are very unlikely to occur among a few fast SMD pullings, clearly the sought PMF cannot be determined by the direct application of JE, except when the pulling paths are close to equilibrium (i.e., when $\overline{W}_d < k_B T$).

We have developed a new method for calculating the PMF and the position dependent diffusion coefficient from forward and reverse SMD pullings with a stiff spring along the reaction coordinate in both F and R directions (the FR method). We assume that the pullings are done with a stiff spring in which case the F work distribution $P_F(W)$ is Gaussian, and according to Crooks fluctuation theorem the R work distribution $P_R(W)$ is also Gaussian. In this chapter we describe different methods for calculating PMFs, the FR method and its application to PMF calculation in the case of water transport through SWNTs. A stochastic model that describes the transport process at microscopic level and
The first step in calculating the PMF is the identification of the reaction coordinate, whose change in time should describe the evolution of the state of the system [73]. A proper reaction coordinate can be the projection of the center of mass of a molecule (or of a part of the molecule) on the axis of the pore in the case of molecular transport through nanopores (e.g., channel protein or SWNT), the angle between two units of a complex protein that undergoes conformational changes or the distance between two ends of a protein in the case of protein unfolding. In our study the reaction coordinate is the projection of the center of mass of a molecule moving through a nanopore on the z axis which coincides with
the axis of the nanopore. The PMF $U(z)$ is by definition determined from the equilibrium distribution function of the system by integrating out all degrees of freedom except the reaction coordinate $z$ [73], i.e.,

$$e^{-\beta U(z)} \equiv p_0(z) = \int d\Gamma e^{-\beta H_0(\Gamma)} \frac{1}{Z_0} \delta[z - \tilde{z}(\Gamma)] ,$$  \hspace{1cm} (4.1)

where $p_0(z)$ is the equilibrium distribution function of the reaction coordinate, $Z_0$ is the partition function, $z$ represents the target value of the reaction coordinate, $\tilde{z} \equiv \tilde{z}(\Gamma)$ is the actual value of the reaction coordinate, and $\delta(z)$ is the Dirac-delta function whose filtering property guarantees that the integrand in Eq.(4.1) is nonzero only when the reaction coordinate has the desired value, i.e., when $\tilde{z}(\Gamma) = z$. Throughout the rest of the chapter, for the simplicity of mathematical calculations, $\beta$ has been set to 1.

The ideal way of computing the equilibrium distribution function $p_0(z)$ is by sampling the configuration space through equilibrium MD simulations. $p_0(z)$ is proportional to the logarithm of the binned histogram of the reaction coordinate sampled along the MD trajectory and the PMF can be calculated directly using

$$U(z) = -\log[p_0(z)] .$$  \hspace{1cm} (4.2)

The limitation of this method is the lack of sampling of the reaction coordinate in the domain of interest, even in the case of very long equilibrium MD simulations. This is due to the short time scale available through the present all atom MD techniques.
4.3 Harmonic guiding potential, Umbrella Sampling method

Sampling energetical regions that are not accessible during free equilibration simulations, e.g., potential barriers, can be achieved by inserting an external potential that will guide the system towards those regions. A feasible solution is the use of an harmonic guiding potential that will force the system to evolve in such a way that at all times the reaction coordinate $z$ follows its target value $\tilde{z}$

$$V_{\tilde{z}}(\tilde{z}) \equiv V(\tilde{z}(\Gamma) | z) = \frac{k}{2} [\tilde{z}(\Gamma) - z]^2,$$  \hspace{1em} (4.3)

where $k \equiv k_z$ is the stiffness (elastic constant) of the harmonic guiding potential. The corresponding force acting on atom “$j$” is

$$F_j = -\frac{\partial V_{\tilde{z}}}{\partial r_j} = -k [\tilde{z}(\Gamma) - z] \frac{\partial \tilde{z}(\Gamma)}{\partial r_j}.$$ \hspace{1em} (4.4)

With the additional harmonic guiding potential, the Hamiltonian of the new biased system becomes $H_{\tilde{z}} = H_0 + V_{\tilde{z}}(\tilde{z})$. The free energy difference between the equilibrium states of the systems described by the Hamiltonians $H_{\tilde{z}}$ and $H_0$ can be written as a Gaussian convolution of $\exp[-U(z)]$

$$e^{-\delta F_{\tilde{z}}} = \int d\Gamma \frac{e^{-\beta H_0(\Gamma)}}{Z_0} e^{-V_{\tilde{z}}(\tilde{z}(\Gamma))} = \left\langle e^{-V_{\tilde{z}}(\tilde{z})} \right\rangle_0$$
$$= \int d\tilde{z} e^{-U(\tilde{z})} e^{-\frac{k}{2} [\tilde{z} - z']^2}.$$ \hspace{1em} (4.5)

The idea of HGP allowed the development of another method for calculating the PMF from equilibrium simulations, the Umbrella Sampling (US) [73, 76, 77, 82]. In US the range of reaction coordinate values of interest $(z_{min}, z_{max})$ is divided into $N_w$ sampling windows centered about conveniently chosen values $z_i$, $i = 1, \ldots, N_w$. The reaction coordinate is sampled
in each window separately by preparing identical replicas of the system and applying the harmonic guiding potential \( V_{zi}(\bar{z}) \). As a result, the biased distribution functions \( p_i(z) \) can be readily obtained by direct sampling of the reaction coordinate for the biased system. The standard method for efficiently stitching together the biased \( p_i(z) \)'s in order to obtain the equilibrium \( p_0(z) \), and therefore the sought PMF, is the so called \textit{weighted histogram analysis method} or WHAM [82]. When applicable, umbrella sampling combined with WHAM is perhaps the best choice for calculating PMFs. In practice, however, one often encounters situations in which the minimum number of sampling windows required to properly cover the range of reaction coordinate values of interest is excessively large and the application of the method may become computationally unattainable. Molecular transport in channel proteins is a good example. In these cases non equilibrium methods for calculating the PMF, such as the FR method, can be very efficient. In the next two sections we introduce the stiff spring approximation and Crook’s transient fluctuation theorem which are key points in the development of the FR method.

### 4.4 Stiff spring approximation

By using a large value of \( k \) in the harmonic guiding potential, i.e., stiff spring approximation, in Eq. (4.5) the reaction coordinate will follow closely the target value \( z \). One can determine the PMF with a spatial resolution \( \delta z \) by choosing a value of \( k \gg 2/(\delta z)^2 \). In this case the main contribution to the integral in Eq. (4.5) will come from the region \( |z - z'| \ll \delta z \) and the right term in Eq. (4.5) can be expanded in Taylor series about \( z \) [83]

\[
e^{-\delta F_z} = \int dz' e^{-\frac{k}{2}(z-z')^2} e^{-U(z)} \left\{ 1 - U'(z)(z-z') - \frac{1}{2} \left[ U''(z) - U'(z)^2 \right] (z-z')^2 + ... \right\}.
\]
Using the first order term and integrating, the above equation becomes
\[ e^{-\delta F_z} \approx e^{-U(z)} \int dz' e^{-\frac{k}{2}(z-z')^2} = \sqrt{\frac{2\pi}{k}} e^{-U(z)}. \] (4.6)

which is the first order correction to the stiff spring approximation. By taking the logarithm of both sides, we find that the PMF of the unbiased system is well approximated by the free energy difference of the system biased by the harmonic guiding potential

\[ \Delta U = U(z) - U(z_0) \approx \Delta F = F_z - F_{z_0}. \] (4.7)

where \( U(z_0) \) is a constant.

