Computational Studies of Acidic Destabilization and Membrane Association of Diphtheria Toxin Translocation (T) Domain

Submitted to the Graduate School of Chemistry in partial
fulfillment of the requirements for the degree of
Doctor of Philosophy

by

Jose C. Flores-Canales
B.Sc. Mechatronic Engineering

Carnegie Mellon University
September, 2014

Carnegie Mellon University

Mellon College of Science

Dissertation was presented

By

Jose C. Flores-Canales

Defended on

September 10, 2014

and approved by

Dave Yaron, PhD, Carnegie Mellon University Department of Chemistry

Rongchao Jin, PhD, Carnegie Mellon University Department of Chemistry

Daniel Zuckerman, PhD, University of Pittsburgh Department of Computational & Systems Biology, School of Medicine

Dissertation Advisor: Maria Kurnikova, PhD Carnegie Mellon University Department of Chemistry

Abstract

Diphtheria toxin translocation (T) domain is a water soluble protein that consists of ten alpha-helices in high pH solution. Experimental studies have determined that T-domain undergoes conformational changes upon decrease of solution pH, which shifts the protein population from its initial water soluble to a membrane-competent in solution. It was hypothesized that conformational changes of the latter state prepare the protein structure for subsequent membrane binding. After binding, refolding of T-domain on the membrane results in the formation of a transmembrane state, which is characterized by the permeation of the lipid bilayer. The function of transmembrane state is to help the translocation of a catalytic domain attached to the protein N-terminal across the lipid bilayer. The first goal of this work is to study the pH-dependent destabilization of T-domain structure in solution and understand the role of protonation of key residues using a variety of computational and experimental methods. The second goal is to study the subsequent membrane binding of T-domain to lipid bilayers and propose a theoretical model of the early steps of T-domain refolding on the membrane interface using a multiscale approach.

Modeling of the low pH induced conformational changes and membrane association of T-domain is performed in two stages. In the first stage, protonation of N-terminal histidines triggers conformational changes of the N-terminal helices of T-domain in solution. The role of histidines was confirmed by thermodynamic integration, continuum electrostatic calculations, and microsecond long molecular dynamics (MD) simulations. Two microsecond MD simulations of a low pH model of T-domain sampled similar destabilized protein conformations; however, an N-terminal helix showed

dissimilar degree of refolding. To improve the sampling of conformational changes, we proposed and implement a sampling method based on the accelerated molecular dynamics simulations method. Our proposed implementation accelerates the sampling of the conformational landscape of the low pH T-domain model by boosting the direct-space electrostatic interactions of solute-solute atom pairs. In general, this implementation accelerates the sampling of the conformational space of alanine-dipeptide in comparison to the original implementation of accelerated molecular dynamics.

In the second stage, a multiscale approach is used to model the membrane association of the low pH destabilized model of T-domain to lipid bilayers of different compositions. Two preferable membrane-bound conformations of the low pH T-domain model are predicted by equilibrium and free energy calculations. The most frequently observed membrane-bound conformation is stabilized by electrostatic interactions between the protein and the lipid headgroups. In contrast, the less frequently observed membrane-bound conformation is stabilized by hydrophobic interactions between the protein and lipid headgroups. These interactions allow for a deeper insertion of T-domain in the membrane interface. The predicted membrane-bound conformations were refined by atomistic molecular dynamics simulations, which show that membrane-bound conformations are stable for several microseconds. Furthermore, atomistic MD simulations suggested that neutralization of glutamate and aspartate sidechains favored a deeper inserted state of T-domain in the membrane interface. This observation is in good agreement with reported pH-dependent insertion of T-domain in the membrane interface.

To study the assembly of transmembrane helices, a coarse-grained model based on a residue level of representation and a rigid-body Monte-Carlo sampling method is

developed. The scoring energy function is constructed using a knowledge based potential extracted from water soluble protein structures. To compensate the protein interior packing and the solvation differences, an experimentally determined membrane partition scale for all residues was used. This scoring function was tested in a set of three transmembrane homodimers. The proposed scoring function and the associated rigid-body Monte-Carlo sampling method can be applied in the first steps of prediction of near-native structures of transmembrane proteins followed by structural refinement using atomistic MD simulations.

Table of Content

List of Tables

List of Figures

List of Abbreviations

T-domain – Diphtheria Toxin Translocation Domain

GPU - Graphical Processing Unit

MD – Molecular Dynamics

CG-MD – Coarse Grained Molecular Dynamics

COM – Center of Mass

RMSD – Root Mean Square Deviation

RMSF – Root Mean Square Fluctuation

PB – Poisson-Boltzmann

PMF – Potential of Mean Force

PCA – Principal Component Analysis

SASA –Solvent Access Surface Area

TM – Transmembrane

HP – Hydrophobic

AMD – Accelerated Molecular Dynamics

POPG – 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol

POPC – 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine

MC – Monte-Carlo

Acknowledgements

I would like to thank you to my advisor Dr. Maria Kurnikova for her support and guidance during my research work. Dr. Kurnikova and I have worked together on the development and application of innovative computational methods applied to the understanding of protein-membrane interactions. Also, I have to thank you Dr. Kurnikova for the opportunity of studying the membrane insertion pathway of diphtheria toxin. I must also thank you Dr. Igor Kurnikov for sharing his knowledge on how to approach and understand complex biological phenomena using physical models. Also, I thank you the members of Kurnikova's lab, for their insightful suggestions and comments on research discussions. Finally, I would like to thank you to Dr. Nikolay Simakov for rewarding friendship, research collaboration and thoughtful suggestions.

I would like to thank you to Dr. Alberto Coronado. Under his guidance, I began my research interest on computational chemistry. My stance in Pittsburgh has been very fruitful thanks to many friends and I must mention Julian Ramos for his friendship and interesting talks on a variety of subjects. I also want to thank you to my friend Dr. Boris Aguilar for collaborative work and discussion of ideas since days in our undergrad institution in Lima.

Finally, I have to thank you to my dear mother Teresa Canales-Espinoza and my family for their permanent support of all of my endeavors. In a country affected by political violence, my mother made the best to provide my siblings and me of a positive environment, education and a desire for understanding.

Introduction

Diphtheria toxin is a bacterial protein that targets cell surface receptors of eukaryotic cells, penetrates the cell membrane, inhibits the protein synthesis process, and as a result leads to cell death. Despite extensive research on the cell insertion process of diphtheria toxin over the last decades, this insertion process is not well understood.[1-3] Cell insertion of diphtheria toxin is carried out by three of domains: a receptor, a translocation, and a catalytic domain, which are associated to specific roles. The receptor domain binds to a specific cell surface receptor, which is followed by a receptor-induced endocytosis process. The resulting endosome encapsulates the toxin-receptor complex followed by decrease of the endosomal pH. Consequent acidification of endosome interior triggers conformational changes of the toxin translocation (T) domain, which binds and refolds in the membrane interface. Significant refolding of T-domain in the membrane interface leads to the transmembrane insertion of two hydrophobic core helices. This transmembrane state of T-domain facilitates the translocation of a catalytic domain through the endosome membrane. Subsequently, the catalytic domain is released in the cytosol and modifies a key protein of protein synthesis machinery, which leads to the eventual death of the cell. This ability of diphtheria toxin to deliver a catalytic domain has been proposed for treatment against cancer cells.[4] Understanding of the physico-chemical principles guiding the translocation function of T-domain will provide a framework for the potential development of drug-delivery systems.

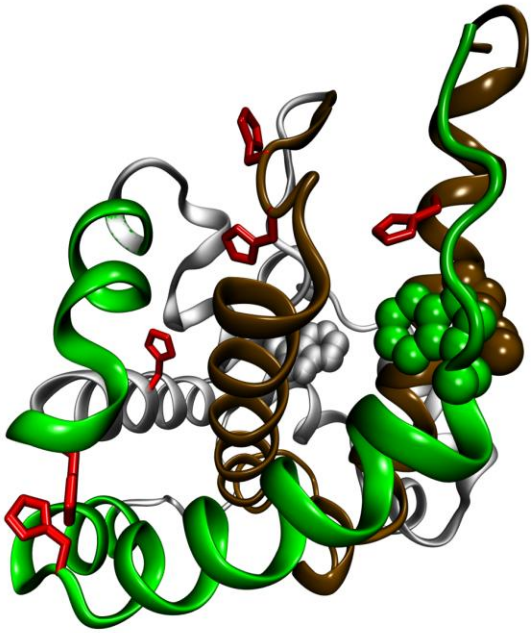

Figure 0.1. Structure of diphtheria toxin T-domain from protein data bank (PDB) code 1FL0.[5] Helices TH1, TH2, and TH4 are shown in green ribbon representation. Helices TH8-9 and are shown in brown ribbon representation. All other helices are shown in grey ribbon representation. Histidine side-chains are shown in red licorice representation. W206 and W281 side-chains are shown in green and grey space-filled representation, respectively. Q369 side-chain is shown in brown space-filled representation.

Structure of T-domain

High resolution structure of T-domain was obtained as part of the whole toxin structure obtained at pH 7.5.[5] Figure 0.1 shows that T-domain consists of ten alpha-helices, which are named TH1-9, TH5'. T-domain structures shows that the hydrophobic helices TH8-TH9 are completely and partially buried in the core of the protein, respectively. Amphiphilic helices TH1-4 and TH5-7 surround the hydrophobic core, as

shown in Figure 0.1. Furthermore, T-domain contains six histidines located in loops of the N-terminal and C-terminal helices. Among them, H223 and H257 are close to each other in the N-terminal region of the protein. Histidine H257 is stabilized by hydrogen bonds with residues S219 and E259. Also, T-domain contains two tryptophan residues located in helices TH1 and TH5, which are used for measurement of intrinsic tryptophan fluorescence. Changes in the maximum wavelength and intensity can be interpreted as changes of the environment surrounding these tryptophans.

pH-Dependent Behavior of T-domain in Solution

Early mutagenesis and fluorescence experiments provided details of the pH-dependent behavior of T-domain in solution. Zhang et al.[6] reported that decrease of solution pH caused the solvent exposure of hydrophobic sites of T-domain in solution, while no separation of the amphiphilic helices TH1-4 and TH5-7 from the hydrophobic core was observed. This was interpreted as a pH-dependent increase of tertiary structure fluctuations of the protein in solution. A different study reported pH-dependent local conformational changes of helix TH1, which exposed hydrophobic residues of helix TH1 to the solvent.[7] Overall, these studies suggested that acidification of the solution induces changes in the structure of T-domain, and as a result, protein-membrane interactions are enhanced through hydrophobic interactions.

Based on mutation and intrinsic fluorescence studies, Chenal et al.[8] proposed that decrease of solution pH triggered the formation of a molten-globule state of T-domain. This molten-globule was proposed to exhibit stable secondary structure, a fluctuating tertiary structure, exposure of hydrophobic regions and a propensity to aggregate due to solvent exposure of hydrophobic surfaces.[8] However, there are no high-resolution

structures to corroborate this hypothesis. Furthermore, it was proposed that histidines were involved in the pH-dependent conformational changes of T-domain in solution, but the atomistic details are unknown.

To identify the role of histidines in the structural changes of T-domain, a mutagenesis study of single or sets of histidines in conjunction to intrinsic tryptophan fluorescence study was performed.[9] This study proposed a sequential model of unfolding and insertion guided by concerted protonation of histidines upon decrease of pH. It was suggested that concerted protonation of a set of histidines induces the formation of a molten-globule intermediate state in solution (histidines H223, H251 and H257). This transition was observed within the pH range of 5 to 7.

A recent kinetics analysis of the membrane binding process of T-domain has been reported by Ladokhin and coworkers.[10] Titration experiments of the wild-type protein in solution showed that the transition pH is 6.3, which was characterized as the transition between the native structure and a membrane-competent state. The formation of the latter state was shown to be independent of the lipid-bilayer composition. Furthermore, circular dichroism studies and single-site mutants of T-domain showed that H257 triggered the formation of the membrane-competent state, which was characterized by decrease of secondary structure content.[11] To better understand the role of H257 in the formation of the membrane-competent state, high resolution structures of T-domain in low pH solution and computational modeling are needed.

Membrane Association and Insertion of T-domain

The structures of membrane-bound and transmembrane states of T-domain are currently unknown. Thus, different biophysical techniques have been used to understand

the process of binding and insertion. A fluorescence spectroscopy and mutation study reported that concerted protonation of histidines triggered the membrane-association of the protein to anionic membranes (histidines H251, H322 and H323). This transition was observed within the range of pH 6 to 7.[9] Further acidification to pH 4 showed protein refolding and deeply insertion in the membrane.

A kinetics study of T-domain has proposed a staggered pH-dependent formation of a membrane-competent state in solution and at least one insertion-competent state in the membrane interface.[10] However, the physico-chemical principles that guide the population of the overlapping membrane-competent and the insertion-competent state are still unknown.[3] Furthermore, neutralization of glutamate and aspartate sidechains in the membrane interface was suggested to be involved in the process of T-domain transmembrane insertion.[3, 12, 13] It was also shown that lipid bilayers composition enhances T-domain binding and insertion. The increase of anionic content in bilayers leads to the increase of the fraction of membrane-bound state of T-domain and accelerates its membrane insertion.[10] The molecular details of the interactions between the protein and anionic lipid bilayers are still unknown.

The thermodynamics of membrane association of T-domain to anionic lipid bilayers was recently studied.[12] A favorable free energy of membrane binding of approximately -10 to -7 kcal/mol was determined at pH 4.3 and pH 7, respectively (bilayers composed of POPC/POPG with ratio 1:3). However, it was not clear if a molten globule conformation was part of the membrane binding process or if the protein was not inserted in the lipid bilayer. Due to the difficulties of crystallization of T-domain in a

membrane environment, computational simulations of the process of binding will help the understanding of the cell insertion mechanism of this protein.

Thesis Outline

This thesis is organized in five chapters, which describe the computational approach used to study the pH-dependent membrane association of T-domain. Chapter one describes microsecond long MD simulations of a neutral and a low pH model of T-domain. Chapter two presents the implementation and testing of a sampling method. This sampling method is applied to the study of conformational changes of T-domain in solution. Chapter three describes a coarse-grained study of T-domain membrane association. Two membrane-bound conformations are predicted and free energy calculations are performed to determine their equilibrium properties. Chapter four presents the parameterization of an improved force-field of anionic lipids. This force-field is used in atomistic MD simulations of membrane bound conformations of T-domain. Finally, Chapter five presents a scoring energy function for prediction of transmembrane helices structures.

Chapter one presents a study that will be submitted to the journal Proteins, entitled "*Microsecond Molecular Dynamics Simulations of Diphtheria Toxin Translocation T-Domain pH-Dependent Unfolding in Solution*". A previous analysis of part of the data presented in chapter one was published in Journal of Molecular Biology, entitled "*pH-Triggered Conformational Switching of the Diphtheria Toxin T-Domain: The Roles of N-Terminal Histidines*". The computational method and results of chapter two are in preparation for submission to the Journal of Chemical Theory and Computation, entitled "*Acceleration of Conformational Changes of Proteins in Explicit Solvent*". Chapter three

describes results that are in preparation for submission to Biophysical Journal, entitled "*Membrane Association of Diphtheria Toxin T-domain Studied by Coarse-Grained Simulations*" and a follow up article based on chapter 4, entitled "*Atomistic Simulations of Membrane Bound Conformations of a low pH refolded Model of Diphtheria Toxin T-domain*". Finally, chapter five describes results that will be submitted to the journal of BMC-Bioinformatics, entitled "*An Empirical Scoring Function for the Transmembrane Helical Protein Assembly*". Also, the sampling method described in chapter five is part of a manuscript to be submitted to Journal of Mathematical Biology, entitled "*Fast and Flexible Geometric Method For Enhancing MC Sampling of Compact Configurations For Protein Docking Problem*".

Chapter 1. Role of Histidines in the Destabilization of Diphtheria Toxin Translocation (T) domain Structure in Solution

1.1 Introduction

Diphtheria toxin translocation (T)-domain insertion into lipid bilayers is triggered by decrease of pH without the assistance from other molecules.[3] T-domain insertion pathway involves the formation of a membrane-competent state in solution initiated by an unknown sequence of conformational changes, which are triggered by protonation of key residues. The low pH insertion mechanism of T-domain and its associated function of translocating a molecular cargo across membranes are suitable features for development of anti-cancer drugs;[4] hence the understanding of the atomistic details of T-domain refolding can expand the set of rules for rational design of pH-dependent molecular switches.

Numerous experimental studies were performed to elucidate the structure[14] and mechanisms[2, 3] of the entire diphtheria toxin and the stand-alone T-domain. However, atomistic details of the T-domain transformation along the membrane insertion pathway are yet to be established. In neutral solution pH, T-domain is an all α-helical protein formed with ten helices TH1-TH9 and TH5'. The solvent accessible helices TH1-4 and TH5-7 encircle two non-polar helices TH8-9 buried in the protein core (see Figure 1.1 for illustration of the protein structure).[5] Lowering pH in solution triggers formation of T-domain intermediate states that brings the system from a soluble to a transmembrane state, in which the hydrophobic hairpin TH8-9 spans the membrane. Since the interior of an endosome has a pH value close to pH 5,[15] protonation of all histidines was implicated

in triggering pH-dependent structural changes.[8, 10, 16, 17] In addition, pH-dependent local conformational changes of the protein in solution were indicated via indirect experimental observations.[18 ,19] Thus, it is important to understand the role of protonation of histidines in the triggering of conformational changes of T-domain in solution.

We have recently reported on the shifts of T-domain histidine pK_as using molecular dynamics (MD) free energy calculations (Thermodynamic Integration).[20] Our results indicated that protonation of certain histidines (e.g. His 257 or His 251) can significantly destabilize T-domain globular structure. Also, accumulated 9 microsecond long MD simulations suggested that upon protonation of histidines in solution T-domain undergoes partial unfolding resulting in structures in good agreement with spectroscopic experiments.[20]

Recent advances in the field of molecular dynamics simulations have made it possible to study the dynamics of disordered structures and processes of partial unfolding and re-folding of proteins in solution.[21-23] However, significantly larger simulated data sets are needed to fully characterize a partially unfolded state of a protein. At present such fully converged simulation remains an unattainable goal.[21].

To gain an insight into structure, diversity and stability of a low pH conformational space of T-domain, we analyzed an accumulated 18 μs long MD simulations of a low and a neutral pH model of T-domain in solution. In this work, we gain insight into T-domain structure in solution at low pH by analyzing similarities and differences in two independent μs long simulations each resulting in protein local unfolding and refolding. The ensemble of these partially unfolded structures has characteristics of a membrane-competent state[20] and remains in a compact conformation.

This chapter is further structured into Methods, Results, and Discussion and Conclusions sections. The Results section first describes simulated models (subsection 1.2.1), then it is further structured into three main complementary subsections (subsections 1.2.2-1.2.4) that together provide comprehensive analysis of the structural changes and dynamics of the protein observed in the simulations. In subsection 1.2.2 positional deviation and fluctuation of individual residues and native contacts are discussed in detail; in subsection 1.2.3 folding/unfolding of the secondary structure is analyzed using a helicity measure[24] and its Principal Component Analysis (PCA); and, in subsection 1.2.4, we look at the resulting structure and dynamics of the hydrophobic core of the protein and a water exposure of the hydrophobic sites, which are essential characteristics of a protein structure preparation for subsequent membrane association.

1.2 Methods

1.2.1 Molecular dynamics simulations

T-domain structure models were constructed from a high resolution structure obtained at conditions of pH 7.5 (residues 201 to 380 from PDB ID code 1F0L).[5] Hydrogen atoms were added by tLeap (AMBER 9 package).[25] Two models with different protonation states for all histidines were constructed, a neutral and a low pH T-domain model. These models were used in MD simulations T1 and T2, respectively. Previous continuum electrostatic calculations showed that the epsilon tautomer of neutral histidine sidechains is preferred for all six of them.[20] N_ε atoms were protonated and N_δ were not protonated in all histidine side chains in the neutral pH T-domain structure. The simulation boxes of the neutral and first low pH T-domain model were created by adding 4587 TIP3P explicit water molecules such that the distance between the protein and the

simulation box edge was 8.0 Å. The total number of atoms is 16521 atoms in these boxes. A detailed description of the simulation setup of T1 and T2 has been published in a recent report.[20] In the case of the second low pH T-domain model (T3) reported here, we added 13215 TIP3P explicit water molecules such that the distance between the protein and the edges of the simulation box was 16 Å. This box contains 42405 atoms. Sodium ions were added to the simulation box to neutralize the simulation box. We equilibrated this system with Desmond 2.0 with the ff99SB force field.[26] The simulation time step was 2 fs and all hydrogen bonds were constrained via SHAKE. Periodic boundary conditions were set up, cutoff radius was set to 10 Å and electrostatic calculations were performed using Particle Mesh Ewald (PME) method. The protein was restrained and the solvent was minimized for 200 steps of steepest descent followed by 50 steps of conjugate gradient descent minimization method. Then the solvent was restrained and the protein with backbone atoms restrained was minimized by a total number of 250 steps as explained above. Anisotropic constant pressure MD equilibration was performed using Martyna-Tobias-Klein (MTK) barostat at temperature of T = 300 K and pressure of 1 atm with positional restraints applied to heavy atoms (1.0 kcal/mol·Å^2) over the first 600 ps. Then, the restraint constants were decreased linearly to 0.0 kcal/mol·Å^2, over the following 400 ps. This was followed by unrestrained constant pressure MD equilibration over 8 ns. Production MD simulations were performed on Anton supercomputer. We used the same force field in the equilibration, except that the electrostatic interactions were computed by Gaussian Split Ewalds method[27] with a grid size of 64 × 64 × 64 and a cutoff radius of 12.0 Å. The time step was 2 fs and all hydrogen bonds were constrained. Berendsen thermostat and barostat were set to T = 310 K (tau = 1 ps) and pressure of 1

atm (tau = 2 ps), respectively. Atom coordinates were saved every 60 ps. MD simulations were carried out up to 2060 ns, 6723 ns and 9668 ns for the MD runs T1, T2, and T3, respectively.

1.2.2 Analysis

MD structures were aligned to the X-ray structure excluding C_α atoms in the tails or those in the loops and tails. The C_α-RMSD is calculated excluding C_α-atoms from the tails and loops identified in the crystal structure. Root mean square deviation (RMSD), distances between individual atoms, averaging of protein structures and secondary structure analysis were calculated using the ptraj program available in AmberTools 1.4. Solvent accessible surface area (SASA) was determined using a probe sphere of 1.4 Å. Molecular figures were prepared using VMD 1.8.7.[28] Secondary structure assignments are calculated using DSSP.[29] The cumulative average of the helicity content per helix in Figure 1.6 is obtained from the DSSP output over the MD trajectories where the window average is 600 ps. We use a measure of native similarity:

$$Q = \frac{1}{N_{pairs}} \sum_{i>j} \exp\left(-\frac{(r_{ij} - r_{ij}^{native})}{2\sigma^2}\right),$$

where N_{pairs} is the number of pairs of native contacts between all C_α atoms in residues 206-375. r_{ij} and r_{ij}^{native} are the distances between C_α atoms in a MD frame and the crystal structure, respectively. σ has a value of 3.0 Å. Contact maps are calculated using the fraction of non-local contacts between the number of pairs N_{ij} of C_α atoms separated by at least 8 Å and N_{max} the maximum number of pairs ij. Residues containing atoms i and j are separated by at least 3 residues along the chain.

1.2.3 Principal components Analysis

In this work, given the relatively large size of T-domain (180 residues) and the variety of conformational changes, we choose a metric able to characterize coarse-grained features of the T-domain folded and the partially unfolded ensembles. Based in our previous work on characterization of similarities and differences of different ensembles of unstructured peptides,[24] we use a helicity measure along the protein sequence for each MD frame (details are described in Ref. 35). To reduce the high-dimensional data obtained, principal component analysis (PCA) is a suitable and widely implemented method. This method is based on the assumption that important protein motions are described by a linear combination of the main principal components of the covariance matrix, however, non-linear features of MD simulations may not faithfully described by this method.[30]

We reduce the conformational space of a protein by principal components, which are the best linear approximations of the covariance matrix of the multidimensional protein space. This dimensional reduction method can be applied to different metrics *i.e.*: position of backbone atoms, backbone dihedral angles or helicity measure (triplets).

1. All MD trajectories are stripped of water and ions, followed by translation of the center of mass and orientation alignment relative to the Cα atoms of the crystal structure using ptraj program from AmberTools.

2. The set of dimensions $\boldsymbol{x} = \{x_1, x_2, \ldots x_i\}$ where i = 1…p, p is the number of dimensions, x_i is a vector of size M, and M is the number of MD frames.

3. A p × p covariance matrix R is defined as follows:

$$R = \frac{1}{M} D D^T$$

where D is a p × M matrix of elements $D_{ij} = x_{ij} - < x_i >$, and $< x_i >$ is the average of x_i over an ensemble of sampled protein conformations.

4. The covariance matrix R can be constructed using two different metrics:

Helicity measure: A peptide is considered α-helical if at least three sequential residues have their backbone dihedrals (φ,ψ) in the α-helical region of the Ramachandran plot, where -100° ≤ φ ≤ -30° and -80° ≤ ψ ≤ -5°. The helicity for a sequence of three residues is defined as follows:

$$h_k = \sum_{i=1}^{3} a_i$$

where:

$$a_i = \begin{cases} 1, & -100° \leq \varphi \leq -30°, \quad -80° \leq ø \leq -5 \\ 0, & otherwise \end{cases}$$

where i is the residue index on a *k*th sequential triplet. The helicity measure is calculated for all consecutive triplet of residues for a protein of N residues. This results in a matrix R of dimensions $N-2 \times N-2$. Another metric is based on the *sin* and *cos* functions of the backbone dihedral angles (φ, ψ) of N residues results in a matrix R of dimensions $4N \times 4N$. In this study PCA analysis is performed over residues 206-375.

5. The covariance matrix R is diagonalized to calculate the eigenvalues $\lambda^{(i)}$ and eigenvectors $\boldsymbol{v}^{(i)}$. Those with the largest eigenvalues account for the largest variance in the multidimensional data. The principal components $\boldsymbol{V}^{(i)} = \boldsymbol{v}^{(i)} \cdot \boldsymbol{x} = v_1^{(i)} x_1 + v_2^{(i)} x_2 + v_3^{(i)} x_3 + \cdots v_p^{(i)} x_p$. For example, the component $v_1^{(i)}$ represents the influence of the first triplet comprising residues 206-208 on the *i*th principal component is represented by $\Delta_1^{(i)} = (v_1^{(i)})^2$. Moreover, the influence of the ϕ_1 backbone dihedral angle on the *i*th

principal component $\boldsymbol{V}^{(i)} = v_1^{(i)} cos\phi_1 + v_2^{(i)} sin\phi_2 + v_3^{(i)} cos\psi_1 + \cdots v_N^{(i)} cos\psi_N$ is $\Delta_1^{(i)} = (v_1^{(i)})^2 + (v_2^{(i)})^2$.[31]

1.3 Results

1.3.1 Simulated Models of T-domain

The initial coordinates for all T-domain simulations were obtained from the crystal structure of the whole diphtheria toxin protein at neutral pH [PDB ID code 1F0L].[5] A stand-alone T-domain protein was created by modifying as appropriate the N and C-terminals of T-domain extracted from the full protein. Two models of T-domain were prepared: one with all six histidine residues in their standard protonation states (further referred to as the neutral pH T-domain model); the second one with the histidines in the protonated state (further referred to as the low pH T-domain model). The T-domain protein structure and location of histidine residues are shown in Figure 1.1. In all simulations all other residues were in their standard ionization states. Three individual MD simulations are presented and analyzed in this chapter.

Two of these simulations of the neutral and low pH T-domain were performed in water box with dimensions $58 \times 59 \times 62$ Å^3.[20] These simulations are referred to as T1 and T2, respectively. The third simulation reported here is a low pH T-domain solvated in a larger water box of the size $78 \times 79 \times 82$ Å^3, further referred to as T3. This larger system is created in part to assess whether confinement of the protein in a smaller water box (as in T2) affects simulated conformational changes. For parameters of the simulations please refer

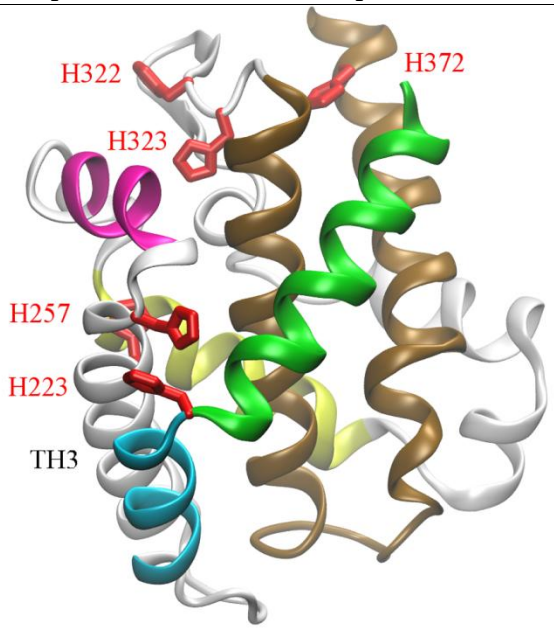

Figure 1.1. Structure of diphtheria toxin T-domain. Helices TH1, TH2, TH4, TH5, and TH8-TH9 are shown in green, cyan, magenta, yellow and brown ribbon representation, respectively. All other helices are shown in light grey ribbons. Histidine sidechains are shown in red licorice representation.

to section **1.2** ***Methods***. 18 μs of atomistic MD simulations of T-domain in explicit water were performed and analyzed in this study.

1.3.2 Simulated Structures and Fluctuations

Root mean squared deviation (RMSD) of C_α-atoms is a rough measure of structure stability or deformation in the simulations. Figure 1.2 shows time-evolution of an average RMSD value for the protein excluding terminal atoms (top panels, black lines) in each simulation (T1, T2 and T3). RMSD of the neutral pH trajectory T1 (Figure 1.2A) increased slightly with time to an average value of 1.9 Å, which is somewhat higher than ones typically reported in stable protein MD simulations on nanosecond time-scales.

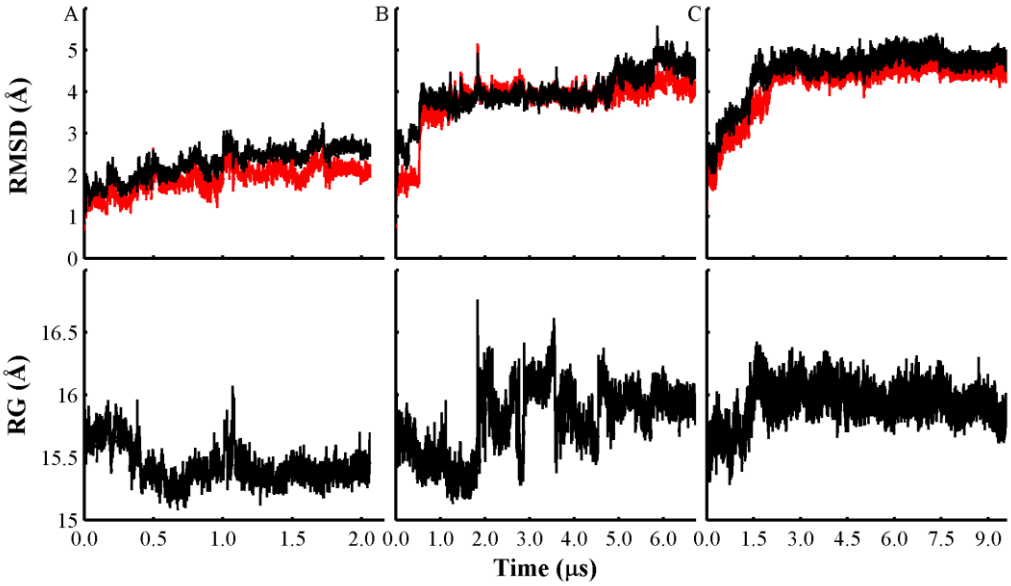

Figure 1.2. RMSD and radius of gyration plots. Panels (A), (B) and (C) display C_α root mean squared deviation (RMSD) and radius of gyration (RG) of trajectories T1, T2, and T3, respectively. C_α-RMSD obtained after C_α alignment of MD frames to the initial X-ray structure excluding C_α terminal atoms (RMSD computed on residues 206-375 as shown in black lines), and excluding C_α atoms in the protein loops and terminals (red lines). Radius of gyration over time for each MD trajectory is shown in lower panels (C_α atoms of the residues 206-375).

However, similarly high RMSD values were reported, *e.g.*, in a millisecond long simulation of a stable natively folded BPTI protein.[32] The RMSD value in the range of 2-3 Å can be indicative of populating sub-states of the native structure that are inaccessible in shorter simulations. Specifically, such relatively high RMSD may be due to repositioning of the loops in solution, which were confined in the crystal structure of the whole toxin. This can be demonstrated by leaving out the loops and computing average

RMSD of atoms that comprise α-helixes only. Such "truncated" RMSD traces are also shown in Figure 1.2 in top panels for all three simulations.

For the neutral pH simulation (T1) the truncated RMSD is noticeably lower than the full RMSD indicating that the structure is stable and maintains its native fold. This observation is further corroborated by decrease in the radius of gyration in T1 (see Figure 1.2A lower panel), which is a measure of the overall compactness of the protein structure. C_α-RMSD traces for both low pH T-domain trajectories T2 and T3 show an abrupt increase to approximately 3Å in the first microsecond of the simulation, followed by gradual increase to *ca.* 5Å over the last microsecond (see, Figures 1.2B, C, top panels). However, the dynamics of the RMSD evolution towards higher values is different in these trajectories. In T2 the initial deformation of the protein is mainly due to the loop deformation (compare black and red lines in Figure 1.2B in the first 0.5 μs), followed by a sharp increase in the "truncated" RMSD (see *ibid.*), indicating a change within the helical structure of the protein. In T3, the RMSD changes gradually over a similar period of time (see Figure 1.2C). Both low pH trajectories exhibit increase in the radius of gyration (Figures 1.2B, C, lower panels) indicating unfolding. The radius of gyration in T2 decreases initially in a manner similar to T1, then increases abruptly, fluctuates further on and stabilizes at a slightly elevated value at the very end of the trajectory. In T3 the increase is sharp over the initial 1.5 μs, but it remains stable, and decreases slightly by the end of the trajectory, which may be indicative of formation of a new structure.

To determine which residues contribute most to the observed average RMSD increase discussed above , it is instructive to plot RMSD values per C_α atom averaged

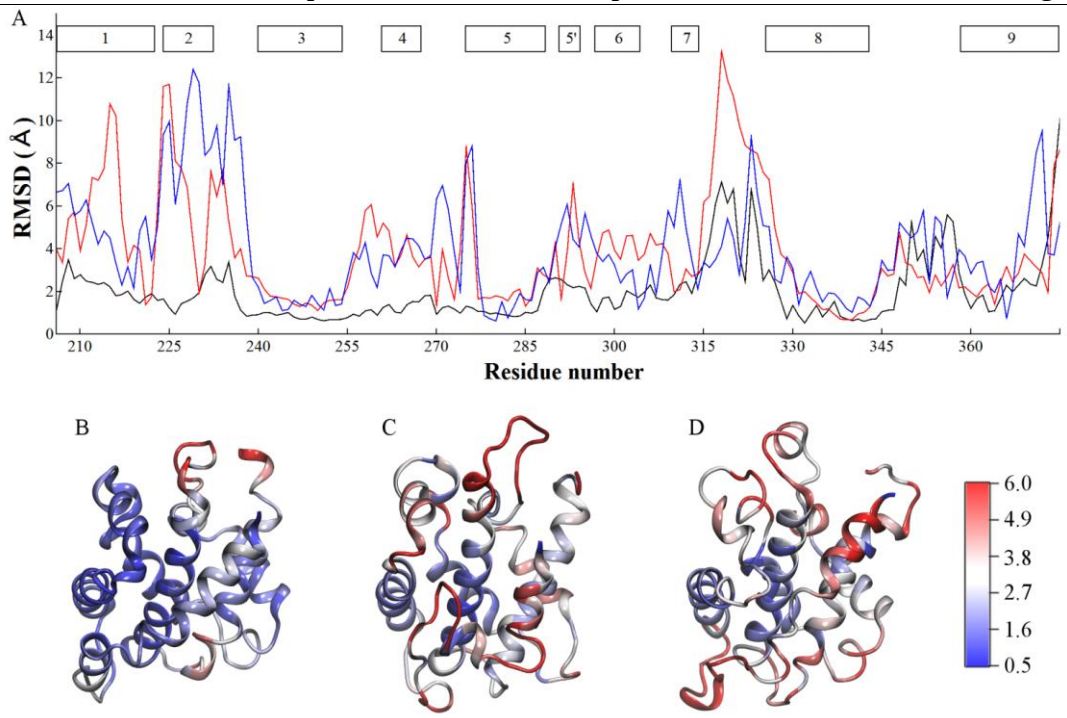

Figure 1.3. Average RMSD curve per residue**.** C_α-Root mean square deviation (RMSD) versus residue number calculated after translation and best alignment of each MD frame relative to the C_α atom coordinates of the crystal structure (residues 206 to 375). (A) C_α-RMSD traces obtained from the last 1 μs of trajectories T1 (black line), T2 (red line), and the last 2 μs of T3 (blue line). In addition, helices TH1-9 and TH5' are indicated by grey filled rectangles (top of the C_α-RMSD plot). Ribbon representation of representative structures colored according to their C_α-RMSD, see the color bar. Structures are obtained from: (B) The last 1 μs of trajectory T1. (C) The last 1 μs of trajectory T2. (D) The last 2 μs of trajectory T3. A representative structure has the lowest C_α-RMSD relative to the average structure over the last segment considered of each trajectory.

over the last stable trajectory segments, further termed at-RMSD (see Figure 1.3A). In Figures 1.3B-D representative protein structures from each trajectory are colored

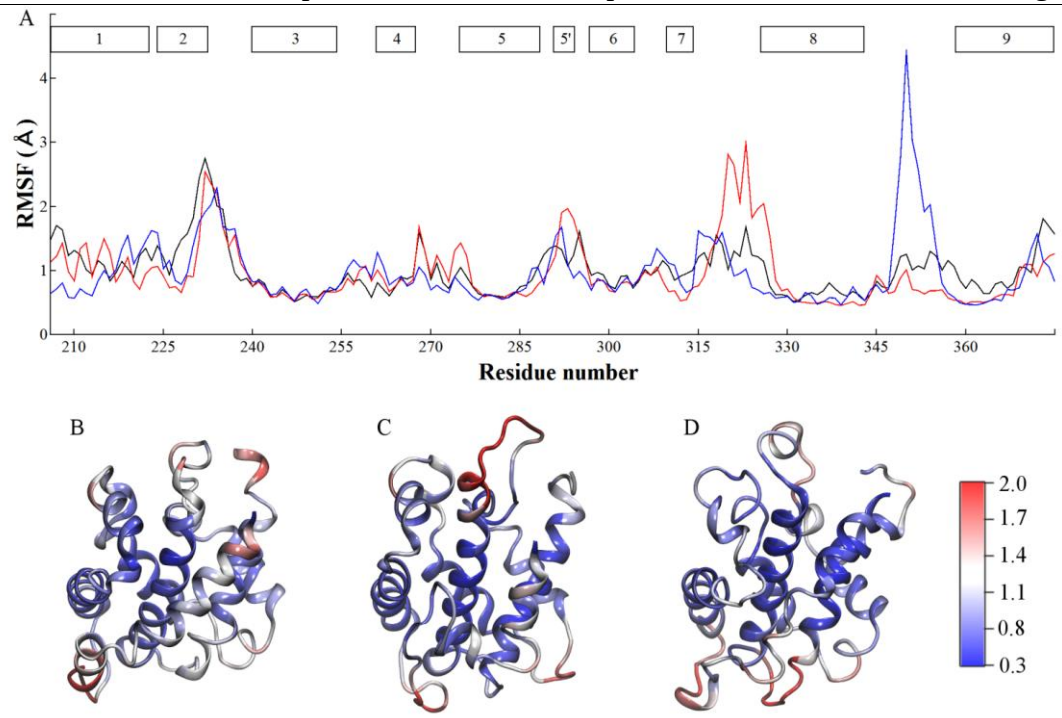

Figure 1.4. Average RMSF per residue. C_α-Root mean square fluctuations (RMSF) versus residue number obtained after translation and best alignment of MD structures relative to the C_α atom coordinates of each representative structure (residues 206-375). A representative structure is the closer to the average structure calculated over the last segment of each trajectory. (A) C_α-RMSF traces obtained from the last 1μs segments of trajectory T1 (black lines), T2 (red lines), and the last 2 μs of T3 (blue lines). Helices TH1-9 and TH5' are represented by filled grey rectangles (top of C_α-RMSF plot). Ribbon representation of representative structures colored according to their C_α-RMSF values, see the color bar. Structures are obtained from: (B) The last 1 μs of trajectory T1. (C) The last 1 μs of trajectory T2. (D) The last 2 μs of trajectory T3.

according to the corresponding at-RMSD values in Figure 1.3A. In T1 at-RMSDs are generally small, with the largest C_α atom displacements in two inter-helical loops TH7-8

and TH8-9. This is a typical result for a stably folded protein in a simulation. In contrast, in T2 and T3 a significant fraction of at-RMSDs have large values due to overall structural changes. However, the preserved protein core is clearly indicated by the low at-RMSDs, which comprises helixes TH3, the TH5 C-terminus and TH8. The largest deformation of the structure in T2 (C_α-RMSD > 4.0 Å) occurs in helices TH1-2, in the N-terminus and in the inter-helical loops TH7-8, TH8-9. In addition to these protein regions, T3 has C_α-RMSD values larger than 4.0 Å in the TH9 C-terminus.

Mobility of atoms with respect to their average position (determined here as root mean squared fluctuations (RMSF) of C_α atoms in the final stable segments of the trajectories, see Figure 1.4) is a useful characteristic for distinguishing unstructured and poorly structured (floppy) from folded conformations. Overall in both low pH simulations the protein adopts a fairly stable structure characterized by low RMSF of the majority of the residues, comparable to the RMSF of the natively folded structure simulation T1. The main difference in local mobility of residues in T2 and T3 simulations arises from the inter-helical loop TH7-8. In T2 its RMSF values reach 3.0 Å as a result of an unstructured conformation (Figure 1.4A, C). In T3 RMSFs are lower than 2 Å in this region, which is partially folded in a helical structure. Note that protonated H322 and H323 are located in this loop. SASA calculations show that side-chains of H322-H323 are largely solvated due to their unstructured conformation in all MD trajectories except in the trajectory T1 where H322 has smaller solvent exposure. The inter-helical loop TH8-9 (Figure 1.4A, D) is more mobile in T2 than in T3.

Disruption of a native structure is best characterized with an order parameter Q, whose value reflects on a number of intact native contacts in the structure (see section **1.2**

Methods for definition of Q). In Figure 1.5, a probability density of the order parameter Q is shown for all trajectories. T1 shows two peaks at Q = 0.91 and Q = 0.86. The second peak is populated by the structures in which one salt bridge R210-E362 is disrupted. Both low pH T-domain trajectories show a similar peak around Q = 0.79 populated during the first 0.5 μs of the trajectories (and before unfolding of the helices TH1, TH2, see further analysis in subsection **1.3.3** ***Secondary Structure Dynamics***). After unfolding of these helices, T2 and T3 populate peaks centered at Q = 0.67 and Q = 0.73, respectively. Further unfolding of T-domain in T2 results in sampling structures with Q = 0.61. An effective number of lost native contacts in T2 is higher than that of T3 trajectory, indicating that the smaller amount of solvating water and consequently smaller simulation box were not severe limiting factors for structural changes of the protein during these simulations. A more detailed analysis of the individual residue-residue contacts lost and formed while the protein underwent structural reformation can be done via analyzing two-dimensional contact maps. These are presented in Supporting Information in Figures 1.S1A-C for all trajectories (see also description of these calculations *ibid.*). In both T2 and T3 simulations, the residues of the helix TH1 form non-native contacts with residues in the inter-helical loop TH8-9 and in the TH9 N-terminus (Figure 1.S1 B, C). In T2, helix TH1 also lost native contacts with residues in the inter-helical loop TH3-4 and TH8 (Figure 1.S1B). In both T2 and T3 helixes TH2 and TH3 loose a number of native contacts (compare to T1 in Figure 1.S1A). Finally, in both T2 and T3 helix TH4 lost contacts of the residue E362 (TH4) with the residues E327 and Q331 in TH8.

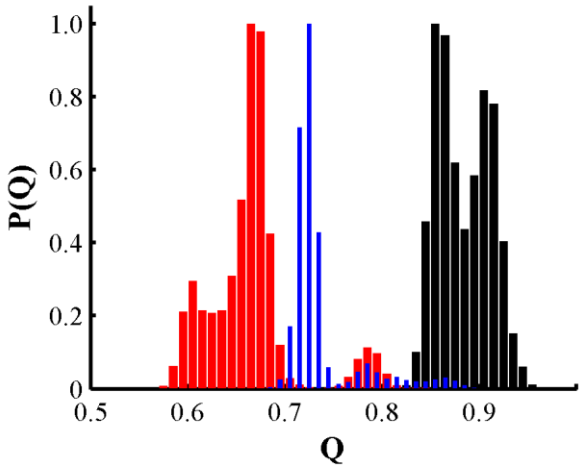

Figure 1.5. Probability density of native contacts. Native contacts Q are calculated over C_α atoms of residues 206-375. P(Q) is calculated over trajectory T1 (black line), T2 (red line), and T3 (blue line). A value of Q = 1 represents the crystal structure and Q = 0 represents an unfolded state.

As indicated by the contact map plots in Figure 1.S1 and at-RMSD plots in Figure 1.3, helices TH1 and TH2 undergo backbone conformational changes in both low pH trajectories T2 and T3 (structures shown in supporting Figure 1.S2). Protonation-dependent transition of the helix TH1 is initiated by a drastic change of the backbone dihedral angles of K216 in both low pH protein trajectories (Figure 1.S3). In T2 unfolding of TH1 is substantial, characterized by the breaking of inter-helical salt bridges K216-E259, K212-E326, and K212-E327 (Figure 1.S4). In T3 there is a disruption of the first two salt-bridges but the third salt-bridge remains intact. Following this partial unfolding TH1 adopts a helix-kink-helix structural motif in T2 (Figure 1.S2 and Ref. 26). Helix TH2 undergoes a helix to coil transformation in both low pH MD simulations. Such similar behavior in two independent microsecond long simulations highlights

importance of the finely tuned electrostatic interactions in the protein N-terminus disrupted by protonated H223 and H257. Protonation of histidines induces disruption of a structure stabilizing hydrogen bond network (H257-S219 and H257-E259) in both trajectories T2 and T3. Furthermore, conformational changes of helices TH1 and TH2 facilitate solvent exposure of the histidine side-chains. Their solvent accessible surface areas are 100 $Å^2$ (H223) and 45 $Å^2$ (H257) after the structural transitions in both low pH trajectories, white they are 28.2 $Å^2$ (H223) and 17.2 $Å^2$ (H257) in T1. Upon transition, the side-chain of H257 is stabilized by a hydrogen bond with the backbone oxygen of A254 in both T2 and T3. In addition, the C-terminus of TH9 is unfolded in both simulations T2 and T3, see Figure 1.S5. Finally, an α-helical turn in TH8 C-terminus and TH9 N-terminus is formed in all MD simulations.