### 4.5 Transient Fluctuation Theorem, JE

Considering a system in equilibrium, through SMD simulations one can modify in time the target value of the reaction coordinate according to a prescribed protocol, e.g., increase it with constant velocity \( z(t) = z(0) + vt, \quad 0 \leq t \leq \tau \), where \( v \) is the constant pulling speed. We denote as forward pulling paths the trajectories for which the time increases from 0 to \( \tau \). For time reverse pulling paths, the reaction coordinate is decreasing in value from \( z(\tau) \) to \( z(0) \) according to \( z(t) = z(\tau) - vt \). Thus by SMD one can sample fast, depending on the value of the pulling velocity, the reaction coordinate by driving the system out of equilibrium. The work done in the pulling process is given by integrating the force introduced by the harmonic guiding potential along the reaction coordinate

\[ W_t \equiv W_z = \int_{z_0}^{z(t)} dz \left[ \frac{\partial V_z(z)}{\partial z} \right] = k \int_{z_0}^{z(t)} dz (z - \tilde{z}). \] (4.8)

In order to relate the work of pulling to the free energy of the system, we use Crooks fluctuation theorem. According to Crooks, if a system fulfills several conditions (i.e., is
in thermal equilibrium, it is driven from that equilibrium by an external perturbation, the energy of the system is finite, the dynamics are Markovian and the dynamics preserve the equilibrium ensemble if the system is unperturbed), then the following relation holds [84]

\[ \langle f(W)e^{-WdF} \rangle_F = \langle f(-W) \rangle_R, \quad (4.9a) \]

or

\[ \langle f(W) \rangle_F = \langle f(-W)e^{-WdR} \rangle_R. \quad (4.9b) \]

Here \( f(W) \) is an arbitrary function of the work \( W \), and

\[ \langle \ldots \rangle_{F/R} = \int dW P_{F/R}(W) \ldots, \quad (4.10) \]

represents the average over forward/reverse paths or, equivalently, the average with respect to the forward/reverse work distribution functions \( P_{F/R}(W) \). The dissipative work in a F/R process is given by

\[ W_{dF/R} = W_{F/R} \mp \Delta F, \quad (4.11) \]

with \( \Delta F = F_z(\tau) - F_z(0) \).

By setting \( f(W) = 1 \) in Eqs. (4.9), one can derive the JE which relates the free energy difference between two equilibrium states of the system to the work of pulling between the two states

\[ \langle \exp(-W_{dF}) \rangle_F = \langle \exp(-W_{dR}) \rangle_R = 1, \quad (4.12a) \]

\[ \langle \exp(-W) \rangle_F = e^{-\Delta F}, \langle \exp(-W) \rangle_R = e^{\Delta F}. \quad (4.12b) \]

JE is most frequently used with the cumulant approximation [78, 79, 83] for calculating the free energy difference from repeated unidirectional nonequilibrium SMD simulations. By taking the logarithm of both parts in Eq. (4.12b) and using both the stiff spring and the
cumulant approximations one obtains the PMF

\[ \Delta U(z) \approx \Delta F = -\log \langle \exp(-W_z) \rangle \approx \bar{W}_z - \sigma_z^2 / 2 , \quad (4.13) \]

where \( \bar{W}_z \) is the mean work of pulling in either F or R directions and \( \sigma_z^2 = \bar{W}_z^2 - \bar{W}_z^2 \) is the variance (second cumulant) of the work. Despite its simplicity, the above relation is valid only close to equilibrium because SMD pulling paths can sample only a narrow region about the peak of the Gaussian \( P_F(W) \). This allows for a fairly accurate determination of the mean work \( \langle W_z \rangle \) but, in general seriously underestimates the variance \( \sigma_z^2 \).

By setting the arbitrary function as \( f(W') = \delta(W - W') \) in Eqs. (4.9) we obtain another useful relation that connects the F and R work distribution functions known as Crooks transient fluctuation theorem

\[ \frac{P_F(W)}{P_R(-W)} = e^{W_{df}} . \quad (4.14) \]

### 4.6 PMF from forward and reverse SMD pullings with a stiff spring

In this section we present our new method for calculating PMFs from few fast SMD pullings along the reaction coordinate in both F and R directions, hereafter referred to as the FR method. We assume that the pullings are done with a spring for which the stiff-spring approximation holds. In this case, the F work distribution \( P_F(W) \) is Gaussian, and according to Crooks fluctuation theorem (4.14) it follows that the R work distribution \( P_R(W) \) is also Gaussian. Thus one can write

\[ P_{F/R}(W) = \left( 2\pi \sigma_{F/R}^2 \right)^{\frac{1}{2}} \exp \left[ -\frac{(W - \bar{W}_{F/R})^2}{2\sigma_{F/R}^2} \right] \quad (4.15) \]
where $\overline{W}_{F/R}$ and $\sigma_{F/R}^2$ are the mean work and variance corresponding to the F and R pulling directions, respectively. The mean dissipative work in the two distinct pulling direction is

$$\overline{W}_{dF/R} = \int dW (W \mp \Delta F) P_{F/R}(W) = \overline{W}_{F/R} \mp \Delta F .$$

Inserting (4.15) into (4.14) and taking into account that $W_{dF} = W - \Delta F$, after little algebra it follows that Crooks fluctuation theorem can hold only if

$$\sigma^2 \equiv \sigma_{F}^2 = \sigma_{R}^2 = \overline{W}_{F} + \overline{W}_{R} \quad (4.17a)$$

and

$$\Delta F = (\overline{W}_{F} - \overline{W}_{R})/2 . \quad (4.17b)$$

Using Eq. (4.17b) in Eq. (4.16) we find that the mean dissipative work is the same in both directions and is

$$\overline{W}_{d} \equiv \overline{W}_{dF} = \overline{W}_{dR} = (\overline{W}_{F} + \overline{W}_{R})/2 . \quad (4.17c)$$

The above equations yield with the same accuracy both the PMF and the mean dissipative work. Unlike the cumulant approximation that needs a high number of unidirectional pullings and can only determine the mean total work $\overline{W}_{F}$ with some statistical correction, the FR method requires a few SMD pullings that can sample reasonably the work about the peak position $\overline{W}_{F/R}$ of $P_{F/R}(W)$, as indicated by the shaded regions in Fig. 4.1. From the mean dissipative work, one can easily determine the diffusion coefficient by assuming that $\overline{W}_{d}$ is proportional to the pulling speed $v$. The slope of $\overline{W}_{d}$ is the position dependent friction coefficient $\gamma(z)$ and the corresponding diffusion coefficient is given by the Einstein relation (in $k_B T$ energy units)

$$D(z) = \gamma(z)^{-1} = v \left( d\overline{W}_{d}(z)/dz \right)^{-1} . \quad (4.18)$$
The calculated $U(z)$ and $D(z)$ can be further used in the equation of motion of the reaction coordinate on a meso (or macro) time scale, given by the Langevin equation corresponding to an overdamped Brownian particle [71]

$$\gamma(z)\dot{z} = -dU(z)/dz + \xi(t),$$

or equivalently, the corresponding Fokker-Planck equation for the probability distribution function $p(z,t)$ of the reaction coordinate

$$\partial_t p(z,t) = -\partial_z j(z,t) = \partial_z D(z)\partial_z p(z,t) + \partial_z U'(z)p(z,t),$$

where $\xi(t)$ is the Langevin force (modeled as a Gaussian white noise) and $j(z,t)$ is the probability current density.

In the following we calculate the PMF that guides the translocation of water molecules across a periodic structure of densely-packed SWNTs, and the corresponding position dependent diffusion coefficient. First we use the equilibrium distribution function and the US methods to calculate the PMF that will be later compared with the one obtained from the FR method.

### 4.7 SWNT as model for channel protein

SWNTs are hydrophobic nanopores that can be regarded as simplified models for the much more complex channel proteins. Thus, they are ideal for testing new computational methods and hypothesis that later can be applied to protein channels. The water filled SWNTs are nontrivial many particle systems comprising thousands of atoms, yet they are easy to simulate and the PMF of waters inside the SWNTs can be easily tuned by changing the van der Waals interaction parameters between the carbon and water molecules [85]. These simulations revealed that hexagonally packed (6,6) SWNTs, with diameter of 8.1 Å, spon-
taneously fill with a single file of water molecules when connecting two water reservoirs. We consider a periodic system (see Fig. 4.2) of 4 hexagonally-packed identical SWNTs of (6,6) armchair type. Each SWNT (156 atoms) has a C–C diameter of 8.2 Å and length 14.7 Å. On both sides of the SWNTs there is a water layer of width 18.9 Å. The system contains 556 water molecules in total. The unit cell has dimensions $23 \times 20 \times 52.5$ Å$^3$ and contains a total of 2292 atoms. All MD simulations were performed in the NpT ensemble ($T = 300$ K and $p = 1$ atm), using periodic boundary conditions and the PME method for full electrostatics [86]. Water molecules were modeled as TIP3P [56]. To facilitate the comparison between the PMFs obtained with different methods, the Van der Waals parameters of the C atoms (of type CA for benzene in the CHARMM force field) [14] were changed (from $\varepsilon = 0.10$ to $\varepsilon = 0.13$ kcal/mol, and from $R_0 = 3.76$ to $R_0 = 4.81$ Å, respectively) to artificially increase the size of the potential barriers in the PMF from 0.35 to 2 $k_B T$. All simulations were performed with the program NAMD2 [54], with a perfor-
mance of $\sim 1$ day/ns on 8 CPUs of a G4 Beowulf cluster (preferred for repeated SMD pullings), or $\sim 12$ hours/ns on 24 CPUs (preferred for long equilibrium MD simulations).

Just like in previously reported simulations [85, 87–89], the initially empty SWNTs filled up completely with water (i.e., 5 molecules per nanotube) in the first few hundreds of ps. Also, the arrangement of the SWNTs prevented water molecules from entering the space between them. Water molecules diffuse across the tubes in a concerted fashion, with a diffusion rate close to the corresponding bulk value. This correlated motion can be described rather well with a continuous-time random walk (CTRW) model [89] but in this case we use a more general stochastic model in which the motion of each water molecule along the $z$–axis of a SWNT is characterized by an effective (position dependent) diffusion coefficient $D(z)$ and a PMF, $U(z)$.

### 4.8 PMF of water in SWCNT from equilibrium MD simulations

First we calculated the PMF from equilibrium simulation using both the equilibrium distribution of waters in the SWNT and the umbrella sampling with WHAM. Given the small size of the system one affords longer equilibrium simulations that allow the calculation of an accurate PMF that can be used later as a comparison test for the PMF obtained with the FR Method. The PMF $U_0(z)$ from Eq. (4.2) was determined from a 9 ns long equilibrium MD trajectory recorded after the system was equilibrated. The histogram $p_0(z)$ was constructed by binning the $z$-coordinate of the O-atoms of all water molecules. No visible change in the normalized distribution $p_0(z)$ could be noticed when the first 7 ns part of the equilibrium MD trajectory was used to build it, indicating that the sampling was complete. Inside the SWNTs (see Fig. 4.3a) $U_0(z)$ has five equidistant minima (water binding sites) with separation distance 2.8 Å and almost identical potential barriers of height $2 k_B T$. It is convenient to label these minima from 1 to 5 along the positive $z$-direction. On both sides,
Figure 4.3: (a) PMF $U_0(z)$ of a water molecule along the $z$-axis of one of the SWNTs obtained through equilibrium MD simulations. The included snapshot illustrates a completely filled SWNT with five water molecules located about the corresponding PMF minima. (b) Comparison between $U_0(z)$ (thin line) and the same PMF $U_{US}(z)$ (thick line) obtained from umbrella sampling. Graphics rendered with the program VMD [1].
moving away from the SWNTs into the bulk water the PMF exhibits three more minima (labeled \(0, -1, -2\) and \(6, 7, 8\), respectively) before it flattens out. Water molecules to move in and out the SWNTs, i.e., to hop between minima \((0, 1)\) and \((5, 6)\), must overcome roughly the same energy barrier as the ones located inside the tubes. However, there is a strong spatial inhomogeneity of the water distribution right outside the nanotubes that is related to the large asymmetry of the energy barrier connecting minima \((-1, 0)\) and \((6, 7)\), respectively. The PMF profile is reflected by the snapshot of the water molecules in Fig. 4.3a and is compatible with the observation that single-file water transport through SWNTs usually occurs in unidirectional bursts. We have also determined the PMF, \(U_{US}(z)\), inside the SWNTs by using umbrella sampling and WHAM. A total of six sampling windows were used. For convenience, these were centered, by means of harmonic guiding potentials with \(k = 1.2\) kcal/mol·Å\(^2\), on the six maxima within the SWNTs of \(U_0(z)\). The samplings of the biased systems were carried out through 5 ns long equilibrium MD simulations. To speed up the computation, the guiding potentials in the four SWNTs were centered on different maxima. Thus each equilibrium MD trajectory provided four biased distribution histograms \(p_i(z)\). The fact that these were properly sampled was tested by making sure that the histograms corresponding to the first 4 ns part of the equilibrium MD trajectory coincided with the one obtained from the entire trajectory. Finally, \(U_{US}(z)\) was determined by using WHAM. As shown in Fig. 4.3b, the agreement between the calculated \(U_0(z)\) and \(U_{US}(z)\) is rather good, though not perfect. While the reconstructed PMF is valid only for a completely filled SWNT, the umbrella sampling method can also be used to determine the PMF of water in partially filled SWNTs [90].
4.9 PMF of water in SWCNT from nonequilibrium SMD pullings