Final structures in T2 and T3 show structural difference in the inter-helical loop connecting helices TH7-8 in which final T2 structure exhibits an extended loop conformation and TH7 adopts a 3_{10} secondary structure (after unwinding of residues 313-314). Instead, T3 representative final structure shows a short α-helix (residues 318-321) and a 3_{10} turn (residues 318-320) in the loop between helices TH7 and TH8. Moreover, this loop is stabilized by docking of the protein C-terminal, which is not observed in T2. This is characterized by the stable side-chain distance of Y380 to H323 and formation of a salt-bridge between D377 - E326 along trajectory T3 in contrast to trajectory T2. Correspondingly, the short helix TH7 adopts a 3_{10} secondary structure in both low pH T-domain representative structures (backbone hydrogen bonds are disrupted in residue 313-314). T3 structure shows partial refolding of TH2 into a 3_{10} turn (residues 229-231).

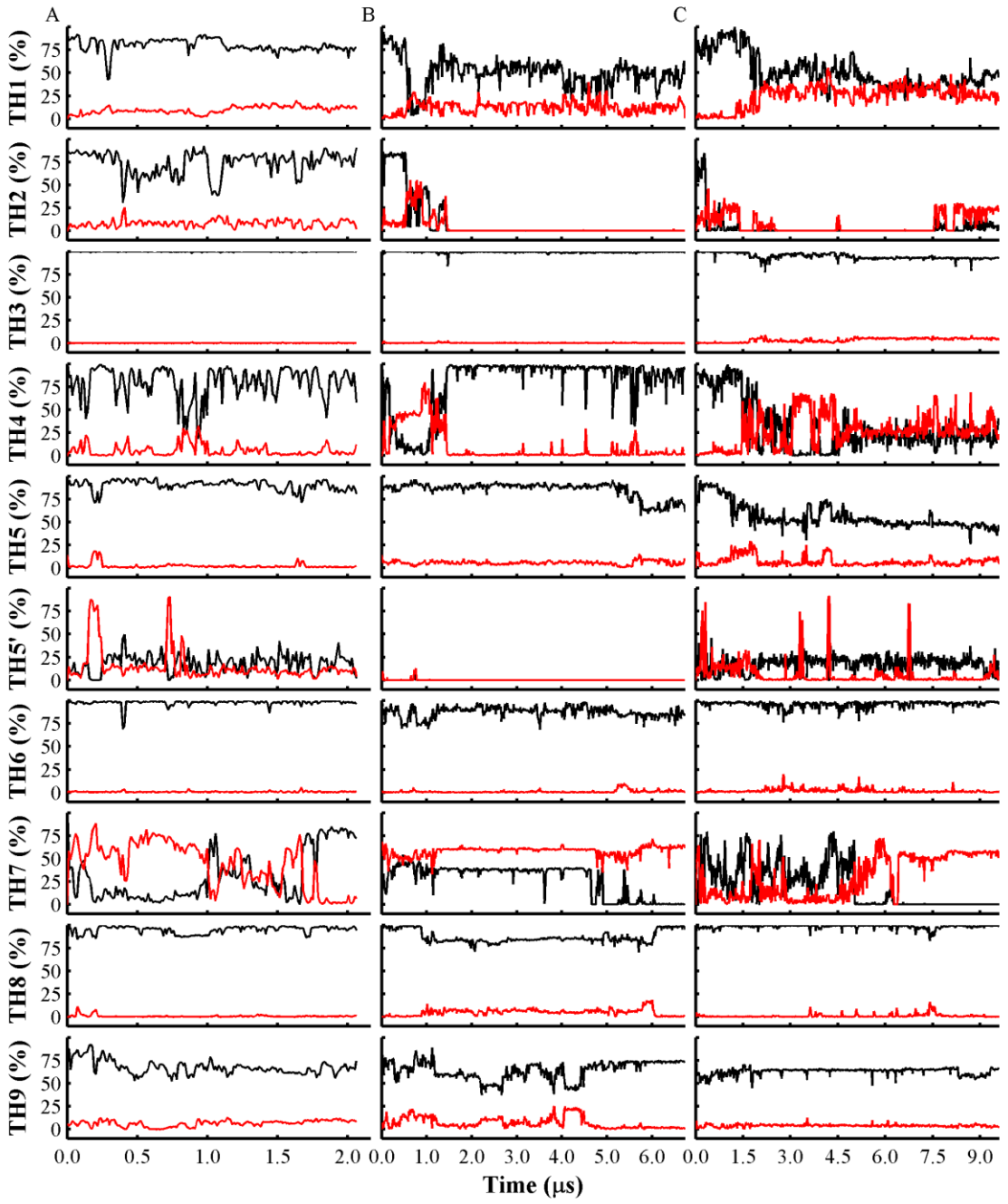

Figure 1.6. Helical content versus simulation time. Alpha-helical content (black lines) and 3_{10} content (red lines) are plotted for each alpha-helix identified in the crystal structure. (A) Secondary structure along the trajectory T1. (B) Secondary structure along trajectory T2. (C) Secondary structure over the course of trajectory T3. Helices identified

in the T-domain crystal structure. Analysis was performed with DSSP program and cumulative average was calculated with an averaging window of 1.2 ns.

1.3.3 Secondary Structure Dynamics

As shown in the previous section the structure of the T-domain at neutral pH (simulation T1) remained mainly unchanged and was stable in a two microsecond long simulation. An overall shift of residue positions from their initial states, as well as their mobility remains low for the duration of this fairly long simulation. However its secondary structure turns out to be rather dynamic on these time-scales. T-domain is an α-helical protein, and therefore its secondary structure folding and unfolding dynamics is well characterized by monitoring to what extend each α-helix maintains its helical form. The helicity content is calculated as a time - averaged fraction of helical residues that comprise an α-helix identified in the crystal structure (Figure 1.6). It is somewhat unexpected that the secondary structure of a natively folded stable T-domain (simulation T1) undergoes such significant dynamic transformations in the course of the simulation as shown in Figure 1.6A. Only three helixes TH3, TH6 and TH8 remain fully helical for 2μs of the simulation. Helixes TH1 and TH2 are both dynamic, i.e. they fluctuate to unfolded states at this time-scale (note the overall structure of the protein remains in the natively folded state, see *e.g.* Figure 1.5). The most common transition is alternating between α-helix and 3_{10} helix forms (also shown in Figure 1.6 in red lines), see *e.g.* TH4, TH7 and TH9 in Figure 1.6A. A noticeable degree of unfolding into coil structures and refolding is also observed in these helixes.

Unlike T1, trajectories T2 and T3 are of a protein that undergoes significant structural transformation in the course of the simulations including irreversible unfolding

of some helixes (see Figure 1.6B, C). Still, there are apparent similarities in the secondary structure dynamics of all three trajectories. In all trajectories helixes TH3, TH6 and TH8 preserve their α-helical content. Early in all trajectories, short helix TH5' and TH9 C-terminus unfold significantly. Helix TH1 begins to unfold at *ca.* 500 ns and *ca.* 2000 ns in trajectories T2 and T3, respectively. Complete unfolding of TH2 occurs approximately at 1400 ns in both T2 and T3. It is possible that the dynamic behavior of TH2 in trajectory T1 is indicative of its propensity to undergo unfolding upon HIS protonation (observed in T2 and T3). Moreover, according to a protein structure disorder prediction meta server[33] the sequence comprising helix TH2 , *i.e.* residues 226-233, is disordered. The meta server employed eight sequence-based disorder predictors, of which at least six predictors qualified this sequence as intrinsically disordered (with disorder prediction consensus larger than 0.8). The TH4 C-terminus unfolds early at 160 ns in T2 and refolds at 1420 ns. In T3, it unfolds at 1530 ns. Despite differences in unfolding behavior of TH4, its unfolding precedes structural changes of TH1 in both low pH T-domain trajectories. Also, TH7 transitions to a 3_{10} conformation at similar times (4630 ns and 5090 ns) in T2 and T3. Conversely, the TH5 N-terminus undergoes a slow transition after 5730 ns in T2 but a rapid one at 1180 ns in T3. (Note that H251 is located in TH3 and within 8 Å of N-terminal residues 275-279 of TH5).

Both similarities and differences are observed in the unfolding paths and the resulting protein structures in trajectories T2 and T3 (Figure 1.5). These are difficult to characterize, especially, given an incomplete and non-stationary sampling of the protein conformational space. Yet a joint analysis of the available trajectories is useful for inferring persistent properties of the underlying structural ensembles. To achieve that, a

principal component analysis (PCA) of the joint simulated structural ensembles is designed here. This consists on the projection of trajectories onto the two lowest principal components (PC) derived using only stationary basins of folded and unfolded structures. The metric used to characterize the structural order is a local helicity measure along the protein sequence introduced in a study of unstructured peptides by Speranskiy and Kurnikova.[24] The measure is defined for triplets of consecutive residues that is sufficient to determine whether the secondary structure is locally α-helical (a detailed description is provided in **1.2** ***Methods*** section). Principal components are initially determined using a joint dataset (T1, T2) that contains samples from the last stable segments of simulations T1 and T2 (1 μs each), *i.e.* using samples corresponding to quasi-equilibrium trajectories of a folded and a refolded protein. By design, first several principal components expose the major structural differences between these two protein states. Indeed, Figure 1.7A shows that the first two principal components obtained from the first dataset differentiate a folded (T1) and a partially unfolded ensemble (T2). The projection of the last 1 μs segment of T3 (also a sample from a quasi-equilibrium trajectory after protein unfolding and reorganization) partially overlaps conformational subspace of T2, (also shown in Figure 1.7A). This illustrates structural similarities in the ensembles of partially unfolded trajectories from two independent simulations. Furthermore, projection of the entire trajectories T2 and T3 on the same PCA vectors (Figure 1.7C) shows that the conformational subspace covered by trajectory T2 contains that of trajectory T3. Projection of the whole T1 onto the same subspace shows only small variation near its initial folded conformation.

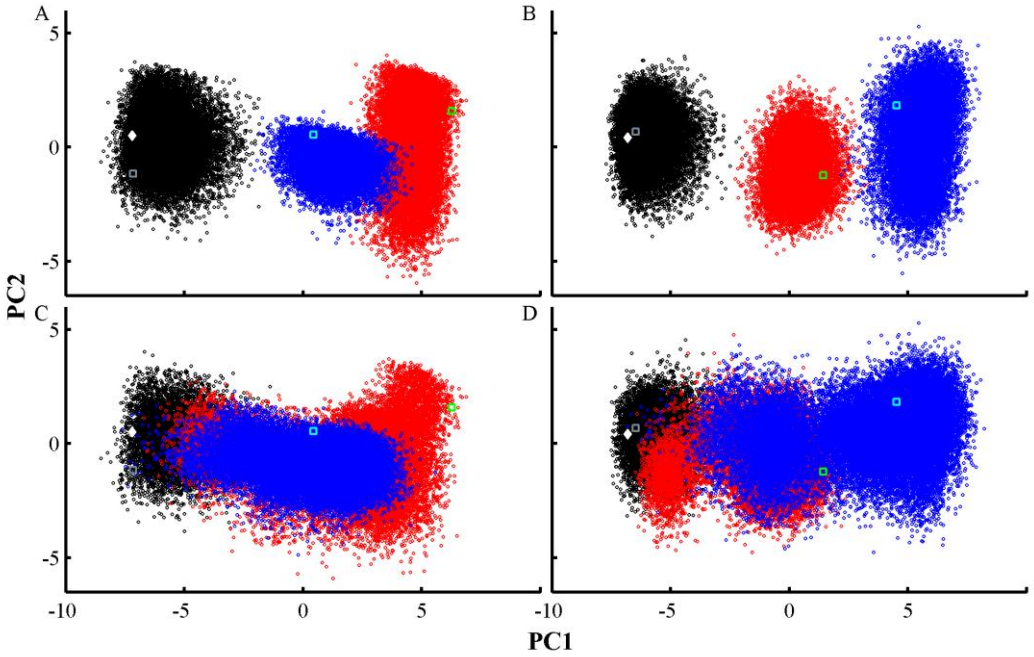

Figure 1.7. PCA projection of MD trajectories. Projection of MD frames on the two main principal components calculated from two datasets comprising the last 1 μs segments of trajectories T1 (black circles), T2 (red circles), and T3 (blue circles). Principal component analysis (PCA) was carried out on the triplet measure for each MD frame (residues 206-375). (A) Two-dimensional projection of the last 1 μs of all trajectories on principal components obtained from a dataset containing the last 1 μs of (T1, T2). (B) Two-dimensional projection of the last 1 μs of all trajectories on principal components obtained from a dataset containing the last 1 μs of (T1, T3.) (C) All trajectories are projected on principal components obtained from a dataset containing the last 1 μs of (T1, T2). (D) All trajectories are projected on principal components obtained from a dataset containing the last 1 μs of (T1, T3). Projection of the X-ray structure is shown in filled white diamond. Final MD snapshots of trajectories T1, T2, and T3 are shown in grey, cyan, and green empty squares, respectively.

In Figure 1.7B the same procedures are applied to the dataset formed first by joining the last equilibrium segments of trajectories T1 and T3. The samples from the last 1μs of trajectory T2 is then projected onto the two lowest PCA components of this (T1, T3) dataset. In this case the first two principal components of the second dataset show no overlap of the last MD segments of T2 and T3; however, projection of the complete MD trajectories on these principal components (Figure 1.7D) shows, once again, a partial overlap of structural ensembles of T2 and T3. MD frames from T2 and T3 that sample the same region of the projected conformational space correspond to the trajectory segment after the unfolding of TH1 at *ca.* 0.5 μs of T2 and to the first 5 μs of T3. Thus, the final segment of T3 samples a new set of conformations that are structured differently from ensemble in T2. The first principal component contributes *ca.* 44 – 47% of the variance in both datasets, see Appendix Figure 1.S6. Furthermore, triplets containing residue H257 have the largest influence on PC1 in both datasets, shown in Figure 1.S7. On the other hand, PC2 is influenced by triplets in helices TH6 or TH7 and by triplets close to residue H223 (residues 218-220 and residues 224-226 on each dataset, respectively).

The helicity measure used above is one way to characterize the secondary structure of a protein. Another structural measure commonly used to characterize subtle differences in protein conformations[35-37] is via individual values of the backbone dihedral angles (see **1.2** ***Methods*** for description). Such measure reports on a slightly different set of structural descriptors and is more sensitive to the specific local conformation of the backbone. For completeness of the structural analysis PCA was also performed in the manner described

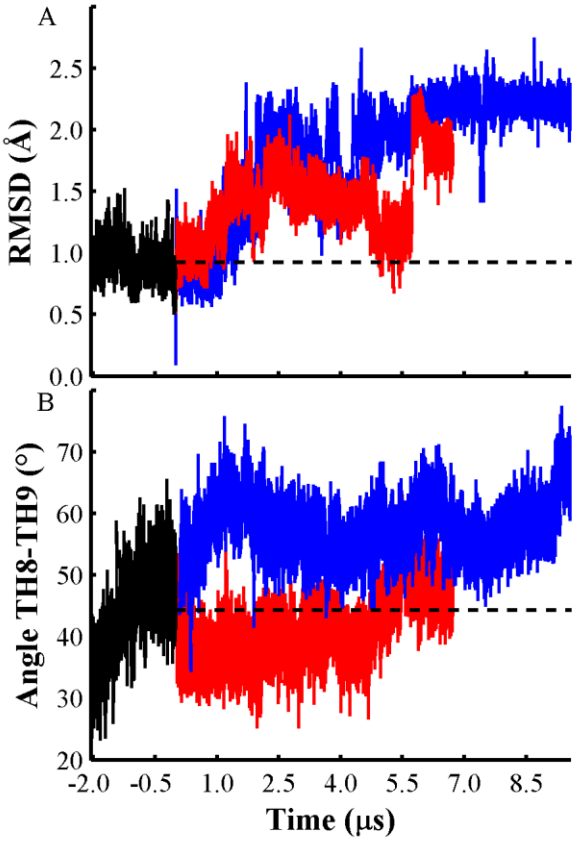

Figure 1.8. RMSD and interhelical angles of protein core helices. (A) C_α-RMSD of residues involved in the hydrophobic core formed by TH3, TH5 and TH8. RMSD traces trajectory T1 (black line), T2 (red line) and T3 (blue). (B) Interhelical angle between helices TH8 and TH9 along trajectories T1 (black), T2 (red) and T3 (blue). Average values from T1 are shown in dashed black lines.

above for the backbone dihedral angle measure. The results are presented in the subsection **1.5** ***Appendix***.

1.3.4 Protein Hydrophobic Core and Surfaces

Preserving a hydrophobic core and formation of a balanced water exposed surfaces is one of a signature properties of water-soluble natively folded protein structure.

Unfolding and subsequent protein-protein association in solution is typically accompanied to exposure of hydrophobic core to the solvent. In proteins that associate with membranes, exposure of hydrophobic sites to the solvent is expected to increase a protein membrane-association propensity. In diphtheria toxin increase of the membrane propensity is assumed at low pH due to partial unfolding and refolding of a protein into a postulated membrane-competent state. Such a state may be characterized by increased exposure of hydrophobic sites and partial loss of hydrophobic core. In this subsection we analyze how structural changes observed in the simulations and characterized in previous subsections affect solvent exposure of hydrophobic site. From analysis of T1 simulation in the previous sections we identified the non-polar helix TH8 as the T-domain core surrounded by hydrophobic surfaces of helices TH3, TH5 and TH9. The neutral pH MD simulation shows a stable core comprised of helices TH3, TH5 and TH8 (RMSF and RMSD less than 1Å, see Figure 1.5 and Figure 1.8A respectively). In contrast, C_α-RMSD traces increase to *ca.* 1.3 and 2.2 Å over both low pH trajectories T2 and T3, respectively. In both trajectories this increase is due to N-terminal unfolding of TH5. This unfolding occurs in the vicinity of protonated H251 (helix TH3), which has a brief increase of its side-chain solvent exposure in both trajectories T2 and T3. In trajectory T3, unfolding of a TH5 helical turn results in polarity changes surrounding W281 side-chain, which is contained in TH5 as well as Y278 (see Figure 1.S12). Furthermore, the new preferred state of Y278 involves side-chain/side-chain interactions with protonated H251, which is accompanied by the disruption of interhelical hydrogen bonds of Y278 to (S332-S336). Previous sections have shown that TH9 undergoes bending and C-terminal unfolding in both low pH T-domain MD runs. Therefore, to characterize the dynamics of TH9 relative

to the stable helix TH8, we determined their inter-helical crossing angle. This is calculated using C_α atoms of the stable TH9 N-terminus. Trajectory T3 shows a larger average angle (57 °) than the average angle over trajectory T2 (41 °) and the one obtained over the neutral pH protein trajectory (44 °). The interhelical angle in trajectory T3 is characterized by an early transition to around 60 °, in contrast to the late transition to larger angles over the last 400 ns of trajectory T2, see Figure 1.8B. This angle increase coincides with the C_α-RMSD increase of the hydrophobic core formed by helices TH3, TH5 and TH8, as shown in Figure 1.8A. Helix axes were determined using C_α atoms with RMSF < 1.3 Å in all MD simulations. Details of the interhelical angle calculation are explained in the **1.2.2** ***Analysis*** subsection. Similar calculations show that helical packing of TH6 and TH9 is highly dynamic in all MD trajectories. The interhelical angles vary between 48 ° to 90 ° with an approximate average angle of 72 ° in all MD simulations. This core is maintained with an average C_α RMSD < 1.7 Å in all MD simulations. The crossing angle fluctuations of helix TH9 highlights the loosening of hydrophobic contacts and solvent exposure of hydrophobic regions of the putative T-domain transmembrane alpha-helices in conditions that model low pH solution. The following section describes these changes observed in our T-domain simulations.

To identify changes on the protein surface hydrophobicity we calculated the hydrophobic solvent accessible surface area (HP-SASA). There is an average increase of approximately 400 $Å^2$ of HP-SASA in both low pH MD trajectories relative to the neutral pH MD simulation (see Figure 1.S13A). In particular, helices TH8 and TH9 have more hydrophobic sites exposed to the solvent in the low pH T-domain simulations than the rest of the protein. The HP-SASA of TH9 has an average value of 116 $Å^2$ through the T1

trajectory. This average increases to 187 and 213 Å^2 over T2 and T3 trajectories, respectively. HP-SASA of TH8 has an average value of 39.3 Å^2 calculated over the T1 trajectory. In contrast, this average value increases to 97.3 over the T2 trajectory and it remains around 40.7 Å^2 over the T3 trajectory.

Several hydrophobic sites in helix TH9 that are located in the hydrophobic interface of helices TH8 and TH9 become solvent exposed. For example, the F368 side-chain is significantly buried in trajectory T1 (8.4 Å^2), but its solvent accessibility increases along T2 (16.6 Å^2) and T3 (36.6 Å^2); the V371 side-chain remains partially buried in trajectory T1 (10.2 Å^2), but it is greatly exposed to the solvent over both low pH T domain trajectories T2 (40.2 Å^2) and T3 (78.5 Å^2); and I365 side-chain has a higher average SASA value in trajectory T2 (22.8 Å^2) than in T1 (11.0 Å^2) (yet, is it almost buried in T3 (1.3 Å^2).

In addition to the increased solvent exposure of the hydrophobic regions in helices TH8 and TH9 in T2 and T3 other protein regions become more solvent accessible. The SASA of the entire protein remains 9161 Å^2 on average over the neutral pH T-domain simulation (T1) and increases to 9639 and 9552 Å^2 in the trajectories T2 and T3, respectively (see Figure 1.S13B). Analysis of the different contributions to the protein solvent exposure shows that refolding of helices TH1 and TH2, as well as loop fluctuations contribute to the increased SASA in both low pH T-domain MD simulations.

1.4 Discussion and Conclusions

Partially unfolded ensembles of a diphtheria toxin T-domain model at low pH conditions were generated in two independent multi-microsecond MD simulations. The

third simulation presented here is of a protein in a folded form at neutral pH. In both low pH simulations protein unfolding transitions occurred in the time-frame between 1 μs to 5 μs. Further simulations led to formation of partially refolded structures with overall compact structure. These partially unfolded (from a native structure) and refolded structures may be attributed to formation of a membrane-competent state of the protein, which has been identified in previous work based on experiments of Ladokhin and co-workers.[10, 11]

The neutral pH T-domain simulation shows that the T-domain structure is stable, but it exhibits the unfolding and refolding of several helices. It also shows two different neutral pH T-domain states, which are characterized by the disruption of a salt-bridge in helices TH1 and TH9. Low pH T-domain simulations show partially unfolding in 2 μs, where trajectory T2 samples a faster refolding route possibly different from that sampled in trajectory T3. Both T2 and T3 refolded structures are stabilized over the last microseconds of the simulations with low RMSFs over the protein sequence. PCA analysis shows that T2 and T3 sample similar conformational subspace in their refolding routes from folded to unfolded T-domain states, but T3 may be sampling an additional subset of structures. The later is probably due to its longer simulation time and its larger simulation box than those of T2. The same analysis highlights the relative importance of His 257 in the unfolding-refolding process of both protonated T-domain MD trajectories. The simulated meta-stable state of a protein shows increased protein core fluctuations and solvent exposure of hydrophobic regions, which may facilitate protein-membrane interactions through the T-domain membrane insertion pathway.

PCA analysis was applied to various metrics that characterize conformational structural ensembles. The similarities of the conformational space sampled in our independent trajectories were best resolved using the helicity measure, earlier introduced by Speransky et al.,[24] in which the minimal helicity along the primary sequence is determined using three consecutive residues. The helicity measure shows that trajectories T2 and T3 sample similar conformational subspace in the space described by the two lowest principal components of the joint dataset (T1, T2). In these principal components the structures show partial subspace overlap of the last stable segments of T2 and T3 trajectories. Analysis of individual contribution of each residue triplet to the slowest principal component PC1 shows that in both unfolding trajectories T2 and T3 the unfolding transition and the resulting re-folded conformation is mainly influenced by triplets containing H257. In general, the observed importance of H257 is in good agreement with its suggested role in destabilization of T-domain structure at low pH.[11, 20] The second most important principal component, PC2 is influenced by the residues in helices TH6 or TH7 and by those near H223, which is located in the partially refolded N-terminal region. Thus, PCA analysis using the helicity measure based on triplets of residues is useful in the description of the dynamics of partially unfolded proteins. In contrast, the PCA analysis on the backbone dihedral space (a metric that is often used in characterization of protein structural ensembles) is unable to capture relative importance of different amino acids in the slow conformational changes observed in T-domain. Overall, the PCA analysis of trajectories T2 and T3 describes two similar transition routes from folded to partially unfolded T-domain ensembles.

Our simulations suggest that protonation of histidines triggers the refolding/unfolding of N-terminal helices TH1 and TH2.[20] Local conformational changes in helix TH1 were also detected in titration experiments by Wang *et al.*,[18] in which single fluorescence labels spanning helix TH1 suggested solvent exposure of hydrophobic residues in TH1. To rotate this helix, which was proposed by the authors, it may be necessary to break inter-helical salt bridges. As a result of histidines protonation, several interhelical salt-bridges located in the protein N-terminal are disrupted in our simulations of low pH T-domain in solution. For example, the salt bridge involving residue K216, which indeed occurs in both trajectories T2 and T3. Conformational flip of the K216 backbone dihedral angle initiates unfolding of TH1. In presence of lipid membranes, disruption of these salt-bridges may facilitate protein-membrane interactions.

Helix TH2 unfolds rapidly in both low pH protein simulations. This helix is predicted to be disordered by a consensus of disorder prediction algorithms.[33] Thus, our molecular dynamics studies coupled with disorder predictors show the role of protonation in triggering disordered regions of otherwise structurally stable proteins at neutral pH. The protonation-dependent disorder of TH2 may have implications for the membrane association and subsequent translocation of the protein N-terminus across the lipid bilayer. A similar case of locally disordered protein triggered by protonation of a single residue has been recently observed in multiple microsecond long dynamics of EGFR.[22]

Unfolding transition of helix TH1 follows conformational changes of TH4 in both low pH simulations. This can be explained by the loss of multiple tertiary contacts of these helices and helix TH8 and the disruption of a hydrogen bond network between residues in TH1 and H257 in the interhelical loop of TH3-4. Previously, we have reported a

significant free energy change of protonation of H257 in the protein native state (*ca.* 6.7 kcal/mol),[20] which may explain significant backbone rearrangements required to accommodate a positive charge in H257 side-chain. Similarly, positive free energy of protonation of H251 in the native state (around 3.7 kcal/mol),[20] may indicate conformational changes in helix TH5 N-terminus observed in the simulation in presence of the positive charge in H251. A partially unfolded full toxin crystal structure[19] shows helix TH4 unfolded, partial unfolding of TH1 and TH5 and the lack of electron density of TH2, which are in agreement with our simulations. Also, partial loss of helicity of TH1 at K216 is in agreement with the simulations.

In this chapter, we used a pre-determined fixed protonation state of the protein to mimic its state at a specific pH, the neutral pH and the low pH. This is an approximation which allows us to simulate protein behavior over microsecond long time-scales. A potentially better approach would be to perform a constant pH molecular dynamics, in which protonation states of protonatable residues can be modified over simulation time according to their current pK_a values. However, such approach would require extensive MD simulations of a relatively large protein solvated in explicit solvent, which precludes equilibrium sampling of partially unfolded and protonation states of all titratable residues. Development of biased molecular dynamics simulations may enhance the generation of ensembles of partially unfolded proteins in solution and this is a subject of future work.

In summary, we have presented extensive MD simulations and detailed analysis of partial unfolding of diphtheria toxin T-domain upon protonation of histidine residues in solution. Thorough comparison of simulated structures with experimental observations indicates that the results of the simulations complement experimental investigation to

provide unattainable microscopic details of structures that undergo transient modifications. Molecular modeling thus proves to be a valuable tool for such systems.

1.5 Appendix

1.5.1 Principal Component Analysis on Dihedral Space.

In addition to the principal component analysis (PCA) on the helical triplet space, we performed PCA analysis on the backbone dihedral angles of the ten α-helices identified in the crystal structure (TH1-9, TH5'). We defined two datasets containing the last 1 μs segments of trajectories (T1, T2) and (T1, T3). Thus, these datasets contain an ensemble of folded and refolded conformations of T-domain. We first calculated the principal components (PC) from the dataset (T1, T2). Figure 1.S8A shows that the last 1 μs segments of T2 and T3 have no overlap in a reduced conformational subspace described by the two lowest principal components. Figure 1.S8C shows the projection of all MD frames. T2 contains the majority of T3 MD frames on the two principal components of dataset (T1, T2), which is similar to our observations in the triplet space (helicity), see Figure 1.7C. Secondly, we also performed PCA analysis on the dataset (T1, T3). Figure 1.S8B shows no overlap of last segments of T2 and T3 on the first two principal components. Figure 1.S8D shows partial overlapping of the entire trajectories T2 and T3. The first principal component of both datasets (T1, T2) and (T1, T3) has a significant variance contribution ca. 54 – 56 %, see Figure 1.S9. This contribution of PC1 is similar to that observed in the triplet space. However, residues with the largest influence in PC1 are different for the dihedral and triple space. For example, Fig 1.S10A shows that a backbone dihedral in residue T267 has the largest influence in PC1 for the dataset (T1, T2), while a triplet containing residue H257 in PC1 obtained from the same dataset (see Figure 1.S7A). In general this is also observed for dataset (T1, T3). For example, Figure 1.S10B shows that a backbone dihedral in E298 has the largest influence in PC1, while a

triplet containing residue H257 has the largest influence for the same dataset (see Figure 1.S7B).

We also performed dihedral PCA analysis over the complete trajectories T2 and T3. We conclude that the projection of T2 and T3 have no significant overlap on the lowest PC vectors from T2 or T3 (see Figures 1.S11A, B). Notice that PC1 has a smaller variance contribution ca. 18 – 22 % (see Figure 1.S12), which is lower than the ones calculated for datasets (T1, T2) and (T1, T3).

1.5.2 Supplementary Figures

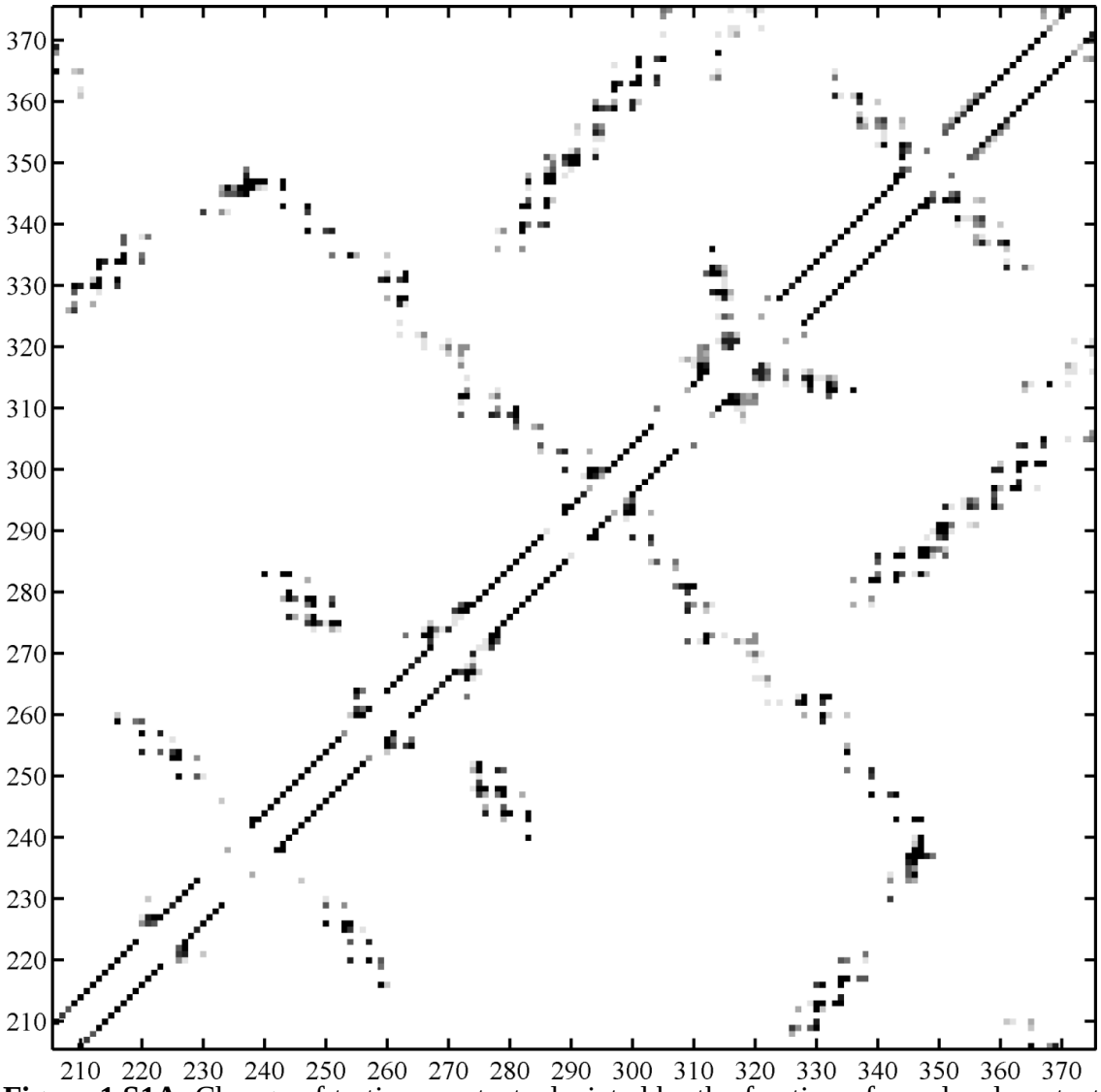

Figure 1.S1A. Change of tertiary contacts depicted by the fraction of non-local contacts, calculated over Cα atoms of residues 206-375 in trajectory T1. Fraction values in the range of 0 and 1 are represented by grey scale. Fraction of non-local contacts between pairs ij of Cα atoms separated by at least 8 Å and residues containing atoms i and j are separated by at least 3 residues.

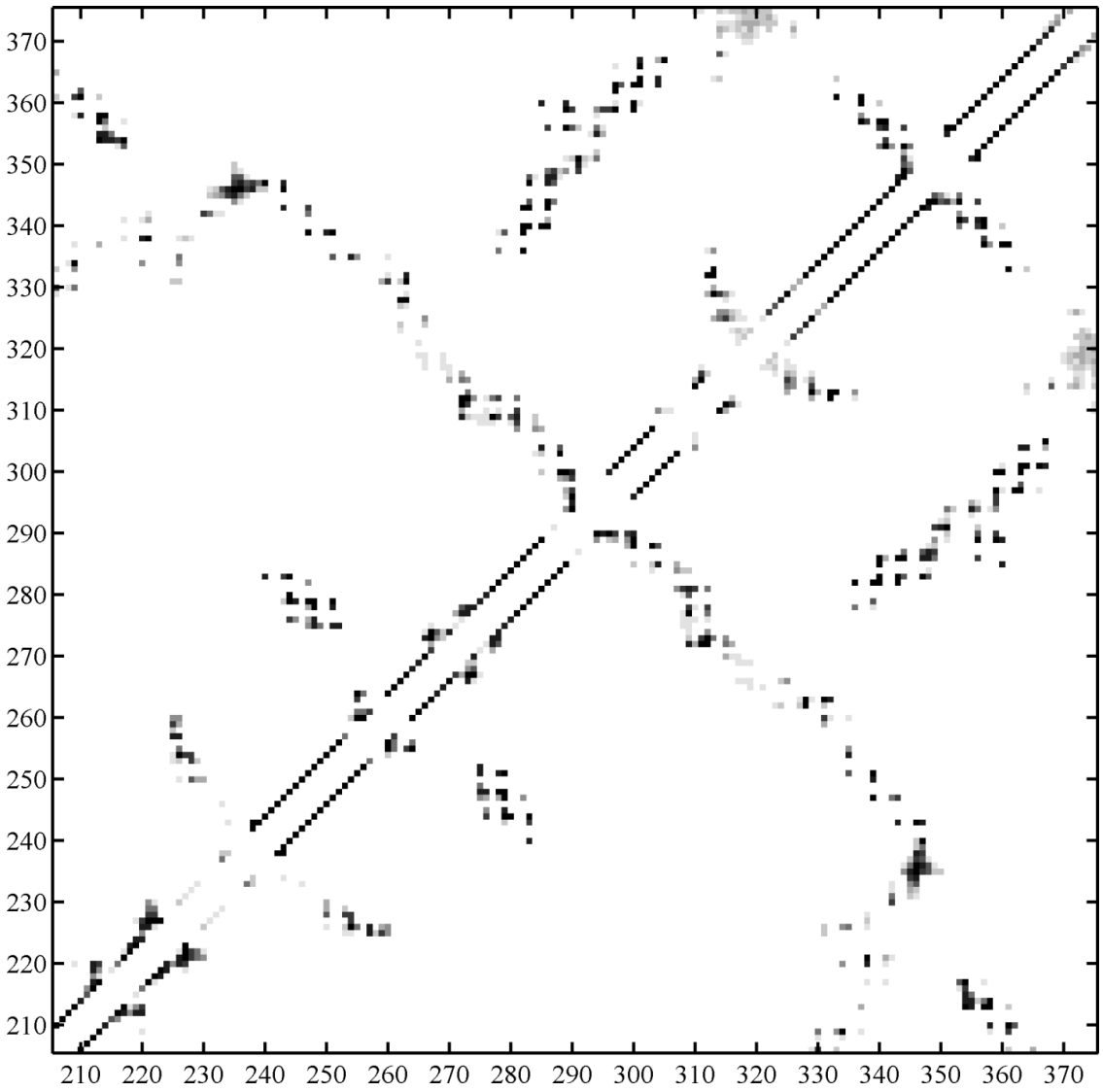

Figure 1.S1B. Change of tertiary contacts depicted by the fraction of non-local contacts, calculated over Cα atoms of residues 206-375 in trajectory T2. Fraction values in the range of 0 and 1 are represented by grey scale. Fraction of non-local contacts between pairs ij of Cα atoms separated by at least 8 Å and residues containing atoms i and j are separated by at least 3 residues.

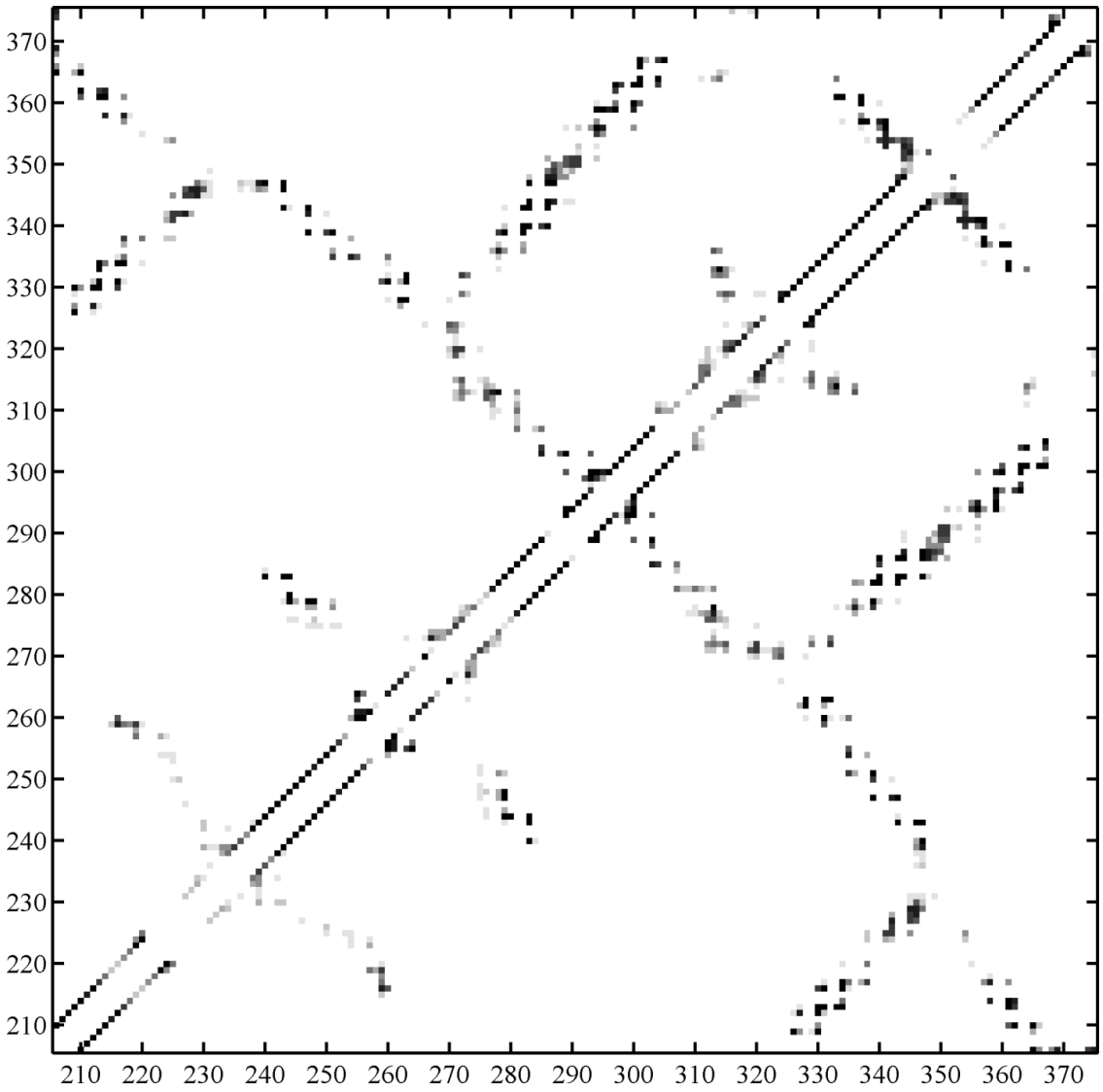

Figure 1.S1C. Change of tertiary contacts depicted by the fraction of non-local contacts, calculated over Cα atoms of residues 206-375 in trajectory T3. Fraction values in the range of 0 and 1 are represented by grey scale. Fraction of non-local contacts between pairs ij of Cα atoms separated by at least 8 Å and residues containing atoms i and j are separated by at least 3 residues.

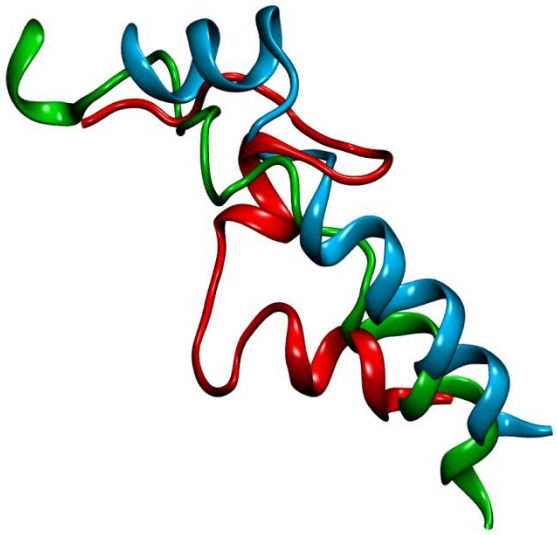

Figure 1.S2. Overlay of helices TH1-2 from representative structures obtained from the last 1 μs of trajectories T1 (cyan), T2 (red) and the last 2 μs of T3 (green) in cartoon representation. Helices TH1 and TH2 are partially unfolded in both representatives T2 and T3 after the conformational changes described in the results section.

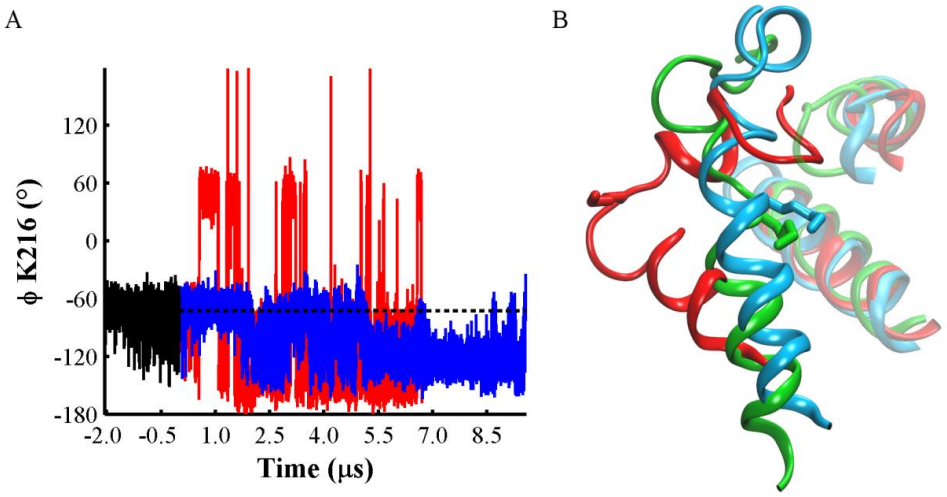

Figure 1.S3. (A) φ dihedral angle traces of K216 over trajectories T1 (black line), T2 (red) and T3 (blue). Broken black lines represent the average value of φ K216 (-73 °) in trajectory T1. (B) Overlay of helix TH1 from representative structures of T1 (cyan ribbon), T2 (red) and T3 (green). K216 is highlighted by stick representation. Helices TH3 and TH8 are shown in transparent representation using the same colors.