By employing our new FR approach described in Section 4.6, the PMF $U_{FR}(z)$ was determined from a small number of fast F and R SMD pullings of water molecules across the SWNTs. In each SMD simulation four water molecules were pulled across the SWNTs (one molecule per nanotube) by applying a stiff ($k = 10$ kcal/mol $\cdot$ Å$^2$) harmonic guiding potential (see Eq. (4.3)) that moved with $v = 20$ Å/ns along the $z$-axis of the nanotubes. Only four such pullings were performed in both F and R directions between the extremities of the interval $z \in [-10, 10]$ Å. Each SMD simulation was started from an equilibrated configuration (in accordance with the applicability of Crooks fluctuation theorem) and was 1 ns long. Out of the $4 \times 4 = 16$ F and R trajectories only those where retained for analysis in which the corresponding SWNT remained filled with water at all times. In several cases, once the pulled water molecule crossed halfway the channel the binding sites behind it remained unoccupied. Since such configurations correspond to a different free energy profile, such trajectories must be dropped in determining the PMF for a completely filled SWNT. Thus, we ended up with 7 F and 14 R paths for calculating the PMF. Because we already know the “exact” PMF $U_0(z)$, we deliberately did not choose to add more trajectories from extra simulations. Indeed, since in the case of large biomolecules one can afford only a small number of SMD runs, our goal here is to test the viability of the proposed FR method for calculating PMFs under such unfavorable conditions. The external work along the F and R paths, including the mean work $\overline{W}_{F/R}$, are shown in Fig. 4.4a and 4.4b, respectively. Note that in order to display $W_R$ on the same plot with $W_F$, the sign of the former needs to be reversed and shifted to the origin of the latter. As shown in Fig. 4.4c, within the SWNTs (indicated by dashed vertical lines) the PMF $U_{FR} = (\overline{W}_F - \overline{W}_R)/2$ agrees surprisingly well with $U_0$ and $U_{US}$. We have checked (results not shown) that by increasing the pulling speed to $v = 40$ Å and using a similarly small number of F and R trajectories, the
Figure 4.4: (Color online) Work along (a) forward and (b) reverse SMD pullings (dashed lines). The mean work $W_{F/R}$ is shown as thick solid line. (c) Comparison between $U_0(z)$ (dashed line) and $U_{FR}(z) = (W_F - W_R) / 2$ (solid line), obtained from fast forward and reverse SMD pullings. Vertical dashed lines indicate the extremities of SWNTs.
quality of the obtained PMF is very similar to the one shown in Fig. 4.4c. However, in this particular case, the higher the pulling speed the most likely that the SWNTs will partially empty during pulling and, therefore, more runs are necessary to collect a minimum number of paths for calculating the PMF. As discussed in Section 4.5, for Crooks fluctuation theorem to be valid it is necessary that the initial states of both F and R pullings be sampled from an equilibrium distribution. Thus, strictly speaking the above results using the FR method applies only to the two ends of the considered interval. The good agreement between $U_{FR}$ and $U_0$ suggests that our method may give reliable PMFs for all values of the reaction coordinate $z$ in the considered interval. However, it is simple to extend our FR method to cases where this issue may impact negatively on the determination of the PMF. Thus, the reaction coordinate interval was divided into 40 segments of the same length. For each division point, the system was equilibrate for a few hundreds of ps by using the same harmonic guiding potential centered about those points. Starting from statistically independent equilibrium configurations, 4 pullings with the same $v = 10 \ \text{Å/ns}$ in both F and R directions were carried out on each segment. None of the SWNT emptied during these short SMD runs and, therefore, all trajectories were used for analysis. Close inspection of calculated PMF suggests that compared to the “exact” $U_0$, $U_{FR-40}$ is not as good as $U_{FR}$. Thus, one may conclude, that more sampling in the FR method does not necessarily give better results. Indeed, in the FR method we only need a good estimate of the mean F and R work, and not a complete sampling of the corresponding work distribution functions as for methods using JE.

We also calculated the PMF using JE with the cumulant approximation method, separately to F and R trajectories. The calculated PMFs were biased in both directions, behavior recognized in previous work in which the PMF of a glycerol molecule in a GlpF channel was calculated for the first time [78]. This bias is due to the invalidity of JE for few, fast unidirectional pullings.
4.10 Dissipative work and diffusion coefficient

Next, we focus on the determination of the mean dissipative work and the corresponding diffusion coefficient. In Fig. 4.5a the mean dissipative work derived from the individual F/R pullings and from the FR method are plotted. As expected, $\overline{W}_{dF/R} = \sigma_{F/R}^2 / 2$ calculated from the variance of $W_{F/R}$ seriously underestimate $\overline{W}_d$ determined from the FR method by using Eq. (4.17c). This observation has several consequences. First, the fact that $\overline{W}_{dF/R}$ does not increase fast enough with the pulling distance clearly indicates that only a small region about $\overline{W}_{F/R}$ of $P_{F/R}(W)$ is sampled and not the entire work distribution function.

Figure 4.5: (a) Mean dissipative work determined from the FT method (thick-solid) and the cumulant approximation applied separately to forward (thin-solid) and reverse (thin-dashed) pullings. (b) Validity test of JE along forward (solid) and reverse (dashed) processes. The PMF $U$ in the corresponding dissipative work $W_{dF/R} = W_{F/R} \mp U$ is determined from the FR (thick) and the cumulant approximation (thin) methods, respectively.
Second, the strongly biased PMFs $U_{F/R}$, obtained from the cumulant approximation, lead to underestimated dissipative work $W_{dF/R} = W_{F/R} \mp U_{F/R}$ that give the false impression that the JE equation is satisfied along the F/R pullings, as shown in Fig. 4.1b (thin lines). This, of course, is expected because $U_{F/R}$ are calculated based on the assumption that JE holds. The reality is that, in fact, JE fails to hold for both F and R pullings as the system departs from equilibrium. The reason, of course, is that paths with negative dissipative work ($W_d < 0$) that are crucial for the validity of JE (Eq. (4.12)) are not sampled. This is clearly illustrated in Fig. 4.5b where $\langle \exp(-W_{dF/R}) \rangle$, plotted by using the correct expressions $\overline{W}_{dF/R} = W_{F/R} \mp \Delta U$ (thick lines), decay rapidly towards zero as the system is pulled away from equilibrium. Clearly, the larger the deviation from equilibrium the less JE is satisfied.

The position dependent $D(z)$ can be calculated from the slope of $\overline{W}_d$ according to Eq. (4.18). Since the mean dissipative work is almost linear it is not surprising that the diffusion coefficient has an almost constant value $D \approx 71 \text{ Å}^2/\text{ns}$. This is more than three times smaller than the bulk diffusion coefficient of water $D_{\text{bulk}} \approx 250 \text{ Å}^2/\text{ns}$.

### 4.11 Stochastic model of water transport in SWNTs

The determined $U(z) \equiv U_{FR}(z)$ and $D$ provide the input in the FPE Eq. (4.19b) for describing water transport through SWNT on meso/macro time scales. This should be regarded as a generalization of CTRW model of Berezhkovskii and Hummer [89]. In principle, by solving the FPE for the nonequilibrium distribution function $p(z,t)$ for well defined initial and boundary conditions one can completely characterize the single-file transport of water molecules in the considered SWNTs. In the CTRW model single-file water molecules occupy the binding sites (PMF minima) within the SWNT. Since they cannot pass each other, the diffusion of water molecules across the nanotube is brought about by random hops to the empty binding sites right in front or behind them. The waiting (or residence)
time between two consecutive hops is a stochastic Poisson process. Besides the equidistant spacing between two adjacent sites \( a \), the mean waiting time \( \tau \) is the defining parameter of the CTRW model. In terms of \( \tau \) the effective diffusion coefficient is \( D_{\text{eff}} = \frac{a^2}{2\tau} \).

In our stochastic model \( \tau \) is identified with the mean first passage time [91] (MFPT) from one minimum \( (z_i, i = 1, \ldots, 5) \) of the PMF \( U(z) \) into the adjacent one \( z_j \), with \( j = i \pm 1 \), and is given by

\[
\tau_{i,j} = \int_{z_i}^{z_j} dx e^{U(x)} / D(x) \int_{z_i}^{x} dy e^{-U(y)} .
\]

(4.20)

Now, the mean waiting time can be expressed as

\[
\bar{\tau} = \left( \sum_{i=1}^{N-1} \tau_{i,i+1} + \sum_{i=2}^{N} \tau_{i-1,i} \right) / 2(N-1) .
\]

(4.21)

In our case \( N = 5 \) and the corresponding mean waiting time \( \bar{\tau} \approx 84 \) ps. Applying our stochastic model to the pristine SWNT considered in Ref. [89] (for which the barrier height between binding sites is only 0.35 \( k_B T \) compared to 2 \( k_B T \) in our modified SWNTs) one obtains \( \bar{\tau} \approx 12.9 \) ns that compares very well with the reported 13 ns. Furthermore, the effective diffusion coefficient \( D_{\text{eff}} \) of single-file water molecules in SWNTs can be defined as

\[
D_{\text{eff}} = D \left( \frac{\bar{a}^2}{2D\bar{\tau}} \right) ,
\]

(4.22)

where \( \bar{a} = 2.8 \) Å is the mean spacing between two adjacent binding sites. \( D_{\text{eff}} \) describes the diffusion of fictitious particles in the absence of the PMF with the same mean diffusion time on a distance \( \bar{a} \) as the mean waiting time \( \bar{\tau} \). In our case we get \( D_{\text{eff}} \approx 45 \) Å\(^2\)/ns. It is this diffusion coefficient that can be measured from the well known asymptotic formula \( \langle \Delta z^2(t) \rangle = 2D_{\text{eff}}t \) from equilibrium MD simulations. Indeed, from our simulations we obtain \( D_{\text{eff}} \approx 48 \) Å\(^2\)/ns, in very good agreement with the result from our stochastic model.

Finally, one can calculate the mean permeation time \( T \) across the channel in two different ways: (i) as the MFPT from one end of the nanotube to the other, and (ii) as \( L^2 / 2D_{\text{eff}} \).
where $L$ is the length of the SWNT. In both cases one obtains essentially the same result: $T \approx 1.45$ ns between $z_1$ and $z_5$, and $T' \approx 3.2$ ns between $z_0$ and $z_6$ (i.e., between the binding sites right outside the ends of the SWNTs). The observed 12 permeations per nanotube in 9 ns corresponds to a permeation time 1.38 ns that is a good estimate for $T$ but it is considerably shorter than $T'$. Thus, even in this relatively simple case very long equilibrium MD simulations are needed to calculate the unidirectional water flux through the modified SWNTs by simply counting the number of full permeations of water molecules, reinforcing once again the value of our stochastic modeling approach.

4.12 Conclusions

In this chapter we have shown that by employing Crooks fluctuation theorem [84] within the stiff-spring approximation the potential of mean force along a suitably chosen reaction coordinate can be determined (at least semiquantitatively) from combining a few fast forward and time reversed nonequilibrium processes started from an equilibrium configuration and subject to the same evolution protocol of the reaction coordinate. In the proposed FR method one determines simultaneously both the PMF ($U$) and the mean dissipative work ($W_d$) without invoking JE. In fact, JE is not even satisfied for fast F or R pullings simply because processes with negative dissipative work (that transiently violates the second law and are exponentially small in number) are not sampled. The FR method is based on a key observation involving Crooks fluctuation theorem: whenever the F work distribution function $P_F(W)$ is Gaussian (e.g., in the case of the stiff-spring approximation) then $P_R(W)$ is also Gaussian. Furthermore, $P_{F/R}(W)$ have the same width and are shifted by precisely twice the corresponding free energy difference between the equilibrium states connected by the F and R processes. Thus, both $U$ and $W_d$ can be readily determined from the mean F and R work ($\overline{W_{F/R}}$). The practical success of the FR method stems from the fact that the mean work $W_{F/R}$ can be measured rather accurately from only a few fast
F/R pullings. This also explains why previous methods, based on the direct application of JE, fail to work away from equilibrium, making them inefficient for practical applications. Indeed, the width of $P_{F/R}(W)$, which is proportional to $W_d$, cannot be determined even approximately from a few unidirectional pullings, unless these are close to equilibrium and rendering $P_{F/R}(W)$ sufficiently narrow. The FR method works rather well for both small and large (e.g., biomolecular) systems. Although here we applied and tested the FR method in the context of SMD simulations, in principle this can be applied equally well to analyze properly designed single molecule experiments.

To test its viability, we have applied the FR method to determine the PMF and position dependent diffusion coefficient of single-file water molecules in SWNTs. The derived PMF was found to be in good agreement with the one obtained from standard equilibrium MD methods, e.g., umbrella sampling and, together with the diffusion coefficient were used in a stochastic model that provides a generalization of the recently proposed CTRW model for single-file water transport in SWNTs [89].
5 Cellular particle dynamics simulations of multicellular systems
5.1 Introduction

Even though morphogenesis, the complex of processes of structure formation in multicellular living systems, is under strict genetic control, eventually simple physical principles can capture essential aspects of the self-assembly of cells into tissues. Developmental biology data led Steinberg [92, 93] to propose a guiding principle, later used in many models for cell aggregates, known as the differential adhesion hypothesis (DAH). DAH states that structure formation in multicellular systems occurs due to (i) differences between the cell adhesion apparatus of different types of cells and from (ii) cell motility; cells seek positions which allow them to establish the largest possible number of strong bonds with their neighbors. Experiments designed to confirm Steinberg’s DAH, have shown that under certain conditions living tissues and multicellular aggregates behave as highly viscoelastic liquids. Tissue liquidity, brought about by cellular adhesion and motility, forms the basis of the newly developed bioprinting technology [94], which is used to design and build 3D tissue constructs by employing computer-controlled layer-by-layer deposition of bioink (submillimeter size cell aggregate) droplets onto biopaper (biocompatible gel). The simplicity and predictive power of DAH inspired theoretical modeling of structure formation in multicellular systems. There exist computational models that treat the cell as a collection of coupled viscoelastic elements [95], contiguous spins defined on a discrete lattice [96], deformable viscoelastic ellipsoids [97] or as a conglomerate of subcellular particles [98]. However, none of these models was applied for studying the formation of multicellular complex structures that are produced through the bioprinting process. In order to describe and predict the self-assembly process of bioprinted multicellular constructs we have developed a computer simulation method referred to as cellular particle dynamics (CPD). In our model each individual cell in the multicellular system is a collection of so-called cellular particles (CPs) interacting through short range contact forces. In addition the integrity of the cell is maintained by a properly designed confining potential. CPs from different cells
interact through a strong short range potential. The time evolution of the system can be predicted by means of computer simulations (similarly to MD simulations of biomolecules), i.e., by numerically integrating the equation of motion of each cellular particle. Because the inside of a living cell represents a highly viscous and fluctuating environment, the dynamics of the CPs can be best described as an overdamped Brownian motion, i.e., CPs obey an overdamped Langevin equation of motion.

The CPD method can be used to predict biophysical properties of tissues starting from cellular and subcellular characteristics. In order to achieve this we have developed a theoretical and computational framework that uses as input experimental results of basic, easy to reproduce, biological processes. The two processes considered here are the fusion of two multicellular spheroids and the uniaxial compression of a multicellular spheroid placed between two parallel plates. In principle, the force parameters that describe the interactions between the particles in the CPD method can be determined by using the results of biophysical measurements of the studied multicellular system. The experimental and the simulated properties of the system are related by the continuum theory of viscoelastic materials (rheology). The following sections describe in detail the CPD method, our framework for expressing the CPD parameters in SI units and both experimental and simulation results of the fusion and compression processes mentioned above. The experimental and CPD simulation results are connected through continuum theory of viscoelastic materials. The “calibration” procedure of the CPD parameters is illustrated in the case of Chinese hamster ovary (CHO) multicellular spheroidal aggregates.