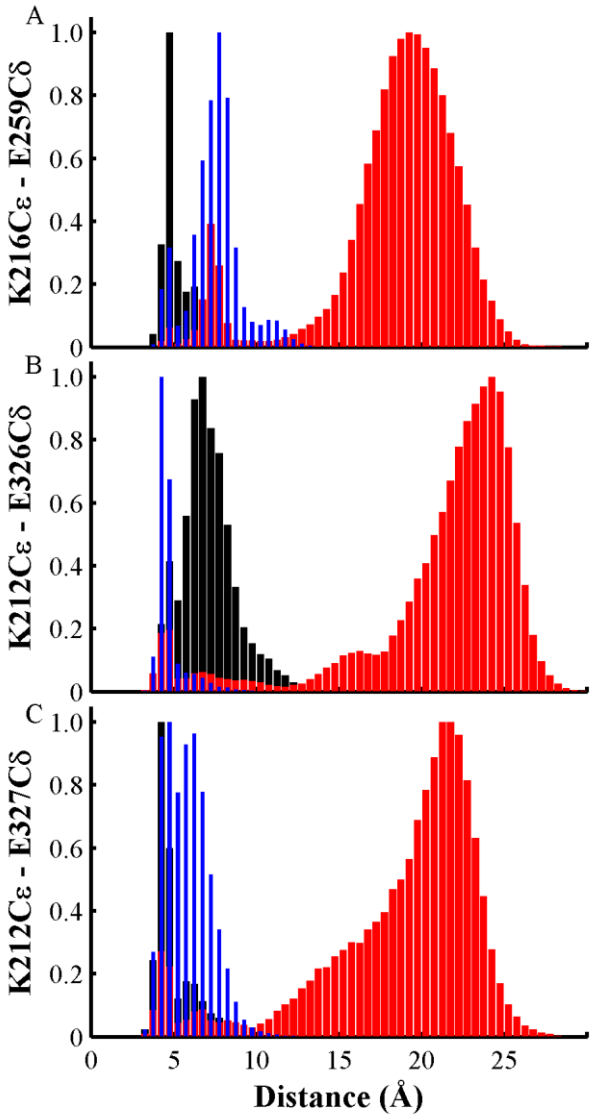

Figure 1.S4. Normalized frequency of distances in salt-bridges obtained over trajectories T1 (black histograms), T2 (red histograms), and T3 (blue histograms). (A) Histogram of distances of atoms in side-chains of K216 – E259. (B) Histogram of distances of atoms in side-chains of K212 – E326. C. Histogram of distances of atoms in side-chains of K212 – E327. K212 is located in helix TH1, E259 is located in the loop between TH3 and TH4, E326 and E327 are located in helix TH8.

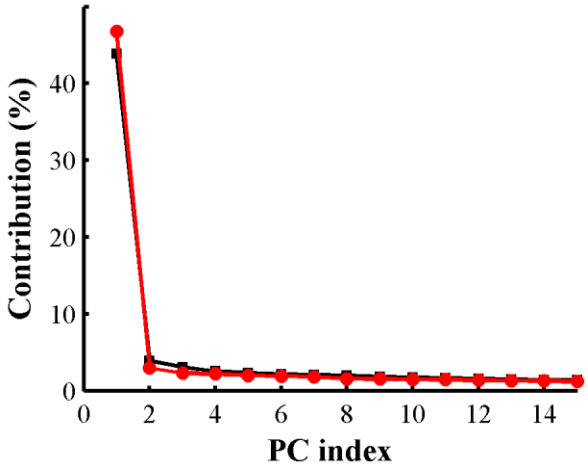

Figure 1.S5. Variance contribution of the first 15 principal components (PC) obtained from datasets containing the last 1 μs segments of trajectories (T1, T2) or (T1, T3) shown by black and red lines, respectively. Helicity measure is calculated for each MD frame (residues 206-375).

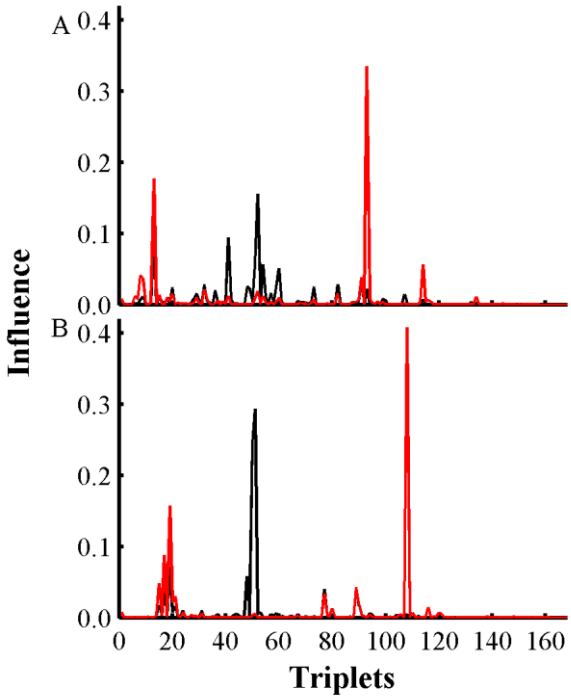

Figure 1.S6. Influence of each triplet on the first principal component (black lines) and second principal component (red lines). (A) Influence of each triplet obtained from datasets containing the last 1 μs segments of trajectories (T1, T2). (B) Influence of each triplet obtained from datasets containing the last 1 μs segments of (T1, T3). Helicity measure is calculated for each MD frame (residues 206-375).

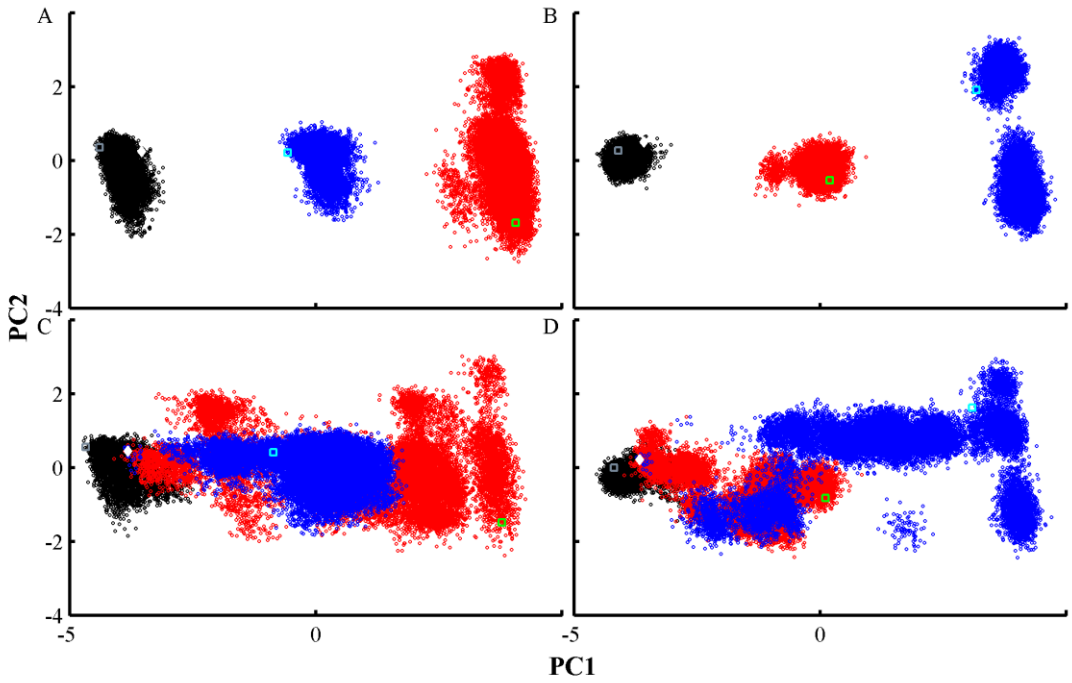

Figure 1.S7. Projection of MD trajectories on the two main principal components calculated for two datasets comprising the last 1 μs segments of trajectories (T1, T2) or (T1, T3). MD frames from trajectories T1, T2, and T3 are represented by black, red, and blue circles, respectively. (A) Two-dimensional projection of the last 1 μs of all trajectories on principal components obtained from a dataset containing the last 1 μs of (T1, T2). (B) Two-dimensional projection of the last 1 μs of all trajectories on principal components obtained from a dataset containing the last 1 μs of (T1, T3). (C) Two-dimensional projection of all trajectories on principal components obtained from a dataset containing the last 1 μs of (T1, T2). (D) Two-dimensional projection of all trajectories on principal components obtained from a dataset containing the last 1 μs of (T1, T3). Projection of the X-ray structure is shown in filled white diamond. Final MD frames of trajectories T1, T2, and T3 are shown in grey, cyan, and green empty rectangles, respectively. Dihedral principal component analysis (dPCA) was carried out over the

backbone dihedral angles of helices TH1-9, TH5' identified in the crystal structure, excluding loops and terminal residues.

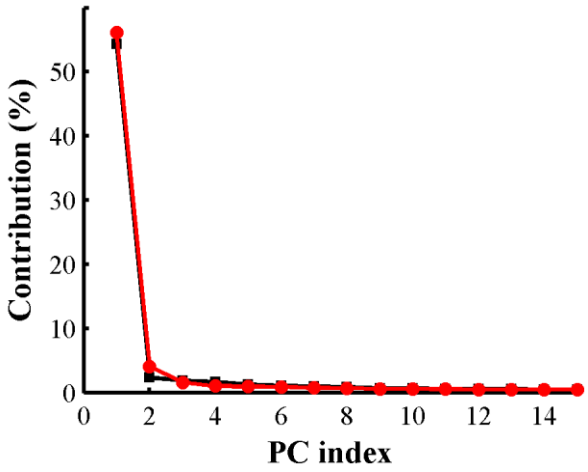

Figure 1.S8. Variance contribution of the first 15 principal components (PC) obtained from datasets containing the last 1 μs segments of trajectories (T1, T2) or (T1, T3), shown in black and red lines, respectively. Dihedral PCA analysis was carried out over backbone dihedral angles of residues from helices (TH1-9 and TH5').

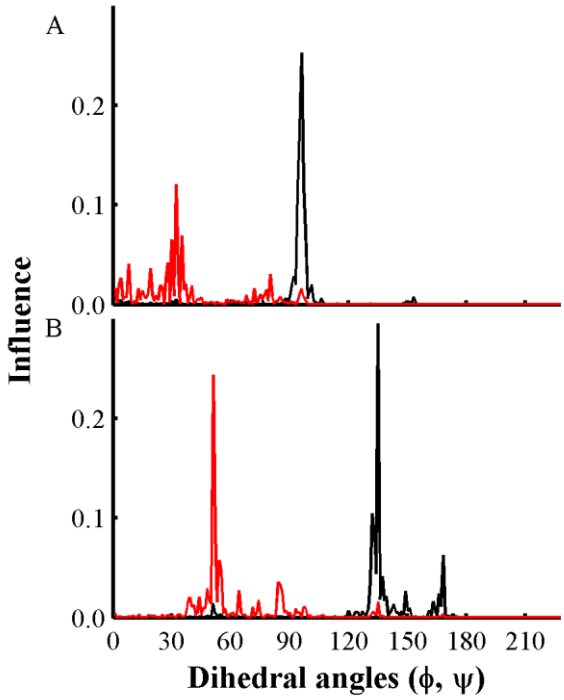

Figure 1.S9. Influence of each backbone dihedral angle on the first principal component (black lines) and second principal component (red line). (A) Influence of dihedral angles (φ, ψ) obtained from datasets containing the last 1 μs segments of trajectories (T1, T2). (B) Influence of dihedral angles (φ, ψ) obtained from datasets containing the last 1 μs segments of (T1, T3). Backbone dihedral angles are obtained from residues from helices (TH1-9 and TH5').

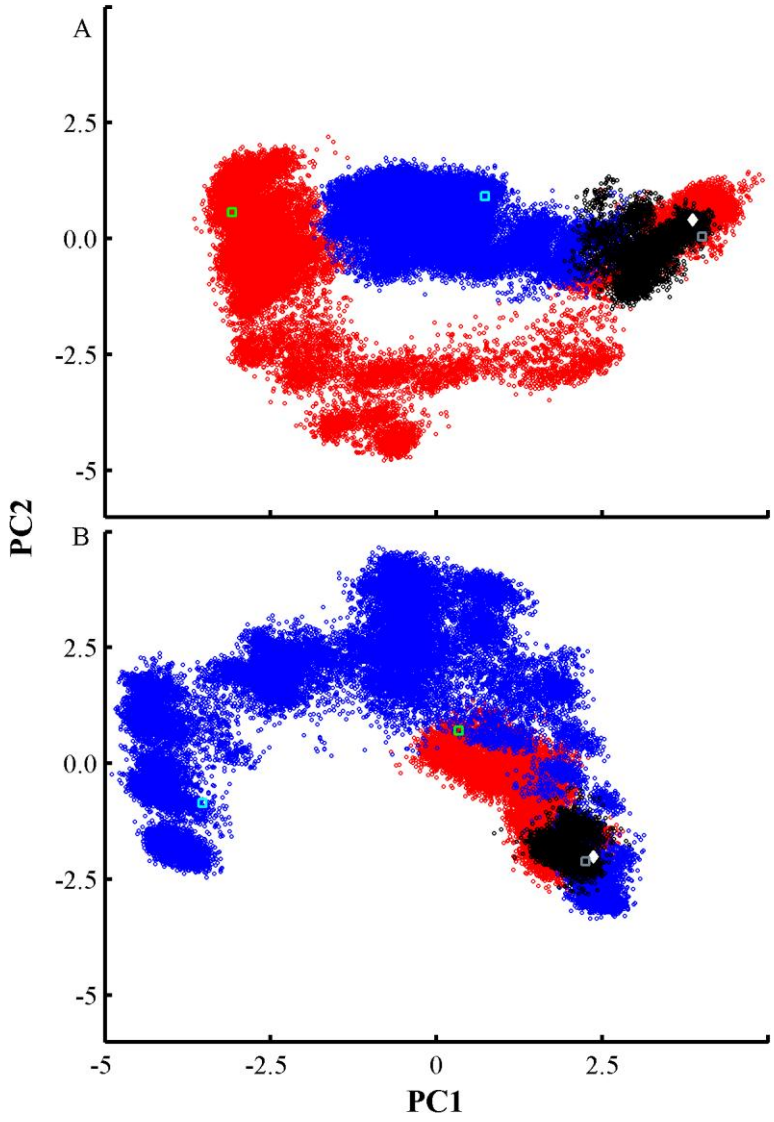

Figure 1.S10. Projection of all MD trajectories on the two first principal components calculated using MD frames from: (A) Trajectory T2. (B) Trajectory T3. Dihedral PCA analysis was carried out over backbone dihedral angles of residues from helices (TH1-9 and TH5'), excluding loops and terminal residues. MD snapshots from trajectory T1, T2, and T3 are represented by black, red, and blue circles, respectively. Projection of the X-ray structure is shown in filled white diamond. Final conformations of trajectories T1, T2, and T3 are shown in grey, cyan, and green empty rectangles, respectively.

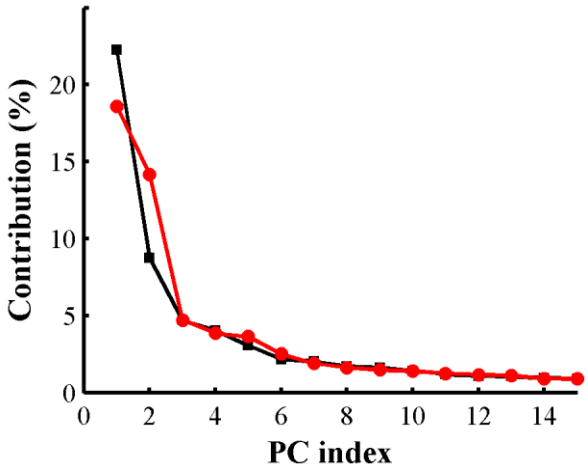

Figure 1.S11. Variance contribution of the first 15 principal components obtained from trajectories T2 (black line) and T3 (red line). Dihedral PCA analysis was carried out over backbone dihedral angles of residues from helices TH1-9 and TH5'.

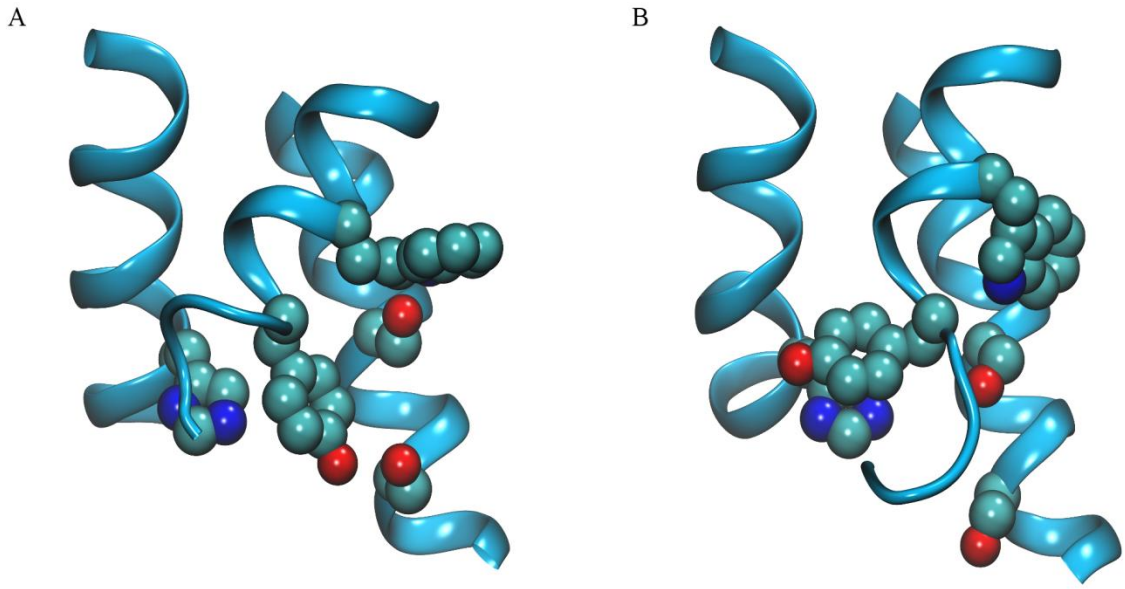

Figure 1.S12. (A) Initial and (B) final frames obtained from MD trajectory T3. Side-chains of H251, S332, S336, Y278 and W281 are highlighted in space-filled representation. Oxygen, nitrogen, and carbon atoms are colored in red, blue, and blue, respectively. Helices TH3, TH5, and TH8 are represented in cyan ribbons.

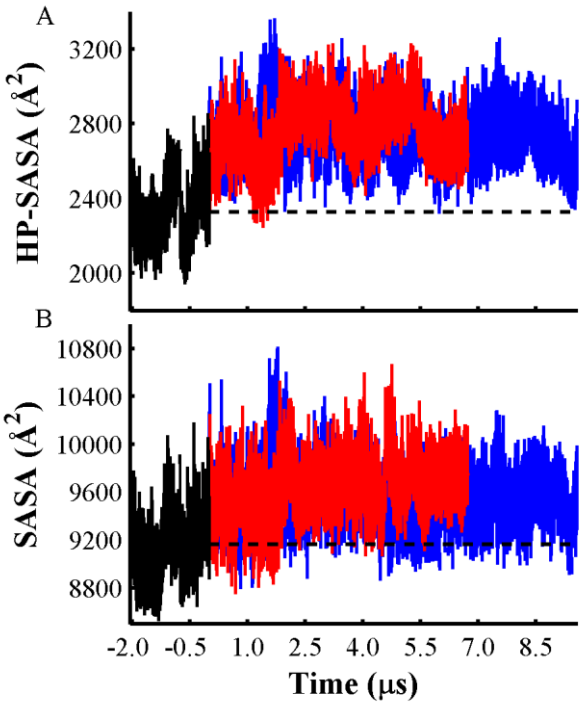

Figure 1.S13. SASA traces versus simulation time of hydrophobic and all side-chains in T-domain. (A) Hydrophobic SASA (HP-SASA) of the entire T-domain for T1, T2, and T3, shown in black, red and blue lines, respectively. (B) SASA of all side-chains in T-domain for T1, T2, and T3, shown in black, red and blue lines, respectively. Broken black lines represent the average values of HP-SASA and SASA for T1, 2328 Å^2 and 9161 Å^2, respectively.

Chapter 2. Implementation of an Accelerated Molecular Dynamics Method. Case of Study: Diphtheria Toxin T-domain

2.1 Introduction

Conformational changes of proteins are often associated with their function in complex processes in cells such as DNA replication, signal transduction and transport across membranes. The mechanics of conformational changes in proteins is increasingly studied by atomistic molecular dynamics (MD) simulations, which can provide insights of the protein behavior upon changes in the protein environment (for example changes of protonation states,[20, 34, 35] ligand-binding,[36] or membrane interactions[37]). These computational studies have benefited from recent improvements of atomistic force fields and the development of a specialized MD machine (Anton), which was used to study the folding and unfolding of relatively small proteins over millisecond time-scales.[38] However, proteins containing more than a hundred residues solvated in explicit water are generally limited to atomistic MD simulations of time-scales of hundreds of nanoseconds to a few microseconds using conventional computational resources. Thus, the study of slow structural changes of proteins of this size requires the application of enhanced sampling methods, which can rely on multiple copies of the system (temperature or Hamiltonian replica exchange)[39, 40] or on single-copy based methods such as dihedral based tempering,[41] and boosting of potential energy terms (accelerated molecular dynamics).[42, 43] In this study, we propose an implementation of accelerated molecular dynamics (aMD) based on boosting electrostatic interactions of solute-solute atom pairs and we apply it to the study of conformational changes of diphtheria toxin translocation T-domain triggered by protonation of key residues. This proposed method can benefit the

conformational sampling of large proteins and possibly influence the understanding of the relationship of conformational changes and protein function upon changes in the environment.

aMD accelerates the conformational sampling of a biomolecular system by decreasing the energy barriers separating energy wells, which facilitates the conversion of the system from one stable state to another one. aMD simulations are based on the modification of the energy landscape in which the energy wells are filled when the current potential energy is lower than a reference value. Otherwise, the potential energy is left unperturbed. In particular, aMD does not require previous information of the conformational landscape. However, aMD simulations depend on the size of the system and the selection of the predefined reference value, which introduce a type of error denominated statistical noise and affect a statistical sampling error (see Markwick et al.[44]). For example, aMD simulations of relatively large systems or the setting of a high reference value can result in the frequent sampling of high energy conformations with little sampling of low energy conformations. The few low energy conformations are associated to large weighing factors, which distort the reweighting of the observed properties of the system. A low predefined reference value can result in slow conversions of the system between stable states. To avoid the sampling of high energy conformations, several modifications of the aMD method have been proposed in order to protect the high energy barriers.[45, 46] Sinko et al.[46] proposed a new boosting potential that modifies the energy wells and the high energy barriers and the degree of acceleration is controlled by four parameters. However, the increment of parameters increases the difficulty of finding

optimal values that accelerate the conformational sampling for each particular biomolecular system.

In this chapter, we propose an implementation of the aMD method that consists of the boosting of electrostatic interactions between solute-solute atom pairs of proteins in explicit water. This implementation aims to accelerate the sampling of conformational states and improve the reweighting of properties of relatively large biomolecules. We test this implementation in the sampling of the free energy landscape of alanine dipeptide and in the sampling of conformational changes of a relatively large protein such as diphtheria toxin translocation (T) domain. T-domain is known to undergo conformational changes upon protonation of histidines in explicit solvent.[20, 47]

2.2 Theory

Accelerated molecular dynamics consists in the modification of the potential energy surface of molecular systems by adding a positive bias potential $\Delta V(r)$ when part or the total potential energy $V(r)$ is below a reference value E_{cut}:[42]

$$V^*(r) = \begin{cases} V(r) & V(r) \geq E_{cut} \\ V(r) + \Delta V(r) & V(r) < E_{cut} \end{cases}, \quad (2.1)$$

$$\Delta V(r) = \frac{\left(E_{cut} - V(r)\right)^2}{\alpha + \left(E_{cut} - V(r)\right)}, \quad (2.2)$$

where $V^*(r)$ is the modified potential energy, $\Delta V(r)$ is the biasing continuous function, r are the coordinates of the system, E_{cut} is a parameter calculated using the average potential energy obtained from short conventional MD simulations and controls the

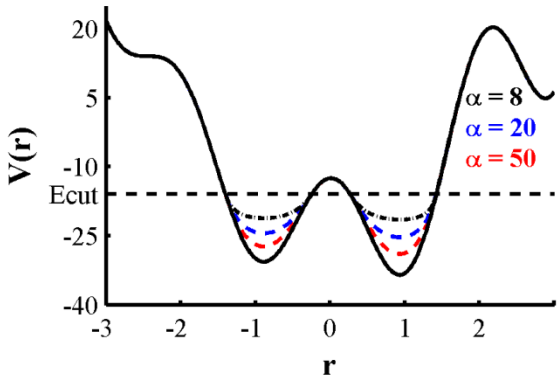

Figure 2.1. Representation of a unidimensional energy potential V(r) (black continuum lines). Modified potential are shown in broken lines for different values of α. The reference E_{cut} value is shown by an horizontal broken line.

degree of perturbation of the potential energy surface, and α is a parameter that controls the shape of the modified potential (see Figure 2.1). Values of α close to zero induce a random walk on the potential energy surface. A similar result is obtained for significantly large values of E_{cut}. The canonical ensemble of a property $A(r)$ can be calculated by reweighting each configuration obtained from a MD simulation performed on the modified potential energy $< A(r) > = < A(r) e^{\Delta V(r)/k_B T} > / < e^{\Delta V(r)/k_B T} >$ where k_B is the Boltzmann's constant and T is the temperature. Currently, there are two variants of the method implemented in PMEMD program,[48] which are called dihedral boosting[42] and dual-boosting aMD.[43] The first approach applies the boosting potential to only dihedrals, while the second one applies an overall boosting potential to the total energy of the system and applies a separate biasing potential to all dihedrals in the system. In this work, we apply the bias potential to part of the electrostatic interactions of solute-solute atoms of a system in explicit solvent and periodic boundary conditions. This sampling method is

denominated aMD_EE. The electrostatic energy of the system is determined by the functional form:

$$V_{elec}(r) = \frac{1}{2}\sum_{\mathbf{n}}{}' \sum_{i,j}^{n_{atoms}} \frac{q_i q_j}{|\boldsymbol{r}_{i,j}+\mathbf{n}|}, \qquad (2.3)$$

where n_{atoms} is the number of atoms in the system, q_i and q_j are the fixed charges of atoms i and j, $\boldsymbol{r}_{i,j}$ is the vector joining atoms i and j within the unit cell, $\mathbf{n} = n_1\mathbf{x} + n_2\mathbf{y} + n_3\mathbf{z}$ is the cell vector and $n_{1,2,3} \in \mathbb{N}$. $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are vectors of length L that form the unit cell. The prime indicates that contributions from atoms $i = j$, excluded atoms and $\mathbf{n} = 0$ are omitted. However, this sum is not appropriate for computational calculations because is not absolutely convergent.[49] Particle mesh Ewald (PME) method solves this problem by partitioning the electrostatic energy into separate contributions $V_{elec}(r) = V_{direct}(r) + V_{reciprocal}(r) + V_{corr}(r)$ denominated direct-space and reciprocal-space summations and correction contributions, respectively.[50, 51] These terms converge rapidly. The direct-space energy term $V_{direct}(r)$ involves the summation over interactions of atomistic charges in the system and Gaussian charge distributions of neutralizing magnitude placed at each individual charge. β is the Ewald parameter, which controls the width of the charge distribution and the rate of convergence. β is also chosen so that direct-space interactions beyond a cutoff radius are negligible. The direct-space energy term $V_{direct}(r)$ can be decomposed into only solute-solute and non-solute interactions $V_{direct}(r) = V_{direct,solute}(r) + V_{direct,*}(r)$ represented by:

$$V_{direct}(r) = \frac{1}{2}\sum_{\mathbf{n}}{}' \sum_{i,j}^{n_{solute}} q_i q_j \frac{\mathrm{erfc}(\beta|\boldsymbol{r}_{i,j}+\mathbf{n}|)}{|\boldsymbol{r}_{i,j}+\mathbf{n}|} + \frac{1}{2}\sum_{\mathbf{n}}{}' \sum_{i,j}^{*} q_i q_j \frac{\mathrm{erfc}(\beta|\boldsymbol{r}_{i,j}+\mathbf{n}|)}{|\boldsymbol{r}_{\mathrm{i,j}}+\mathbf{n}|}, \quad (2.4)$$

where $V_{direct,solute}(r)$ is the direct summation over pairs of solute-solute atoms. $V_{direct,*}(r)$ is the direct space summation over all pairs of atoms, but pairs of solute atoms. erfc is the complementary error function and n_{solute} is the number of solute atoms. The asterisk represents summation over all pairs of atoms, but pairs of solute atoms i, j. In this work, aMD simulations are performed over the modified potential:

$$V^*(r) = \begin{cases} V_0(r) + V_{direct,solute}(r) & V_{direct,solute}(r) \geq E_{cut} \\ V_0(r) + V_{direct,solute}(r) + \Delta V_{direct,solute}(r) & V_{direct,solute}(r) < E_{cut} \end{cases}, \quad (2.5)$$

where $V_0(r)$ is the original energy potential of the system without the direct-space contribution of solute-solute atoms. Therefore $V(r) = V_0(r) + V_{direct,solute}(r)$. The functional form of $\Delta V_{direct,solute}(r)$ is implemented as abovementioned. The modified force is

$$F^* = -\nabla V^*(r) = -\nabla V_0(r) - \nabla V_{direct,solute}(r)\left(\frac{\alpha}{\alpha + (E_{cut} - V_{direct,solute}(r))}\right)^2, \quad (2.6)$$

Note that the complementary function erfc is approximated by a cubic-spline, which permits the calculation of analytical derivatives of the direct-space energy term.

2.3 Methods

2.3.1 Molecular Dynamics Simulations of Alanine Dipeptide

The initial extended structure of alanine dipeptide was created using tleap (AMBER12)[48] and solvated by 606 water molecules TIP3P. The distance between the

peptide and the edge of the simulation box was approximately 10 Å. The total number of atoms in this system was 1840. In all simulations the force field ff99SB was used,[52] simulation time step was 2 fs, and all hydrogen bonds were constrained via SHAKE.[53] Furthermore, periodic boundary conditions were set up, the cutoff radius was set to 8 Å, and electrostatic calculations were performed using Particle Mesh Ewald (PME) method.[50] A brief description of the equilibration and production simulations is described in the following paragraph. The peptide was restrained and the solvent was minimized for 200 steps of steepest descent followed by 50 steps of conjugate gradient descent minimization method. Peptide backbone atoms were restrained and the entire system was minimized by a total number of 250 steps as explained above. The system was linearly heated to T = 300 K using the NVT ensemble and a Langevin thermostat over 20 ps with restraints in all protein atoms (constant force of 10 kcal/mol·Å^2). This was followed by 400 ps of anisotropic NPT ensemble equilibration using Berendsen barostat with a reference pressure of 1atm, taup = 0.5 ps, and T = 300 K with positional restraints applied to heavy atoms (2.5 kcal/mol·Å^2). Unrestrained NPT equilibration was performed over 11 ns with taup = 1.0 ps. Parameters for aMD_EE simulations were estimated using an extended 10 ns NVT ensemble simulation. A similar approach of adjusting the parameters of total boost aMD simulations was used for aMD_EE. The average direct-space electrostatic energy of solute-solute pairs was $< E >_{EEDIR}$= -7.6 kcal/mol. The reference value was determined by $E_{cut} = < E >_{EEDIR} + 0.1 * n_{solute}$, where $n_{solute} =$ 22. The parameter α was calculated using $\alpha = 0.16 * 0.4 * n_{solute}$. Parameters for dual-boost aMD simulations were calculated using energy averages from an extended 10 ns NVT ensemble simulation. The average total energy was $< E >_{total}$= -5827 kcal/mol

and $< E >_{dih}=$ 11.5 kcal/mol. Dual-boost parameters are calculated using $E_{T,cut} = < E >_{total} + 0.175 * n_{atoms}$, $\alpha_T = 0.175 * n_{atoms}$, $E_{dih,cut} = < E >_{dih} + 3.5 * n_{res}$, and $\alpha_{dih} = 0.2 * 3.5 * n_{res}$, where $n_{res} = 2$ is the number of residues. aMD_EE and dual-boost aMD simulations were performed using an in-house modified version of pmemd.MPI and pmemd.cuda, respectively. To obtain a converged simulation of the peptide, a conventional MD simulation was carried out using pmemd.cuda over 1 μs on a GTX-680 card.[54] aMD simulations were carried out up to 12ns. Atom coordinates were saved every 2 ps.

2.3.2 Molecular Dynamics Simulations of T-domain

The initial coordinates of T-domain were obtained from a high resolution structure of the entire diphtheria toxin at pH 7.5 (PDB ID code 1F0L). The protein model contains residues 201-380. Hydrogen atoms were added using tLeap (AMBER12). Previous atomistic MD simulations of the low pH model of T-domain solvated in explicit solvent were performed with all histidines protonated over 6.8 μs and 9.5 μs, which are referred as aMD1 and aMD2, respectively.[20, 47] Details of these simulations are described in previous work. In this work, the low pH T-domain model was created by adding 13215 TIP3P explicit water molecules and the distance between the protein and the simulation box edge was 16.0 Å. All six histidine sidechains were set in a protonated state while all other titratable residues were set in their standard states. Four sodium ions were added to the simulation box to neutralize the system. The total number of atoms was 42405. The system was equilibrated with pmemd.cuda and the force field ff99SB was used in all simulations reported. The simulation time step was 2 fs and all hydrogen bonds were constrained via SHAKE. Periodic boundary conditions were set up, cutoff radius was set

to 12 Å, and electrostatic calculations were performed using Particle Mesh Ewald (PME) method. The protein was restrained and the solvent was minimized for 200 steps of steepest descent followed by 50 steps of conjugate gradient descent minimization method. In the second minimization stage, protein backbone atoms were restrained and the entire system was minimized by a total number of 250 steps as explained above. Afterwards, the system was linearly heated to T = 310 K using the NVT ensemble and a Langevin thermostat over 20 ps with restraints in all protein atoms (constant force of 1 kcal/mol·Å^2).This was followed by a 580 ps anisotropic NPT ensemble equilibration using Berendsen barostat with a reference pressure of 1atm, taup = 1 ps, and T = 310 K with positional restraints applied to heavy atoms (1.0 kcal/mol·Å^2). Then, the restraint constant was set to 0.75 kcal/mol·Å^2 over the following 100 ps. Other 100 ps are performed with a restraint constant to 0.5 kcal/mol·Å^2 and another simulation of the same length with a restraint constant of 0.25 kcal/mol·Å^2. This was followed by unrestrained NPT equilibration over 58 ns. To estimate the parameters for dual-boost aMD simulations, an extended run of 50 ns was performed using the NVT ensemble. The average total energy was $< E >_{total}$= -128974 kcal/mol and $< E >_{dih}$= 1960 kcal/mol. Dual-boost parameters are calculated using $E_{T,cut} = < E >_{total} + 0.16 * n_{atoms}$, $\alpha_T = 0.16 * n_{atoms}$, $E_{dih,cut} = < E >_{dih} + 3.5 * n_{res}$, and $\alpha_{dih} = 0.2 * 3.5 * n_{res}$, where $n_{res} = 180$ is the number of residues in the protein. After the calculation of aMD parameters, multiple independent dual-boost aMD production simulations were performed using pmemd.cuda on GTX-680 cards. Dual-boost aMD simulations were performed using the NVT ensemble. To estimate the parameters for aMD_EE simulations, a second extended run was performed over 10 ns using the NVT ensemble.

Electrostatic interactions were computed by PME method with a cutoff radius of 10 Å. The average direct-space electrostatic energy of solute-solute pairs was $< E >_{EEDIR}$= -3523.4 kcal/mol. The reference value was determined by $E_{cut} =< E >_{EEDIR} + 0.24 * n_{solute}$, where $n_{solute} = 2756$. The parameter α was calculated using $\alpha = 0.16 * 0.2 * n_{solute}$. Parameters used for aMD_EE simulations of alanine dipeptide were slightly modified to accelerate the sampling of T-domain structural changes. Production aMD_EE simulation was carried out using the NVT ensemble, PME method with a cutoff radius of 10 Å. aMD_EE simulation was carried out up to 140 ns, five independent dual-boost aMD up to an accumulated 716 ns and a conventional MD simulation up to 140 ns. Atom coordinates were saved every 2 ps.

2.3.3. Analysis

To calculate the C_α-RMSD curves over simulation time, MD structures were translated and rotated relative to the X-ray structure excluding C_α atoms in the tails or those in the loops and tails. The C_α-RMSD is calculated excluding C_α atoms from the tails and loops identified in the crystal structure. Root mean square deviation (RMSD), distances between individual atoms, averaging of protein structures and secondary structure analysis were calculated using the ptraj program available in AmberTools13.[48] Molecular figures were prepared using VMD 1.9.1.[55] Secondary structure assignments was determined by DSSP program.[56] Reweighting of aMD trajectories was performed using in-house python scripts.

2.3.4 Covariance Matrices

A covariance matrix was calculated by first translating and rotating all MD frames relative to their average structure using C_α atoms of residues 206-375. This selection of

residues did not include the protein terminal residue, which were flexible in all MD simulations. The covariance matrix analysis was performed using the coordinates from C_α atoms of residues 206-375 as follows. The set of dimensions $\boldsymbol{x} = \{x_1, x_2, \ldots x_i\}$ where $i = 1 \ldots p$, p is the number of dimensions, x_i is a vector of size M, and M is the number of MD frames. A $p \times p$ covariance matrix R is defined as follows: $R = 1/M\ DD\hat{}T$ where D is a $p \times M$ matrix of elements $D_{ij} = x_{ij} - < x_i >$, and $< x_i >$ is the average of x_i over an ensemble of sampled protein conformations.

2.4 Results and Discussion

2.4.1 Accelerated Molecular Dynamics Simulations

We test the proposed sampling method on two biomolecular systems and compare its sampling performance relative to conventional MD and dual-boost aMD. Alanine dipeptide solvated in explicit solvent is a convenient system because of its free energy landscape and conformational space can be described by two backbone dihedral angles. The second system tested is diphtheria toxin T-domain, which is composed of 180 residues and adopts an alpha-helical structure at high pH. Acidification of the solution triggers decrease of the secondary structure content, local rearrangements of the N-terminal helices and formation of a membrane-competent state, which were determined by a series of experiments.[20] These observations were in good agreement with microsecond-long MD simulations of T-domain in explicit solvent, which showed that protonation of histidines triggers the partial unfolding of the N-terminal helices and increases of the hydrophobic solvent accessible surface area.

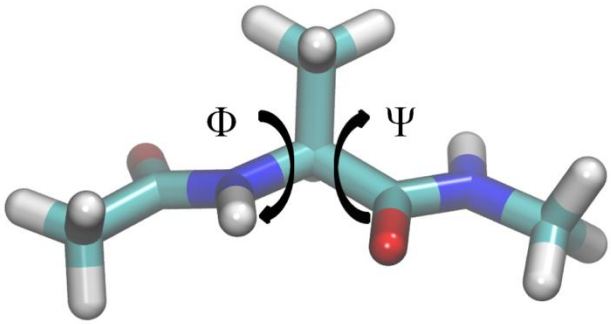

Figure 2.2. Structure of alanine dipeptide. Dihedral angles Φ and Ψ are indicated.

2.4.2 Alanine Dipeptide

The structure of alanine dipeptide and its backbone dihedral angles Φ and Ψ are shown in Figure 2.2. Figure 2.3 shows the free energy landscape projected on these two dihedral angles for conventional MD and aMD simulations. The free energy landscape obtained by a 1 μs conventional MD simulation is shown by Figure 2.3A. It shows the population of three different free energy minima with values of Φ lower than -30 °. Two other less populated minima are shown for values of Φ around 60 °. A 12 ns conventional MD simulation samples the three lowest free energy regions, but shows no sampling of the region for values of Φ around 60 ° (see Figure 2.3B). Free energy surfaces of aMD simulations are directly computed by exponential reweighting of each MD frame. Remarkably, a 12 ns aMD_EE simulation shows the sampling of all low free energy regions, as shown in Figure 2.3D. Notice that a 12 ns dual-boost aMD shows few low free energy conformations (see Figure 2.3C). Overall, the proposed method aMD_EE

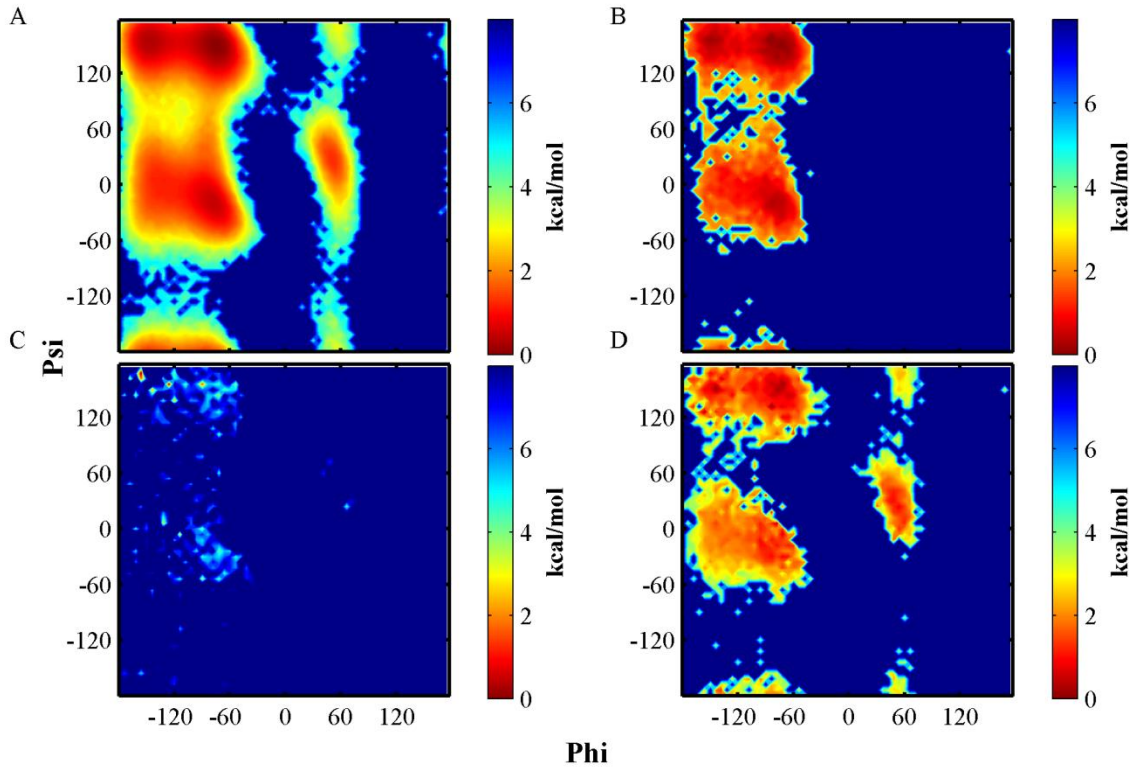

Figure 2.3. Comparison of free energy profiles of alanine dipeptide. A) Results obtained from a 1000 ns conventional MD simulation. Similar results obtained from: (B) a 12 ns conventional MD trajectory. (C) a 12 ns dual-boost aMD. (D) a 12 ns aMD_EE. Reweighting was performed as indicated on methods.

avoids the generation of MD frames with large weighting factors, which are usually generated by the dual-boost aMD simulation.

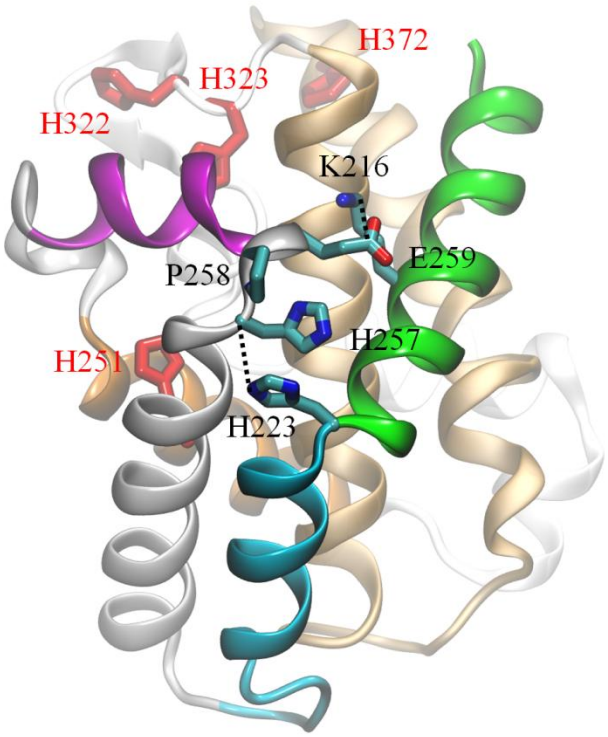

Figure 2.4. X-ray structure of T-domain. Coordinates were extracted from diphtheria toxin structure at pH 7.5 [PDB 1F0L]. Helices TH1, TH2, TH4 and TH5 are shown in green, cyan, magenta and orange ribbon representation, respectively. Helices TH8 and TH9 are shown in brown ribbons. Sidechains of K216, H223, P258, H257 and E259 are shown in licorice representation. Sidechains of H251, H322, H323 and H372 are shown in red licorice representation. The distances between atoms H223@NE2 – H257@CA (4.4 Å) and K216CE – E259CD (5.0 Å) are highlighted by broken lines.

2.4.3 Diphtheria Toxin T-domain

Diphtheria toxin T-domain consists of ten alpha-helices, named TH1-9, TH5', at pH 7.5, as shown in Figure 2.4. Anton MD simulations of T-domain solvated in explicit solvent showed that protonation of histidines triggers partial unfolding of N-terminal helices, exposure of hydrophobic sites while retaining a global compact structure.[20, 47]

These features were verified by circular-dichroism and fluorescence experiments.[20] However, two independent trajectories of length 6.8 μs and 9.5 μs showed some differences in the extension of conformational changes of the N-terminal helices.[47] These conventional MD simulations are referred as cMD1 and cMD2, respectively.