5.2 CPD method

We propose a cell model which makes no assumptions on cell shape. A cell is represented as a set of interacting cellular particles (CPs). In the particular model that we present here, each cell is made of 10 CPs that interact with CPs from the same cell via a potential $V_{\text{intra}}$
Figure 5.1: Left: Interaction potential energies (top) and forces (bottom) between CPs from same (blue) and different (red) cells. Right: Aggregate made of 500 model cells, each with 10 CPs. Two of the neighboring cells are presented in close caption and colored in red and green. CPs are represented as solid spheres. The (transparent) cellular surface is determined by the envelope of the intracellular CPs. The images were rendered with VMD.
and with CPs form neighbor cells via a potential $V_{\text{inter}}$. The integrity of biological cells is maintained by the cell membrane anchored to an interconnected network of protein filaments, the cytoskeleton. To mimic the effect of the cellular membrane, we have subjected all CPs within a cell to a harmonic confining potential. The contact interaction between CPs in a cell was modeled through a 12-6 LJ potential, which contains both an attractive (adhesion) component and a repulsive (excluded volume) component. The interaction between CPs from different cells is described by a modified LJ interaction, $V_{\text{inter}}$, which is strong and has a range of the order of the size of the CPs, thus being able to describe the adhesion between adjacent cells.

$$V_{\text{intra}} = 4\varepsilon_1 \left[ \left( \frac{\sigma_1}{r} \right)^{12} - \left( \frac{\sigma_1}{r} \right)^6 \right] + \theta(r - \alpha) \left[ \frac{k}{2}(r - \alpha)^2 \right], \quad (5.1)$$

$$V_{\text{inter}} = 4\varepsilon_2 \left[ \left( \frac{\sigma_2}{r + \delta} \right)^{12} - \left( \frac{\sigma_2}{r + \delta} \right)^6 \right], \quad (5.2)$$

where $\varepsilon_1, \varepsilon_2$ are the depths of the LJ potential wells, $\sigma_1$ is the effective linear size of a CP, $\sigma_2$ and $\delta$ are related through $\delta = 2^{1/6}(\sigma_2 - \sigma_1)$ and determine the shape of $V_{\text{inter}}$, $r$ is the distance between the two CPs, $\theta(r)$ is the Heaviside step function, $\alpha$ is a distance of the size of the cell, and $k$ is the elastic constant of the confining potential (and is related to the cell stiffness). Both the interaction potentials and the corresponding forces are displayed in Fig. 5.1. The time evolution of the CPs is obtained by integrating the following overdamped Langevin equations of motion

$$\mu \frac{d\vec{r}_{i,n}}{dt} = \vec{\zeta}_{i,n} + \vec{F}_{\text{intra},i,n} + \vec{F}_{\text{inter},i,n} \quad (5.3)$$
where \( \mu \) is the friction coefficient, \( \tilde{\xi} \) is the random force, and \( \vec{F}_{\text{intra}} \) and \( \vec{F}_{\text{inter}} \) are forces of interaction between CP \( i \) and CP \( j \) and are given by

\[
\vec{F}_{\text{intra},i,n} = -\sum_j \partial V_{\text{intra}}\left(|\vec{r}_{i,n} - \vec{r}_{j,n}|\right) / \partial \vec{r}_{i,n} \quad (5.4a)
\]

\[
\vec{F}_{\text{inter},i,n} = -\sum_{m,j} \partial V_{\text{inter}}\left(|\vec{r}_{i,n} - \vec{r}_{j,m}|\right) / \partial \vec{r}_{i,n} \quad (5.4b)
\]

In Eqs. (5.3), (5.4a) and (5.4b) indices \( i \) and \( j \) refer to CPs and indices \( m \) and \( n \) refer to cells. We model \( \tilde{\xi}(t) \) as a Gaussian white noise with zero mean and variance \( \langle \xi_i(t)\xi_j(0) \rangle = 2D\mu^2\delta(t)\delta_{ij} \), where \( D \) is the self diffusion coefficient, \( \delta \) is the Dirac delta function, and \( \delta_{ij} \) is equal to 1(0) for \( i = j(i \neq j) \).

### 5.2.1 Implementation in LAMMPS

The CPD method was implemented in the parallel code Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [99]. LAMMPS (http://lammps.sandia.gov) is a modular software package used for simulating a wide range of materials, from biomolecules and polymers to coarse-grained, mesoscopic systems.

The overdamped Brownian dynamics (OBD) integrator was included in LAMMPS by modifying one of the existing integrators. The interaction potential \( V_{\text{intra}} \) was implemented as a special type of bond (identified as “cell” type) by modifying one of the existing bond modules. The cell bond was applied only between CPs with the same molecule index. In our case a molecule represents a CPD cell. Therefore CPs from the same cell are identified with the same molecule index. \( V_{\text{intra}} \) was tuned to allow flexibility for the shape of the cell. The small value of \( \varepsilon_1 = 1.48 \) gives the CPs within the cell high mobility. However, the \( \alpha = 2 \) and \( k = 10 \) prevents the CPs from escaping the cell through “evaporation”. The effective linear size of a CP is \( \sigma = \sigma_1 \). The CPD (computer) units were defined by setting \( D = \mu = \sigma = 1 \).
The interaction potential $V_{\text{inter}}$ was implemented as a pair-wise interaction of type LJ/expand, which is already present in LAMMPS and is given by Eq. (5.2). $V_{\text{inter}}$ was applied as a pair-wise interaction between CP $i$ and all the other CPs, excluding the ones with the same molecule index as particle $i$. The shape of $V_{\text{inter}}$ from Fig. 5.1 is given by $\varepsilon_2 = 40$ and $\sigma_2 = 5$. The potential was switched to 0 when the distance between two CPs from different cells, $r$, became larger than $1.5 \sigma$.

In order to start a CPD simulation, LAMMPS requires two input files: a configuration file, and a data file with the coordinates of the CPs. The configuration file contains a list of commands that are executed by LAMMPS. These commands specify all the characteristics of the simulation. Below are included some of the commands used in our CPD simulations with their values in brackets: 1) Type of units (LJ type - sets the fundamental quantities $\sigma$, $\varepsilon$ and the Boltzmann constant equal to 1), 2) Type of atom (bond - specific to bonded particles), 3) Boundary conditions, 4) Name of data file, 5) Type of bond (“cell”) with values for bond parameters (parameters of $V_{\text{intra}}$ defined above), 6) Type of pair-wise potential (LJ/expand) with values for potential parameters (parameters of $V_{\text{inter}}$ defined above), 7) Integration timestep expressed in time units ($tu$), 8) Type of integrator (OBD) with values for $\mu$ and $D$ ($\mu = D = 1$), 9) Name of output file for writing information about the energy of the system and frequency of output, 10) Name of trajectory (DCD) file for writing out the coordinates and frequency of output, 11) Minimization command including number of steps, and 12) Run command which is the actual integration of the equation of motion including the number of timesteps.

The data file is divided into several sections. The first section contains information about the number of particles, number of particle types, number of bonds, number of bond types, dimension of the simulation box and mass of particle. The second section contains (in six columns) the index of the particle (from 1 to $N_{\text{CPs}}$), the index of the molecule (from 1 to $N_{\text{cells}}$), the index of the particle type and the $x$, $y$ and $z$ coordinates of each CP. The third section contains all the bonds between the CPs described on four columns as follows:
the index of the bond, the index of the bond type and the indices of the two particles that are bonded.

Once the configuration and data files were built, the total energy of the initial system was minimized using the conjugate gradient algorithm. Then the equation of motion was integrated with a time step $10^{-4}\ t u$. The boundary conditions, when not periodic (PBC), were set to shrink-wrapped (SBC) allowing a flexible simulation box that would encompass all the CPs. During the simulations, the coordinates of all the CPs were saved into a dcd file every $tu$.