To test the ability of our proposed method to reproduce conformational changes observed on microsecond long simulations, we perform an aMD_EE simulation applying a boost potential to the solute-solute electrostatic interactions over approximately 140 ns. In addition, five aMD simulations using the dual-boost approach are performed for an accumulated time of 716 ns. A 140 ns conventional MD simulation is also performed. To determine the degree of conformational changes in all MD simulations, we calculate the root mean square deviation (RMSD) of C_α atoms of helices TH1-9 obtained from all trajectories relative to their positions in the crystal structure (see Figure 2.5). aMD_EE simulation generated protein conformations with the largest RMSD values compared to all other five dual-boost aMD simulations and the 140 ns conventional MD simulation. Furthermore, aMD_EE generates structures with an average RMSD value of 4.1 Å over the last 20 ns, which is similar to the averages values obtained from two microsecond-long MD simulations cMD1 and cMD2, as shown in Figure 2.5. Averages RSMD over the last 20 ns of the dual-boost aMD trajectories are within 1.6 to 3.2 Å. aMD_EE simulation generated similar degree of overall structural changes observed in extensive conventional MD simulations of a low pH model of T-domain in explicit solvent.

Protonation of histidines was demonstrated to induce decrease of secondary structure using long classical MD simulations performed on Anton.[20] Figure 2.6A shows the average helicity content as a function of residue for the aMD_EE and a representative

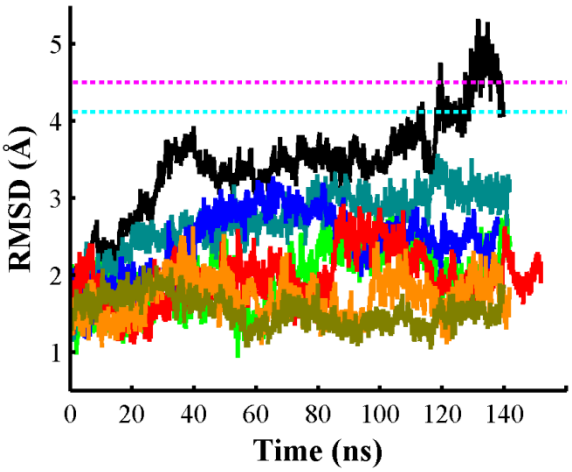

Figure 2.5. Root mean squared deviation (RMSD) traces obtained by different MD methods. Each curve is calculated using C_α-atoms in helices relative to the crystal structure. RMSD obtained from aMD_EE is shown in black line and five other independent trajectories of dual boost aMD are shown in dark cyan, blue, green, red and orange lines. Representative dual-boost aMD is shown in blue line. Conventional MD simulation is shown in olive line. Cyan and magenta broken lines represent the average RMSD obtained from the last 2 μs of conventional MD simulations cMD1 (4.1 Å) and cMD2 (4.5 Å), respectively.

dual-boost aMD trajectory. aMD_EE shows that protonation of histidines induces decrease of helical content of N-terminal helices TH1, TH2, C-terminus of TH3, TH4. Figure 2.6A also shows that aMD_EE sampled partial unfolding of helices TH5, TH6 and TH7. In contrast, only a single dual-boost aMD trajectory shows partial unfolding of helix TH2 (see Figure 2.6A), while in general all dual-boost aMD trajectories show decrease of helical content of helices TH4 and TH7 (see Figure 2.S1). As a reference, a 140ns conventional MD simulation shows no significant changes of the average helicity

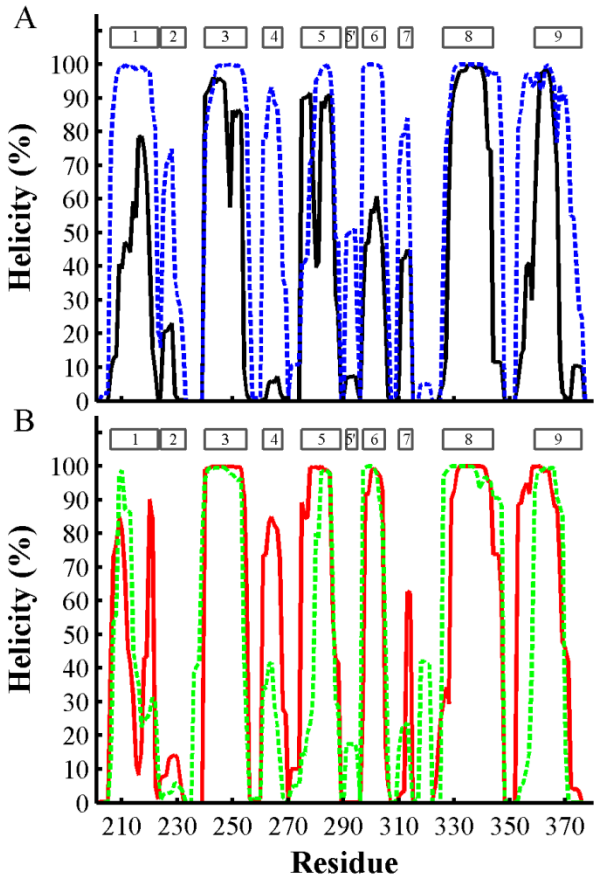

Figure 2.6. Average helicity content per residue obtained by different MD methods. (A) aMD_EE (black line) and a representative trajectory obtained by dual-boost aMD (broken blue line). (B) Two conventional MD simulations of length 6.8 μs (red line) and 9.5 μs (broken green line). Helices TH1-9, TH5' are represented by rectangles on the top of the figure.

per residue (see Figure 2.S1). Microsecond long MD trajectories cMD1 and cMD2 shows partial unfolding of N-terminal helices TH1-2, TH4 and TH7, as shown in Figure 2.5B.

2.4.4 Protonation of H257 Triggers Conformational Changes

Protonation of N-terminal histidines triggers significant changes in residues and helices near their initial positions in the neutral pH structure. In particular, histidine H257 forms a structural motif of sequence HPE, which is characterized by interactions between N_δ atom of H257 and the backbone nitrogen atom of E259 in the X-ray structure, as shown in Figure 2.7A. This close interaction is mediated by the backbone torsional angles of P258 and is located in the loop joining helices TH3 and TH4. Protonation of H257 disrupts the structure of residues P258 and E259, which is shown by the two MD simulations cMD1 and cMD2, as shown in Figures 2.7B, C. aMD_EE simulation samples similar local structural changes around H257 (see Figure 2.7D). Dual-boost aMD also generates structures with local destabilization around H257; however, the ψ backbone torsion angle of P258 samples multiple conformations, as shown in Figure 2.7E. This can be a consequence of the boost applied to all dihedral angles in the protein system. In general, protonation of H257 destabilize backbone torsional angles of successive aminoacids and also disrupts stabilizing hydrogen bonds between between H257 – E259 and H257 – S219.

In addition to changes in the local structure, repulsive electrostatic interactions between the positively charged sidechains of H257 and H223 has been predicted by pK_a calculations to destabilize the neutral pH structure upon decrease of pH in the solution.[3, 20] The destabilizing role of histidines was demonstrated by microsecond long MD, which triggered the disruption of interhelical bridges, for example K216 (helix TH1) and E259. In this work, we find useful to use the distance between both pairs of residues to describe conformational changes around the N-terminal helices. The initial separation between the

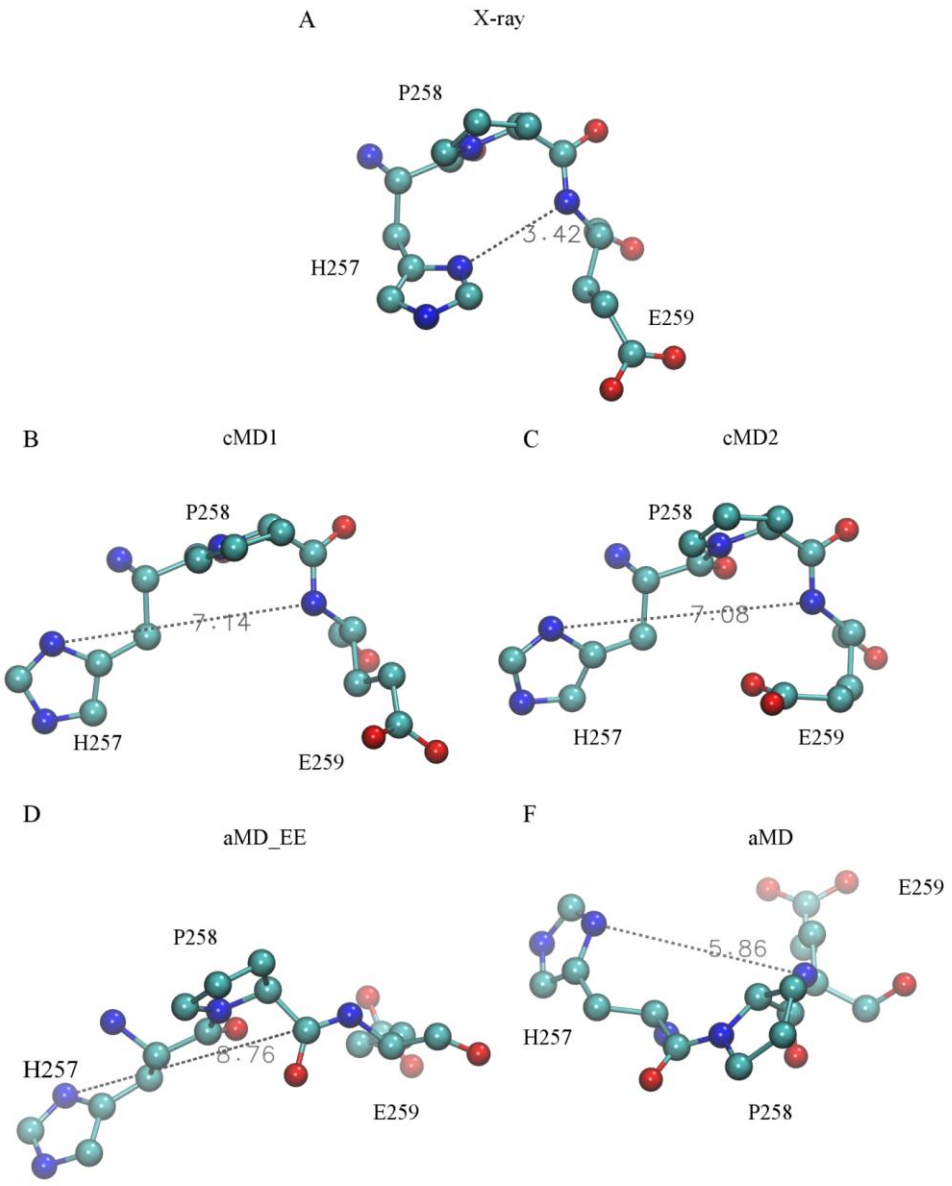

Figure 2.7. Structures of the sequence H257-P258-E259 generated by different MD methods. (A) X-ray structure. (B) Generated by conventional MD simulation called cMD1 of length 6.8 μs**.** (C) Conventional MD simulation called cMD2 of length 9.5 μs. (D) Generated by aMD_EE. (E) Obtained from a dual-boost aMD. Residues H257, P258 and E259 are shown in balls and sticks representation and oxygen, nitrogen and carbon atoms are shown in red, blue and cyan colors, respectively. The distances between atoms

H257@ND1 – E259@N are shown by broken lines on each structure (Angstroms units). Structures are aligned relative to backbone atoms CA, N, C of residues 257-258 in the crystal structure at pH 7.5 [PDB ID 1F0L]

N_ε atom of H223 relative to the C_α atom of H257 in the crystal structure is shown in Figure 2.4. This figure also shows the distance between the atoms C_ε and C_δ of K216 and E259, respectively. As a reference, MD frames of the simulations cMD1 and cMD2 show the increase of the distances of these pairs and the partial unfolding of helices TH1 and TH2 (see Figures 2.8A, B). aMD_EE conformations show that the pair of histidines are separated from each other and the interhelical salt-bridge is disrupted (see Figure 2.8C). This figure also shows partial or total unfolding of helices TH1-4 and rotation of helix TH1 relative to the crystal structure. It also shows the solvent exposure of H251 as a result of the partial unfolding of helix TH3 C-terminus. A representative dual-boost aMD trajectory shows similar disruption of the pair of histidines and the interhelical salt-bridge; however, it only shows unfolding of helix TH2 (see Figure 2.8D). All other dual-boost aMD trajectories did not show unfolding of helix TH2 (see Figure 2.S1).

Previous MD simulation studies of T-domain have demonstrated that protonation of N-terminal histidines triggers conformational changes of the N-terminal helices TH1-2 of atomistic models of T-domain in explicit solvent.[20, 47] Particularly, cMD1 trajectory showed that protonation of histidines triggered the formation of a kink in helix TH1 (see Figure 2.8A), which was characterized by the increased separation of C_α atoms from residues W206 and Q369 relative to the crystal structure. This observation was in good

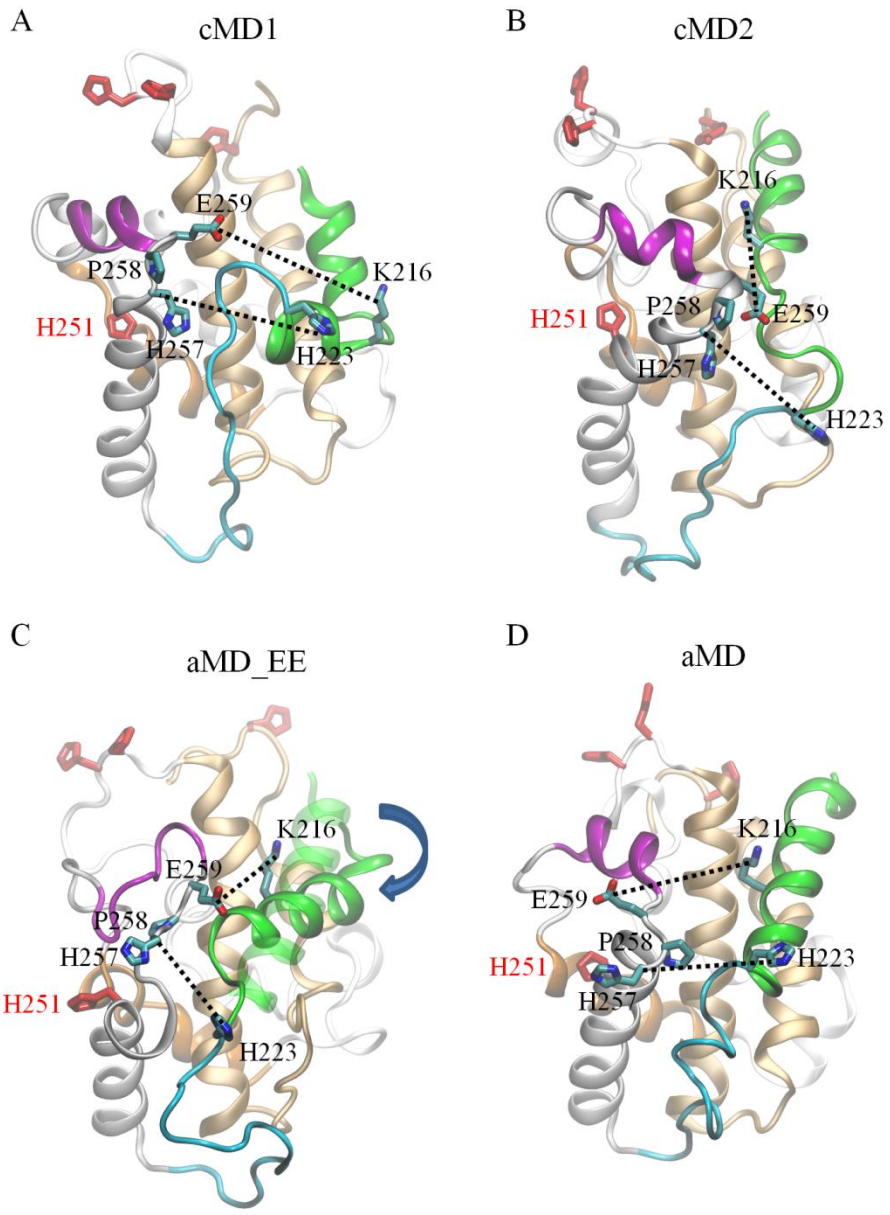

Figure 2.8. Models of partially unfolded T-domain generated by different MD methods. (A) MD frame obtained from a conventional MD simulation of length 6.8 μs**.** (B) Structure obtained from a conventional MD simulation of length 9.5 μs. (C) T-domain structure generated by aMD_EE. The curved arrow indicates the rotational displacement of helix TH1 relative to its initial orientation in the crystal structure (transparent green ribbon representation). (D) Protein structure obtained from a trajectory generated by dual-

boost aMD. Helices TH1, TH2, TH4 and TH5 are shown in green, cyan, magenta and orange ribbon representation, respectively. Helices TH8 and TH9 are shown in brown ribbons. Residues K216, P258, E259 and all histidines are shown in licorice representation. The increase of the distances between atoms H223@NE2 – H257@CA and K216CE – E259CD relative to the X-ray structure is highlighted by broken lines on each structure. The respective distances are for cMD1 (16.5 Å, 18.6 Å), cMD2 (14.8 Å, 11.2 Å), aMD_EE (12.1 Å, 13.3 Å) and aMD (14.0 Å, 14.5 Å).

agreement with the pH-dependent separation between W206 and a single site bimane label in the mutant form Q369C observed by changes of fluorescence.[20] However, the separation of these residues was not observed in cMD2 trajectory, which highlights the difficulties of conventional MD simulations to sample pH-dependent structural changes in large proteins.[47] To show the ability of aMD_EE method to sample the abovementioned structural changes, Figure 2.9 displays the distance traces between C_α atoms of residues W206 and Q369 for all aMD simulations. aMD_EE generates conformations with an average distance (11.7 Å) similar to the observed in the 6.8 μs cMD1 trajectory (12.5Å), while the average distances generated by five dual-boost aMD are within 7.5 – 9.3 Å. The latter average distances are within the reference distance (8.9 Å) observed in the X-ray structure at high pH and the 140 ns conventional MD trajectory (9.0 Å). Inspection of generated structures by aMD_EE shows a partial unfolding of helix TH1 coupled to a rotation relative to the initial crystal structure (see Figure 2.8C).

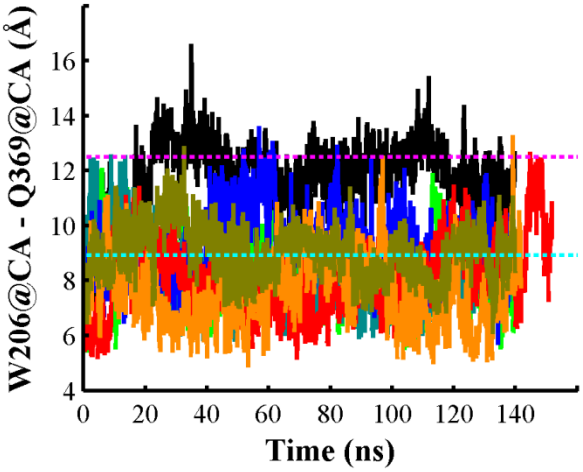

Figure 2.9. Distance between C_α atoms of residues W206 and Q369 as a function of time for different MD methods. Distance values obtained from aMD_EE are shown by a black line. Distance traces obtained from five dual-boost aMD independent trajectory are shown in blue, gree, red and orange lines. A conventional MD simulation is shown in olive line. Representative dual-boost aMD trajectory is shown by a blue line. Magenta broken line represents the average distance (12.5 Å) obtained from the MD simulation cMD2. Cyan broken line represents the distance (8.9 Å) obtained from the crystal structure.

2.4.5 Comparison of Correlated Backbone Displacement

To assess similarities between the ensembles generated by aMD_EE and different set of MD simulations, we calculated the covariance matrix. A covariance matrix provides information of the variance and the correlations among the position coordinates of C_α atoms of the protein backbone. We calculated the covariance matrices for aMD_EE, all dual-boost aMD, cMD1, and cMD2, as shown in Figure 2.10. For each set of MD frames, the covariance matrix was calculated using the coordinates of Cα atoms

(residues 206-375) after translation and rotation relative to their respective average structure. Figure 2.10A shows that aMD_EE similar correlated displacements observed in dual-boost aMD simulations (see Figure 2.10B), but helix TH2 shows a different behavior not observed in the set of dual-boost aMD simulations. This is related to the unfolding of helix TH2 sampled by aMD_EE and the less frequent TH2 unfolding generated by dual-boost aMD simulations. Anton generated MD trajectory aMD1 shows similar degree of correlated displacements observed in aMD_EE; however, the loop between TH7-8 shows coordinated displacement with helices TH1 and TH4 (see Figure 2.10C). Also, aMD1 is characterized by coordinated displacement among atoms in helix TH1, which is due to the formation of a kink in this helix. Anton MD trajectory aMD2 shows similar behavior observed in aMD_EE, except by the correlated movement of helix TH4 with helix TH2 and the loop between TH7-TH8 (see Figure 2.10D). aMD2 also shows correlated displacement among coordinates of helix TH1. aMD2 shows that helix TH2 is correlated with the rest of protein and has similar magnitude of variance observed in aMD_EE. This variation of helix TH2 backbone is not observed in the covariance matrix of dual-boost aMD simulations. In general, aMD_EE generates an ensemble of protein structures with similar behavior observed in microsecond-long conventional MD simulations. Dual-boost aMD simulations show correlated displacement among interhelical loops located in the C-terminal, but no significant correlation among the N-terminal region of the protein. The latter is a common feature observed in aMD_EE and microsecond long conventional MD simulations. To generate an accurate free energy landscape of T-domain, two options are available: run a longer

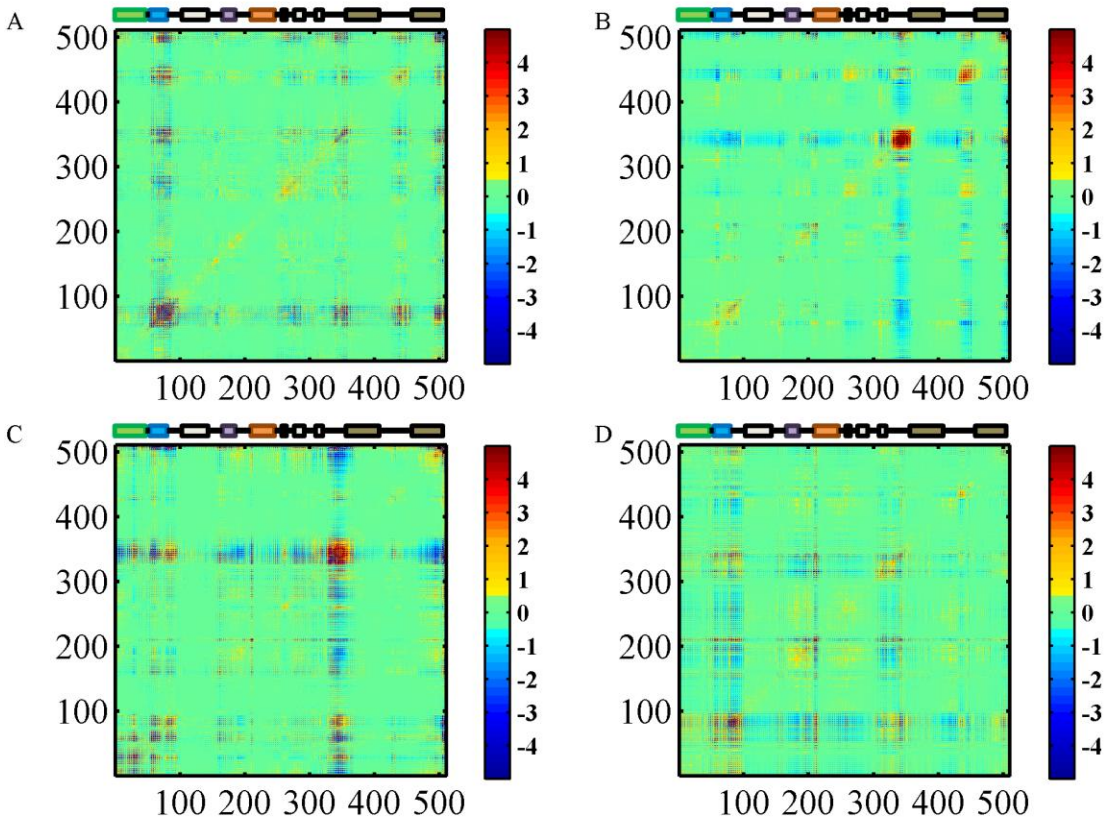

Figure 2.10. Covariance matrices of T-domain obtained from different MD methods. (A) 140ns aMD_EE, (B) accumulated 716 ns dual-boost a_MD trajectories, (C) 6.8 μs cMD1, and (D) 9.5 μs cMD2. For each figure, MD trajectories were translated and rotated relative to the average structure using C_α atoms from residues 206-375. Secondary structure elements are represented by boxes on top of each graph and helices are colored according to Figure 2.4.

aMD_EE simulation or to perform an ensemble of short aMD_EE simulations. The latter can be used in conjunction with a replica exchange method, which is a similar approach previously used for accelerating convergence of free energy calculations using dual-boost aMD applications.[57]

2.5 Conclusions

Sampling of conformational changes associated to changes of protonation is a challenging problem for conventional MD simulations methods. Accelerated molecular dynamics method is a promising approach for sampling of large conformational changes and their associated free energy surfaces of proteins with a few hundred of aminocids. In this work, we implemented a modified aMD method in which the direct-space electrostatic interactions of only solute-solute atom pairs are biased (aMD_EE). Our results suggest that boosting of the electrostatic interactions of solute-solute atoms accelerates the sampling of alanine-dipeptide and a low pH model of T-domain. For alanine-dipeptide, reweighting of each MD frame generated by aMD_EE results in a better estimation of the free energy landscape than that obtained using dual-boost aMD. Furthermore, aMD_EE generated structures of a relatively large protein (T-domain) in good aggrement with microsecond long MD trajectories and reported experimental observations. This approach of boosting electrostatic interactions of solute-solute atom pairs (aMD_EE) has the advantage of generating MD frames with a relatively small weight factor compared to dual-boost aMD applications. Thus, aMD_EE is a promising accelerating method for sampling conformational changes of relatively large proteins. A possible application of aMD_EE is the study of intrinsically disordered peptides and proteins in which aMD_EE could be used to generate conformational ensembles and to calculate a free energy landscape.

2.6 Appendix

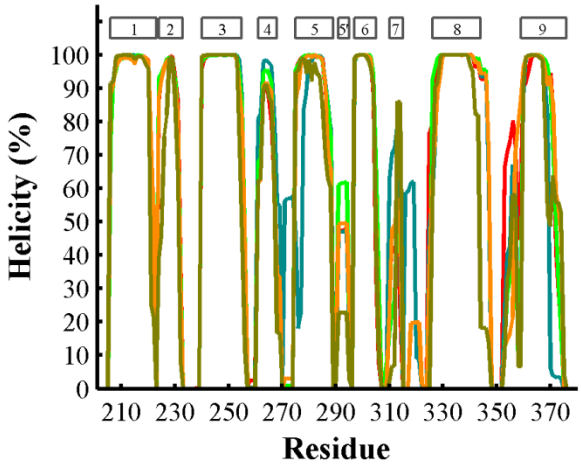

Figure 2.S1. Average helicity content per residue shown as a percentage. Four independent trajectories generated by dual-boost aMD are represented by orange, dark cyan, green and red lines. Conventional MD is represented by an olive line. Helices TH1-9, TH5' are represented by rectangles on the top of the figure. Simulation lengths of all trajectories vary within 140 – 150 ns.

Chapter 3. T-domain Associates to Anionic Lipid Bilayers with Two Predicted Membrane Binding Modes

3.1 Introduction

Diphtheria toxin is a bacterial protein that penetrates eukaryotic cells, where it disrupts protein synthesis and results in cell death. The process of cell entry involves the internalization of the toxin bound to a cell-surface receptor followed by acidification of the endosome interior.[2] The full toxin consists of three domains, each associated with a specific function.[58] Diphtheria toxin translocation (T) domain inserts into the membrane upon decrease of pH in the endosome interior. T-domain aids the membrane insertion and translocation of a toxin catalytic domain across the endosome membrane into the cytosol. In the cytosol, catalytic domain facilitates cell death.[59] Thus, diphtheria toxin relies on a series of structural rearrangements of T-domain that enable delivery of a catalytic fragment through cell membranes.

A stand-alone T-domain in solution adopts a monomeric globular form at neutral pH, comprised of ten α-helices.[60] In the presence of anionic bilayers and acidic solution the protein undergoes major structural rearrangements forming a membrane inserted state, in which its hydrophobic helices span the bilayer.[10] Detailed understanding of T-domain membrane insertion has been precluded by the lack of high resolution structural studies, e.g. X-ray crystallography. Such approaches have been problematic due to the protein tendency to aggregate, and due to the existence of multiple conformations of destabilized protein in low pH solution, as well as in the membrane-associated states.[8, 10, 12, 61-63] Understanding of the membrane association process of T-domain will facilitate the

initial steps towards a complete characterization of its folding in membranes and its translocation function.

Kinetic analysis of T-domain membrane insertion showed that the protein forms a membrane-competent state in low pH solution followed by its membrane association and formation of a insertion competent intermediate.[10] Furthermore, the membrane-competent state of the T-domain and its insertion competent states exist at overlapping ranges of low pH region, which indicates that staggered protonation of several amino acid side-chains may accompany protein refolding at the membrane interface.[10, 64] It was also found that as fraction of anionic lipids was increased, there was an enhanced binding and insertion of T-domain into bilayers.[10]

Protonation of some T-domain histidine side-chains has been recognized to play an important role in various stages of the membrane insertion process, e.g. his 257 and 223 were implicated to act as a molecular switch of the protein partial unfolding in low pH solution.[3, 9-11, 20, 47] It was also suggested that histidine protonation plays a role in the membrane binding,[9] as well as in the final stages of the membrane insertion of the isolated T-domain.[64, 65] Recently, we have performed atomistic MD simulations of destabilized low pH T-domain structures in solution with features of a membrane-competent state.[20, 47] Our extensive atomistic simulations in conjunction with spectroscopic experiments[20] have directly demonstrated, for the first time, the destabilizing role of N-terminal histidines in the partial unfolding of N-terminal helices in solution, and the solvent exposure of hydrophobic sites, while the protein retained its compact structure. These features were interpreted as initial stages of the formation of a membrane-competent state of the T-domain in solution.[20, 47] Recent X-ray structures of

unfolded diphtheria toxin,[66] have indicated possibility of refolding of N-terminal helices as predicted by Kurnikov et al.[20]

The question of how the membrane-competent form of T-domain interacts with the lipid bilayer has not been addressed yet. It has been studied in kinetic experiments that provide a low-resolution information on dependency of the protein-lipid association on the solution pH and the membrane lipid composition.[10] This work aims to model how composition of the lipid bilayer as well as the structural and protonation states of the T-domain affect its association with the membrane. To achieve this goal, we performed MD simulations using the coarse-grained representation for the protein, lipid and water. The coarse-grained protein structure in this study is based on our previously generated atomistic models of the T-domain membrane-competent state.[20]

Coarse-grained molecular dynamics (CG-MD) simulations and potential of mean force (PMF) calculations are used for this purpose. CG-MD simulations allow for efficient computations of membrane association of proteins on microsecond time-scales using a reduced representation of the molecules.[67, 68] CG-MD simulations have successfully demonstrated the role of both electrostatic interactions between key residues and anionic lipids, and as well as hydrophobic contacts in the membrane association of proteins.[69, 70]

3.2 Methods

3.2.1 Coarse-grained Models and Standard CG-MD Simulations

Two coarse-grained models of T-domain were constructed using atomistic structures at neutral and low pH.[20] The neutral pH T-domain model was built from a high resolution structure of diphtheria toxin at pH 7.5 (PDB ID code: 1F0L) and histidines

were set to a neutral state (see Figure 3.1A).[5] The low pH T-domain model was constructed from the last MD generated structure of a 6.8 μs long all-atom MD simulation of a refolded state of T-domain, which was triggered by protonation of histidine side-chains (see Figure 3.1B).[20] All of the six histidines were also set to protonated state in the coarse-grained model of the low pH T-domain structure. MARTINI v2.1 force field was used to model both the protein and the lipid in this work.[71] In order to maintain the tertiary structure of the CG protein models, an elastic network model ElneDyn was applied with a force constant 10 kJ mol^{-1} $Å^{-2}$ and a cutoff radius of 9 Å.[72]

To study T-domain membrane association with bilayers, our protein model was placed 100 Å away from an equilibrated lipid bilayer and solvated by approximately 11000 CG water beads. Preformed bilayers were composed of 256 phospholipids with different mixtures of anionic and neutral phospholipids and the total charge of the system was neutralized with counterions. We performed five independent MD runs for each T-domain folded state and lipid bilayers composed of 1-palmitoyl-2-oleoyl phosphatidyl-choline (POPC) and 1-palmitoyl-2-oleoyl phosphatidyl-glycerol (POPG) with the following ratios: POPC, POPC/POPG 3:1, and POPC/POPG 1:3. In total, 30 × 2 μs long CG-MD simulations were performed. In the following discussion, such equilibrium MD simulations performed using a typical protocol and in the absence of any biases will be referred to as standard MD.

3.2.2 Potential of Mean Force Calculations

Umbrella sampling calculations were performed to obtain biased distributions along a reaction coordinate z in which the protein position is defined by the center of

mass (COM) distance of the protein and the lipid bilayer. The potential of mean force $w(z)$ along the reaction coordinate z is defined:

$$w(z) - w(z^*) = -k_B T \ln \langle \rho(z) \rangle / \langle \rho(z^*) \rangle, \quad (3.1)$$

$$w(z) - w(z^*) = -k_B T \ln \frac{\int dR \delta(z'(R) - z) e^{-U(R)/k_B T}}{\int dR \delta(z'(R) - z^*) e^{-U(R)/k_B T}}, \quad (3.2)$$

where $w(z^*)$ and z^* are defined as arbitrary, $\rho(z)$ is the distribution function along z, R are the coordinates of the system, $z'(R)$ is a function that describes z as a function of R, $U(R)$ is the potential energy as a function of R, k_B is the Boltzmann's constant, T is the temperature.[73]

Three different initial orientations of the membrane-competent state model of T-domain were set every 1 Å within the range of 32 Å to 58 Å along the reaction coordinate. Thus, a total of 81 restrained simulations were performed. To insert the protein in the preformed bilayer, we used the perl script InflateGro.[74] The protein and a pre-equilibrated bilayer containing 512 phospholipids were solvated by approximately 19800 CG water beads. The protein tertiary structure was restrained by an elastic network as indicated in subsection **3.2.1** of ***Methods***. Afterwards, neutralizing counterions were added. Each MD trajectory was equilibrated over 50 ps with time step of 1 fs with positional restraints applied to the entire protein, followed by 400 ns with time step of 20 ps with positional restraints on the protein backbone atoms. All positional restraints were removed for production simulations of length 800 ns and a time step of 20 fs. Total simulation time for PMF calculations accounts for a total of 64.8 μs. A harmonic restraint was applied to the center of mass distance between the protein and the bilayer with a force constant of K = 10 kJ mol^{-1} $Å^{-2}$. The COM of the bilayer is calculated by using those lipids inside a cylinder of 38 Å. The weighted contribution of lipids was switched

off between 33 Å and 38 Å. This cylinder was centered at the protein and was aligned in a parallel orientation relative to the membrane normal axis. MD frames were saved every 100 ps and position restraints data was saved every 1 ps. We constructed the PMF curves by combining and unbiasing the restrained CG-MD simulations using the weighted histogram analysis method (WHAM),[75] as implemented in g_wham.[76] Equilibration of the protein orientation was assessed by the equilibration of angles formed by helices TH8 and TH9 relative to the membrane normal. These angles equilibrated over the last 600 ns of each restrained CG-MD trajectory, and data was extracted from these segments. The standard deviations were calculated using the Bayesian bootstrap of complete histograms. The free energy of binding $\Delta G_{binding}$ was calculated from the PMF curve as:

$$e^{-\Delta G_{binding}/k_BT} = \frac{\int_{bound} dz e^{-w(z)/k_BT}}{\int_{unbound} dz e^{-w(z)/k_BT}}, \qquad (3.3)$$

where the integral over the bound macrostate goes from the minimum value of z to approximately z = 51 Å. For the unbound macrostate, we assume $w(z) = 0$ and it is integrated over the same volume as that of the bound state.[77, 78] Integrals are approximated using the trapezoidal rule and the error is estimated from standard deviations along the PMF curve.

3.2.3 Simulation Protocols and Analysis

All CG-MD simulations were performed using GROMACS 4.5.3.[79] A semi-isotropic NPT ensemble was used for all simulations with Berendsen thermostat and barostat set to a reference of T = 310 K and P = 1 bar, respectively. The following constants were set to τ_T = 1 ps, τ_T = 1 ps and a compressibility value of 3.0×10^{-5} bar^{-1}. The Coulomb interactions were shifted to zero between 0 and 12 Å and the Lennard-Jones interactions between 9 to 12 Å. The neighbor list was updated every 5 step with a

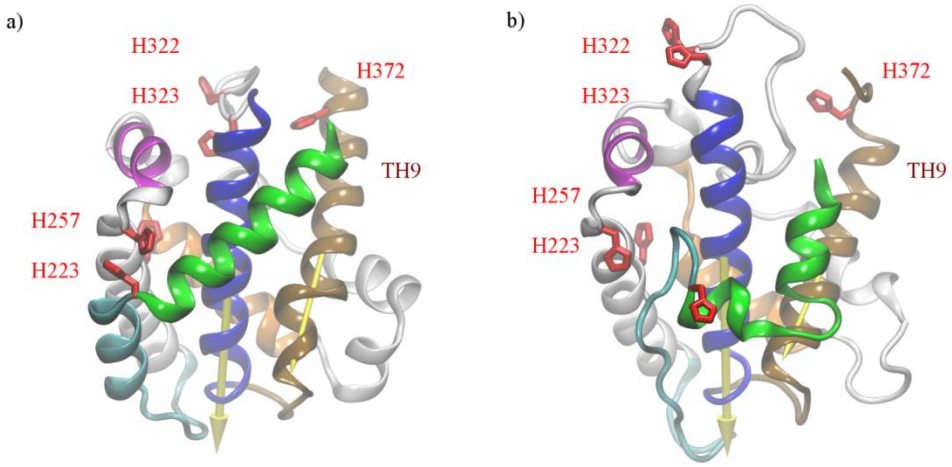

Figure 3.1. Neutral pH folded structure and a low pH unfolded structure of T-domain. a) The neutral pH structure is obtained from the diphtheria multi-domain structure obtained at neutral pH conditions [PDB 1F0L].[5] All side-chains are set to their standard protonation state at pH 7.5. b) The low pH partially unfolded T-domain structure is generated by fixing histidines in their protonated state while the other residues are set to their standard protonation state. Explicit solvent molecular dynamics simulation were performed over *ca.* 6.8 μs on ANTON supercomputer.[20] Helices TH1, TH2, TH4, and TH5 are represented by green, cyan, magenta and orange ribbon representation, respectively. Helices TH5', TH6, TH7 are represented by grey ribbon representation. Helices TH8 and TH9 are represented by blue and brown ribbon representation. An arrow is drawn in each of the latter helices. The angle formed by these arrows and the membrane normal axis is calculated in all CG-MD simulations (the normal axis direction is towards the membrane core). All histidines are highlighted by red licorice representation.

cutoff radius of 14 Å. The dielectric constant was set to 15. Each protein-lipid system was first equilibrated for 50 ps with a time step of 1 fs and with the protein restrained in its position in space. This was followed by a 5 ns equilibration simulation with a time step of 20 fs and the protein still restricted in space. The protein was allowed to move freely in the production simulations that followed equilibration. A production trajectory was saved every 100 ps. Trajectory analysis was performed using GROMACS analysis tools and in-house scripts; visualization was performed using VMD 1.9.1.[55]

3.3 Results

3.3.1 Simulations of the T-Domain Approaching the Bilayers

To investigate the membrane association of T-domain and the effects of anionic lipids, we modeled the interactions of the neutral pH and the low pH structures of T-domain with pre-formed phospholipid bilayers composed of mixtures of POPC and POPG (see detailed description of protein structures in Figure 3.1). The top and bottom panels in Figure 3.2 display the COM distance between the neutral and low pH protein structures, and bilayers of different compositions as a function of time, respectively. As can be seen, larger fractions of anionic lipids (POPC/POPG 1:3) promote steady association of either neutral or low pH T-domain structures to bilayer interfaces in ten CG-MD trajectories (see Figure 3.2C, F). In these trajectories, the protein binds to the bilayer within 0.3 μs to 1 μs. Following the initial protein association, the protein remains attached to the bilayer interface over the rest of the simulation length. Initial protein-membrane contacts are formed at residues in the N-terminal helices. For example, the low pH T-domain structure initiates protein-membrane contacts with residues located in

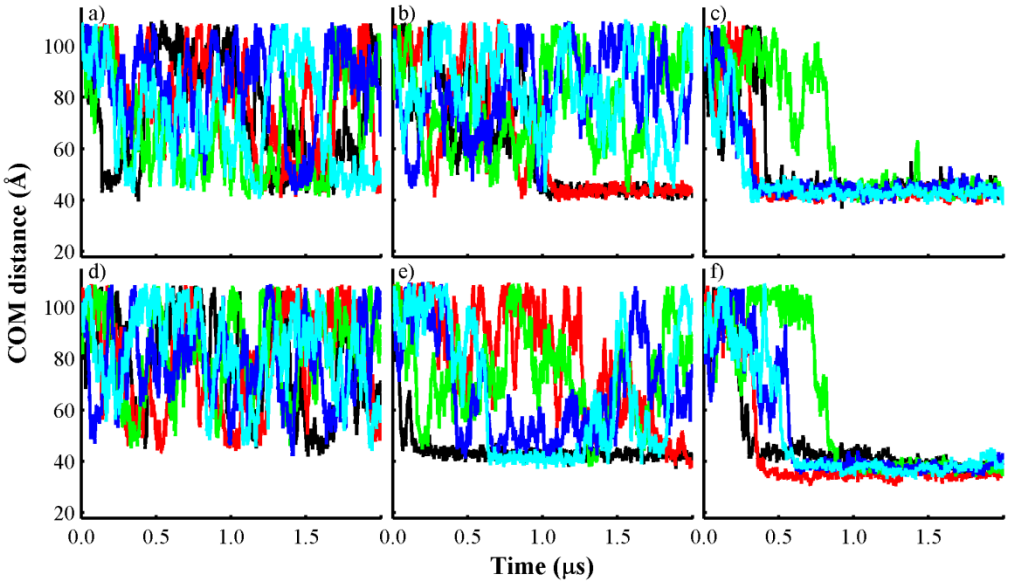

Figure 3.2. Spontaneous membrane binding/unbinding events of T-domain. Distance (COM) between the neutral pH structure and bilayers of the following composition: a) POPC. b) POPC/POPG 3:1. c) POPC/POPG 1:3. Distance of the low pH unfolded T-domain structure and bilayers of the following composition: d) POPC. e) POPC/POPG 3:1. f) POPC/POPG 1:3. Each line color represents an independent CG-MD production simulation, which is initiated with the same protein orientation and different seed number. Accumulated CG-MD production simulation time is 60 μs.

the loop between TH2 and TH3, the loop between TH3 and TH4, or in the protein terminals. Also, the neutral pH structure forms the initial protein-membrane contacts with the residues located in the loop between TH2 and TH3, or with the residues of the helix TH2 and the loop TH3-TH4.

Simulations of T-domain in the presence of bilayers containing a smaller ratio of anionic phospholipids POPC/POPG 3:1 exhibit a somewhat different behavior. Figure

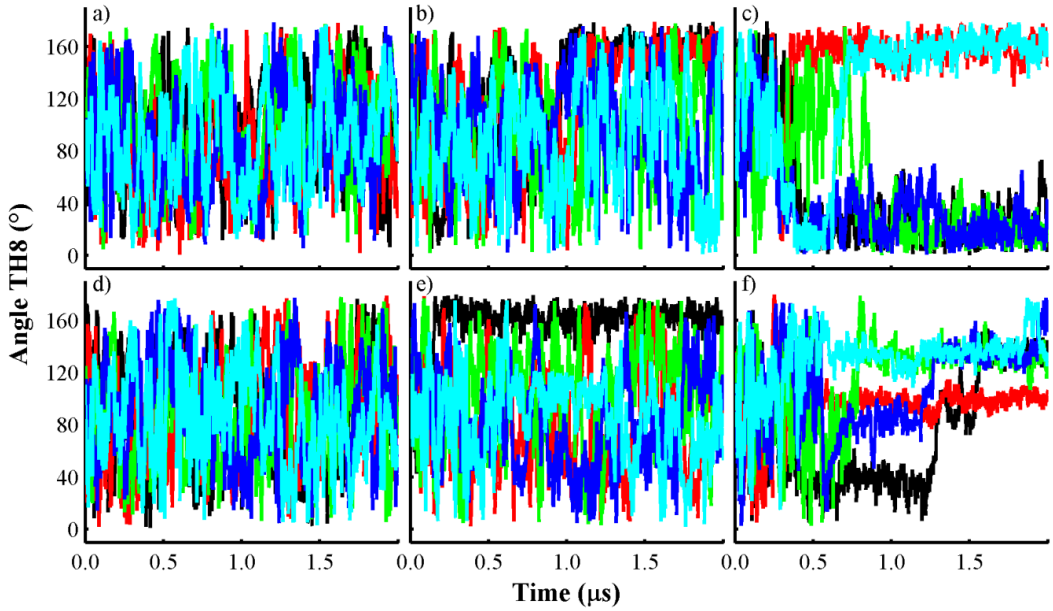

Figure 3.3. Angle formed by the axis of helix TH8. The angle is formed between (C_α atoms of residues 323-343) and the normal axis to the membrane plane versus simulation time (the normal axis direction is towards the membrane core). For example, helix TH8 is parallel to the membrane plane if the angle value is close to 90 °. Angles obtained from CG-MD simulations of the neutral pH T-domain structure and lipid bilayers of the following composition: a) POPC. b) POPC/POPG 3:1. c) POPC/POPG 1:3. Data obtained from CG-MD simulations of the low pH unfolded T-domain structure and lipid bilayers of the following composition: d) POPC. e) POPC/POPG 3:1. f) POPC/POPG 1:3. Each line color represents an independent CG-MD production simulation, which is initiated with the same protein orientation and different seed number. Accumulated CG-MD production simulation time is 60 μs.

3.2E shows that the low pH T-domain structure rapidly associates and forms a stable membrane-associated state in a single trajectory. The neutral pH structure binds to the membrane interface over the last 1 μs of two trajectories (see Figure 3.2B). Most of the

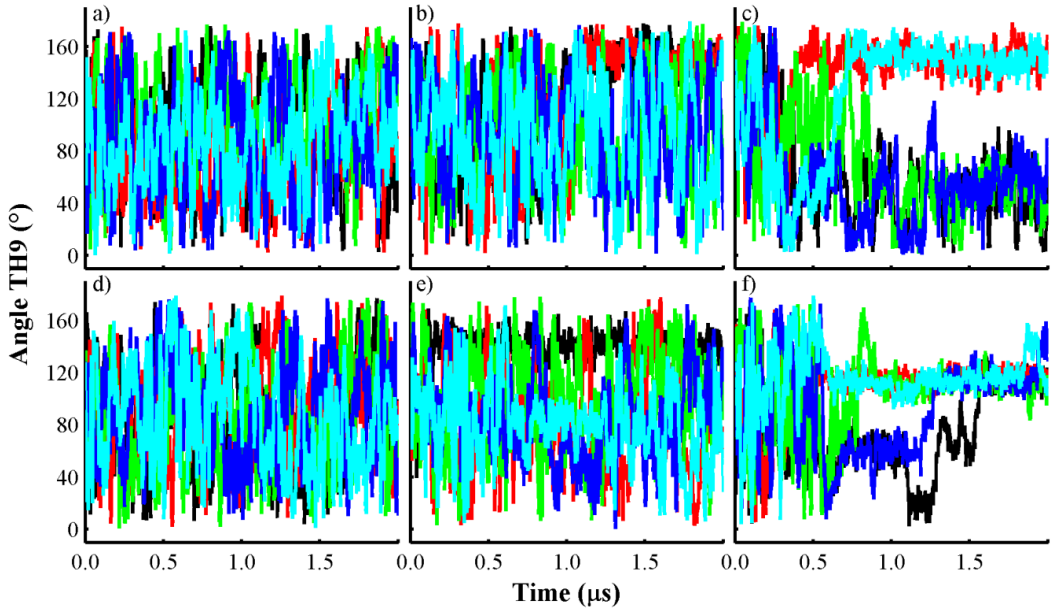

Figure 3.4. Angle formed by the axis of helix TH9. The angle is formed between (C_α atoms of residues 359-375) and the normal axis to the membrane plane versus simulation time (the normal axis direction is towards the membrane core). For example, helix TH9 is parallel to the membrane plane if the angle value is close to 90 °. Angles obtained from CG-MD simulations of the neutral pH T-domain structure and lipid bilayers of the following composition: a) POPC. b) POPC/POPG 3:1. c) POPC/POPG 1:3. Data obtained from CG-MD simulations of the low pH unfolded T-domain structure and lipid bilayers of the following composition: d) POPC. e) POPC/POPG 3:1. f) POPC/POPG 1:3. Each line color represents an independent CG-MD production simulation, which is initiated with the same protein orientation and different seed number. Accumulated CG-MD production simulation time is 60 μs.

initial protein-membrane contacts are formed in the N-terminal helices of the protein. For instance, the low pH structure has initial protein-membrane contacts in the loop TH2-

TH3. The neutral pH structure initiates lipid contacts with the residues of the helix TH2 or with the residues 295-296. In the presence of pure POPC bilayers (see Figure 3.2A, D) both the neutral pH and the low pH T-domain structures bind only transiently to the membrane interface. Therefore, addition of negatively charged phospholipids to bilayers promotes rapid and stable membrane association of T-domain. Bilayers composed mainly of anionic lipids (POPC/POPG 1:3) promote deeper insertion of the low pH T-domain structure than that observed for the neutral pH T-domain structure. Decreasing the anionic content in the bilayers POPC/POPG 3:1 promotes shallow inserted states of the neutral and low pH T-domain structures.

To monitor the protein orientation dynamics, we found it useful to calculate the angles between the axes of helices TH8 and TH9, and the normal axis to the membrane plane as a function of time (see Figures 3.3, 3.4, and Figure 3.1 for structure notation). Overall, the protein orientation relative to the membrane plane is stabilized as the fraction of anionic lipids is increased. Figure 3.3F shows that the low pH T-domain adopts two distinct membrane-associated conformations with bilayers containing POPC/POPG 1:3. These two orientations have helices TH8 and TH9 in a near parallel or at an oblique angle to the membrane plane (see Figures 3.3F and 3.4F). Helix TH8 shows similar angles (120°) in four of the five trajectories and is stabilized within 0.6 μs to 1.5 μs. The neutral pH T-domain structure adopts two membrane-bound conformations in which helices TH8 and TH9 are near perpendicular to the membrane plane upon binding to the mostly anionic bilayer (POPC/POPG 1:3) (see Figures 3.3C and 3.4C). One of these orientations is observed in three of the five independent trajectories. In simulations with a small ratio

of anionic lipids (POPC/POPG 3:1) or no anionic lipids (POPC) the protein orientation with respect to bilayer was highly dynamic (see Figures 3.3A, D and 3.4A, D).

Overall, following the initial association of T-domain with the bilayer helices TH8 and TH9 may require up to 1 μs to stabilize, see *e.g.* a trajectory shown in black lines in Figure 3.2F, (and Figures 3.3F, 3.4F). Note that the time-scales in the coarse-grained simulations are faster than those observed in atomistic simulations because, in part, of the lower viscosity of the simulated media.[80] A typical factor of 4 is used to scale time in such simulations; this factor corresponds to the ratio of the diffusion constant of the coarse-grained water relative to the real water. Using this scaling factor we estimate that the effective time-scale of the protein orientational dynamics in the membrane interface was within the 2.4 μs - 6 μs range. Overall, these simulations indicate microsecond time-scales for the T-domain membrane association and reorientation in the membrane interface.

3.3.2 T-domain Equilibrated at the Bilayer Interface

T-domain binding to the bilayer interface is analyzed using equilibrium fragments (the last 500 ns of each trajectory) from all simulations of the bilayers containing anionic phospholipids (see Figure 3.S1). To investigate the similarities and differences of membrane association of the neutral and low pH T-domain structures, we calculated

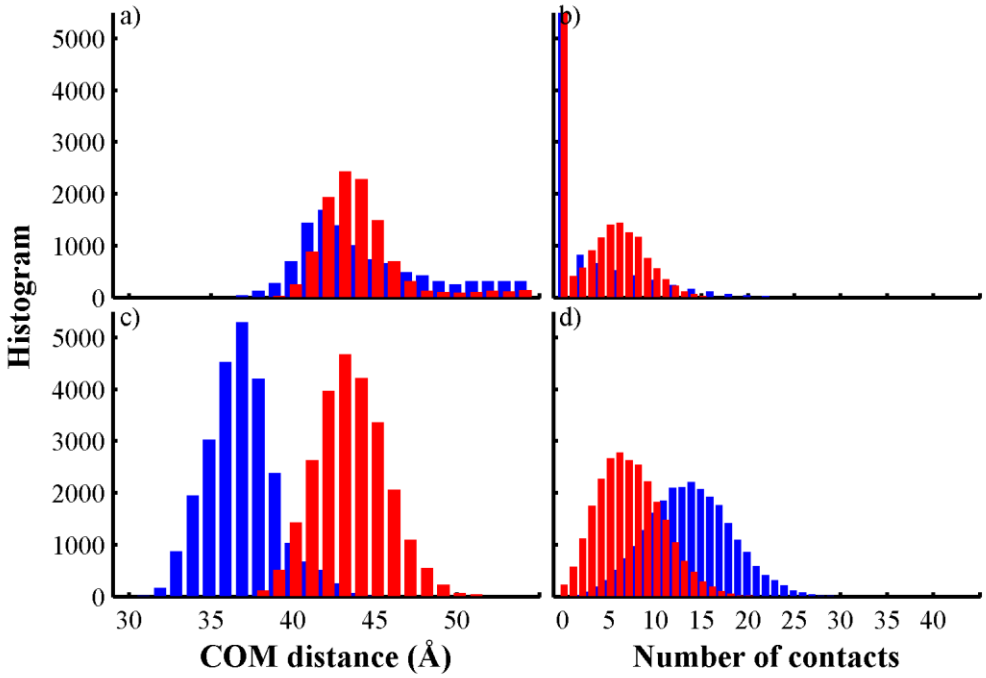

Figure 3.5. Histograms of COM distances and number of contacts between T-domain and bilayers. Data obtained from simulations of neutral pH and low pH T-domain structures are represented by red and blue bars, respectively. a) Histogram of the COM distance between T-domain and bilayers containing POPC/POPG 3:1. b) Histogram of contacts between T-domain and bilayer interface containing POPC/POPG 3:1. c) Histogram of the COM distance between T-domain bound to POPC/POPG 1:3. d) Histogram of contacts between T-domain and the membrane interface containing POPC/POPG 1:3. Data is obtained from the last 500 ns segments of all standard CG-MD simulations of anionic membranes. The number of contacts is calculated between the protein and membrane interface beads

histograms of the COM distances between the protein and bilayers and histograms of the number of contacts between residues and head-group beads from data obtained of the abovementioned equilibrium ensembles, (see Figure 3.5). Figure 3.5C shows the

histogram of COM distances of between protein and bilayers composed of mainly anionic lipids (POPC/POPG 1:3). This highly anionic bilayer promotes deeper insertion of the low pH structure (major peak at 37 Å) than that of the neutral pH structure (major peak at 43 Å). Furthermore, this observation is supported by the significant number of contacts between the low pH structure and the membrane interface than that of the neutral pH structure, see histograms of contacts in Figure 3.5D. Similar analysis is performed on simulations of T-domain and bilayers containing a smaller fraction of anionic lipid POPC/POPG 3:1. Figure 3.5A shows that the neutral and low pH conformations of the T-domain are inserted in similar shallow positions with major peaks at 43 Å and 42 Å, respectively. This shallow inserted state has a smaller number of protein-membrane contacts than that of the low pH structure, as shown in Figure 3.5B. Furthermore, both T-domain structures have a predominant peak at zero protein-membrane interface contacts, which shows the decreased affinity of the protein to associate with membranes of lower fraction of negatively charged lipids (see Figure 3.5B). The following subsections provide detailed description of CG-MD simulations of T-domain and lipid bilayers of higher anionic content (POPC/POPG 1:3).

3.3.3 Protein Orientations at the Bilayers with the Higher Anionic Content

To analyze orientations of representative membrane bound conformations of T-domain, we computed the normalized density profiles of coarse-grained atoms as a function of their location in the membrane normal axis. Figure 3.6 shows density profiles of cationic, hydrophobic and histidine residues and bilayer components for two distinct protein orientations of the low pH T-domain structure and bilayers of higher anionic content (POPC/POPG 1:3). Density profiles were calculated over the equilibrated

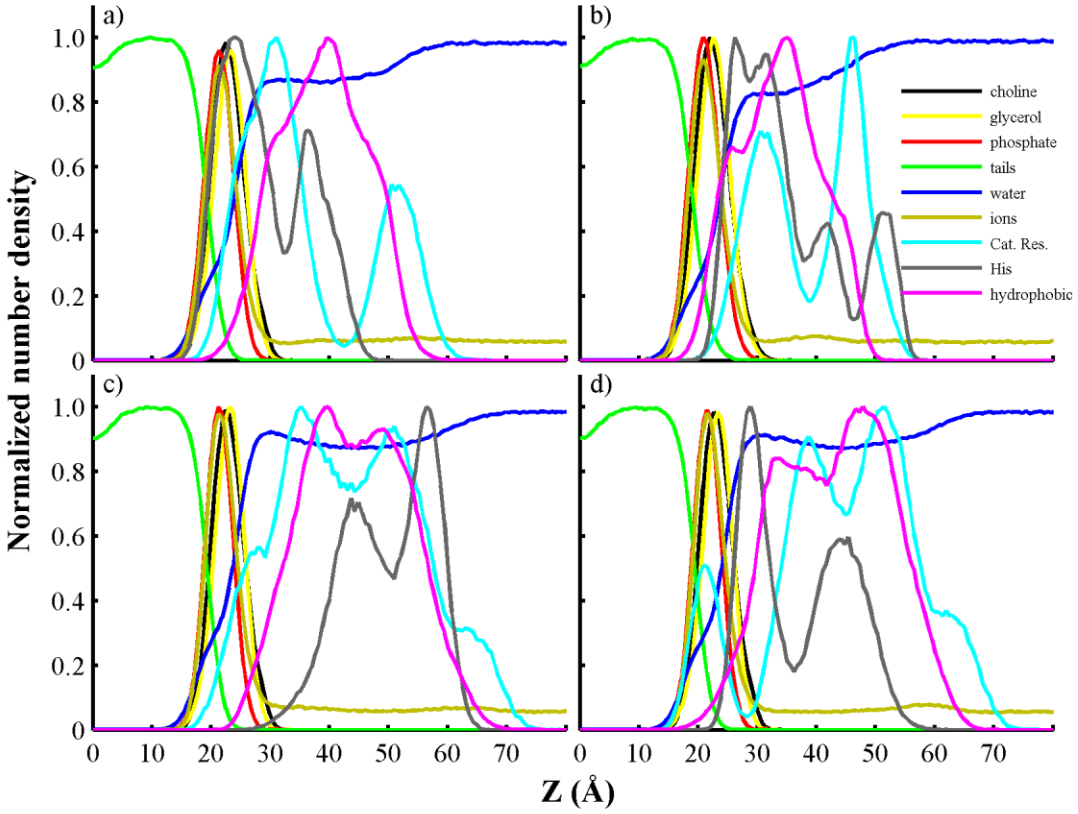

Figure 3.6. Normalized number densities of chemical groups as a function of their location in the normal membrane axis. Data is obtained from the equilibrated segments (last 500 ns) of representative standard CG-MD simulations of T-domain structures with lipid bilayers containing POPC/POPG 1:3. a) Membrane bound conformation B1 of low pH T-domain structure. b) Membrane bound conformation B2 of low pH T-domain structure. c) Neutral pH T-domain structure associated to a bilayer in three simulations of a set of five independent standard CG-MD simulations. d) Neutral pH T-domain structure attached to a lipid bilayer in two trajectories of a set of five independent standard CG-MD simulations. Density profiles of lysine residues are similar to those of cationic residues in all plots, except the last one. In this plot, the density peak of cationic residues located between the phosphate regions corresponds to the N-terminal and arginine residues. Density profiles of headgroup atoms of POPC (choline) and POPG (glycerol)

are shown in black and yellow lines, respectively. Density profiles of phosphate and tails groups from all lipids are shown in red and green lines, respectively. Density profiles of water and sodium counterions (ions) are shown in blue and dark yellow, respectively. Profiles of cationic residues (Lys, Arg, and N-terminal), histidines and hydrophobic residues are shown in cyan, grey and magenta lines, respectively.

segments of representative trajectories (see protein orientation stabilization over the last 500 ns, shown in Figures 3.3F and 3.4F). These orientations are further referred to as conformations B1 and B2. The former one is observed in four of five independent trajectories. Conformation B1 shows that cationic residues reside between the lipid head-groups and water regions and populate a pronounced peak close to the membrane interface at 30 Å in the normal membrane axis (see Figure 3.6A). Furthermore, protonated histidines have a predominant density peak between the lipid head-group and phosphate regions. Hydrophobic residues show less penetration in the membrane interface, compared to their distribution in conformation B2 (see Figure 3.6B). Moreover, conformation B2 exhibits a decreased penetration of cationic and histidine residues and an increased insertion of hydrophobic residues in the lipid head-group and tail regions in contrast to conformation B1 (see Figure 3.6B). It also shows a shoulder in the density profile of hydrophobic residues located between the lipid head-group and water regions.

In addition, five independent trajectories of the low pH T-domain structure with histidines in their neutral state and bilayers composed of POPC/POPG 1:3 results in the protein binding to the bilayer interface (data not shown). Strikingly, the protein adopts a

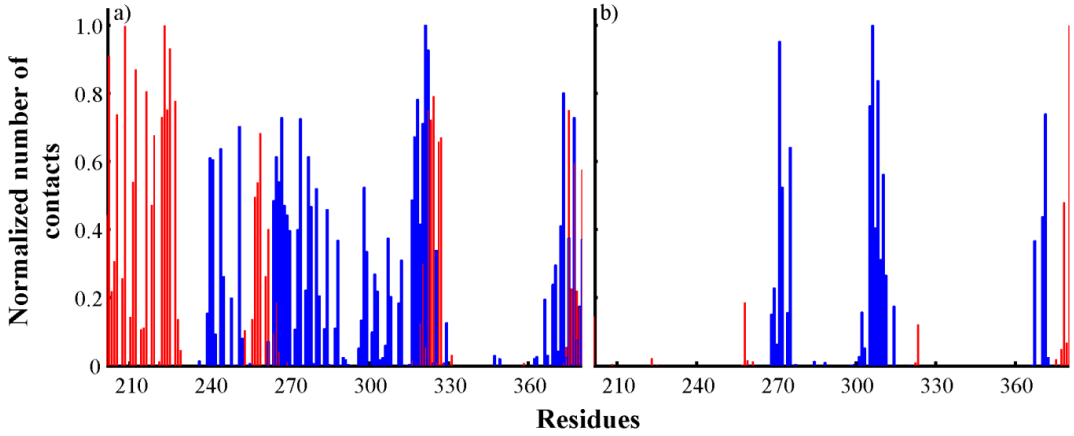

Figure 3.7. Normalized number of protein-membrane interface contacts as a function of residue number. Data is obtained from the last 200 ns of two independent standard CG-MD simulations of the low pH T-domain structure and a preformed membrane containing POPC/POPG 1:3. Blue bars correspond to membrane bound state B1 and red bars correspond to membrane bound state B2. a) Protein and membrane interface contacts. b) Protein and lipid tails contacts. All protein and lipid atoms separated by a distance lower than 5 Å are considered to be in contact.

stable conformation B2 in a single trajectory. The other four trajectories show less deep insertion of the protein. Two of these trajectories show exchange of conformations between B1 and similar orientations observed in the simulations of the neutral pH T-domain model and bilayers (POPC/POPG 1:3).

CG-MD simulations of the neutral pH T-domain structure and bilayers of mixture POPC/POPG 1:3 display two different membrane bound orientations, *e.g.* see Figures 3.3C and 3.4C. The most frequently observed orientation displays penetration of cationic residues in the head-group region, as shown in Figure 3.6C. The same figure shows that histidines and hydrophobic residues have no significant penetration in the membrane

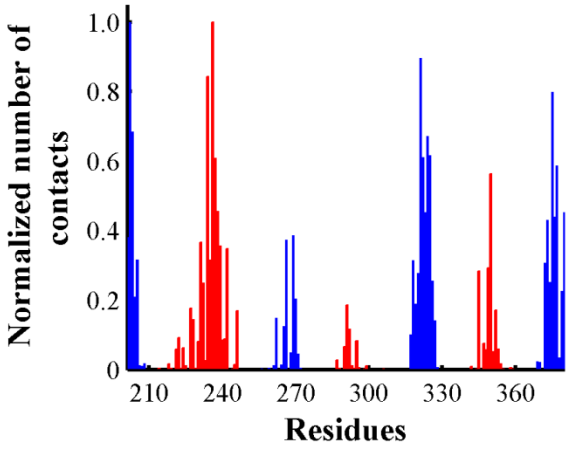

Figure 3.8. Normalized number of protein-membrane interface contacts as a function of residue number. Data is obtained from the last 200 ns of two independent standard CG-MD simulations of the neutral pH T-domain structure and a preformed membrane composed of POPC/POPG 1:3. Blue and red bars correspond to two representative trajectories, shown in black and red lines in Fig. S1 (c), S2 (c). All protein and head-group, phosphate beads separated by a distance lower than 5 Å are considered to be in contact.

interface region. In contrast Figure 3.6D shows that the less frequently observed orientation has hydrophobic and histidine residues residing in the head-group region.

The low pH T-domain conformation exhibits more membrane interacting residues than those of the neutral pH structure bound to the bilayers with high anionic content (POPC/POPG 1:3), (see Figures 3.7 and 3.8). Conformations B1 and B2 have two different patterns of the protein-membrane contacts. Conformation B1 has significant number of contacts with the residues located in the partially unfolded helices TH1 and TH2, the loop TH3-4, the N-terminus of TH8, and the C-terminus of TH9, as shown in Figure 3.7A. Conformation B2 has protein-membrane contacts located in the helices

TH3-5, TH6-7, N-terminus of TH8 and C-terminus of TH9 (Figure 3.7A). Furthermore, B2 has a significant number of the protein - aliphatic tail contacts with the residues P271, I306, P308 and V371, while these contacts are not present in the orientation B1, as shown in Figure 3.7B.

The neutral pH T-domain structure displays two different membrane bound orientations with respect to bilayers with the high anionic lipid content. Histograms of the protein-membrane interface contacts show that the most frequently observed orientation has the majority of contacts localized in the loop TH8-9 and helix TH3 (see Figure 3.8). The second membrane bound orientation shows contacts at the TH1 N-terminus, TH8 N-terminus and TH9 C-terminus, also shown in Figure 3.8.

3.3.5 Protein Orientations at the Bilayers with the Lower Anionic Content

Decrease of anionic content (POPC/POPG 3:1) decreases the propensity of the membrane binding for both the neutral and low pH T-domain structures. Both neutral and low pH structures have the majority of the membrane interactions with the residues of the TH1 N-terminus, TH8 N-terminus and TH9 C-terminus (see Figure 3.S2). In general, both membrane bound conformations are similar to the one described above in the simulations of the neutral pH T-domain with highly anionic bilayers (POPC/POPG 1:3). Compare to contact histograms in Figure 3.8. There are slight differences in conformations of the neutral and low pH T-domains. In the case of the low pH T-domain structure, histidine and cationic residues reside in the glycerol group region (see Figure 3.S3A). The neutral pH structure have a decreased penetration of histidines, while cationic residues have a significantly increased density peak between the glycerol groups, as shown in Figure 3.S3B.

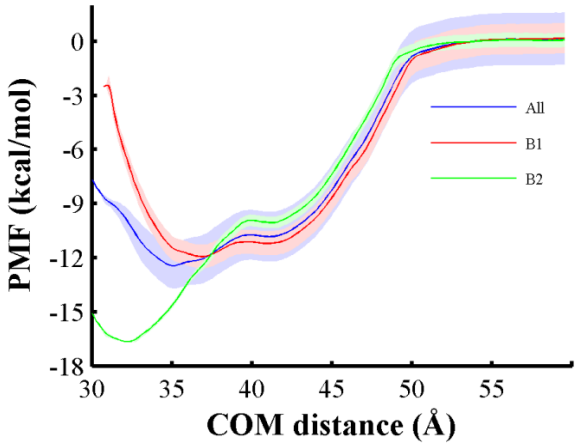

Figure 3.9. Potential of mean force (PMF) profile along the COM distance reaction coordinate (z axis). The membrane contains POPC/POPG 1:3. We perform three independent umbrella window simulations at each position along the reaction coordinate. These umbrella simulations start from protein orientations B1, B2 and an orientation observed at a lower fraction of POPG lipids. a) PMF curve and standard deviations calculated from the combined data of all umbrella CG-MD simulations are shown by a blue line and shades, respectively. b) PMF profiles of orientations B1 and B2 are shown in red and green lines, respectively. Standard deviations are shown by shades in their respective color. We performed 27 × 3 umbrella CG-MD simulations equally spaced every 1 Å over the reaction coordinate and the PMF profiles are calculated with WHAM method.[76] Standard deviations are obtained from bootstrapping calculations. Additional computational details are described in **3.2** ***Methods***.

3.3.6 Potential of Mean Force Calculations

We performed umbrella sampling calculations to determine the free energy profile of the low pH structure of T-domain across a preformed bilayer of higher anionic content (POPC/POPG 1:3). To enhance the conformational sampling, we performed three sets of

umbrella sampling calculations with different initial protein orientations. These protein orientations are the binding modes B1, B2 and the non-preferable orientation. This non-preferable orientation is not observed in equilibrium simulations of spontaneous binding of T-domain to bilayer with a high anionic content (POPC/POPG 1:3). The non-equilibrium orientation is obtained from simulations of the low pH protein structure and bilayers containing POPC/POPG 3:1 (see Figure 3.S2B). Figure 3.9 shows the free energy profiles for the complete set of simulations, orientation B1, and B2. The free energy profile obtained for the entire set of simulations has two minima; the relative global minimum is located at 35 Å and an intermediate at 41 Å. The free energy difference between these minima is 1.6 kcal/mol. The free energy barrier from the intermediate to the relative global minimum is within the thermal noise. Furthermore, the calculated free energy of binding from this curve is $\Delta G_{binding}$= = -11.3 ± 0.12 kcal/mol, which is calculated as indicated in the **3.2 *Methods*** section.

Inspection of each set of umbrella simulations shows that orientation B1 and the non-preferable protein orientation have similar stabilized orientations and PMF profiles. Thus we combine their MD frames into a single free energy profile, which is referred as PMF profile of conformation B1 (see red line in Figure 3.9). Free energy profile of B1 has a free energy minimum located at COM distance 37 Å and a local minimum located at COM distance of 41 Å, which represents an intermediate state. In contrast, the B2 free energy profile exhibits a deeply inserted global minimum at COM distance of 32 Å and an intermediate state at 41 Å. The free energy difference between the minima of each profile B1 and B2 is around 0.7 and 6.6 kcal/mol, respectively. Both profiles B1 and B2 have a free energy barrier between the intermediate to the relative global minimum,

which is within the thermal noise and is located near 40 Å. Overall, the lowest free energy states of membrane bound conformations B1 and B2 are inserted at similar COM positions observed in our standard CG-MD simulations, which sample the majority of conformations around the COM distances of 37 Å and 34 Å, respectively.

The convergence of protein orientation in our umbrella sampling simulations is monitored by the angles between helices TH8 and TH9, and the membrane normal. For example, Figure 3.S4 displays data from restrained simulations within the COM distance range of 35 Å to 45 Å in which the protein is in contact with the bilayer interface. The protein orientation converges after the first 200 ns of each window simulation (see Figure 3.S4); however, some simulations show some fluctuation of the angle values after this initial segment. Angle traces are shown for membrane-associated states with COM distance lower than 39 Å, at the transition state at 39 Å and for intermediate states with COM distance larger than 39 Å. The latter restrained simulations sample protein conformations with helices TH8 and TH9 in a near perpendicular orientation to the membrane plane (see cyan lines in Figure 3.S4).

3.3.7 Membrane-associated Conformations at Equilibrium

Protein conformations associated with the bilayer are extracted from the global minimums of PMF profiles B1 and B2. Figure 3.10A displays a structure of conformation B1 obtained from the last MD frame of the restrained simulation at COM distance of 37 Å. The partially unfolded helix TH1 and the unfolded TH2 form contacts with the membrane interface. The majority of contacts in this region involve residues 223-227, which includes the protonated sidechain of H223. The N-terminus of helix TH8 forms contacts involving residues 322-327; residue E259 also forms contacts with the

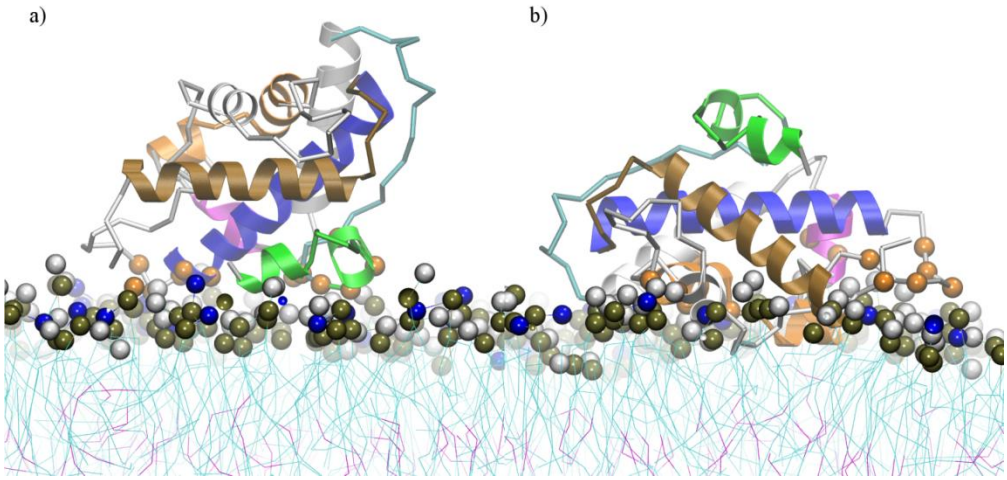

Figure 3.10. Models of the lowest free energy conformations B1 and B2. Structural representation of the lowest free energy binding modes of the low pH T-domain structure associated to an anionic lipid bilayer composed of POPC/POPG 1:3. a) Ribbon representation of membrane bound state B1 obtained from a restrained simulation at center of mass (COM) distance of 37 Å. b) Ribbon representation of membrane bound state B2 obtained from a restrained simulation at COM distance of 32 Å. Helices TH1, TH2, TH4, TH5, TH8 and TH9 are shown in green, cyan, magenta, orange, blue and brown ribbon representation, respectively. Other helices are shown in grey ribbons. Residues with normalized number of contacts larger than 0.6 are highlighted in orange space filled representation. Residues 202, 222-227, 259, 322-327 are highlighted in B1 representation and residues 240-241, 244, 248, 264-265, 277, 280, 288, 299, 316, 318, 320-322, 373, 377 (not shown) are highlighted in B2 representation. Head-group beads of POPC and POPG are highlighted in grey and blue space filled representation. Phosphorous atoms are shown in dark brown space filled representation. Residues more likely to be in contact with the membrane interface are shown in orange space-filled representation.

membrane interface. Figure 3.10B shows conformation B2, which is the last MD frame of a restrained simulation at COM distance of 32 Å. Remarkably, the partially unfolded helices TH1-2 are located away from the membrane interface in contrast to conformation B1. Helix TH5 is in a deeper inserted state and residues of helices TH3, TH4 and TH9 C-terminus form contacts with the membrane interface. Note that conformations B1 and B2 share a similar region of protein-membrane contacts, which is located around the N-terminus of helix TH8 (residues 320-322). Moreover, H322 is a common protein-membrane contact in both conformations. The histograms of the protein-membrane contacts are similar with the ones obtained from the standard CG-MD simulations.

Protein structures at the free energy barrier of each PMF profile B1 and B2 are shown in Figure 3.11. Overall, the protein orientations are slightly different from their respective global minima (see Figure 3.10). At the free energy barrier of the conformation B1 the region near the N-termini of helices TH1 and TH8 forms the majority of contacts with the membrane interface (see Figure 3.11A). In contrast, a transition structure of the conformation B2 displays the majority of the protein – membrane contacts in residues of the helices TH3, TH4, and near the N-terminus of the helix TH8 (as shown in Figure 3.11B). Furthermore, we observe that protonated H322 forms the membrane interface contacts in both transition states. This observation suggests that H322 plays a role in the protein membrane association as well in the partial unfolding around this region. The histograms of the contacts for these transition states are shown in Figure 3.S5.

The protein structures at the intermediate states (the second minimum in each PMF profile) are similar to those observed in the standard CG-MD simulations of the low

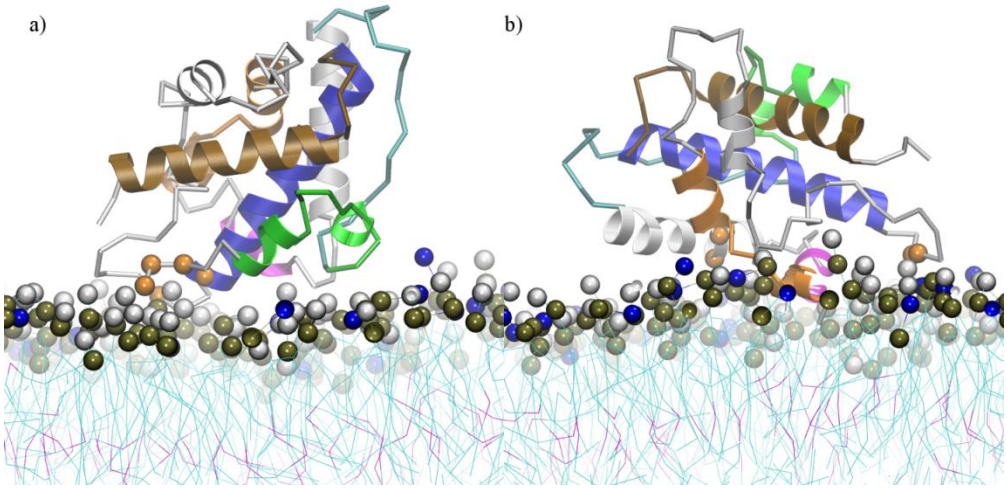

Figure 3.11. Models of the intermediate states of membrane bound conformations B1 and B2. Structural representation of the intermediate membrane bound states along the free energy profile curves of the low pH T-domain structure in the presence of a preformed membrane containing POPC/POPG 1:3. Each representation corresponds to the last MD snapshot of restrained simulations at center of mass distance of 39 Å. a) Ribbon representation of intermediate membrane bound state B1. b) Ribbon representation of membrane bound state B2. Helices TH1, TH2, TH4, TH5, TH8 and TH9 are shown in green, cyan, magenta, orange, blue and brown ribbon representation, respectively. Other helices are shown in grey ribbons. Residues with normalized number of contacts larger than 0.6 are highlighted in orange space filled representation. Residues 202-204, 322, 326, 375 (not shown) are highlighted in B1 representation and residues 251-252, 264, 267, 321-322 are highlighted in B2 representation. Head-group beads of POPC and POPG are highlighted in grey and blue space filled representation. Phosphorous atoms are shown in dark brown space filled representation.

pH structure with bilayers containing a small fraction of anionic lipids POPC/POPG 3:1 (see Figures 3.S2B and 3.S6). In this intermediate state, the residues 203-204 near the N-terminus of the helix TH1, the residues H322 and H323 near the N-terminus of TH8, the residue Y375 in the C-terminus of TH9 form stable contacts with the membrane interface, as shown in Figure 3.S6. Also, the residue E262 of the helix TH4 forms protein-membrane contacts, but less frequently.

3.4. Discussion

Spectroscopic studies of the pH-dependent membrane binding of T-domain and its membrane-associated states required a minimum amount of anionic lipids, e.g. POPG,[10, 12] and this observation was also reported in a solid state NMR study of the protein membrane insertion.[13] A similar observation is drawn from our CG-MD simulations. Anionic bilayers promote formation of stable membrane-associated states over the time-scales simulated. In contrast, neutral lipids (pure POPC) alone allow for only transient binding of the protein.

Simulations of T-domain association with lipids with low anionic content (POPC/POPG 3:1) resulted in partial association with the membranes, namely, three of ten simulations resulted in formation of the long-lived lipid-protein complexes. Neutral and low pH T-domain structures show similar membrane–associated conformations in anionic bilayers composed of POPC/POPG 3:1. These similar membrane–associated states involve protein-membrane interactions of cationic residues located in the protein terminals and residues near the N-terminus of the helix TH8 (residues H323 and E326). Note that the neutral pH T-domain structure has histidines in their neutral state, resulting in a decreased penetration in the membrane interface compared to the low pH T-domain

structure. At the same time the membrane-associated neutral pH structure has a slightly larger fraction of its terminal cationic residues embedded between lipid head-groups (see Figure 3.S3). This observation suggests that the membrane association of T-domain with bilayers containing a small fraction of anionic POPG phospholipids depends on specific distribution of positive charge on the protein. Protonation of all histidines not only decreases the overall charge of the T-domain from -10 electron units to -4, letting it to insert deeper into a negatively charged bilayer but also re-distributes positive charge on the protein, thus promoting a slightly different orientation at the lipid interface.

T-domain associates with the membranes strongly and relatively rapidly in all simulations with bilayers predominantly composed of anionic lipids (POPC/POPG 1:3). Our simulated results correlate well with the previously observed binding of T-domain to large unilamellar vesicles (LUV) composed of various compositions of POPC/POPG at pH 6.5.[10] In these experiments the fraction of the protein bound was significantly higher when the fraction of the anionic lipid POPC/POPG increased from 3:1 to 1:3.

The low pH T-domain model shows two predominant membrane-associated conformations B1 and B2 in bilayers composed of POPC/POPG 1:3 (see Figure 3.10 and results for description). B1 is the most frequently observed membrane bound conformation, which forms stable protein-membrane contacts at residues of the partially unfolded helices TH1-2 and the loop between TH3-4 (see Figure 3.10A). The conformation B2 has more protein-membrane interface contacts at residues of helices TH3, TH4-5, TH6-7 and is stabilized by hydrophobic interactions with the lipid tails (see Figure 3.10B). Furthermore, B2 shows partially unfolded N-terminal helices exposed to the solvent. However, conformations B1 and B2 have a common binding surface

composed by residues in the N-terminus of TH8 and C-terminus of TH9. Furthermore, these conformations have similar orientations of helical axes of TH8-9, which are near parallel or at an oblique angle relative to the membrane plane. Finally, some trajectories show residues of TH2 or residues of the loop between TH2-3 as initial protein-membrane contacts.

Equilibrium properties of the low pH T-domain structure as a function of the insertion depth in the bilayer (POPC/POPG 1:3) are best characterized using potential of mean force calculations. The estimated $\Delta G_{binding}$ is -11.3 ± 0.12 kcal/mol, which is consistent with the observation that the protein binds spontaneously to bilayers composed mainly of negatively charged lipids in low pH solution.[10] Comparison of the PMF profiles shows that B2 has a global minimum deeply inserted in the membrane interface, while B1 has the second minimum located in a relatively shallow position. The lowest minimum of conformation B2 is favored by 4.7 kcal/mol relative to the equivalent minimum of B1. However, a different trend is calculated for the intermediate and transition states of both conformations. B1 intermediate state is preferable by 1.2 kcal/mol, which is relative to the intermediate state of B2. Similarly, B1 transition state is lower by 1.2 kcal/mol than that of B2.

Based on our previous atomistic modeling of the membrane-competent state of T-domain in explicit solvent followed by the present membrane binding study, we can propose the following mechanism of the initial stages of membrane association of T-domain. First, protonation of histidines induces solvent exposure of hydrophobic residues of the helixes TH8 N-terminus and TH9 C-terminus and a partial unfolding/refolding of N-terminal helices, which facilitates disruption of the inter-helical salt-bridges.[20, 47]

Secondly, the low pH T-domain structure binds to the membrane with conformation B1, which is guided by electrostatic interactions between the exposed cationic residues located in the partially unfolded N-terminal helixes, and the anionic lipids. B1 is frequently observed in our CG-MD simulations, which is probably due to its lower relative free energy barrier. Note that the deeply inserted state B2 is only sampled by a single CG-MD trajectory, which is probably due to its non-preferable transition state. However, at equilibrium, B2 is favored by hydrophobic interactions and the solvent exposure of the N-terminal helices. Additional CG-MD simulations of the low pH protein structure with histidines set in their neutral state provide further indication that conformation B1 is favored by protonation of histidines. The rate or probability of formation of B2 in our standard CG-MD simulations were not affected by protonation of histidines but was driven by pre-unfolding the protein. The predicted membrane–associated states B1 and B2 share some similarities with reported pH-dependent conformational states of T-domain by neutral reflectometry and solid state NMR experiments[13]. For example, conformation B2 has similar features to those observed in the protein bound state at pH 6 such as: solvation of the N-terminal helices and the near parallel orientation of the hydrophobic helices relative to the membrane normal axis (see Figure 3.10B and Figures 3.3F, 3.4F). Conformation B1 has some similar features observed in the protein bound conformations at pH 4; however, B1 is less deeply inserted in the membrane interface. Chenal et al.[13] suggested that neutralization of acidic side-chains in the N-terminal helices may drive further insertion of the refolded structure of T-domain in the membrane interface. The authors reported a majority of parallel conformations, which may be explained by the small fraction of anionic lipids in the

membrane (POPC/POPG 4:1).[3] There is some experimental evidence that acidic residues located in the hydrophobic hairpin formed by TH8-9 also play a role in the process of the membrane association of T-Domain (A.S. Ladokhin, unpublished data). In future studies, we will explore the role of protonation of the acidic side-chains in the predicted membrane bound conformations of T-domain.

3.5. Conclusions

In summary, this work presents a coarse-grained modeling of the membrane association process of neutral and low pH T-domain models. We show that T-domain associates more rapidly and steadily to bilayers as the fraction of anionic lipids increases. In particular, our simulations show that the low pH T-domain model binds deeply in the membrane interface of predominantly anionic bilayers and forms extensive protein-membrane contacts. Furthermore, combined approaches of standard and umbrella sampling simulations suggest two membrane-associated states of the low pH T-domain structure, which are characterized by distinctive patterns of electrostatic and hydrophobic interactions between the protein and the bilayer. Based on these results, we propose an initial membrane association pathway of T-domain. This study supports previous findings of the role of protonation of histidines in the formation of a refolded state of T-domain in solution with properties of a membrane-competent state. Our results indicate the micro-second time-scales are involved in the membrane insertion process of proteins of a few hundred of residues such as the T-domain. In future work, we will refine the identified membrane bound conformations by atomistic MD simulations. The role of neutralization of acidic side-chains in the membrane–associated states of T-domain will be also studied.

3.6 Appendix

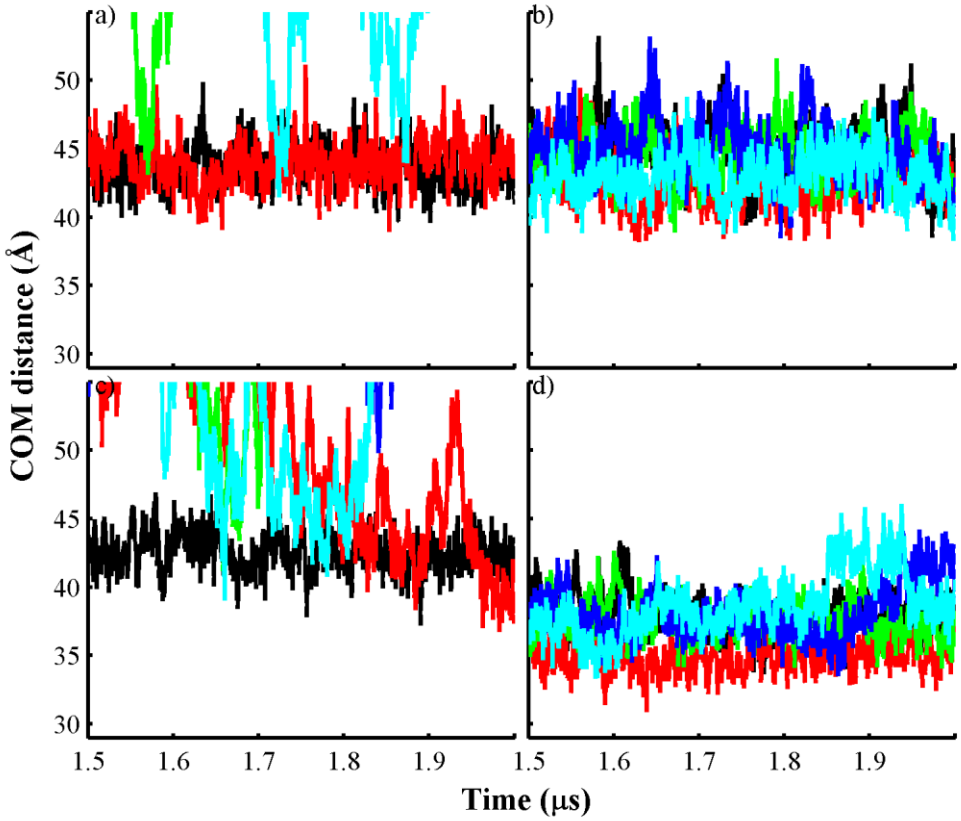

Figure 3.S1. Plots of the center of mass (COM) distance between T-domain and anionic lipid bilayers of different composition obtained from the last 500 ns of standard CG-MD simulations. Distance between the neutral pH structure and bilayers of the following composition: a) POPC/POPG 3:1. b) POPC/POPG 1:3. Distance of the low pH unfolded T-domain structure and lipid bilayers of the following composition: c) POPC/POPG 3:1. d) POPC/POPG 1:3. Each line color represents an independent CG-MD production simulation, which is initiated with the same protein orientation and different seed number.

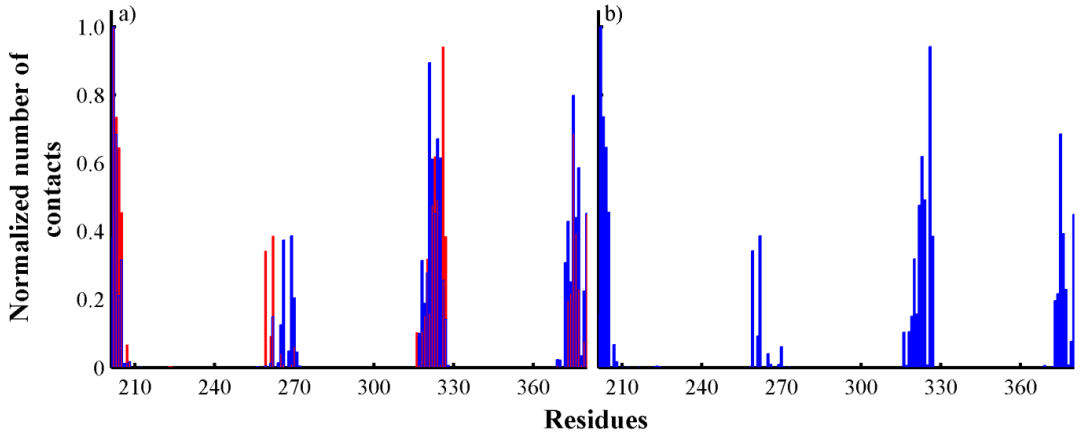

Figure 3.S2. Normalized number of protein-membrane interface contacts as a function of residue number. Data is obtained from the last 200 ns segments from: a) Two independent standard CG-MD simulations of the neutral pH T-domain structure. b) One simulation trajectory of the low pH T-domain structure. The preformed membrane contains POPC/POPG 3:1. All protein and head-group, phosphate beads separated by a distance lower than 5 Å are considered to be in contact.