5.2.2 CPD simulations - equilibration, fusion and compression

The CPD spheroid with 500 cells and 10 CPs/cell was created from 5000 CPs arranged in a close-packed lattice. The xyz file with the coordinates of each CP was modified into an input data file. Each 10 consecutive CPs were assigned the same molecule (cell) index. The final data file was used as input for the LAMMPS code. The energy of the CPD aggregate was minimized and then a free equilibration was run for 2500 $tu$ with SBC.

For the fusion simulation, two identical copies of the equilibrated aggregate with 500 cells were placed next to each other along the $x$ axis. Their centers were placed at a distance equal to one aggregate diameter. This system was then equilibrated with SBC for 2500 $tu$.

The uniaxial compression of the multicellular spheroid was implemented in LAMMPS by defining a virtual moving wall. The wall was modeled by applying an external elastic force to all the CPs found beyond the current position of the wall. The force was oriented perpendicular to the plane of the wall, pointing in the direction of motion of the wall. For the compression simulation, two moving walls were placed at the top and the bottom of the spherical aggregate perpendicular to the $z$ axis. Once the equilibration started, the two walls started moving towards the center of the aggregate with constant velocity of $0.394 \sigma/tu$. After 10 $tu$ the two moving walls were stopped. At this point the diameter of the initial aggregate was reduced by 40%. With the two walls fixed at their current positions,
we carried out a free equilibration for 2000 $tu$. We recorded the total force exerted by the two aggregates on the two walls during the equilibration simulation.

5.3 Predicting biophysical properties of tissues using the CPD method

Finding the right values for the parameters in the interaction potentials $V_{\text{intra}}$ and $V_{\text{inter}}$ from Eqs. (5.1) and (5.2) is not a trivial task. The simulations with optimized parameters need to reproduce both fusion and compression processes. Once the right combination of parameters has been found, the basic length, time and energy units, i.e., $L = \sigma$, $T = \sigma^2/D$ and $E = E_T = \mu D$, need to be determined. These units are needed to relate dimensionless quantities obtained in the computer simulations to the corresponding experimental quantities with proper SI units. To this end, we compare the results of independent biomechanical experiments (fusion and compression, performed on multicellular systems) with those of the corresponding CPD simulations. This comparison is done in terms of viscoelastic quantities, which appear in the analytic expressions we derive from the continuum or phenomenological theory of viscoelasticity for the appropriate experimental situation. We describe both simulation and experimental results in terms of tissue-level viscoelastic parameters: the surface tension, $\gamma$, the viscosity, $\eta$, and the shear modulus, $G$.

For each of the three experimental biomechanical quantities $(\gamma, \eta, G)$, we have corresponding simulated ones $(\tilde{\gamma}, \tilde{\eta}, \tilde{G})$. The experimental value is measured directly in SI units while the simulated value is expressed in the proper CPD unit. Therefore:

$$\gamma = \tilde{\gamma} \left( \mu D / \sigma^2 \right), \quad \eta = \tilde{\eta} \left( \mu / \sigma \right), \quad G = \tilde{G} \left( \mu D / \sigma^3 \right).$$

(5.5)

Solving these equations we obtain the CPD units expressed in SI units:
\[
\sigma = \left(\frac{\gamma}{\tilde{\gamma}}\right) \left(\frac{G}{\tilde{G}}\right), \quad \mu = \sigma \left(\frac{\eta}{\tilde{\eta}}\right), \quad D = \left(\frac{\sigma^2}{\mu}\right) \left(\frac{\gamma}{\tilde{\gamma}}\right).
\] (5.6)

In order to determine \(\gamma\), \(\eta\) and \(G\) we apply continuum and phenomenological models to the fusion and compression experiments which yield three relations between these parameters. A relation between \(\gamma\) and \(\eta\) is determined from the fusion of two spherical aggregates through the characteristic fusion time. The surface tension \(\gamma\) and a second relation between \(\eta\) and \(G\) are obtained from the uniaxial compression of the spherical cellular aggregate. In the following two sections we illustrate how the experimental and CPD simulation results, for both the fusion and compression processes, can be related through continuum theory.

### 5.4 Fusion of two aggregates

#### 5.4.1 Continuum model

Cell aggregate fusion is used to estimate the ratio between tissue viscosity and surface tension [100] by following the time course of early stages of the fusion. The method relies on the analogy between cell aggregates and droplets of highly viscous liquids, for which the early stages of fusion (sintering) has been described analytically by Frenkel [101]. Recently, fusion experiments and Frenkel’s continuum approach were extended to the full time span of cell aggregate fusion [102].

The fusion process of two viscous liquid drops is driven by surface tension and is resisted by viscosity. The continuum model for the fusion of the two identical aggregates has been treated in a previous paper [102]. Based on this model, we use the angle \(\theta\), defined in Fig. 5.2, to describe the evolution in time of the geometrical shape of the aggregate during fusion. Since the time dependence of the angle \(\theta(t)\) does not depend on the volume conservation, this can be used to describe the fusion of multicellular aggregates where cells proliferate or die as long as the density of the aggregate remains (approximately) constant.
Figure 5.2: Fusion of highly viscous polymer droplets [2]. The red circles represent schematics of the spherical caps that model the shape of the fusing droplets.

In Fig. 5.2 are shown several snapshots from the coalescence process of two highly viscous droplets. During the fusion process, the two droplets are approximated by two identical spherical caps of radius $R(t)$. The shape of the two fusing sphere caps is characterized by the instantaneous angle $\theta$ defined as $\sin(\theta) = r(t)/R(t)$. Based on the continuum theory [102] this quantity can be approximated with

$$\sin^2(\theta) = \left(\frac{r}{R}\right)^2 \approx 1 - e\left(-\frac{t}{\tau}\right)$$  \hfill (5.7)

By fitting this expression to experimental and simulated data we obtain characteristic fusion times $\tau$ for both cases. The characteristic fusion time is related to the biomechanical properties of the tissue, i.e., surface tension $\gamma$ and viscosity $\eta$, by: $\tau = \eta R_0/\gamma$.

During fusion, the volume conservation can be expressed in terms of $\rho(\theta)$, the ratio of the instantaneous radius $R$ and the initial radius $R_0$ of the spheroid

$$\rho(\theta) = \frac{R}{R_0} = \left[\frac{(1 + \cos\theta)^2(2 - \cos\theta)}{4}\right]^{-\frac{1}{3}}.$$  \hfill (5.8)
5.4.2 Fusion experiment and simulation

A series of fusion experiments were available through our collaboration with Dr. Forgacs’s group. In these experiments, two multicellular spherical aggregates made of CHO cells, initially in contact, fused during 288 hours. Using snapshots of this system taken every 24 hours we measured the instantaneous radius of each aggregate and the radius of the interface between the two aggregates. For this we superimposed two identical circles on the two aggregates as shown in Fig. 5.3a. Using the diameter of the initial aggregate as length unit we recorded the values of $R$ and $r$ for each snapshot. Next we calculated both the change in volume $\rho(\theta) = R/R_0$ and the $\sin^2(\theta) = (r/R)^2$ and plotted them in Fig. 5.4a,b. In Fig. 5.4a the experimental $\rho(\theta)$ (dotted line) is situated above the theoretical curve (continuous line) implying that the volume of the cellular aggregates during fusion increases. This is an expected result because the cell division time of CHO cells is much shorter than the fusion time. By fitting the measured $\sin^2(\theta)$ with Eq. (5.7) we calculated the characteristic fusion time $\tau$ for CHO aggregates. For 500 micron aggregates $\tau$ was $\sim 60$ hours. The experimental points and the fitting curve for $\sin^2(\theta)$ are shown in Fig. 81.