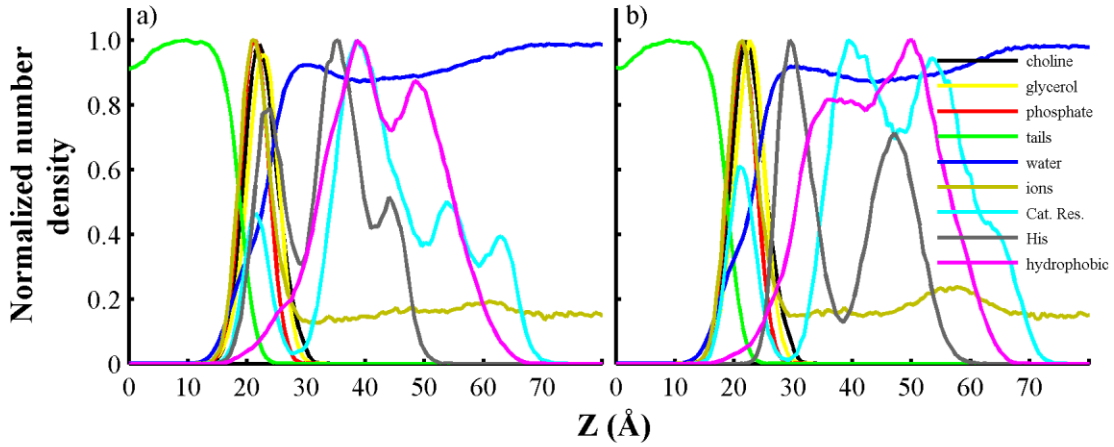

Figure 3.S3. Normalized number densities of coarse-grained groups as a function of their location in the normal membrane axis. The maximum of each number density profile is scaled to one. Data is obtained from the last 500 ns of representative standard CG-MD simulations of T-domain structures with lipid bilayers containing POPC/POPG 3:1. a) Membrane bound conformation of low pH T-domain structure. b) Membrane bound conformation of neutral pH T-domain structure. Density profiles of lysine residues are similar to those of cationic residues in all plots, but the density peak between the phosphate region corresponds to the N-terminal and arginine residues. The center of the bilayer coincides with z = 0. Density profiles of headgroup atoms of POPC (choline) and POPG (glycerol) are shown in black and yellow lines, respectively. Density profiles of phosphate and tails groups from all lipids are shown in red and green lines, respectively. Density profiles of water and sodium counterions (ions) are shown in blue and dark yellow, respectively. Profiles of cationic residues (Lys, Arg, and N-terminal), histidines and hydrophobic residues are shown in cyan, grey and magenta lines, respectively.

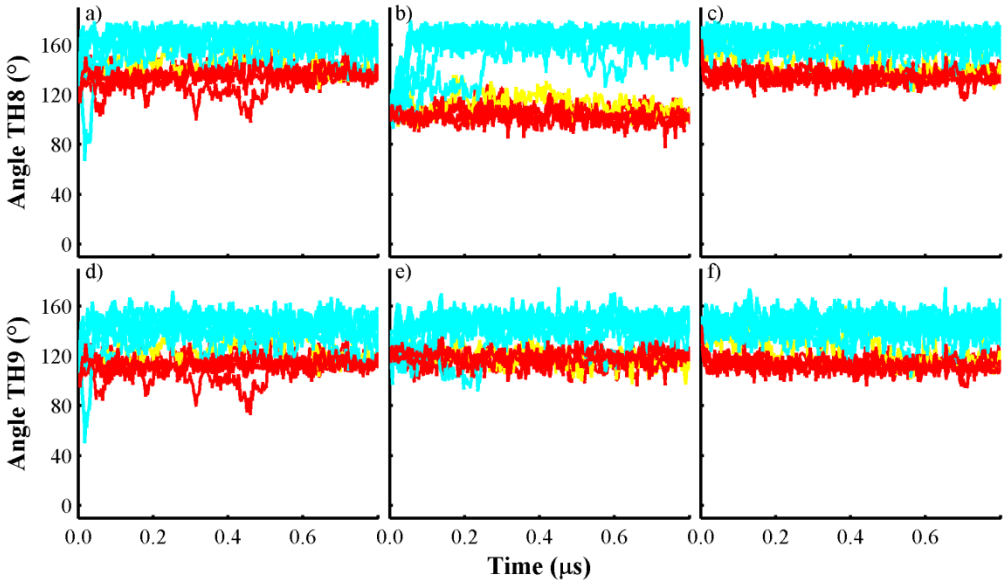

Figure 3.S4. Angle formed by the axes of helices (TH8 and TH9) and the normal axis to the membrane plane versus simulation time (the normal axis direction is towards the membrane core). For example, helix TH9 is perpendicular to the membrane plane if the angle value is close to 0° or 180°. We present results from umbrella CG-MD simulations in which the low pH T-domain structure begins to associate to the membrane interface, which spans the range of 35 Å to 45 Å along the reaction coordinate, as shown by free energy curves in Figure 3.9. a) and d) display the angles formed by helices TH8 and TH9 obtained from the PMF curve of the membrane bound state B1, respectively. b) and e) display the angles formed by helices TH8 and TH9 obtained from the PMF curve of the membrane bound state B2, respectively. c) and f) show the angles formed by helices TH8 and TH9 obtained from the PMF curve obtained from an initial non-preferable protein orientation, see text for details. Protein conformations where helices TH8 and TH9 are approximately perpendicular to the membrane plane are shown in cyan lines. These

correspond to umbrella simulations spanning the range of 40 Å to 45 Å along the reaction coordinate. Protein conformations where helices TH8 and TH9 are approximately parallel to the membrane plane are shown in red. These correspond to umbrella simulations spanning the range of 35 Å to 38 Å along the reaction coordinate. The intermediate states are shown in yellow lines, corresponding to umbrella simulations restrained at center of mass distance of 39 Å. Each single trace represents an independent umbrella CG-MD production simulation, which are initiated with the same initial protein orientation and different seed number. All plots show data obtained from umbrella sampling simulations of the low pH T-domain structure and a preformed membrane of mixture POPC/POPG 1:3.

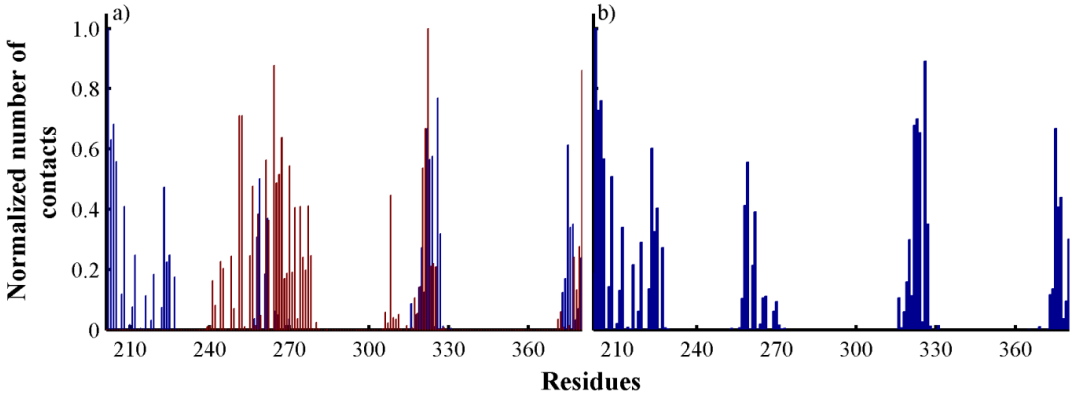

Figure 3.S5. Normalized number of protein-membrane interface contacts as a function of residue number. Data is obtained from the last 600 ns of three umbrella window simulations of the low pH T-domain structure and a preformed membrane composed of POPC/POPG 1:3, shown in Figure 3.9. a) Blue and red bars correspond to the transition conformations of PMF profiles of B1 and B2, respectively. Data is obtained from umbrella simulations restrained at COM distance of 39 Å. b) Blue bars correspond to the transition conformation of an umbrella simulation restrained at COM distance of 39 Å and with an initial non-preferable protein orientation. All protein and head-group, phosphate beads separated by a distance lower than 5 Å are considered to be in contact.

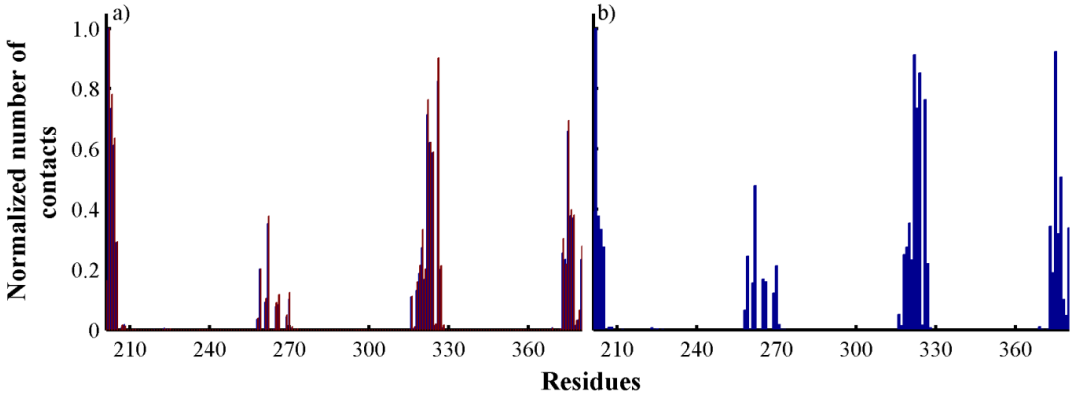

Figure 3.S6. Normalized number of protein-membrane interface contacts as a function of residue number. Data is obtained from the last 600 ns of three umbrella window simulations of the low pH T-domain structure and a preformed membrane containing POPC/POPG 1:3, the center of mass distance is restrained at 42 Å. a) Blue and red bars correspond to the intermediate state of PMF profiles of B1 and B2, respectively. b) Blue bars correspond to the intermediate state obtained from an umbrella simulation with an initial non-preferable protein orientation at COM distance of 42 Å. All protein and head-group, phosphate atoms separated by a distance lower than 5 Å are considered to be in contact.

Chapter 4. Influence of Acidic Amino-acids in the Orientation and Insertion of T-domain Membrane Bound States.

4.1 Introduction

Diphtheria toxin, a soluble protein at neutral pH, binds to a cell surface receptor and block protein synthesis of sensitive cells by delivering an enzymatic fragment using its own endocytic pathway.[2] Endosomal acidification triggers conformational changes in a translocation (T)-domain facilitating its membrane insertion and permeation of its N-terminus attached to an enzymatic domain.[2] The membrane insertion and function of the stand-alone T-domain has been extensively studied; however, different experimental conditions have suggested a diverse set of membrane bound and inserted states of T-domain.[3, 8, 10, 12, 61, 81] A complete understanding of the unassisted pH-dependent insertion of T-domain may facilitate the development of improved drug delivery methods for therapeutic applications such as oncogenic treatment.[4]

Stand-alone T-domain adopts a monomeric alpha-helical structure at neutral pH.[11, 62] In response to acidification, the protein structure undergoes conformational changes facilitating its membrane association followed by the membrane insertion of two hydrophobic helices (named TH8 and TH9), which facilitate the translocation of a protein fragment attached to the T-domain N-terminus.[82, 83] A series of experiments, free energy calculations and extensive atomistic molecular dynamics (MD) simulations have shown that T-domain partially unfolds into a membrane-competent state upon protonation of histidines.[10, 11, 20, 47] This membrane-competent state is characterized by refolding of N-

terminal helices and solvent exposure of hydrophobic and charged sites.[20, 47] Recently, we have performed coarse-grained MD simulations to study the membrane association process of the MD generated model of the membrane-competent state to anionic bilayers. Two preferable membrane bound conformations were identified by extensive unbiased and umbrella sampling calculations.[84] Both predicted membrane bound conformations showed helices TH8-TH9 with a parallel orientation relative to the membrane plane, which was in agreement with low resolution information reported by solid state nuclear magnetic resonance (NMR).[13] These NMR experiments showed pH-dependent conformational changes of T-domain conformations bound to lipid bilayers.[13] It was hypothesized that the protonation state of acidic side-chains may affect the refolding of the protein in the membrane interface.[12, 13, 84] A separate study of the kinetics of T-domain insertion to anionic lipid bilayers showed the formation of a membrane interface intermediate state and an insertion competent state.[10] However, atomic details of the degree of insertion of specific protein regions and possible structural rearrangements towards deep inserted states are not understood.

Our goal is to characterize the initial stages of T-domain binding on the membrane interface and explore the possible role of neutralization of acidic side-chains in the protein conformation by using atomistic molecular dynamics. Previously, we have reported[84] two possible protein-membrane binding orientations by multiple microsecond long MD simulations using MARTINI coarse-grained models of a low pH T-domain structure and mixtures of lipid bilayers. In this chapter, we refine these two frequently observed protein-membrane binding modes by all-atom MD simulations using a modified force field for anionic lipids compatible with the force field for proteins Amber ff99SB.[52]

Our MD simulations show that the two predicted membrane bound conformations stabilize on the membrane interface, while both conformations insert deeply in the membrane interface. Atomistic simulations of the two predicted membrane bound conformations suggest protein regions that are most likely to interact with the lipid headgroups and tails. Neutralization of acidic residues facilitates reorientation and deep insertion of membrane bound conformation.

4.2 Methods

4.2.1 Force Field Modifications

Atomistic MD simulations of proteins using Amber FF99SB force field have resulted in good agreement with NMR experimental data, folding of small α-helical or β-sheet proteins and limited agreement with temperature dependent helix-coil transitions.[38, 85-87] Moreover, we have recently reported the formation of a T-domain membrane-competent state in solution using FF99SB, which has features in good agreement with fluorescence experiments.[13, 14] However, development of force field parameters for anionic lipids compatible with both tensionless MD simulations and Amber force fields is still an ongoing process.[88, 89] In order to study T-domain association to lipid bilayers composed of the anionic 1-palmitoyl-2-oleoyl phosphatidyl-glycerol (POPG), we propose a number of modifications in phospholipids parameters based on two lipid force fields. The schematic representation of POPG is shown in Figure 4.S1. Recently, two related lipids force field have been reported, which are based on the General Amber force field (GAFF).[90] One of them, Lipid11 provides a systematic charge derivation methodology compatible with Amber Restrained Electrostatic Potential (RESP).[89, 91] However, Lipid11 requires an external surface tension NPT ensemble to successfully reproduce the

properties of lipid bilayers. To circumvent the use of external surface tension NPT ensemble, GAFF parameters of the acyl chains have been re-parameterized by fitting the torsional and Lennard-Jones parameters to reproduce properties of pentadecane liquid.[88] Thus, GaffLipid successfully reproduced a range of properties including the area per lipid of POPC lipid bilayers. However, MD simulations of POPG lipid bilayers were not reported for this recent force field. We used GaffLipid parameters to model POPG lipids; however, the average area per lipid is around 55 $Å^2$ while the experimental value is 67.3 $Å^2$ for POPG at T = 310 K (see Figure 4.S2).[92] To solve this problem we have modified the Lennard-Jones parameters of the head-group atoms using parameters from CHARMM36 as an initial guess.[93] However, the area per lipid was around 57 $Å^2$ (data not shown). The next step was to replace bond, angle bond, and torsional parameters by those from GaffLipid. MD simulations using the modified force field resulted in an area per lipid around 60 Å2, which is closer to the experimental value of 67.3 $Å^2$. The force field modifications are shown in Table 4.S1 (see Supporting Information). Fixed charges of the lipid headgroup atoms were fitted using the multi-conformer RESP method,[94] which used 19 different lipids conformations. All lipid conformations were optimized in the gas phase using Hartree-Fock self consistent field (HF-SCF) with a basis set HF/6-31g* and Gaussian03.[95] Atomic charges for the lipids tails were the same as described in Lipid11. The test system consisted of 72 POPG lipids solvated by approximately 3600 TIP3P water molecules, sodium counterions were added to neutralize the system. We also report the order parameters $S_{CD} =< 3cos^2\theta - 1 >/2$, where θ is the angle between the vector joining the lipid tails C-H and the membrane normal. Large values of S_{CD} represents ordered lipid tails and smaller ones represent less ordered. The order

parameters S_{CD} of the lipid tails are similar to those reported in the original force field (see Figure 4.S3). The density profile of pure POPG bilayers at two different temperatures shows that sodium counterions reside mostly in the phosphate group region, which is in good agreement with previous MD simulations (see Figure 4.S4).[96, 97] Future work will require optimization of the dihedral angles of the headgroup and phosphate group atoms.

4.2.2 Conversion of Coarse-grained to Atomistic Models

A previous study of the membrane association of T-domain predicted two membrane binding modes B1 and B2, which were determined by unbiased MD simulations and umbrella sampling calculation of coarse-grained models of the protein and lipid bilayers.[84] In this study, a representative conformation from each binding mode was extracted from an umbrella window simulation corresponding to the lowest free energy profile. Each coarse-grained model of T-domain bound to a bilayer was converted into an all-atom model using the conversion protocol of Cojocaru et al.[98] The original number of phospholipids in the coarse-grained model was 512, which was reduced to 192 lipids in the atomistic model. Furthermore, the coarse-grained contained a mixture of POPC/POPG 1:3, which was changed to only POPG lipids in our atomistic models.

4.2.3 Equilibration protocols

Hydrogen atoms were added to the initial low pH T-domain structures using tLeap available in AMBER12 package.[48] The simulation boxes of the membrane bound states of the low pH T-domain model were created by adding approximately 14800 TIP3P explicit water molecules such that the distance between the protein/bilayer and the simulation box edge was 12.0 Å. The total number of atoms is approximately 72000

atoms. The size of the simulation boxes were 120 Å × 121 Å × 116 Å and 123 Å × 112 Å × 122 Å for the first and second protein membrane orientation, respectively. Sodium counter-ions were added to neutralize the protein/membrane charge. Equilibration simulations were performed using PMEMD[99] on GPU cards with the FF99SB-ILDN[100] force field. The simulation time step was 2 fs and all hydrogen bonds were constrained via SHAKE.[53] Periodic boundary conditions were set up, cutoff radius was set to 10 Å and electrostatic calculations were performed using Particle Mesh Ewald (PME) method.[50] We equilibrated each system as follows. The protein and lipid heavy atoms were restrained and the solvent was minimized for 250 of steepest descent minimization method followed by 750 steps of conjugate gradient descent. Then the solvent was restrained and the protein was minimized by a total number of 5000 steps with the first 2500 steps of steepest descent. Each system was slowly heated in five stages of 5 ps from T = 0.1 K up to T = 310 K with NVT ensemble using Langevin thermostat. On each stage restrains were applied as follows: a) Harmonic restraints with a force constant of 10 kcal mol^{-1} $Å^2$ were applied to protein and lipid atoms. b) Harmonic restraints with a force constant of 5 kcal mol^{-1} $Å^2$ were applied to protein and lipid atoms. c) Harmonic restraints with a force constant of 2.5 kcal mol^{-1} $Å^2$ were applied to protein atoms. Harmonic restraints with a force constant of 1.0 kcal mol^{-1} $Å^2$ were applied to protein atoms over the last two heating stages. Anisotropic NPT ensemble equilibrations using Berendsen barostat at 1 atm were carried out in three stages. In the first stage protein atoms were restrained with a force constant of 0.5 kcal mol^{-1} $Å^2$ over 250 ps, the following stage has restraints on backbone atoms with a force constant of 0.25 kcal mol^{-1} $Å^2$ over 250 ps and all restraints were removed over the last stage of 170 ps. Then, each

system is equilibrated over 200 ns using the NPT ensemble with semi-isotropic scaling available in PMEMD.

4.2.4 Production Protocols on Specialized Computer Anton

We used the final structures of equilibration MD simulations. The force field parameters for MD simulations on the MD specialized hardware Anton[101] were the same as for the equilibration simulations, except that the Gaussian Split Ewald method[27] with a grid size of 64 × 64 × 64 was used and a cutoff radius of 12.36 Å for the Van der Waals and short-range electrostatic interactions. A reversible multiple-time step algorithm was used with a time step of 2 fs for bonded and short range non-bonded interactions and 6 fs for the long range non-bonded forces. All hydrogen bonds were constrained with SHAKE algorithm.[53] Semiisotropic NPT conditions were applied with a multigrator integrator, which has a Noose-Hoover[102] thermostat with an interval of 24 steps and a Martyna-Tobias-Klein[103] barostat with an interval of 240 steps. The temperature was set to 310 K and pressure to 1.0 atm. MD simulations of conformation B1 and B2 were carried out over 3541 ns and 810 ns, respectively. Extended MD simulation of conformation B1 at T = 323 K was carried out over 2124 ns. A second extended MD simulation of conformation B1 at T = 323 K and with acidic side-chains neutralized was carried out over 3459 ns.

4.2.5 Analysis

Analysis and molecular visualization was carried out with in-house tcl/tk scripts using VMD 1.9.1[55] and AmberTools12.[48]

4.3 Results

4.3.1 Description of Molecular Models

Extensive unbiased MD simulations and umbrella sampling calculations of coarse-grained models of T-domain and lipid bilayers of different compositions have suggested two predominant binding modes of a low pH T-domain model bound to anionic bilayers.[84] We refine representative snapshots of these two membrane bound conformations by atomistic MD simulations. To model the lipid bilayer, we use POPG. In this study, a pure anionic bilayer is used because negatively charged lipids promote membrane binding and facilitate faster transmembrane insertion of T-domain.[84]

4.3.2 Simulated Membrane Bound Conformations

The two predominant conformations observed in our CG-MD simulations are denominated B1 and B2.[84] After conversion from CG to atomistic representation and equilibration, we perform approximately 3.5 μs and 0.8 μs atomistic MD simulations of each orientation B1 and B2 at T = 310 K, respectively. Structural rearrangements of T-domain in our atomistic simulations are roughly estimated by the root mean square deviation (RMSD) of C_α atoms relative to their initial position in helices TH1-9, TH5' (see Figure 4.1). This is calculated after translation and overlay of each MD frame relative to the initial protein structure. RMSD gradually increases to an average value of 2.0 Å over the initial 2.0 μs of the MD trajectory of conformation B1 (see Figure 4.1A), while conformation B2 shows an abrupt change of RMSD to 1.7 Å after the first 100 ns (see Figure 4.1A). In order to observe possible refolding of the less deeply inserted state B1, we extend the simulation of this conformation at a slightly higher temperature of T = 323 K. Extended simulation of conformation B1 results in a slightly increase of RMSD to 1.7 Å over the last 500 ns, as shown in Figure 4.1B. Partial unfolding of the solvent

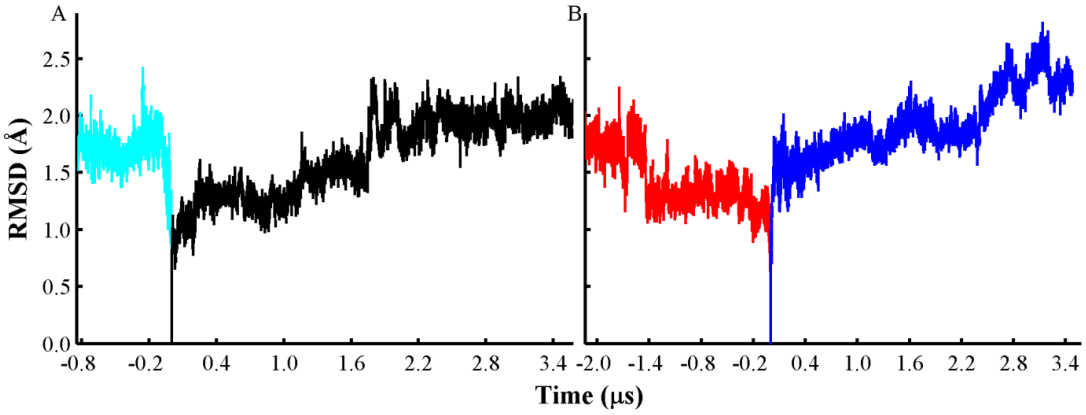

Figure 4.1. RMSD curves as a function of time obtained from conformations B1 and B2. Changes of the C_α root mean squared deviation (RMSD) of C_α from helices TH1-9, TH5' relative to the protein initial coordinates. (A) RMSD traces obtained from orientation B1 (black line) and orientation B2 (cyan). Simulations were performed at T = 310 K. (B) RMSD curves obtained from orientation B1 (red) and with acidic side-chains neutralized (blue lines). Simulations were performed at T = 323 K. All histidines are set in their protonated state in all simulations and other acidic residues are set in their standard state, unless indicated.

exposed helix TH5 occurs over this segment. To explore the possible role of neutralization of acidic side-chains in the protein conformational changes, we retrieve an MD frame from the simulation of conformation B1 at T = 310 K at simulation time *ca.* 1.5 us. All acidic side-chains are neutralized, while histidines are kept in their protonated state. Simulation of the neutralized conformation B1 results in an increase of RMSD to 2.4 Å over the last 500ns. Helix TH5 is partially unfolded over this segment (see Figure 4.1B). These MD simulations show that the protein retains its compact structure, while it remains bound to the bilayer interface.

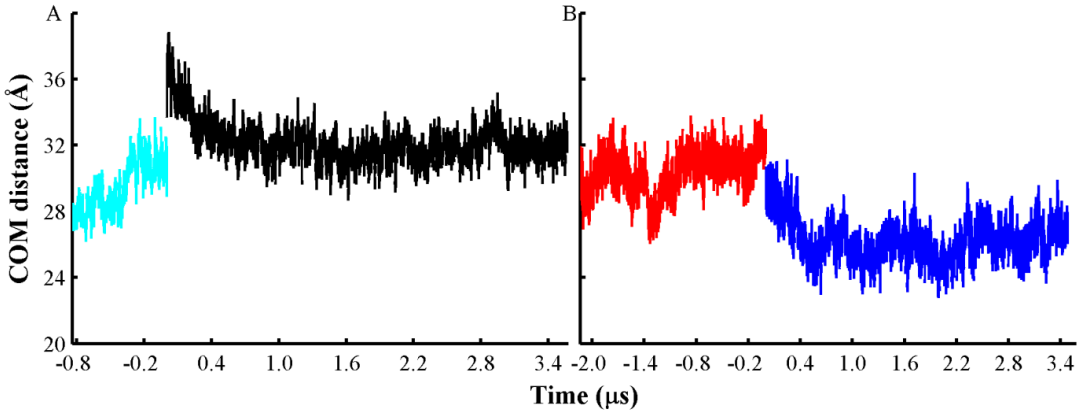

Figure 4.2. COM distance as a function of time obtained from conformations B1 and B2. Changes of the center of mass (COM) distance between the low pH T-domain models and lipid bilayers containing POPG. (A) COM distances obtained from orientation B1 (black line) and orientation B2 (cyan). Simulations were performed at T = 310 K. (B) COM distances obtained from orientation B1 (red) and acidic side-chains neutralized (blue lines). Simulations were performed at T = 323 K. All histidines are set in their protonated state in all simulations and other acidic residues are set in their standard state, unless indicated.

4.3.3 Degree of Insertion of Membrane Bound Conformations

We monitor the degree of insertion of T-domain in the membrane interface by calculating the center of mass (COM) distance between the protein and the bilayer as a function of time in Figure 4.2. For example, Figure 4.2A shows that the membrane bound conformation B2 penetrates deeper in the bilayer interface, which is initially placed at *ca.* 32 Å and then rapidly moving to a position near 28 Å. In contrast, membrane bound conformation B1 is initially positioned at *ca.* 37 Å and gradually adopts a stable inserted

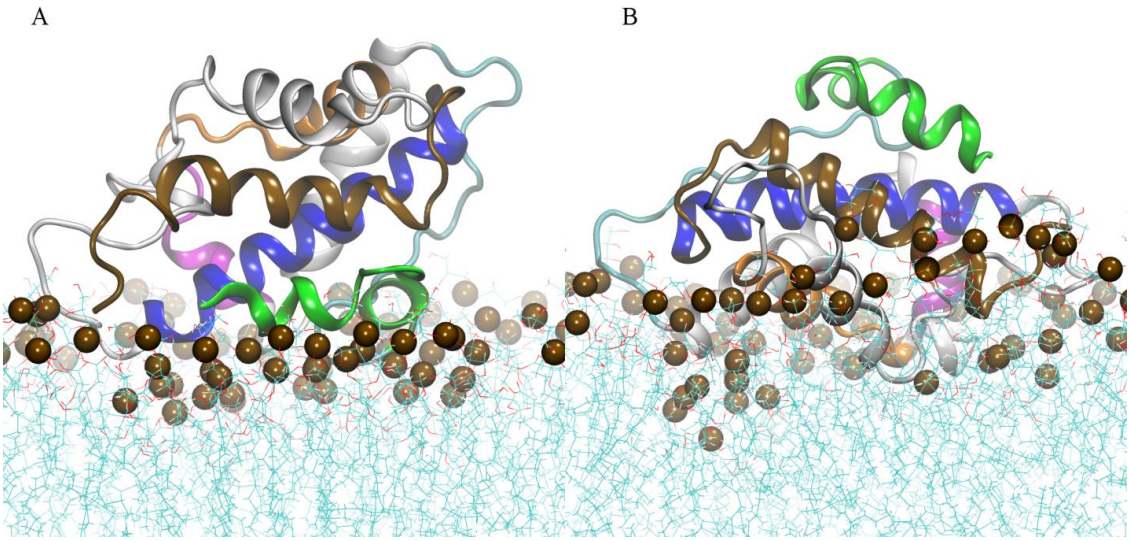

Figure 4.3. Atomistic models of membrane bound conformations B1 and B2. Final MD frames of atomistic simulations of low pH T-domain with fixed protonation state for histidines and standard state for the acidic side-chains. (A) orientation B1 and (B) orientation B2. Helices TH1, TH2, TH4, TH5, TH8 and TH9 are shown in green, cyan, magenta, orange, blue and brown ribbon representation, respectively. Other helices are shown in grey ribbons. Phosphorous atoms are shown in dark brown space filled representation. Water molecules are not shown.

state around 32 Å over the 1.0 μs of simulation time. At a higher temperature, protein conformation B1 penetrates slightly deeper the membrane interface to an average COM distance of 30 Å (see Figure 4.2B). Neutralization of acidic side-chains induces deeper insertion of the protein to an average COM distance of 26 Å. The following paragraph describes the last MD generated structures for each simulation.

Figure 4.3A shows the most frequently observed orientation B1, which has protein-membrane contacts localized at residues in helices TH1-2, TH4-5, and the N-terminus of TH8 and C-terminus of TH9. In contrast, Figure 4.3B shows the less

frequently observed orientation B2, which forms membrane contacts on residues of the loop TH2-3, TH3, TH5-6, and the N-terminus of TH8 and C-terminus of TH9. The latter orientation penetrates deeper the membrane interface than orientation B1 at simulation temperature T = 310 K.

Increase of temperature in the extended simulations and neutralization of acidic side-chain facilitates insertion of the protein in the membrane interface, as shown in Figure 4.4. This shows the last MD frames of simulations of membrane bound conformation B1 with different protonation states of acidic side-chains. Figure 4.4B shows that neutralization of acidic side-chains induced an overall rotation of the protein compared to the protein structure with acidic residues in their standard state (see Figure 4. 4A). As a result, the refolded helix TH1 penetrates the phosphate atoms region and forms contacts with the lipid tails. Furthermore, loops between TH5'-6 and TH8-9 form contacts with the membrane interface. The following paragraphs will provide a detailed description of the changes in the orientation and structure of the membrane bound conformations.

4.3.4 Changes of Orientation of Membrane Bound Conformations

To determine changes in the overall orientation of the protein, we find useful to compute the orientation of hydrophobic helices TH8 and TH9 relative to the membrane normal axis (see Figure 4.5). Figures 4.5A, C show that both membrane bound conformations B1 and B2 retain their overall orientation, in which helices TH8-9 exhibit a near parallel or oblique conformation relative to the membrane plane. In particular helix TH8 in conformation B2 remains parallel to the membrane plane. MD simulation of membrane conformation B1 at a slightly higher temperature of T = 323 K display similar

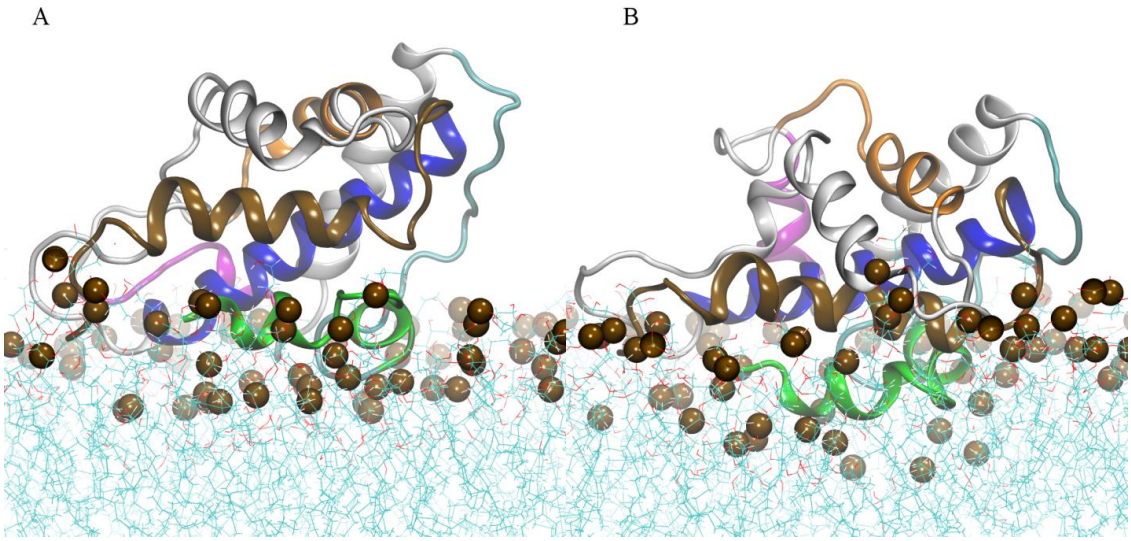

Figure 4.4. Different atomistic models of membrane bound conformation B1. Final MD frames of atomistic simulations of membrane bound conformation B1 with fixed protonated state for all histidines. (A) Low pH T-domain structure with acidic residues set in their standard protonation state. (B) Low pH T-domain structure with acidic residues set in their neutralized state. Helices TH1, TH2, TH4, TH5, TH8 and TH9 are shown in green, cyan, magenta, orange, blue and brown ribbon representation, respectively. Other helices are shown in grey ribbons. Phosphorous atoms are shown in dark brown space filled representation. Water molecules are not shown.

oblique conformations of helices TH8-9; however, neutralization of acidic side-chains drives helices TH8-9 to a near parallel orientation relative to the membrane plane (see Figure 4.5B, D).

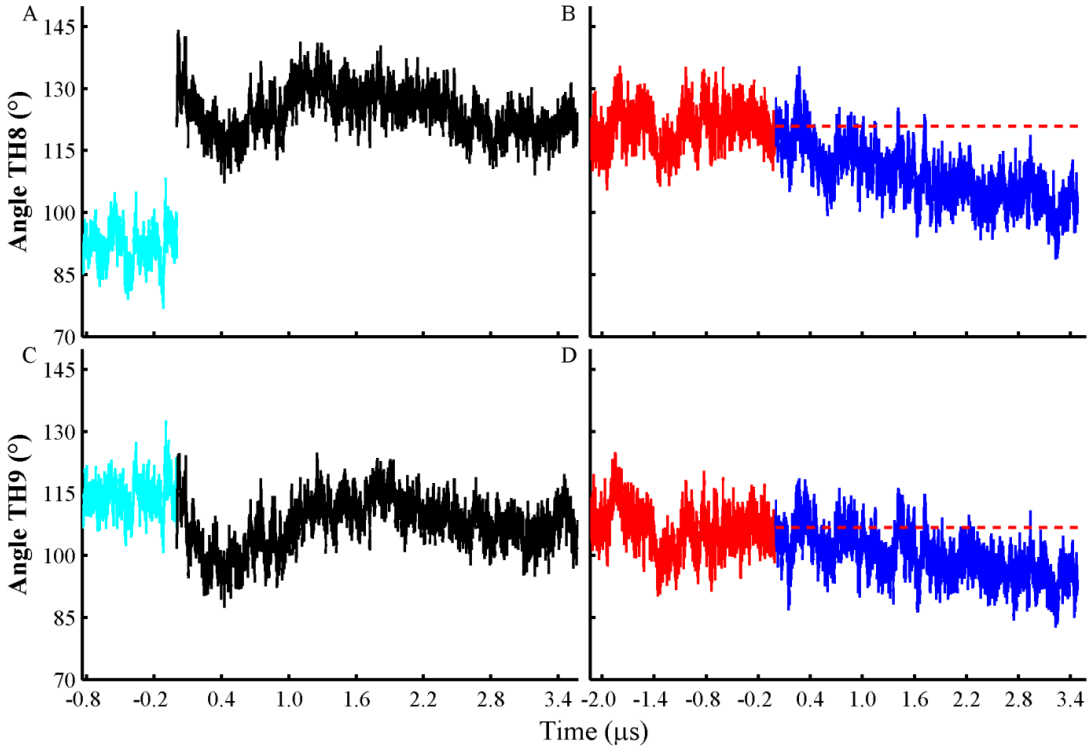

Figure 4.5. Orientation of helices TH8 and TH9. Changes of the angles formed by the axes of helices TH8 (A, B) and TH9 (C, D) relative to the membrane normal axis. Angle traces obtained from membrane bound conformations B1 and B2, at T = 310 K, are shown in cyan and black color, respectively. Simulations at T = 323 K of membrane bound conformation B1 with acidic side-chains in their standard and neutralized state are shown in red and blue colors, respectively. The average values for helices TH8 (121 °) and TH9 (107 °) obtained from the former MD simulation are shown in red broken lines. A value of 90 ° corresponds to a parallel orientation of a helix relative to the membrane plane. All histidines are set in their protonated state in all simulations and other acidic residues are set in their standard state, unless indicated.

4.3.5 Protein-Membrane Interactions

To analyze specific differences of the membrane bound conformations of T-domain in the simulations, we determine the normalized histograms of protein and membrane interface contacts. Figure 4.6 shows the patterns of protein-membrane contacts obtained from the last 500 ns of all MD simulations. Figures 4.6A, B shows that membrane bound conformation B1 and B2 have different patterns of membrane contacts. Particularly, conformation B1 shows that residues in helices TH1-2 and in the loop between TH3-4 form stable contacts with the membrane interface, while the membrane bound conformation B2 forms stable contacts at residues in the loop between TH2-3, TH3, loop TH4-5, TH5, TH6 and loop TH5'-6. Notice that both conformations show similar contacts are observed in the N-terminus of TH8 and in helix TH9. Extended MD simulations of conformation B1 at slightly higher temperature retains most of the protein-membrane contacts (see Figure 4.6C). However, neutralization of acidic side-chains results in the formation of new protein-interface contacts in residues located in the unfolded helix TH5', loop TH5'-6, loop TH8-9 and in helix TH9, as shown in Figure 4.6D. This reorganization of protein-interface of contacts can be depicted by the increased bilayer insertion of residues D295, N296 located in the loop between helices TH5'-6, as shown in Figures 4.S5A, B. Furthermore, residues D352, F355 located in the loop between TH8-9 penetrate the region of lipid tails groups (see Figures 4.S5C, D). Penetration of residues in the lipid tails groups is described in the following paragraph.

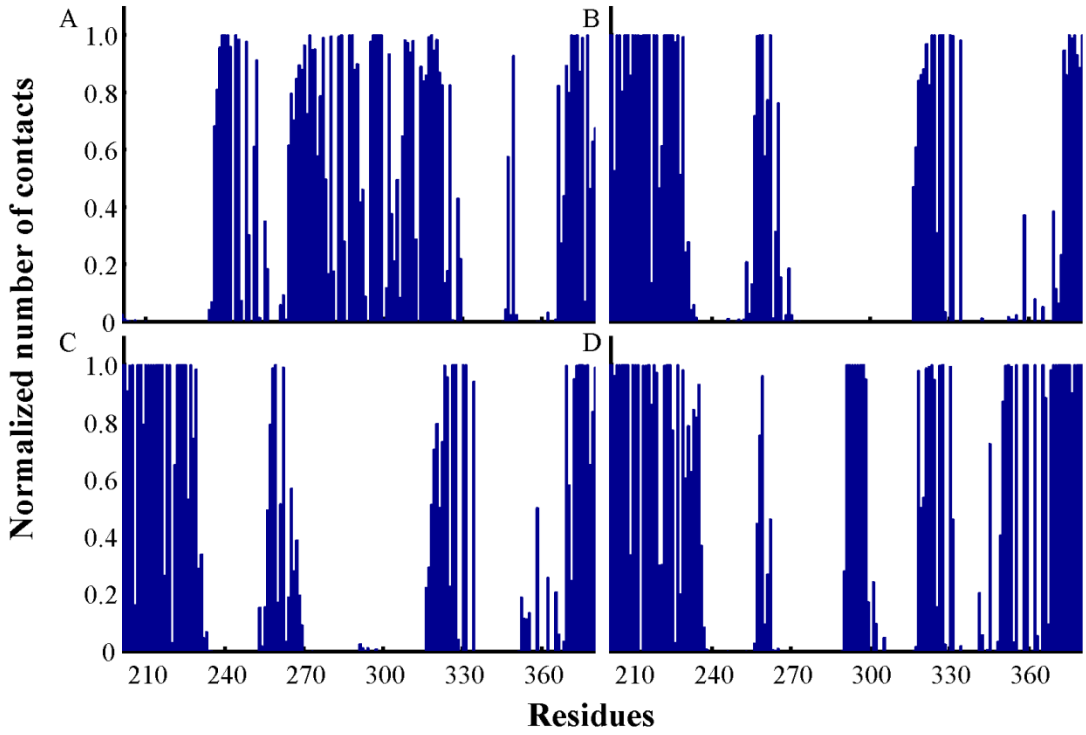

Figure 4.6. Normalized number of protein and membrane interface contacts as a function of residue number. (A) Membrane bound state B2 at T = 310 K. b) Membrane bound state B1 at T = 310 K. (C) Membrane bound state B1 at T = 323 K. (D) Membrane bound state B1 with acidic side-chains neutralized at T = 323 K. Data is obtained from the last 500 ns of each MD simulation. Heavy atoms of protein and headgroup/phosphate separated by a distance lower than 5 Å are considered to be in contact.

Both membrane bound conformations show different extent of interactions with lipid tails, as shown by histograms obtained from the last 500 ns of each simulation. Figure 4.7A shows that conformation B2 forms stable contacts at residues in the loop TH4-5 (residue P271), helix TH6 (residues E298, A302), loop TH6-7 (residues S305, I306, P308), and helix TH9 (residues N366, L367, V370, V371). In contrast, conformation B1 shows smaller number of contacts between the protein and lipid tails

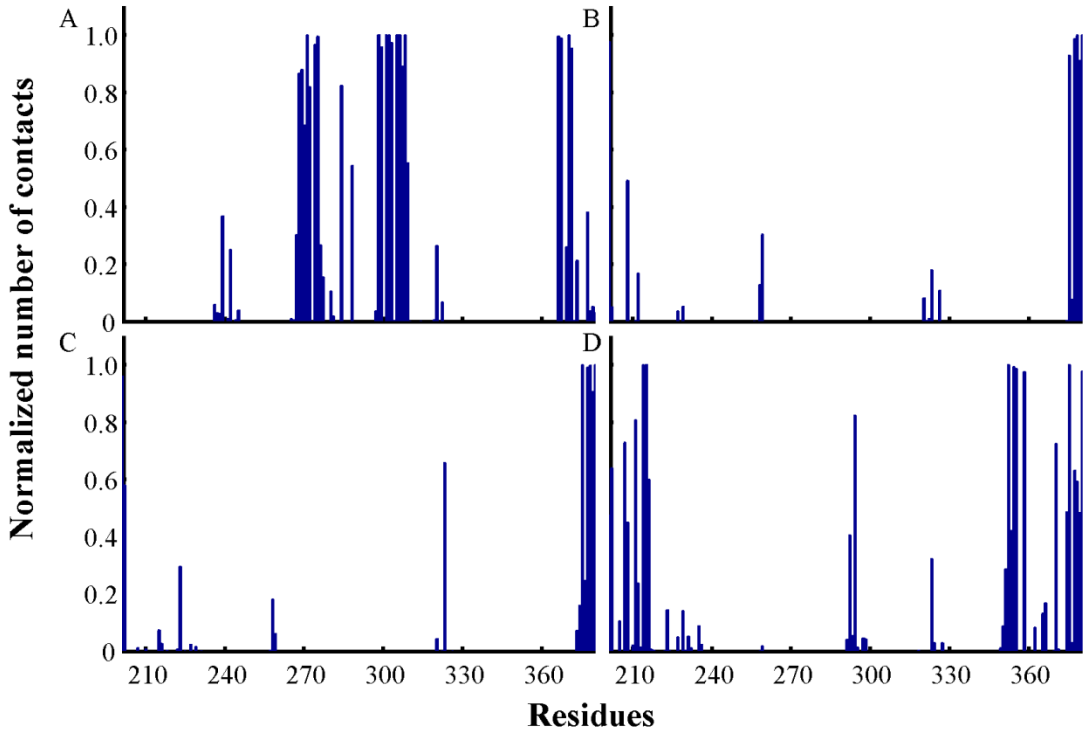

Figure 4.7. Normalized number of protein and lipid tails contacts as a function of residue number. (A) Membrane bound state B2 at T = 310 K. b) Membrane bound state B1 at T = 310 K. (C) Membrane bound state B1 at T = 323 K. (D) Membrane bound state B1 with acidic side-chains neutralized at T = 323 K. Data is obtained from the last 500 ns of each MD simulation. Heavy atoms of protein and lipid tails separated by a distance lower than 5 Å are considered to be in contact.

(see Figures 4.7B, C). However, Figure 4.7D shows that neutralization of acidic side-chains in conformation B1 facilitates the increase of the number of contacts between the protein and the bilayer hydrophobic core increases. As a result, stable membrane contacts are formed in helix TH1 (residues K214, T215), in the loop between TH8-9 (residues D352, G354, F355) and helix TH9 (Y358, Y375). For example, the increased insertion of K214 in the bilayer interior is shown by a normalized density profile in Figure 4.S6A.

The degree of insertion of K214 is similar to that observed in residues P271 and P308 involved in the formation of protein and lipids tails in conformation B2, as shown in Figures 4.S6B, C. Furthermore, neutralization of acidic side-chains in conformation B1 induces increased insertion of V370 in helix TH9 (see Figure 4.S6D). In contrast, conformation B2 shows a slightly deeper insertion of V370 in the bilayer interior (see Figure 4.S6D).

The association of T-domain to the lipid bilayer can induce changes in the behavior of the membrane. Figure 4.S7 shows the distance of the average positions of phosphate atoms (DPP) in the upper and lower leaflets as a function of time. Conformations B1 and B2 induce none or smaller average thickness than that of a pure bilayer at T =310 K, respectively (see Figure 4.S7A). In particular, B2 induces a decrease of ΔDPP = -2 Å relative to a pure bilayer. MD simulation of conformation B1 at T = 323 K reduces the average thickness of the bilayer (ΔDPP = -1 Å), as shown in Figure 4.S7B. Notice that neutralization of acidic side-chains further reduces the membrane thickness (ΔDPP = -2 Å) accompanied by further insertion of the protein (see Figure 4.S7B).