Figure 5.3: Fusion of two multicellular spherical aggregates made of CHO cells (a) and CPD cells (b). The white circles are used for measuring the instantaneous radius of the fusing aggregate and the radius of the interface between the two aggregates. Original images were produced with the confocal microscope (a) and VMD (b). The overlapping with the circles was done in CanvasX.
Figure 5.4: Fusion of spherical aggregates made of CHO cells and CPD cells. The volume conservation is verified through angle dependence of $\rho(\theta) = R/R_0$ during the fusion process in experiment (a) and simulation (c). The time dependence of $\sin^2(\theta) = (r/R)^2$ during fusion from experiment and simulation are displayed in (c) and (d), respectively. By fitting the experimental data with the theoretical curve we calculated the characteristic fusion time $\tau$ and the time corresponding to the initial state $t_0$. 
The same procedure was applied to the fusion of two CPD spheroids. The fusion simulation was described in Section 5.2.2. In Fig. 5.3c,d are presented several snapshots of the two fusing spheroids at different instances of time. As in the case of the experimental fusion, we superimposed identical circles over the two spheroids in each snapshot and calculated the values of $r$, $R$ and angle $\theta$ every $50 \, tu$. By fitting the $\sin^2(\theta)$ with the theoretical function from Eq. (5.7) we calculated a characteristic fusion time for CPD aggregates of $\sim 520 \, tu$. As shown in Fig. 5.4c, the volume of the CPD aggregate was conserved during the fusion process.

The good agreement between the theoretical curve and both experimental and simulation data from Fig. 5.4b and Fig. 5.4d is a confirmation that our CPD method reproduces very well the viscous behavior of multicellular aggregates.

## 5.5 Uniaxial compression of spherical aggregates

### 5.5.1 Phenomenological Model

Biomechanical properties of both living and simulated cellular aggregates are observed by investigating the deformation of the aggregate and the post compression relaxation of the force. The kinetics of force relaxation in uniaxial compression of artificial tissue aggregates probe tissue viscoelasticity [103] and provide the second relation that connects the viscosity $\eta$ and the shear stress modulus $G$. The relaxation of the force $F(t)$, exerted by the aggregate on the compression plates has been proposed through a phenomenological viscoelasticity model [103] and is well approximated by a double exponential expression

$$F(t) = F_1 e^{-t/\tau_1} + F_2 e^{-t/\tau_2} + F_0.$$

(5.9)
The two key parameters in Eq. (5.9), are the equilibrium compression force \( F_0 \) and the viscoelastic (Maxwell) relaxation time, \( \tau_2 = \eta/G \) [104]. The shorter time constant \( \tau_1 \) describes the local, cellular level relaxation of the aggregate [103].

The compression is also used for measuring the effective tissue surface tension \( \gamma \), defined as the free energy of the unit area of the interface between an aggregate and cell culture medium [105, 106]. The surface tension of the aggregate can be determined based on the equilibrium force (\( F_0 \)) and geometric parameters (\( H, R_1 \)) of the compressed aggregate [105]. Recently, a more precise method of extracting \( \gamma \) has been developed by employing the exact solution of Laplace equation [3]. According to this study the surface tension, \( \gamma \) is given by

\[
\gamma = \frac{F_0}{2\pi R_1 f_{-1}(H/2R_1)}
\]  

(5.10)

\( F_0 \) is the equilibrium compression force, \( H \) the distance between the compressing plates and \( R_1 \) the radius of the circle from the equatorial plane of the compressed aggregate.

The phenomenological model and the method for calculating the surface tension provide the third and fourth relations needed for measuring \( \gamma, \eta \) and \( G \). They are next applied to both experiments and CPD simulations.

### 5.5.2 Compression experiment and simulation

Compression experiments of spherical aggregates between parallel plates have been treated previously and are presented in detail in [3]. The experimental and simulated values of \( F_0 \) and \( \tau_2 \) can be determined by fitting the relaxation force with Eq. (5.9). In Fig. 5.5a,b are shown examples of uncompressed and compressed cushion tissue aggregates. The force relaxation curve and the theoretical fit are shown in Fig. 5.5c. Similar plots for the compression of the CPD spheroid are shown in Fig. 5.5d,e,f. With \( F_0 \) and the parameters, \( H \) and \( R_1 \), that characterize the shape of the equilibrated compressed spheroid we could determine the surface tension \( \gamma \) of the multicellular aggregate.
With $\gamma$, $\eta$ and $G$ calculated from the fusion and compression experiments, we have all the viscoelastic quantities needed to “calibrate” the CPD method. By expressing the CPD units $L$, $T$ and $E$ in SI units using Eqs. (5.6), the CPD method is able now to predict the time evolution of more complex systems.

### 5.6 Application: fusion of 10 multicellular spheroids

The “calibrated” CPD method was tested in the case of the fusion of 10 spheroids. The goal was to predict the fusion time of 10 CHO multicellular aggregates by using the fusion of 10 CPD spheroids. The experimental and simulated systems before and after the fusion are shown in Fig. 5.6. Initially arranged along a circle, the spheroids fused and eventually formed a toroidal shape. In experiment were used 500 micron CHO aggregates [4]. In the CPD simulation were used 10 identical copies of the equilibrated spheroid made of 500...
cells.

Figure 5.6: Fusion of 10 cellular aggregates. Initial and final configurations of CHO [4] (CPD) systems are shown in a) and b) (c and d). Different colors were used for neighboring aggregates.

The fusion of the CPD spheroids from Fig. 5.6c,d reproduces very well the experimental fusion of the CHO aggregates from Fig. 5.6a,b. Furthermore, the time parameter ($T = \sigma^2/D$) can be expressed in SI units based on the fusion of two spherical aggregates. Thus, we could calculate the fusion time of the 10 CPD aggregates which was $\sim 186$ h. From experiments the fusion time for the structure in Fig. 5.6a,b was $\sim 168$ h. The simulated fusion time differs only by $\sim 11\%$ from the experimental one.

## 5.7 Conclusions

In this chapter we presented a new computer simulation method that is able to predict the self-assembly process of bioprinted multicellular constructs. Our CPD method reproduces very well the basic fusion and compression experiments. Using theoretical models for the fusion and compression processes we calculated biomechanical properties of multicellular aggregates from both experiments and simulations. These biomechanical properties were then used to “calibrate” our CPD parameters and express the length, time and energy units in terms of SI units. Thus we could employ the CPD simulations to predict the time evolution of a multitude of multicellular systems of different shape and size. To demonstrate its predictive power, we have applied the CPD method to predict the fusion time of 10 CHO aggregates into a ring like structure. The agreement between the CPD prediction and the
experiment was excellent (within ~ 11% of error). The next step is to extend our predictions to more complex geometries (i.e., tubular or branched tubular structures), and predict the time evolution towards the final stationary configuration. These geometries are relevant for bioprinting of blood vessels.
References


Bogdan Barz was born on January 21, 1979 in Sighetu Marmatiei, Romania. He graduated from the Physics Department of “Babes-Bolyai” University in Cluj-Napoca, Romania, in July 2001. He earned his Masters Degree in Applied Mathematics from the “Babes-Bolyai” University in July 2003. He joined Dr. Ioan Kosztin’s group in August 2004 and received his Ph.D. degree in Physics from the University of Missouri in September, 2009. He recently accepted a postdoctoral position at Drexel University in Dr. Brigita Urbanc’s group, starting October 2009.