4.3.6 Changes in Lipid Bilayers Behavior

One of the most important properties of lipid bilayers is the area per lipid. In our MD simulation of conformation B1 at T = 310 K, we find decrease of the area per lipid relative to a pure POPG bilayer, while MD simulation of orientation B2 does not show changes in the average area per lipid (see Figure 4.S8A). Furthermore, MD simulation of conformation B1 at a temperature of T = 323 K still induced a decrease of the area per lipid. In contrast, neutralization of acidic side-chains in T-domain induced fluctuations in the area per lipid with an average area similar to pure POPG bilayers (see Figure 4.S8B).

The observed changes in the area per lipid correspond to an increase of order in the lipid tails, which can be described by the increase of the order parameter S_{CD} of lipid tails in simulations of conformation of B1 at T = 310 K and T = 323 K, as shown in Figure 4.S9.

4.4 Discussion and Conclusions

Our atomistic MD simulations have explored the degree of refolding of T-domain bound to lipid bilayers. Previously, extensive atomistic MD simulations, pK_a calculations and spectroscopy experiments have provided evidence of the role of histidines in the formation of a membrane-competent state of T-domain in solution, which exposes hydrophobic and charged sites in the protein surface in preparation for subsequent membrane binding.[20, 47] Following coarse-grained MD simulations suggested the favorable membrane binding propensity of the partially unfolded structure of T-domain. As a result, two preferable membrane bound conformations, B1 and B2, of the low pH T-domain model were predicted.[84] However, the protein tertiary structure in the coarse grained models was restrained. In this study, atomistic MD simulations at T = 310 K showed that conformations B1 and B2 retained their overall structure, orientation and membrane contacts over several hundreds of nanoseconds. In particular, conformation B1 adopts a shallowly insertion depth in the membrane interface with membrane contacts located in protein regions that underwent conformational changes upon protonation of histidines.[20, 47] These regions involve partially unfolded helices TH1-2, the loop between TH3-4 and N-terminal residues in TH8 and helix TH9. In contrast conformation B2 is deeply inserted in the bilayer interface and partially penetrates the hydrophobic core of the lipid bilayer. Conformation B2 deeply insertion is favored by persistent protein – lipid tails interactions at residues in the loop between TH6-7 and residues in helix TH9. The

resulting final structures of the predicted membrane bound conformations B1 and B2 show an oblique and a near parallel orientation of C-terminal helices TH8-9 relative to the membrane plane, respectively.

We analyze the extreme case in which all acidic side-chains in T-domain are neutralized at the membrane interface, while histidines remained protonated. Extended MD simulations of the shallowly inserted conformation B1, at T = 323 K, resulted in partial unfolding of the solvent exposed helix TH5 and deeper insertion in the membrane interface; however, the overall protein orientation and the pattern of membrane contacts remained similar to those observed in the MD simulation at T = 310 K. Neutralization of acidic side-chains favors deeper insertion of the protein and triggers an overall change in the orientation of helices TH8-9 relative to the membrane normal axis. Note that neutralization of acidic side-chains increases the net positive charge of helix TH1 from +1 to +5, which favors deep insertion of helix TH1. Also, the removal of negative charges in the loops between TH5' – TH6 and TH8 – TH9 favors the insertion of these loops. Neutralization of acidic side-chains E349 and D352 has been suggested to facilitate the transmembrane insertion of the hydrophobic hairpin TH8-9. In general, changes in the protonation state of side-chains at the membrane interface have been suggested to occur during the membrane interface binding of T-domain and peptides.[10, 104] This is in agreement pH-dependent conformational changes of T-domain bound to membranes by solid NMR experiments,[13] which showed near parallel and oblique orientations of the protein C-terminal region bound to lipid bilayers composed of POPC:POPG 4:1. Despite the coarse resolution of these experiments, it was shown that N-terminal helices are exposed to the solvent at pH 6. Further acidification of the solution

to pH 4, showed deeply insertion of the C-terminal helices retaining their parallel orientation relative to the membrane plane, while the N-terminal helices formed contacts with the membrane interface. It was suggested that neutralization of acidic side-chains removed repulsive interactions between negatively charged residues and the anionic head-groups.

In summary, atomistic MD simulations of membrane binding modes B1 and B2 showed that the protein remains in a similar conformation predicted by coarse-grained MD simulations. MD simulations of orientation B2 showed that its deep insertion is facilitated by stable contacts of hydrophobic sites of T-domain with the lipid tails and the solvent exposure of the N-terminal helices TH1-2. Furthermore, MD simulations of the shallowly inserted conformation B1 showed that neutralization of acidic side-chains induce changes in the protein insertion depth and overall orientation relative to the membrane plane. A similar feature of both conformations is the oblique or near parallel orientation relative to the membrane plane of the hydrophobic helices TH8-9. Both membrane bound conformations have been hypothesized to be part of the initial stages of the insertion folding pathway of T-domain.[84] Future studies addressing the refolding of the protein on the membrane will require the use of advanced sampling techniques such as accelerated molecular dynamics and replica exchange methods.

4.5 Appendix

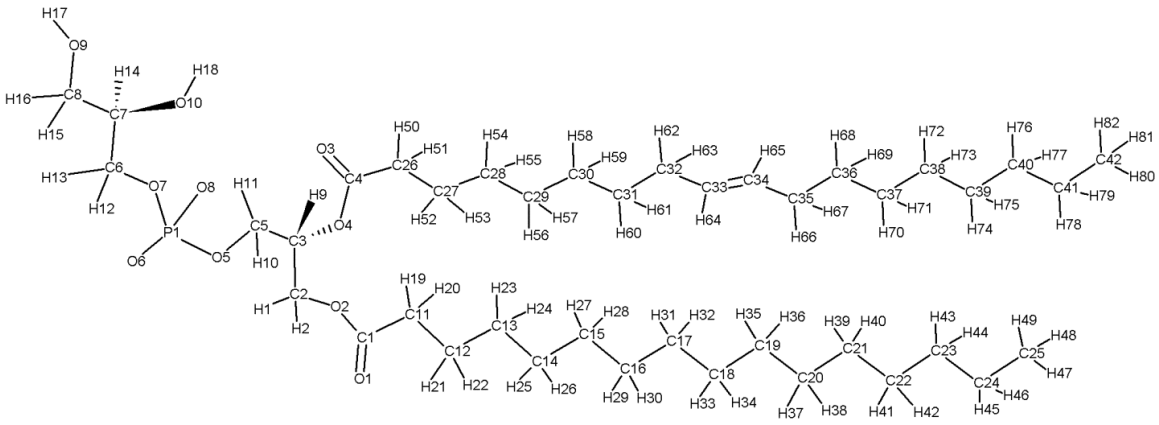

Figure 4.S1. Schematic representation of POPG phospholipids (R-phosphatidylglycerol).

Atom types, charges and Lennard-Jones parameters are shown in Table 4.S1.

Table 4.S1. Atomistic charges of POPG phospholipid (R-phosphatidylglycerol)

Atom name	Atom type	Atom charge
C11	a3	-0.1382
H19	hl	0.0411
H20	hl	0.0411
C12	a3	-0.0119
H21	hl	0.0172
H22	hl	0.0172
C13	a3	-0.0103
H23	hl	0.0163
H24	hl	0.0163
C14	a3	-0.0178
H25	hl	0.014
H26	hl	0.014
C15	a3	-0.0194
H27	hl	0.0015
H28	hl	0.0015
C16	a3	0.0241
H29	hl	-0.0012
H30	hl	-0.0012
C17	a3	-0.0009
H31	hl	-0.0016
H32	hl	-0.0016
C18	a3	0.016
H33	hl	-0.0093
H34	hl	-0.0093
C19	a3	0.0185
H35	hl	-0.0086
H36	hl	-0.0086
C20	a3	0.0062
H37	hl	-0.0023
H38	hl	-0.0023
C21	a3	0.0121
H39	hl	-0.0017
H40	hl	-0.0017
C22	a3	-0.0209
H41	hl	0.0016
H42	hl	0.0016
C23	a3	-0.0109
H43	hl	0.0083

H44	hl	0.0083
C24	a3	0.0319
H45	hl	0.0041
H46	hl	0.0041
C25	a3	-0.094
H47	hl	0.0189
H48	hl	0.0189
H49	hl	0.0189
C1	Cs	0.719
O1	O3	-0.5843
O2	Os	-0.4177
C2	Ct	0.0278
H1	Hh	0.0901
H2	Hh	0.0901
C3	Ct	0.2212
H9	Hh	0.0842
O4	Os	-0.3905
C4	Cs	0.6892
O3	O3	-0.5742
C5	Ct	0.022
H10	Hh	0.0594
H11	Hh	0.0594
O5	Os	-0.4463
O6	O1	-0.7549
P1	Pz	1.2064
O7	Os	-0.504
O8	O1	-0.7549
C6	Ct	0.1262
H12	Hh	0.0287
H13	Hh	0.0287
C7	Ct	0.3137
H14	Hh	-0.023
O10	Oh	-0.6841
H18	Ho	0.4248
C8	Ct	0.1413
H15	Hh	0.0195
H16	Hh	0.0195
O9	Oh	-0.6352
H17	Ho	0.3979
C26	b3	-0.0858
H50	hl	0.0195

H51	hl	0.0195
C27	b3	0.0073
H52	hl	0.0056
H53	hl	0.0056
C28	b3	0.0097
H54	hl	0.008
H55	hl	0.008
C29	b3	-0.0083
H56	hl	0.0113
H57	hl	0.0113
C30	b3	-0.0376
H58	hl	0.0018
H59	hl	0.0018
C31	b3	0.0297
H60	hl	0.0068
H61	hl	0.0068
C32	b3	0.0477
H62	hl	0.0293
H63	hl	0.0293
C33	c2	-0.3071
H64	ha	0.1396
C34	c2	-0.1677
H65	ha	0.1147
C35	b3	0.0513
H66	hl	0.0186
H67	hl	0.0186
C36	b3	-0.0054
H68	hl	-0.0061
H69	hl	-0.0061
C37	b3	0.0183
H70	hl	0.0045
H71	hl	0.0045
C38	b3	-0.0022
H72	hl	0.0077
H73	hl	0.0077
C39	b3	-0.0096
H74	hl	-0.0031
H75	hl	-0.0031
C40	b3	-0.0031
H76	hl	-0.0024
H77	hl	-0.0024

C41	b3	0.0606
H78	hl	-0.0199
H79	hl	-0.0199
C42	b3	-0.0345
H80	hl	0.0064
H81	hl	0.0064
H82	hl	0.0064

Bond, angle and dihedral parameters used in POPG lipids.

BOND
a3-hl 337.30 1.092 same as c3-hc
a3-a3 303.10 1.535 same as c3-c3
b3-hl 337.30 1.092 same as c3-hc
b3-b3 303.10 1.535 same as c3-c3
a3-c2 328.3 1.508 same as c2-c3
b3-c2 328.3 1.508 same as c2-c3
Ct-Hh 340.0 1.090 changed from 331 bsd on NMA nmodes; AA, RIBOSE
Ho-Oh 553.0 0.960 same as HO-OH JCC,7,(1986),230; SUGARS,SER,TYR
O1-Pz 525.0 1.480 same as O2-P JCC,7,(1986),230; NA PHOSPHATES
Os-Pz 230.0 1.610 same as Os-P JCC,7,(1986),230; NA PHOSPHATES
Ct-Os 320.0 1.410 same as CT-OS JCC,7,(1986),230; NUCLEIC ACIDS
Cs-Os 450.0 1.323 same as C -OS Junmei et al, 1999
Cs-O3 570.0 1.229 same as C -O JCC,7,(1986),230; AA,CYT,GUA,THY,URA
Cs-Ct 317.0 1.522 same as C -CT JCC,7,(1986),230; AA
Ct-Oh 320.0 1.410 same as CT-OH JCC,7,(1986),230; SUGARS
Ct-Ct 310.0 1.526 same as CT-CT JCC,7,(1986),230; AA, SUGARS
Cs-a3 317.0 1.522 same as C -CT JCC,7,(1986),230; AA
Cs-b3 317.0 1.522 same as C -CT JCC,7,(1986),230; AA
ANGLE
a3-a3-hl 46.400 110.050 same as c3-c3-hc
a3-a3-a3 63.200 110.630 same as c3-c3-c3
hl-a3-hl 39.400 108.350 same as hc-c3-hc
b3-b3-hl 46.400 110.050 same as c3-c3-hc
b3-b3-b3 63.200 110.630 same as c3-c3-c3
hl-b3-hl 39.400 108.350 same as hc-c3-hc
a3-a3-c2 63.7 110.96 same as c2-c3-c3
a3-c2-c2 64.3 123.42 same as c2-c2-c3
a3-c2-hl 45.1 120.00 same as c3-c2-hc
c2-a3-hl 47.0 110.49 same as c2-c3-hl
b3-b3-c2 63.7 110.96 same as c2-c3-c3
b3-c2-c2 64.3 123.42 same as c2-c2-c3
b3-c2-hl 45.1 120.00 same as c3-c2-hc
c2-b3-hl 47.0 110.49 same as c2-c3-hl
ha-c2-b3 45.660 117.300 same as c3-c2-ha

ha-c2-a3 45.660 117.300 same as c3-c2-ha
Ct-Oh-Ho 55.0 108.50 same as CT-OH-HO
Hh-Ct-Oh 50.0 109.50 same as H1-CT-OH changed based on NMA nmodes
Ct-Ct-Hh 50.0 109.50 same as CT-CT-H1 changed based on NMA nmodes
Hh-Ct-Hh 35.0 109.50 same as H1-CT-H1
Ct-Ct-Oh 50.0 109.50 same as CT-CT-OH
Ct-Ct-Os 50.0 109.50 same as CT-CT-OS
Hh-Ct-Os 50.0 109.50 same as H1-CT-OS changed based on NMA nmodes
Ct-Os-Pz 100.0 120.50 same as CT-OS-P
O1-Pz-O1 140.0 119.90 same as O2-P -O2
O1-Pz-Os 100.0 108.23 same as O2-P -OS
Os-Pz-Os 45.0 102.60 same as OS-O -OS
Cs-Os-Ct 60.0 117.00 same as C -OS-CT Junmei et al, 1999
O3-Cs-Os 80.0 125.00 same as O -C -OS Junmei et al, 1999
a3-Cs-O3 80.0 120.40 same as CT-C -O
b3-Cs-O3 80.0 120.40 same as CT-C -O
a3-Cs-Os 80.0 115.00 same as CT-C -OS Junmei et al, 1999
b3-Cs-Os 80.0 115.00 same as CT-C -OS Junmei et al, 1999
Cs-a3-a3 63.0 111.10 same as C -CT-CT AA general
Cs-b3-b3 63.0 111.10 same as C -CT-CT AA general
Cs-a3-hl 50.0 109.50 same as C -CT-HC AA general changed based on NMA nmodes
Cs-b3-hl 50.0 109.50 same as C -CT-HC AA general changed based on NMA nmodes
Ct-Ct-Ct 40.0 109.50 same as CT-CT-CT
DIHE
hl-a3-a3-hl 1 0.156 0.000 3.000 same as X -c3-c3-X
a3-a3-a3-hl 1 0.156 0.000 3.000 same as X -c3-c3-X
a3-a3-a3-a3 1 -0.0866 180.0001 1.000 paramfit
a3-a3-a3-a3 1 -0.1109 180.0001 2.000 paramfit
a3-a3-a3-a3 1 0.1352 0.000 3.000 paramfit
b3-b3-b3-b3 1 -0.0866 180.0001 1.000 paramfit
b3-b3-b3-b3 1 -0.1109 180.0001 2.000 paramfit
b3-b3-b3-b3 1 0.1352 0.000 3.000 paramfit
hl-b3-b3-hl 1 0.156 0.000 3.000 same as X -c3-c3-X
b3-b3-b3-hl 1 0.156 0.000 3.000 same as X -c3-c3-X
hl-b3-b3-b3 1 0.156 0.000 3.000 same as X -c3-c3-X
c2-b3-b3-b3 1 0.2109 0.000 3.000 paramfit
c2-b3-b3-hl 1 0.156 0.000 3.000 same as X -c3-c3-X
c2-c2-b3-b3 1 0.000 0.000 2.000 same as X -c2-c3-X
c2-c2-b3-hl 1 0.000 0.000 2.000 same as X -c2-c3-X
c2-a3-a3-a3 1 0.2109 0.000 3.000 paramfit
c2-a3-a3-hl 1 0.156 0.000 3.000 same as X -c3-c3-X
c2-c2-a3-a3 1 0.000 0.000 2.000 same as X -c2-c3-X
c2-c2-a3-hl 1 0.000 0.000 2.000 same as X -c2-c3-X
a3-a3-c2-ha 6 0.000 0.000 2.000 same as X -c2-c3-X

```
hl-a3-c2-ha   6    0.000        0.000         2.000      same as X -c2-c3-X
b3-b3-c2-ha   6    0.000        0.000         2.000      same as X -c2-c3-X
ha-c2-b3-hl   6    0.000        0.000         2.000      same as X -c2-c3-X
X -Ct-Oh-X    3    0.50         0.0           3.         same as X -CT-OH-X JCC,7,(1986),230
X -Ct-Ct-X    9    1.40         0.0           3.         same as X -CT-CT-X JCC,7,(1986),230
X -Ct-Os-X    3    1.15         0.0           3.         same as X -CT-OS-X JCC,7,(1986),230
X -Os-Pz-X    3    0.75         0.0           3.         same as X -OS-P -X JCC,7,(1986),230
X -Cs-Os-X    2    5.40         180.0         2.         same as X -C -OS-X Junmei et al, 1999
X -Cs-a3-X    6    0.00         0.0           2.         same as X -C -CT-X JCC,7,(1986),230
X -Cs-b3-X    6    0.00         0.0           2.         same as X -C -CT-X JCC,7,(1986),230
Ct-Ct-Os-Cs   1    0.383        0.0           -3.        same as CT-CT-OS-C Junmei et al, 1999
Ct-Ct-Os-Cs   1    0.80         180.0         1.         same as CT-CT-OS-C Junmei et al,
1999
Os-Ct-Ct-Os   1    0.144        0.0           -3.        same as OS-CT-CT-OS parm98,
TC,PC,PAK
Os-Ct-Ct-Os   1    1.175        0.0           2.         same as OS-CT-CT-OS Piotr et al.
Oh-Ct-Ct-Oh   1    0.144        0.0           -3.        same as OH-CT-CT-OH parm98,
TC,PC,PAK
Oh-Ct-Ct-Oh   1    1.175        0.0           2.         same as OH-CT-CT-OH parm98,
TC,PC,PAK
Os-Ct-Ct-Oh   1    0.144        0.0           -3.        same as OS-CT-CT-OH parm98,
TC,PC,PAK
Os-Ct-Ct-Oh   1    1.175        0.0           2.         same as OS-CT-CT-OH parm98,
TC,PC,PAK
O3-Cs-Os-Ct   1    2.70         180.0         -2.        same as O -C -OS-CT Junmei et al,
1999
O3-Cs-Os-Ct   1    1.40         180.0         1.         same as O -C -OS-CT Junmei et al,
1999
Cs-b3-b3-b3   1    0.156        0.0           3.         same as X-CT-CT-X JCC,7,(1986),230
THIS
Cs-b3-b3-hl   1    0.156        0.0           3.         same as X-CT-CT-X JCC,7,(1986),230
THIS
Cs-a3-a3-a3   1    0.156        0.0           3.         same as X-CT-CT-X JCC,7,(1986),230
THIS
Cs-a3-a3-hl   1    0.156        0.0           3.         same as X-CT-CT-X JCC,7,(1986),230
THIS
IMPROPER
X -X -Cs-O3         10.5         180.          2.         same as X -X -C -O JCC,7,(1986),230

NONBON
  a3          2.01    0.055          fit to exp
  b3          2.01    0.055          fit to exp
  Ct          2.01    0.055          C36
  Cs          2.0     0.07           C36
  hl          1.34    0.024          fit to exp
  Hh          1.34    0.024          C36
```

Ho	0.2245	0.046	C36
O1	1.7	0.12	C36
O3	1.7	0.12	C36
Os	1.65	0.1	C36
Oh	1.77	0.1521	C36
Pz	2.10	0.585	C36

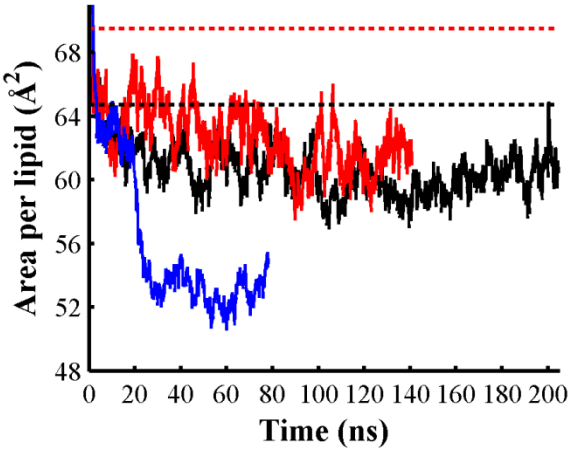

Figure 4.S2. Changes of area per lipid obtained from simulations of POPG alone as a function of time. Area per lipid traces obtained from simulations at T = 310 K (black line) and T = 323 K (red line). Area per lipid curve obtained from POPG using original GaffLipid parameters at T = 310 K is shown in blue line. Broken lines represent the experimental values at T = 310 K (64.7 Å^2) and T = 323 K (69.5 Å^2) shown in black and red broken lines, respectively.

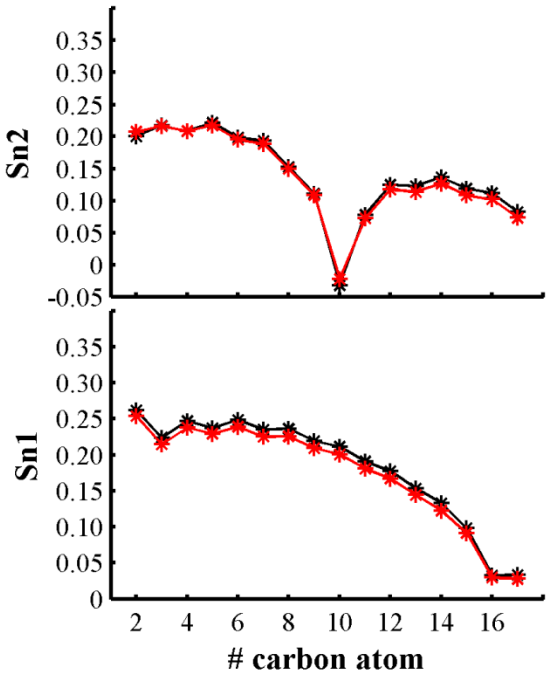

Figure 4.S3. Order parameters S_{CD} of the sn-1 (saturated) and sn-2 (unsaturated) chains of POPG obtained from simulations at T = 310 K (black markers) and T = 323 K (red markers).

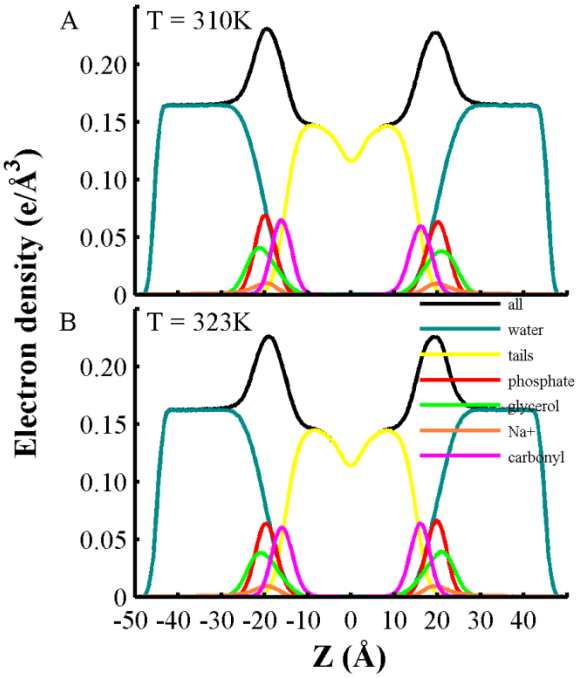

Figure 4.S4. Electron density of all and individual chemical groups obtained from simulations of bilayers of POPG phospholipids at temperatures (A) T = 310 K and (B) T = 323 K. Averages are obtained from the last 150 ns and 90 ns from each simulation, respectively. Density profiles from water, lipid tails, phosphate, headgroup glycerol, counterions and carbonyl groups are shown in cyan, yellow, red, green, orange and magenta lines, respectively.

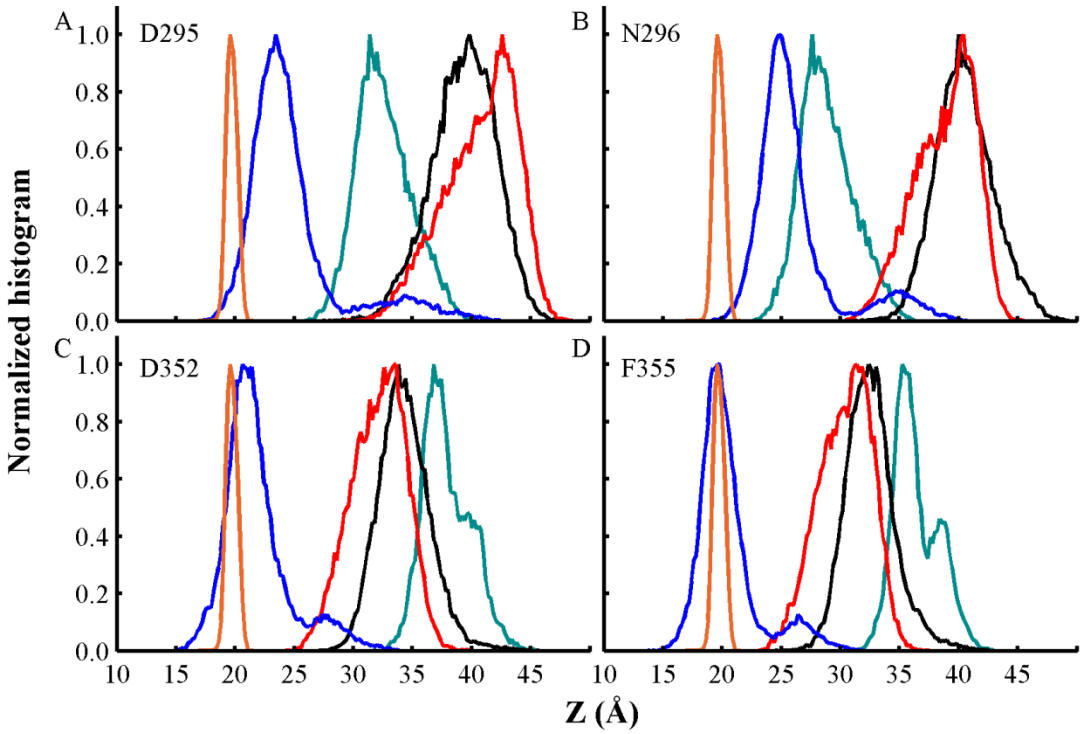

Figure 4.S5. Normalized density of insertion depth of C_α atoms from residues: (A) D295 (B) N296 (C) D352 and (D) F355. Phosphate group density computed over all MD simulations is shown as a reference in orange lines. The bilayer center is located at Z = 0. Histograms from membrane bound conformations B1 and B2 at T = 310 K are shown in dark cyan and black color, respectively. Simulations at T = 323 K of membrane bound conformation B1 with acidic side-chains in their standard and neutralized state are shown in red and blue colors, respectively. Data was obtained from the last 500 ns of each simulation.

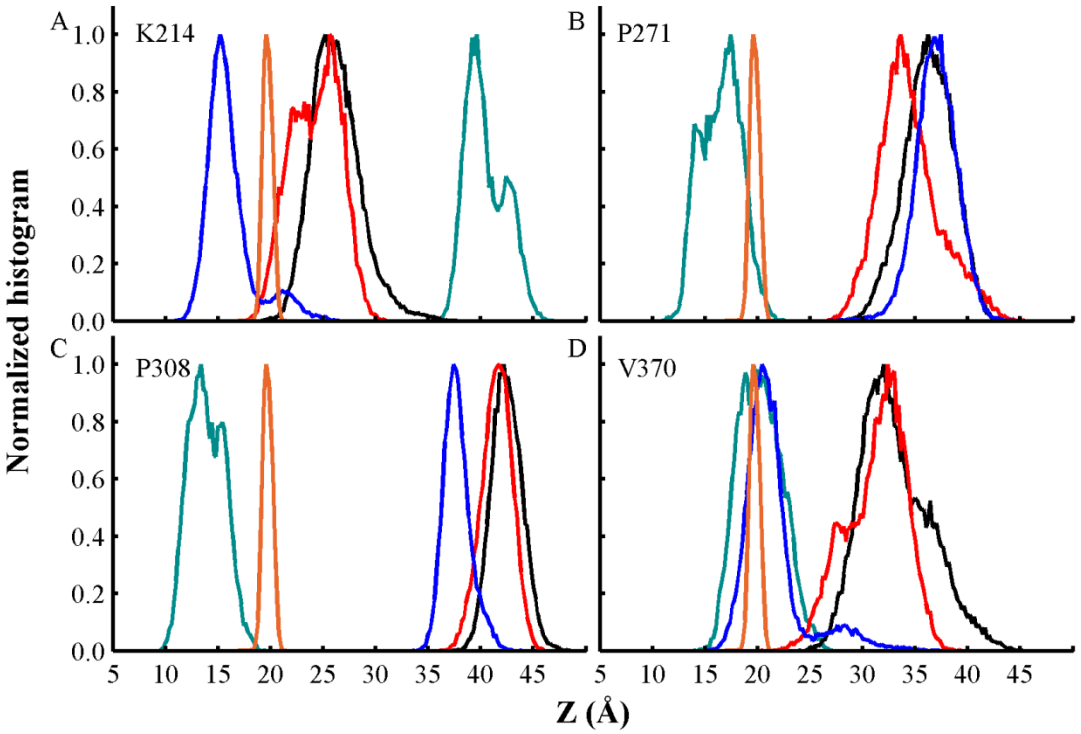

Figure 4.S6. Normalized density of insertion depth of C_α atoms from residues: (A) K214 (B) P271 (C) P308 and (D) V370. Phosphate group density computed over all MD simulations is shown as a reference in orange lines. The bilayer center is located at Z = 0. Histograms from membrane bound conformations B1 and B2 at T = 310 K are shown in dark cyan and black color, respectively. Simulations at T = 323 K of membrane bound conformation B1 with acidic side-chains in their standard and neutralized state are shown in red and blue colors, respectively. Data was obtained from the last 500 ns of each simulation.

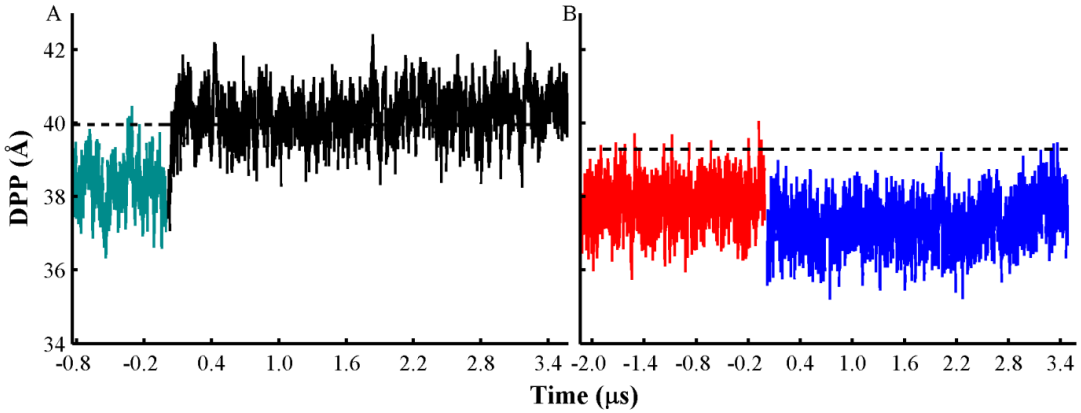

Figure 4.S7. Changes of distance between average positions of phosphorous atoms (DPP) between the top and bottom leaflet. (A) DPP traces obtained from orientation B1 (black line) and orientation B2 (cyan). Simulations were performed at T = 310 K. (B) DPP traces obtained from orientation B1 (red) and with neutralized acidic side-chains (blue lines). Simulations were performed at T = 323 K. All histidines are set in their protonated state in all simulations and other acidic residues are set in their standard state, unless indicated. Averages DPP for each simulation are 38, 40, 38 and 37 Å, respectively. Broken lines represent the average area per lipid obtained from simulations of bilayers containing POPG alone at the corresponding temperatures (40 Å at 310 K and 39 Å at 323 K).

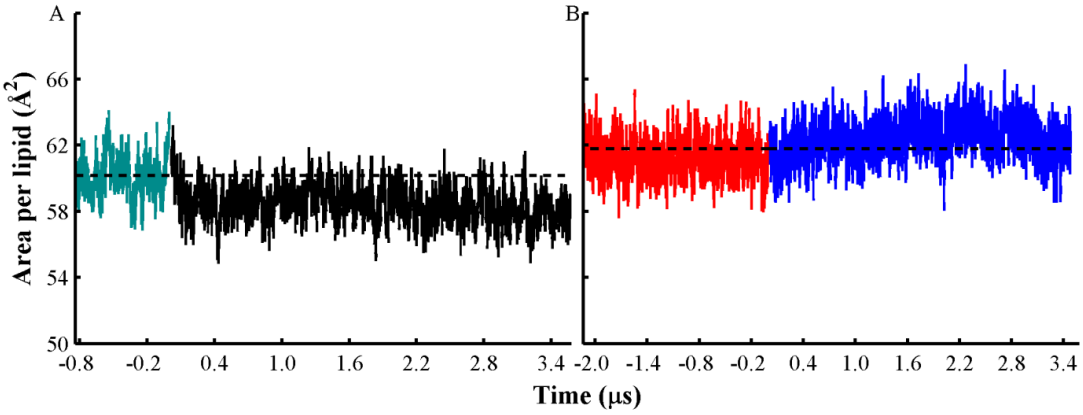

Figure 4.S8. Changes of area per lipid as a function of time: (A) Area per lipid traces obtained from orientation B1 (black line) and orientation B2 (cyan). Simulations were performed at T = 310 K. (B) Area per lipid curves obtained from orientation B1 (red) and with neutralized acidic side-chains (blue lines). Simulations were performed at T = 323 K. All histidines are set in their protonated state in all simulations and other acidic residues are set in their standard state, unless indicated. Averages area per lipid obtained from each simulation are 60, 58, 61 and 62 Å^2, respectively. Broken lines represent the average area per lipid obtained from simulations of bilayers containing POPG alone at the corresponding temperatures (60 Å^2 at 310 K and 62 Å^2 at 323 K).

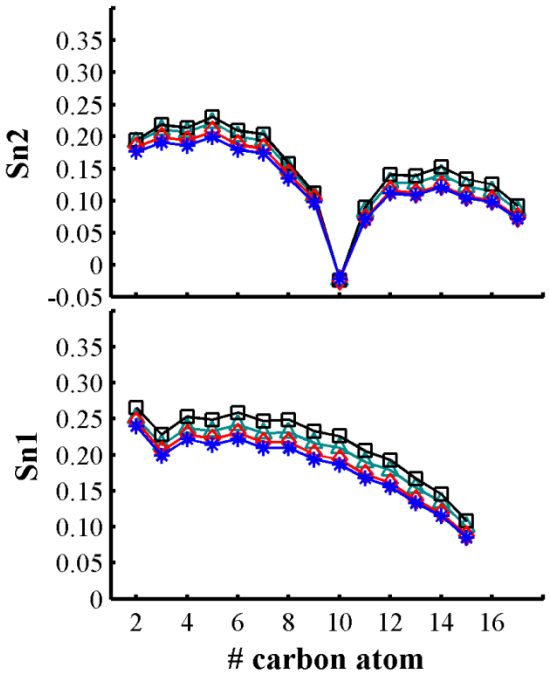

Figure 4.S9. Order parameters S_{CD} of the sn-1 and sn-2 chains of POPG in our simulations of T-domain bound to the membrane. Order parameters obtained from orientation B1 (black line) and orientation B2 (cyan) at T = 310 K. Parameters obtained from orientation B1 (red) and with neutralized acidic side-chains (blue lines) at T = 323 K. All histidines are set in their protonated state in all simulations and other acidic residues are set in their standard state, unless indicated.

Chapter 5. Empirical Prediction Method for Packing of Transmembrane Helices with Rigid Body Monte-Carlo Sampling

5.1 Introduction

Membrane proteins are involved in a number of important processes in living cells such as signal transduction, cell intoxication, and ion transport across lipid bilayers. To understand the relationship of structural dynamics of membrane proteins and their function, high resolution structures of membrane proteins are needed for atomistic molecular dynamics (MD) methods. However, the number of unique known membrane protein structures to date is 495 (accessed on August, 2014),[105] which is around 2 % of protein structures in the PDB database bank.[106, 107] This disparate proportion between the number of known membrane and globular protein structures is due to the difficulties of expression, stabilization, and crystallization of membrane proteins in mimetic environments.[108] Computational methods can offer an alternative pathway for the rapid construction of near-native structures of membrane protein, which could be used for further refinement using experimental constraints and atomistic modeling of the protein dynamics.

The interactions that participate in the assembly of transmembrane (TM) helices can be divided into protein-protein and protein-lipid interactions. The former one is guided by van der Waals, salt-bridges, interhelical hydrogen bonds, and complementary surfaces originated by sequence motifs.[107, 109, 110] Protein-lipid interactions are driven by non-specific interactions between residues exposed to the hydrophobic core, aromatic residues residing in the membrane interface, and the hydrophobic mismatch between the

length of a transmembrane helix and the width of the membrane.[111-113] This mismatch was found to determine the interhelical and tilt angles of transmembrane helices.[113] Overall the relative contribution of all interactions is still unkown and actively discussed in the scientific community.[113] Thus, current computational methods focused to prediction of helical TM structures rely on structural knowledge extracted from the small set of known membrane protein structures,[114, 115] evolutionary constraints obtained by covariance analysis of pairs of amino-acid sequence positions,[116] and on all-atom physics based models.[117, 118] Due to its simplicity and success in low resolution structure prediction, knowledge based potentials are used in several computational structure prediction protocols aimed to predict the structure of small globular proteins from sequence such as Rosetta[119] and I-TASSER.[120] Knowledge based potentials rely on the assumption that the distribution of different pairs of residues along a given reaction coordinate follows a Boltzmann distribution.[121-123] Barth et. al.[115] reported a Rosetta based method able to predict membrane protein structures, which used a pairwise knowledge based potential obtained from water soluble proteins in addition to an implicit solvent model of the membrane.[115] Recently, a promising prediction method used scoring functions obtained from globular proteins in conjunction with knowledge based scoring terms calculated from a small set of membrane protein structures.[114] However, the small number of membrane protein structures with low sequence similarity can affect the calculation of reliable statistical potentials.

To generate near-native structural models of transmembrane proteins, we propose a coarse-grained (CG) scoring energy function based on a residue-based pairwise potential obtained from available globular proteins. In addition, a simplified model of the

membrane is constructed based on geometrical restraints and using a 'biological' partition scale for amino-acids.[124] A rigid-body Monte-Carlo method is implemented to sample exhaustively the conformational space of transmembrane helices.

In this chapter, we describe the energy terms of the proposed scoring function and the rigid-body Monte-Carlo sampling method. Computational details of the coarse-grained model homodimers are presented. The sampling of the conformational space of a set of four homodimers is presented in the section **5.3** ***Results***.

5.2 Methods

5.2.1 Metropolis Monte-Carlo method

Metropolis Monte-Carlo (MC) is an importance sampling method that generates an ensemble according to the Boltzmann factor.[125] MC simulations were performed in order to explore the conformational space that is accessible by translational and rotational random steps of rigid helices. In all simulations reported protein transmembrane alpha-helices were held rigid. The rigid body Metropolis MC simulator was implemented in HARLEM program.[126]

5.2.2 Move Sets

Trial conformations of rigid bodies are generated by a basic move set of a small translational and rotational displacement of a single helix per MC step, see Figure 5.1. The maximum step size for both type of displacements are to follow a well known criterion of the Metropolis MC acceptance ratio. This criterion establishes that for an optimum sampling the acceptance ratio should fluctuate around 50 % of trial moves should be accepted. However, from our experience this high ratio of acceptance will generate

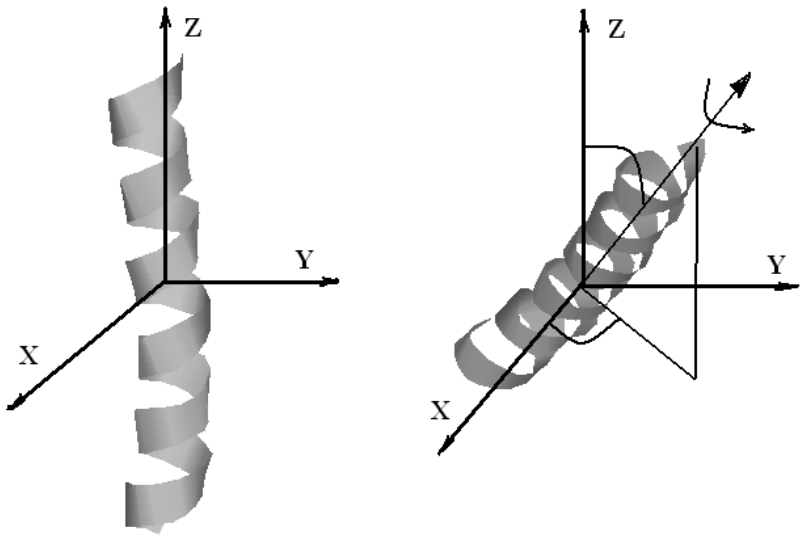

Figure 5.1. Degrees of freedom of rigid transmembrane helices. First helix is kept fixed while the second helix is rotated and translated. Rotation angles are represented by curved lines.

random structures with small conformational fluctuations. Moreover, analysis of MC trajectories showed that low energy conformations of different energy basins are conformational different by the rotation around the principal axis of a rigid helix, see Figure 5.1.

In order to overcome high energy barriers, we implemented a move set with scheduled frequencies in which a single of degree of freedom (DOF) is modified over 70 % of all MC trials, while trial moves involving all DOF are less frequently perturbated (30 % of all MC trials). Table 5.S1 shows the details of frequency of all move trials. The frequency of the move trials were adjusted by achieving uniform sampling of the conformational space of two transmembrane helices with only membrane plane restraints, volume restraint and steric interactions. The acceptance ratio was approximately 50 %.

The purpose of this move set is to randomly generate trial conformations with higher probability to jump over the energy barrier.

5.2.3 Scoring Energy Function

The intermolecular energy E_{int} of a structure model in this work is calculated as

$$E_{int} = w_1 E_{pairwise} + w_2 E_{steric} + w_3 E_{mem} + w_4 E_{vol} + w_5 E_{solv}, \quad (5.1)$$

where $E_{pairwise}$ is a pairwise distance-dependent potential of mean force of interaction between residues *i* and *j*, E_{steric} is the steric overlap energy among super-atoms, E_{mem} is the energy term that constrains the TM helices in the membrane plane and E_{vol} prevents the helix mass center to sample farther than a radius of 15 Å. $E_{solvation}$ accounts for the interaction of amino acids in different regions in the lipid bilayer. A description of each of these terms follows:

A. Pairwise Interaction Energy. $E_{pairwise}$ energy term is the distance dependent interaction energy between geometric centroids of two sidechains dictated by the DFIRE formalism.[127] This distance dependent energy term is extracted from the pairwise residue distribution in water soluble proteins of known structure. This was motivated by the underrepresentation of high resolution membrane protein structures in comparison with the number of known water soluble proteins. We additionally calculated the pairwise residue potential from aminoacids present in the core of soluble proteins based on the assumption that the internal protein structure is similar in both soluble and membrane proteins and small differences are due to solvation energy effects. However, we found

that the number of polar residues in the core of globular proteins is insufficient for significant statistics for the DFIRE procedure (data not shown).

DFIRE knowledge potential has been successful to identify native-like proteins in comparison with similar knowledge based potentials.[127] The reason of the success of DFIRE relative to similar empirical potentials relies on the reference state that is based on ideal gas confined in a finite protein-size sphere. The reference state accounts for the random probability of pairwise interactions that do not give information about specific residue-residue interactions. Larriva and co-workers performed minimization simulations of rigid alpha helices with different coarse grained interaction centers such as alpha carbons, beta carbons or side-chain model (SCM), see Figure 5.2.[128, 129] It was found that DFIRE-SCM has better predictive scoring energies than the other coarse grained interaction centers.

The residue-residue specific interactions are calculated following the DFIRE-SCM formalism, described as follows:

$$E(i,j,r) = -\eta RTln\frac{N_{obs}(i,j,r)}{(r/r_{cut})^{\alpha}(\Delta r/\Delta r_{cut})N_{obs}(i,j,r_{cut})}, \qquad (5.2)$$

where η is a constant to facilitate quantitative comparison with experimental values; R is gas constant and T is the absolute temperature; i and j represent two interaction centers separated by distance r; $N_{obs}(i,j,r)$ is the observed number of pair i,j within distance shell $r - \Delta r/2$ to $r + \Delta r/2$ in the database of folded structured used; r_{cut} is the cut-off distance for the potential, Δr_{cut} is the bin width at r_{cut}; α is an exponent related to the finite size of the folded structures.

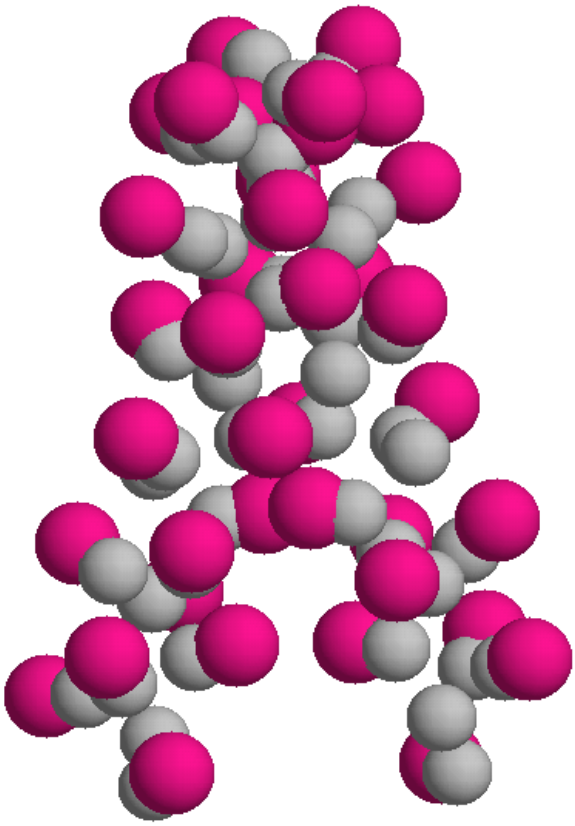

Figure 5.2. Representation of rigid transmembrane helices. Cα atoms are shown in grey space-filled representation. Sidechain atoms are represented by a single super-atom shown in magenta space-filled representation.

The parameters values are: cut-off distance r_{cut} = 14.5 Å, the bin width Δ□ is set to 2 Å for r < 2 Å, 0.5 Å for 2 Å < r < 8 Å, and 1Å for 8 Å < r < 15 Å.[17] The value of α = 1.61 as reported by Zhou et al.[127]

The weight factor w_1 was adjusted until the near native structure of Glycophorin A dimer (GpA) is predicted with the lowest scoring energy. w_1 is set to 0.5 for interactions of Gly, Ser, Ala, Val, Thr, Asp, Gln and Asn. The interactions among these residues are assumed to be stronger to the possibility to form hydrogen bonds and

because of their small size such as Gly, Ala and Val. Otherwise, the weight factor is set to 0.2.

B. Repulsion and restraint energy terms. Residue-residue energy interactions are discretized over a range of 15 Å. In the range of 0 to 2.0 Å the potential lacks of a repulsive interaction that for MC sampling purposes will result in the acceptance of highly overlapped conformations. To prevent the acceptance of these structures by the sampling method, we introduced the energy repulsion E_{steric} as follows:

$$E_{steric} = \begin{cases} 0 \\ \eta \frac{\sigma_{ij}^4}{r^4}, \end{cases} \quad (5.3)$$

where σ_{ij} is the van der Waals diameter for residues *i* and *j*. The van der Waals radii are taken from elsewhere.[130] *r* is the distance between residues *i* and *j*. η is the constant used in DFIRE formalism. The behavior of this functional form will prevent the acceptance of conformations slightly overlapped which for sampling efficiency will be desirable. For all MC simulations the repulsive term given by equation 5.3 will be used in the scoring function, otherwise it will be noted. A hard-sphere energy term is included in the van der Waals repulsion energy term for distances lower than half the value of the sum of the residues radii.

The weight factor w_2 was adjusted similarly to w_1 and it is set to 0.5 for interactions between Cα atoms. It is set to 4.0 for interactions between Cα atom of Gly and the side-chain representative atom of other residues. Otherwise is set to 1.0. Figure 5.3 shows the pairwise interactions of Leu-Leu and Gly-Gly residues. The pairwise interaction of Gly-Gly is of special interest due to its influence in helix-helix interactions of Glycophorin A dimer (GpA). It has been suggested that due to its small size Glycine residues allows a strong packing between the two helices of Glycophorin A.[131] From

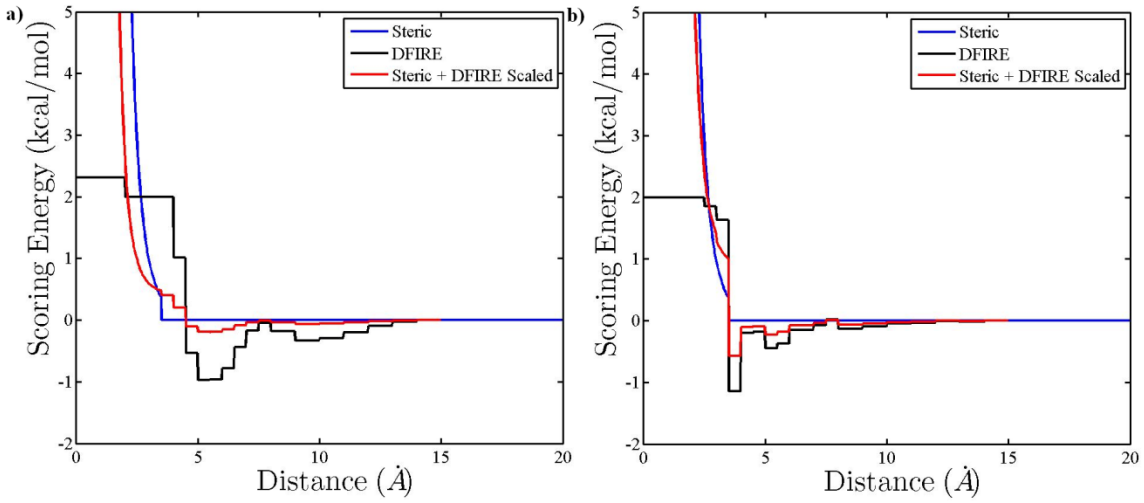

Figure 5.3. Pairwise energy interactions of residues. a) Leu-Leu potential of mean force (black line). b) Gly-Gly potential of mean force (black line). Steric energy term is given by equation 5.3 (blue line). The sum of steric and scaled DFIRE energy term is shown in red line.

Figure 5.3, we see that the pairwise interaction of Gly-Gly has a minima around 3.8 Å and Gly-Leu has a minima at 4.5 Å of the interaction distance. This is in good agreement with the van der Waals diameter of the residue level representation.

The energy term E_{mem} refers to the energy term that restraints the N and C terminals of TM helices to be located in the membrane plane.

$$E_{mem} = \begin{cases} 0 \\ (|z| - z_0)^2, \end{cases} \qquad (5.4)$$

where z is the position of the TM terminals along the axis normal to the plane of the membrane bilayer. z_0 is the equilibrium position of the TM terminals with respect to the

center of the bilayer core. z_0 takes values of 12.0 or 14.0 Å. The value of w_3 is set to 10.0.

$$E_{solv} = \begin{cases} 0 \\ (R_{max} - r_i)^2, \end{cases} \quad (5.5)$$

where R_{max} is the radius of the sphere in which TM helices are able to sample, its value for all simulations is 20.0 Å. r_i is the distance of the center of mass of the helix *i*. The value of w_4 is set to 1.0.

C. Solvation Energy. Recently, Hessa et al proposed a 'biological' hydrophobicity scale for amino acids based on experimental determination of the apparent free energy of insertion of a transmembrane helix into the membrane of the endoplasmatic reticulum.[132] Furthermore, Hessa et al reported a 'biological' apparent free energy of insertion for amino acids that varies with the amino acid position in a transmembrane helix.[124] This proposed free energy of insertion is added to our model to account for the interactions of amino acids within the membrane. This acts as a major driving force during the processes of insertion and folding of transmembrane helices in the lipid bilayer.

The insertion free energy term is added for each amino acid that has solvent accessible surface area (SASA) larger than zero. The probe radius is 1.9 Å.

The insertion free energy term for Tryptophan and Tyrosine:

$$E_{app}^{aa} = a_0^{aa} e^{-\mathrm{a}_1^{\mathrm{aa}}\mathrm{i}^2}, \quad (5.6)$$

Otherwise:

$$E_{app}^{aa} = a_0^{aa} e^{-\mathrm{a}_1^{\mathrm{aa}}\mathrm{i}^2} + \mathrm{a}_2^{\mathrm{aa}}\left(e^{-\mathrm{a}_3^{\mathrm{aa}}(\mathrm{i}-\mathrm{a}_4^{\mathrm{aa}})^2} + e^{-\mathrm{a}_3^{\mathrm{aa}}(\mathrm{i}+\mathrm{a}_4^{\mathrm{aa}})^2}\right), \quad (5.7)$$

where a_0^{aa}, a_1^{aa}, a_2^{aa}, a_3^{aa}, and a_4^{aa} are the set of parameters obtained for each aminoacid type aa,[124] i indicates the position of the aminoacid in the axis perpendicular to the

Table 5.1: Set of four TM homo-dimer proteins

Membrane Protein	PDB Code	Residues used for MC simulations
Glycophorin A	1AFO	$I^{(73)}$TLIIFGVMAGVIGTILLISYGI$^{(95)}$
BNIP3	2J5D	$V^{(164)}$FLPSLLLSHLLAIGLGIYIG$^{(184)}$
EphA1 (pH= 4.3)	2K1K	$E^{(547)}$IVAVIFGLLLGAALLLGIL$^{(566)}$
ErbB2	2JWA	$T^{(52)}$SIISAVVGILLVVVLGVVFGI$^{(73)}$

membrane plane, E_{app}^{aa} is the apparent free energy of insertion for the amino acid type aa.

$$E_{mem} = \begin{cases} 0 \\ \sum_{aa} E_{app}^{aa}, \end{cases} \qquad (5.8)$$

The weight factor w_5 is set to 1.0.

5.2.4 Analysis

The C_α root mean square deviation (RMSD) is obtained for all MC trajectories relative to the native structure of each TM protein.

5.3 Results

We test the scoring function ability to identify near-native models of three known TM structures by ten independent Metropolis Monte-Carlo simulations. The number of MC steps per simulation is 10^5. Table 5.1 shows the PDB codes of Glycophorin A, BNip3, EphA1 and ErbB2 in the second column. The third column contains the amino-acids used in each TM model. These amino-acids are located in the hydrocarbon region of the lipid bilayer according to the OPM database.[133] This set of TM homodimers does not contribute to the amino-acid pair statistics, instead only water soluble protein structures determined by X-ray diffraction contribute to DFIRE residue-pairwise statistics. However, GpA is used to adjust the values of the weighting factors of pairwise

and steric interactions and the other three homodimers are used for testing. A brief description of each TM protein tested and the respective results of the scoring energy are given below.

5.3.1 Glycophorin A

The human red blood cell Glycophorin A protein has been the subject of intensive investigation in order to understand helix-helix interactions in the trans-membrane region of lipid bilayers. Early mutation studies showed a sequence dependent motif formed by residues GxxxG that mediates favorable association of GpA monomers, where x can be other amino-acids.[134] This result was corroborated by the determination of the NMR structure model of GPA solvated on micelles.[131] The GxxxG motif allows the formation of close intermolecular backbone-backbone contacts, which consists of the packing of the surface formed by Gly residues against the one formed by the side-chains of the other helix. The NMR structure of GpA shows a right-handed topology and a crossing angle of -50 degrees (see Figure 5.4 for pictorial representation). Different MD simulations have showed the formation of a loosely packed left-handed topology of GpA dimer, which was characterized by positive angles.[135, 136] Our scoring energy with scaled weighting factors scored the left-handed topology with the lowest scoring energy value and the near-native structure with the second lowest scoring energy. However, optimization of these weighting factors is dependent of the sampling of closed-packed conformations. The effects of improved sampling will be shown in subsection **5.3.5** ***Exhaustive Rigid Body Sampling***. Figure 5.5A shows the scoring energy against the C_α RMSD deviation from the native structure model of GpA. In this figure, the lowest scoring energy value of -7.2 presents a left-handed structure and a RMSD value of 6.8 Å. An inspection of this

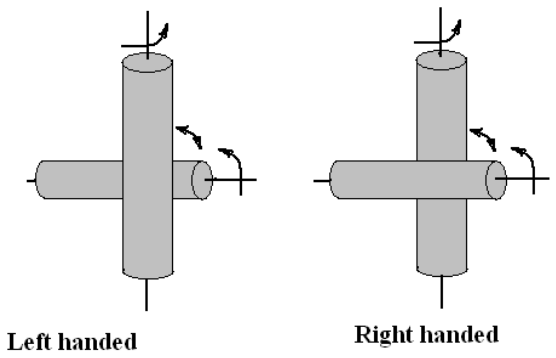

Figure 5.4. Definition of left and right-handed topology for helical dimers.

conformation reveals that the dimer interface contains mainly aliphatic residues such as Leu, Ile, Val and Ala. This packing propensity of aliphatic residues is due to the hydrophobic effect reproduced by the DFIRE interaction energy. The second lowest energy value of -6.8 corresponds to a left-handed

structure and a RMSD value of 1.9 Å. This conformation shows tight packing of Gly residues in the GxxxG motif. We test the modified scoring energy in three other homodimers as shown in the following paragraphs.

5.3.2 BNip3

BNip3 homodimer protein is involved in hypoxia-induced death of normal and malignant cells. Its transmembrane domain favors the dimerization of the protein. BNip3 solution NMR structure model have been determined in both lipid bicelles and micelles.[137, 138] The former structure displays packing of helices mediated by a GxxxG motif and interchain hydrogen bond between side-chains of Ser172 and His173. Sulistijo et al.[138] reported a NMR model containing a clear symmetric interchain hydrogen bonding of Ser172 and His173 and C-Hα-O hydrogen bonding between both helices backbones. The homodimer structure shows a dependence on external conditions such as

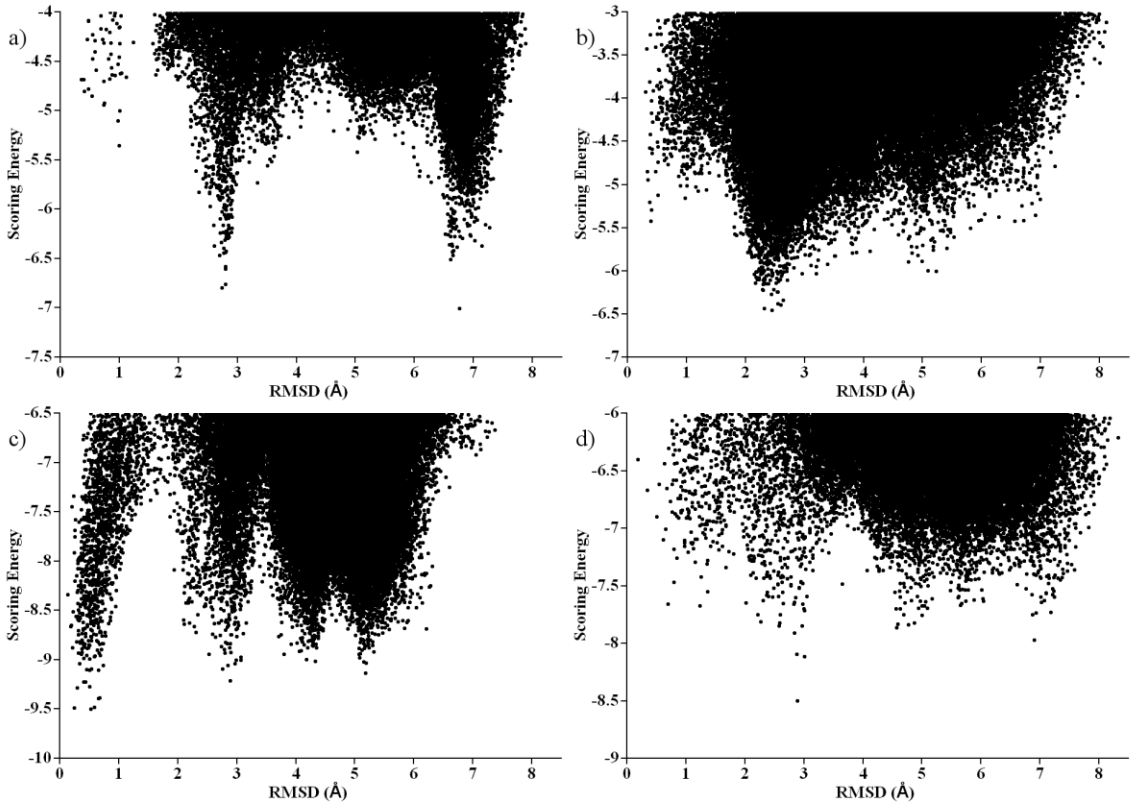

Figure 5.5. Score energy versus RMSD for a set of 4 TM proteins. a) Score energy plot of GpA. b) Score energy plot of BNIP3. c) Score energy plot of EphA1. d) Score energy of erbB2.

lipid contain and lipid titration. BNip3 transmembrane domain has a right-handed structure and a crossing angle of 35.0 degrees. We test the prediction performance of our scoring energy on Bocharov's BNip3 NMR structure model.[137] Figure 5.5B shows that the lowest scoring energy structure corresponds to -6.5 with a RMSD value of 2.4 Å from the native structure.

5.3.3 EphA1

The receptor tyrosine kinase EphA1 is activated through dimerization of the transmembrane domain. It was recently reported a solvent NMR structure model on lipid

bicelles.[139] The EphA1 structure shows a right handed topology and a crossing angle of 44 degrees. The lowest minimum scoring energy conformation corresponds to near-native conformations with a score value of -9.5 and RMSD value of 0.5 Å from the native structure, as shown in Figure 5.5C. The structure contains a double small-xxx-small residue motif, formed by AxxxGxxxG. The small residues, like in the case of GPA, facilitate the close packing of transmembrane domains.

5.3.4 erbB2

erbB2 protein is a member of epidermal growth factor receptor family, which belongs to receptor tyrosine kinase (RTK). RTKs consist of a ligand-binding, transmembrane, and cytoplasmatic phosphorilization domain. Dimerization of erbB2 transmembrane domains triggers a cascade of events in the cell interior. Recently, it was reported a NMR structure model of a possible activated erbB2 dimer on lipid bicelles.[140] This structure is a good test for our scoring energy function. erbB2 shows a right handed topology and a crossing angle of 40 degrees. There are two possible GxxxG motifs that might facilitate close packing. As a result, two different helical interfaces of dimerization have been proposed: GxxxG contact close to the N and C terminal. These two possible interfaces have been associated with conformational changes and heterodimer formation. MC trajectories of the Bocharov structure show the lowest scoring energy conformation with -8.5 and RMSD of 2.9 Å, as shown in Figure 5.5D.

Figure 5.6 shows detailed comparison of the backbone representation of native and lowest scoring energy structures for each TM homodimer. Residues that form the protein-protein surface of contact are shown in different colors. In general, near-native structures are predicted with low scoring energies.

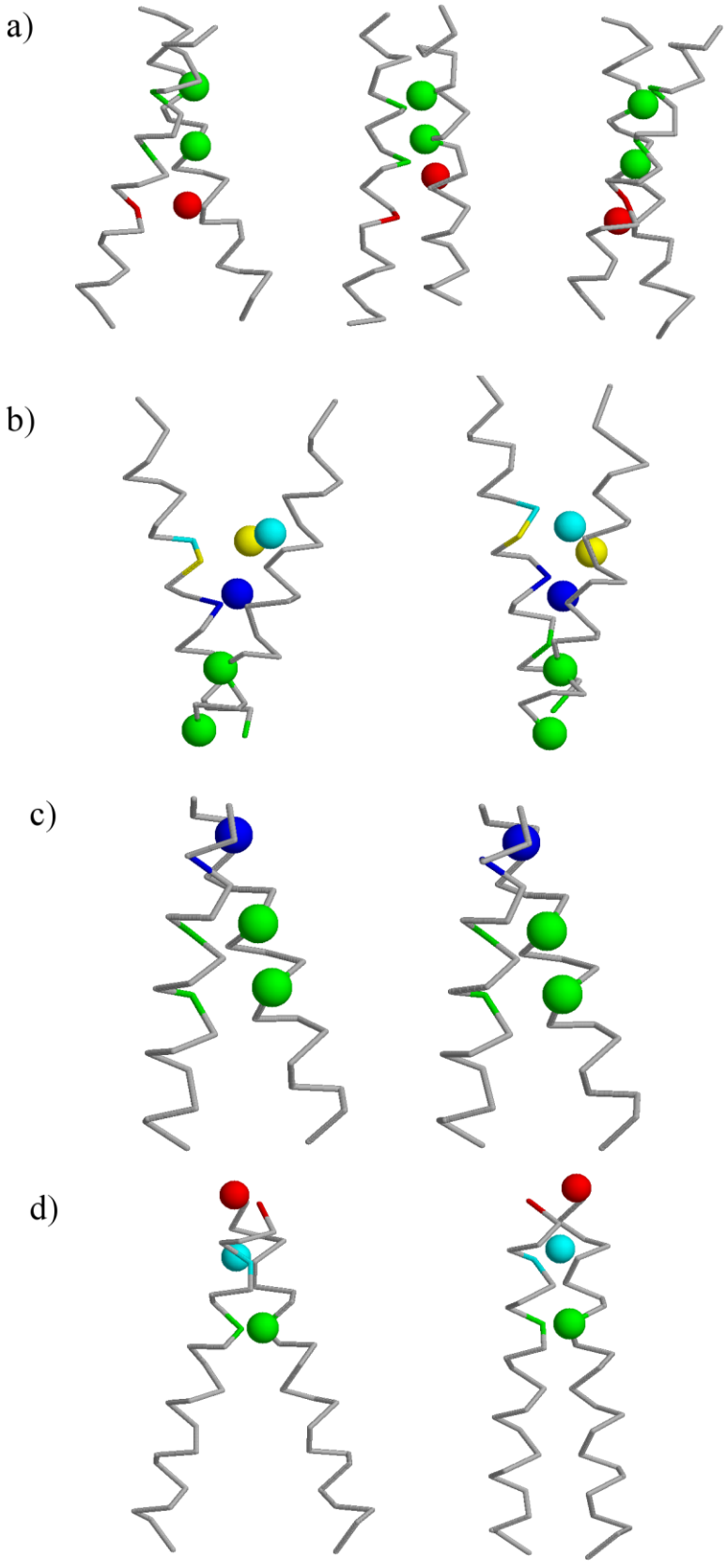
a)
b)
c)
d)

Figure 5.6. Backbone representation of native and low energy structures of 4 TM proteins. CG side-chains in the native dimer interface surface are highlighted by bead representation and colors where Gly, Ala, Thr, Ser and His amino-acids are shown in green, blue, red, cyan and yellow colors. a) CG representation of the native TM NMR structure of GpA is shown in the left. The lowest scoring energy CG model with left-handed topology and RMSD value of 6.8 Å is shown in the middle. The second lowest scoring energy with right handed topology with a RMSD value of 2.7 Å is shown in the right. b) CG representation of the native TM NMR structure of BNip3 is shown in the left. The lowest scoring energy CG model is shown in the right c) CG representation of the native TM NMR structure of EphA1 is shown in the left. The lowest scoring energy CG model is shown in the right. d) CG representation of the native TM NMR structure of ErbB2 is shown in the left. The lowest scoring energy CG model is shown in the right. Low energy structures were aligned to each native TM structure.

5.3.5 Exhaustive Rigid Body Sampling

We revisited the connection between the scoring energy function and sampling by performing a Metropolis MC simulation using the set of move sets described in Table 5.S1 (see description in **Methods**). Figure 5.7 shows that the near-native structure with a RMSD value of 1.9 Å is scored with the lowest scoring energy of -9.8. The left-handed structure with a RMSD value near 6.8 Å is scored with the second lowest energy of -8.1. Furthermore, it shows that the addition of the solvation energy term corrects the energy landscape of GpA, namely correcting the paiwise interactions of residues obtained from globular proteins.

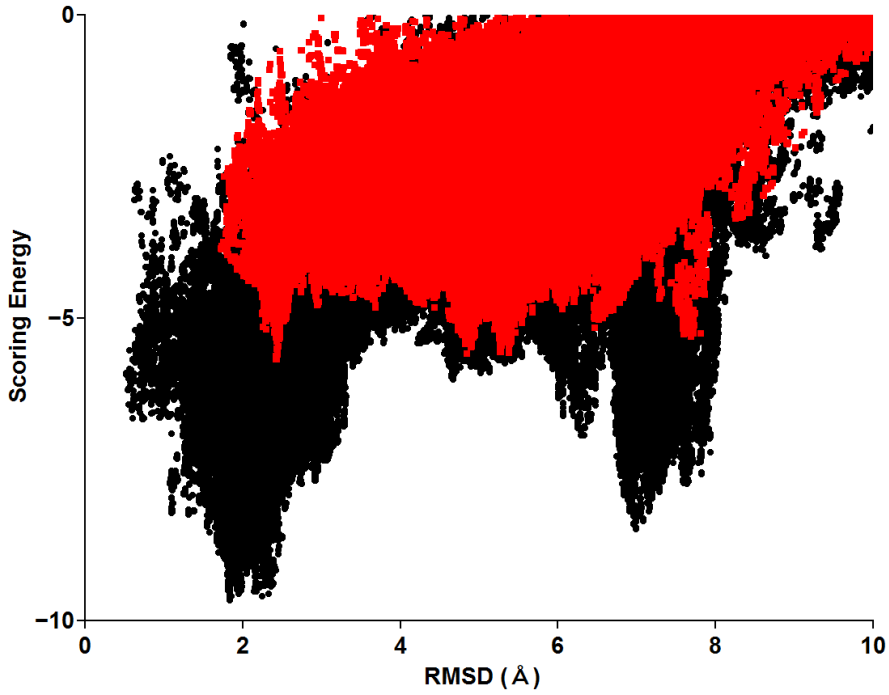

Figure 5.7. Score energy versus RMSD for Glycophorin A. Full scoring energy function was used to generate configurations shown in black dots. Scoring energy without the E_{solv} was used to generate configurations shown in red dots. Both Metropolis MC simulations were performed over 10^6 trial steps.

5.4 Discussion

Our method initially scores the left-handed topology of GpA as the lowest energy structure. The left handed topology of GpA shows no formation of Thr87-Thr87 contact, instead both residues are facing the hydrophobic core. Instead, the non-weighted version of our scoring method favors packing of non-polar residues due to the hydrophobic effect present in the potential of mean force obtained from water soluble proteins. To correct the scoring of the near-native right-handed structure, weighting factors for pairwise and

steric were added, which account for the formation of polar interactions and packing of small residues. As a result, the pairwise interaction between Thr87-Thr87 is strengthened, which accounts for a hydrogen bond suggested to stabilize the dimer association of GpA.[131] Although, our scoring energy strengthens interactions between Thr-Thr amino acids, prediction of hydrogen bond interactions will require addition of an explicit energy term, such as in Rosetta.[115] Our results suggest the role interhelical hydrogen bonds to fine tune the near-native structures of TM helices, as shown by structure prediction of three homodimers of known structure: BNip3, EphA1, and erbB2. Future work will involve the side-chain packing given the helical backbone model obtained from secondary structure prediction algorithms. Furthermore, predicted coarse-grained models will require the conversion to all-atom models for refinement using molecular dynamics simulations.

A novel energy term is included in our empirical scoring energy. This term considers the physico-chemical environment around a residue inserted in the lipid bilayer. Our results suggest the possibility to employ knowledge based potentials obtained from water soluble proteins provided that a correct membrane-like environment is incorporated in the scoring energy. It is noteworthy, that a biological scale obtained from sec-translocation to the membrane interior is capable to correct the scoring energy profile of GpA. This insight corroborates the importance of the chemical environment around residues in the assembly of helical membrane proteins. Our current scoring method is capable to predict structures with no knowledge extracted from membrane protein structures, as shown in a set of three helical TM proteins. Note that prediction of the near-

native structure of GpA is shown to be dependent of the preference of aminocids position in the lipid bilayer and their exposure to lipids.

5.5 Conclusions

This is the first step in the process to model and predict membrane protein structures. Current scoring energy functions based on water soluble proteins can be used to predict near-native structures of membrane proteins as a first step in the process of TM structure prediction. The preferred packing of aminoacids found in globular proteins is compensated with the incorporation of a solvation biological scale for each residue. This residue accessible area energy term includes information about the residue preference of interacting with the membrane and its location along the membrane normal. Our results suggest the importance of interhelical hydrogen bond formation to tune the correct assembly of TM helices. The next step will require further refinement of both the model representation and the energy function.

5.6 Appendix

Table 5.S1. Frequency of different move sets for the rigid-body MC sampling

	(1) A single DOF is modified (small or maximum step size) 70 % of Total MC steps (20% X, 20% Y, 10% Z, 20% Phi, 10% Theta, 20% Psi)	**(2) All DOF are modified (small or maximum step size) 30 % of Total MC steps**
Translational	X, Y = 70 % [-1.0, 1.0] Å 30 % [$-R_{max}$, R_{max}] Maximum step size. Z = 80 % [-1.0, 1.0] Å 20 % [$-R_{max}$, R_{max}] Maximum step size.	33 % [-0.1, 0.1] Å 33 % [-5.0, 5.0] Å 33 % [$-R_{max}$, R_{max}] Maximum step size
Orientational	Phi, Psi = 60 % [-2.0, 2.0] degrees 40 % [-180, 180] Maximum step size. Theta = 90 % [-2.0, 2.0] degrees 10 % [0, 180] Maximum step size.	33 % [-0.2, 0.2] degrees 33 % [-20.0, 20.0] degrees 33 % [-180, 180] Maximum step size

(1) A single degree of freedom (DOF) is modified with the option to sample small or large uniform displacements. (2) All DOF are modified simultaneously using uniform small or large steps. Detailed description of frequencies of movements for each DOF is shown for translational and orientational moves.

Summary of Thesis

In summary, this thesis presented computational studies of protonation-dependent conformational changes of T-domain in solution and its membrane association to lipid bilayers. Finally, a coarse-grained energy function is presented for the prediction of transmembrane helical proteins.

Chapter 1 presents the analysis of extensive atomistic MD simulations of a neutral and a low pH T-domain model. MD simulations of the low pH model demonstrate that protonation of histidines trigger conformational changes of the protein structure in solution. These structural changes involve unfolding/refolding of N-terminal helices TH1-2, partial unfolding of helix TH4 and solvent exposure of hydrophobic sites, which were in agreement with experimental observations of a membrane competent state. The formation of this intermediate state in solution was suggested to facilitate T-domain association to membrane interfaces.

Chapter 2 presents the implementation and testing of an accelerated molecular dynamics method. This method modifies electrostatic interactions of solute-solute atom pairs, which improves the reweighting of free energy landscapes and accelerates the sampling of conformational states of T-domain. It is shown that the proposed method can accelerate the sampling of conformational changes previously observed after several hundreds of nanoseconds of equilibrium MD simulations.

Chapter 3 details extensive equilibrium and free energy calculations of coarse-grained models of T-domain and lipid bilayers of different composition. Partially unfolded T-domain spontaneously binds to bilayers composed of mostly anionic lipids.

Two preferable membrane-bound conformations of the partially unfolded T-domain were predicted.

Chapter 4 describes extensive equilibrium atomistic simulations of predicted membrane-bound conformations of partially unfolded T-domain. These simulations show that both bound conformations stabilize over several microseconds and insert deeper in the membrane interface. Neutralization of glutamate and aspartate sidechains, in addition to protonated histidines, triggers changes in the protein orientation relative to the membrane and deeper insertion in the membrane interface. This suggests that changes of protonation states can be associated to the refolding and insertion of T-domain in lipid bilayers.

Chapter 5 presents an empirical energy function for the prediction of transmembrane helical proteins. The empirical function is constructed using a knowledge based potential obtained from water soluble and a membrane model based on a 'biological' partition scale of amino-acids. A rigid body Metropolis MC method is implemented for the exhaustive sampling of the conformational space of a pair of rigid bodies. Metropolis MC simulations showed that this energy function predicts successfully near-native structures of three known transmembrane homodimers, which will be used for further refinement with atomistic models.

Future Work

In chapter 1, it was demonstrated that protonation of histidines triggers conformational changes of T-domain. Among them, H257 was predicted to have the largest free energy of destabilization of the protein structure at low pH by free energy calculations. However, direct comparison of our atomistic model with experimental titration curves and apparent pH value of transition in solution is lacking. The recently developed Amber force field FF14SB and an improved generalized Born solvent model could be used for constant pH MD simulations of T-domain.

The proposed accelerated molecular dynamics method, denominated aMD_EE in Chapter 2, could be used for folding studies of small proteins with different force fields. Furthermore, it could be applied to the study of conformational changes of relatively large proteins in explicit solvent. Also, this method could be applied to the study of structural refolding of the two predicted membrane-bound conformations using atomistic models.

Further work will be required to understand the possible role of changes of protonation of glutamate, aspartate and histidines in the membrane insertion of T-domain. For example, it is known that the transmembrane state of T-domain requires the movement a glutamate and an aspartate (located in the loop TH8-TH9) across the lipid bilayer. The mechanism of translocation of these two acidic side-chains is currently unknown. Neutralization of acidic side-chains in the membrane interface may decrease the free energy penalty of translocating them across the membrane.

Bibliography

1. Collier, R. J., Understanding the mode of action of diphtheria toxin: a perspective on progress during the 20th century. *Toxicon* **2001,** *39* (11), 1793-803.
2. Murphy, J. R., Mechanism of diphtheria toxin catalytic domain delivery to the eukaryotic cell cytosol and the cellular factors that directly participate in the process. *Toxins (Basel)* **2011,** *3* (3), 294-308.
3. Ladokhin, A. S., pH-Triggered Conformational Switching along the Membrane Insertion Pathway of the Diphtheria Toxin T-Domain. *Toxins (Basel)* **2013,** *5* (8), 1362-80.
4. Prince, H. M.; Duvic, M.; Martin, A.; Sterry, W.; Assaf, C.; Sun, Y.; Straus, D.; Acosta, M.; Negro-Vilar, A., Phase III placebo-controlled trial of denileukin diftitox for patients with cutaneous T-cell lymphoma. *J Clin Oncol* **2010,** *28* (11), 1870-7.
5. Steere, B. Characterization of high-order oligomerization and energetics in Diptheria Toxin. University of California, Los Angeles, 2001.
6. Zhan, H. J.; Choe, S.; Huynh, P. D.; Finkelstein, A.; Eisenberg, D.; Collier, R. J., Dynamic Transitions of the Transmembrane Domain of Diphtheria-Toxin - Disulfide Trapping and Fluorescence Proximity Studies. *Biochemistry* **1994,** *33* (37), 11254-11263.
7. Wang, J.; Rosconi, M. P.; London, E., Topography of the hydrophilic helices of membrane-inserted diphtheria toxin T domain: TH1-TH3 as a hydrophilic tether. *Biochemistry* **2006,** *45* (26), 8124-34.
8. Chenal, A.; Savarin, P.; Nizard, P.; Guillain, F.; Gillet, D.; Forge, V., Membrane protein insertion regulated by bringing electrostatic and hydrophobic interactions into play. A case study with the translocation domain of diphtheria toxin. *J Biol Chem* **2002,** *277* (45), 43425-32.
9. Perier, A.; Chassaing, A.; Raffestin, S.; Pichard, S.; Masella, M.; Menez, A.; Forge, V.; Chenal, A.; Gillet, D., Concerted protonation of key histidines triggers membrane interaction of the diphtheria toxin T domain. *J Biol Chem* **2007,** *282* (33), 24239-45.
10. Kyrychenko, A.; Posokhov, Y. O.; Rodnin, M. V.; Ladokhin, A. S., Kinetic Intermediate Reveals Staggered pH-Dependent Transitions along the Membrane Insertion Pathway of the Diphtheria Toxin T-Domain. *Biochemistry* **2009,** *48* (32), 7584-7594.
11. Rodnin, M. V.; Kyrychenko, A.; Kienker, P.; Sharma, O.; Posokhov, Y. O.; Collier, R. J.; Finkelstein, A.; Ladokhin, A. S., Conformational Switching of the Diphtheria Toxin T Domain. *Journal of Molecular Biology* **2010,** *402* (1), 1-7.
12. Ladokhin, A. S.; Legmann, R.; Collier, R. J.; White, S. H., Reversible refolding of the diphtheria toxin T-domain on lipid membranes. *Biochemistry* **2004,** *43* (23), 7451-8.
13. Chenal, A., et al., Deciphering membrane insertion of the diphtheria toxin T domain by specular neutron reflectometry and solid-state NMR spectroscopy. *J Mol Biol* **2009,** *391* (5), 872-83.
14. Bennett, M. J.; Choe, S.; Eisenberg, D., Refined Structure of Dimeric Diphtheria-Toxin at 2.0-Angstrom Resolution. *Protein Science* **1994,** *3* (9), 1444-1463.
15. Murphy, R. F.; Powers, S.; Cantor, C. R., Endosome Ph Measured in Single Cells by Dual Fluorescence Flow-Cytometry - Rapid Acidification of Insulin to Ph-6. *Journal of Cell Biology* **1984,** *98* (5), 1757-1762.

16. Ladokhin, A. S.; Legmann, R.; Collier, R. J.; White, S. H., Reversible refolding of the diphtheria toxin T-domain on lipid membranes. *Biochemistry* **2004,** *43* (23), 7451-7458.
17. Perier, A.; Chassaing, A.; Raffestin, S.; Pichard, S.; Masella, M.; Menez, A.; Forge, V.; Chenal, A.; Gillet, D., Concerted protonation of key histidines triggers membrane interaction of the diphtheria toxin T domain. *Journal of Biological Chemistry* **2007,** *282* (33), 24239-24245.
18. Wang, J.; Rosconi, M. P.; London, E., Topography of the hydrophilic helices of membrane-inserted diphtheria toxin T domain: TH1-TH3 as a hydrophilic tether. *Biochemistry* **2006,** *45* (26), 8124-8134.
19. Leka, O.; Vallese, F.; Pirazzini, M.; Berto, P.; Montecucco, C.; Zanotti, G., Diphtheria toxin conformational switching at acidic pH. *FEBS J* **2014**.
20. Kurnikov, I. V.; Kyrychenko, A.; Flores-Canales, J. C.; Rodnin, M. V.; Simakov, N.; Vargas-Uribe, M.; Posokhov, Y. O.; Kurnikova, M.; Ladokhin, A. S., pH-Triggered Conformational Switching of the Diphtheria Toxin T-Domain: The Roles of N-Terminal Histidines. *J Mol Biol* **2013,** *425* (15), 2752-64.
21. Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E., Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J Am Chem Soc* **2012,** *134* (8), 3787-91.
22. Shan, Y.; Eastwood, M. P.; Zhang, X.; Kim, E. T.; Arkhipov, A.; Dror, R. O.; Jumper, J.; Kuriyan, J.; Shaw, D. E., Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell* **2012,** *149* (4), 860-70.
23. Shan, Y.; Arkhipov, A.; Kim, E. T.; Pan, A. C.; Shaw, D. E., Transitions to catalytically inactive conformations in EGFR kinase. *Proc Natl Acad Sci U S A* **2013,** *110* (18), 7270-5.
24. Speranskiy, K.; Kurnikova, M. G., Modeling of peptides connecting the ligand-binding and transmembrane domains in the GluR2 glutamate receptor. *Proteins-Structure Function and Bioinformatics* **2009,** *76* (2), 271-280.
25. Case, D. A., et al. *AMBER 9, University of California, San Francisco: San Francisco, CA*, 2006.
26. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-25.
27. Shan, Y. B.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E., Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J Chem Phys* **2005,** *122* (5).
28. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* **1996,** *14* (1), 33-8, 27-8.
29. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-637.
30. Lange, O. F.; Grubmuller, H., Full correlation analysis of conformational protein dynamics. *Proteins* **2008,** *70* (4), 1294-312.
31. Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G., Dihedral angle principal component analysis of molecular dynamics simulations. *J Chem Phys* **2007,** *126* (24), 244111.

32. Shaw, D. E., et al., Atomic-level characterization of the structural dynamics of proteins. *Science* **2010,** *330* (6002), 341-6.
33. Huang, Y. J.; Acton, T. B.; Montelione, G. T., DisMeta: a meta server for construct design and optimization. *Methods Mol Biol* **2014,** *1091*, 3-16.
34. Shan, Y. B.; Seeliger, M. A.; Eastwood, M. P.; Frank, F.; Xu, H. F.; Jensen, M. O.; Dror, R. O.; Kuriyan, J.; Shaw, D. E., A conserved protonation-dependent switch controls drug binding in the Abl kinase. *P Natl Acad Sci USA* **2009,** *106* (1), 139-144.
35. Arkin, I. T., et al., Mechanism of Na+/H+ antiporting. *Science* **2007,** *317* (5839), 799-803.
36. Shan, Y. B.; Eastwood, M. P.; Zhang, X. W.; Kim, E. T.; Arkhipov, A.; Dror, R. O.; Jumper, J.; Kuriyan, J.; Shaw, D. E., Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization. *Cell* **2012,** *149* (4), 860-870.
37. Arkhipov, A.; Shan, Y. B.; Das, R.; Endres, N. F.; Eastwood, M. P.; Wemmer, D. E.; Kuriyan, J.; Shaw, D. E., Architecture and Membrane Interactions of the EGF Receptor. *Cell* **2013,** *152* (3), 557-569.
38. Shaw, D. E., et al., Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010,** *330* (6002), 341-346.
39. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **1999,** *314* (1-2), 141-151.
40. Fukunishi, H.; Watanabe, O.; Takada, S., On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J Chem Phys* **2002,** *116* (20), 9058-9067.
41. Zhang, C.; Ma, J. P., Folding helical proteins in explicit solvent using dihedral-biased tempering. *P Natl Acad Sci USA* **2012,** *109* (40), 16392-16392.
42. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J Chem Phys* **2004,** *120* (24), 11919-11929.
43. Hamelberg, D.; de Oliveira, C. A. F.; McCammon, J. A., Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J Chem Phys* **2007,** *127* (15).
44. Markwick, P. R. L.; McCammon, J. A., Studying functional dynamics in bio-molecules using accelerated molecular dynamics. *Phys Chem Chem Phys* **2011,** *13* (45), 20053-20065.
45. de Oliveira, C. A. F.; Hamelberg, D.; McCammon, J. A., Coupling accelerated molecular dynamics methods with thermodynamic integration simulations. *Journal of Chemical Theory and Computation* **2008,** *4* (9), 1516-1525.
46. Sinko, W.; de Oliveira, C. A. F.; Pierce, L. C. T.; McCammon, J. A., Protecting High Energy Barriers: A New Equation to Regulate Boost Energy in Accelerated Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* **2012,** *8* (1), 17-23.
47. Flores-Canales, J. C.; Simakov, N. A.; Kurnikova, M., *Submitted*.
48. Case, D. A., et al., AMBER 12, University of California, San Francisco. 2012.
49. Case, D. A., et al., The Amber biomolecular simulation programs. *J Comput Chem* **2005,** *26* (16), 1668-1688.

50. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1993,** *98* (12), 10089-10092.
51. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A Smooth Particle Mesh Ewald Method. *J Chem Phys* **1995,** *103* (19), 8577-8593.
52. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics* **2006,** *65* (3), 712-725.
53. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J Comput Phys* **1977,** *23* (3), 327-341.
54. Salomon-Ferrer, R.; Gotz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2013,** *9* (9), 3878-3888.
55. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* **1996,** *14* (1), 33-38.
56. Kabsch, W.; Sander, C., Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983,** *22* (12), 2577-2637.
57. Fajer, M.; Hamelberg, D.; McCammon, J. A., Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration. *Journal of Chemical Theory and Computation* **2008,** *4* (10), 1565-1569.
58. Choe, S.; Bennett, M. J.; Fujii, G.; Curmi, P. M.; Kantardjieff, K. A.; Collier, R. J.; Eisenberg, D., The crystal structure of diphtheria toxin. *Nature* **1992,** *357* (6375), 216-22.
59. Honjo, T.; Nishizuka, Y.; Hayaishi, O., Diphtheria toxin-dependent adenosine diphosphate ribosylation of aminoacyl transferase II and inhibition of protein synthesis. *J Biol Chem* **1968,** *243* (12), 3553-5.
60. Bennett, M. J.; Eisenberg, D., Refined structure of monomeric diphtheria toxin at 2.3 A resolution. *Protein Sci* **1994,** *3* (9), 1464-75.
61. Wang, Y.; Malenbaum, S. E.; Kachel, K.; Zhan, H.; Collier, R. J.; London, E., Identification of shallow and deep membrane-penetrating forms of diphtheria toxin T domain that are regulated by protein concentration and bilayer width. *J Biol Chem* **1997,** *272* (40), 25091-8.
62. Palchevskyy, S. S.; Posokhov, Y. O.; Olivier, B.; Popot, J. L.; Pucci, B.; Ladokhin, A. S., Chaperoning of insertion of membrane proteins into lipid bilayers by hemifluorinated surfactants: application to diphtheria toxin. *Biochemistry* **2006,** *45* (8), 2629-35.
63. Montagner, C.; Perier, A.; Pichard, S.; Vernier, G.; Menez, A.; Gillet, D.; Forge, V.; Chenal, A., Behavior of the N-terminal helices of the diphtheria toxin T domain during the successive steps of membrane interaction. *Biochemistry* **2007,** *46* (7), 1878-87.
64. Vargas-Uribe, M.; Rodnin, M. V.; Kienker, P.; Finkelstein, A.; Ladokhin, A. S., Crucial role of H322 in folding of the diphtheria toxin T-domain into the open-channel state. *Biochemistry* **2013,** *52* (20), 3457-63.

65. Rodnin, M. V.; Kyrychenko, A.; Kienker, P.; Sharma, O.; Vargas-Uribe, M.; Collier, R. J.; Finkelstein, A.; Ladokhin, A. S., Replacement of C-terminal histidines uncouples membrane insertion and translocation in diphtheria toxin T-domain. *Biophys J* **2011,** *101* (10), L41-3.
66. Leka, O.; Vallese, F.; Pirazzini, M.; Berto, P.; Montecucco, C.; Zanotti, G., Diphtheria toxin conformational switching at acidic pH. *FEBS J* **2014,** *281* (9), 2115-22.
67. Stansfeld, P. J.; Sansom, M. S., Molecular simulation approaches to membrane proteins. *Structure* **2011,** *19* (11), 1562-72.
68. Marrink, S. J.; Tieleman, D. P., Perspective on the Martini model. *Chem Soc Rev* **2013,** *42* (16), 6801-22.
69. Kalli, A. C.; Wegener, K. L.; Goult, B. T.; Anthis, N. J.; Campbell, I. D.; Sansom, M. S. P., The Structure of the Talin/Integrin Complex at a Lipid Bilayer: An NMR and MD Simulation Study. *Structure* **2010,** *18* (10), 1280-1288.
70. Kalli, A. C.; Campbell, I. D.; Sansom, M. S., Multiscale simulations suggest a mechanism for integrin inside-out activation. *Proc Natl Acad Sci U S A* **2011,** *108* (29), 11890-5.
71. Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J., The MARTINI coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation* **2008,** *4* (5), 819-834.
72. Periole, X.; Cavalli, M.; Marrink, S. J.; Ceruso, M. A., Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *Journal of Chemical Theory and Computation* **2009,** *5* (9), 2531-2543.
73. Roux, B., The Calculation of the Potential of Mean Force Using Computer-Simulations. *Comput Phys Commun* **1995,** *91* (1-3), 275-282.
74. Kandt, C.; Ash, W. L.; Tieleman, D. P., Setting up and running molecular dynamics simulations of membrane proteins. *Methods* **2007,** *41* (4), 475-88.
75. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M., The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. *J Comput Chem* **1992,** *13* (8), 1011-1021.
76. Hub, J. S.; de Groot, B. L.; van der Spoel, D., g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *Journal of Chemical Theory and Computation* **2010,** *6* (12), 3713-3720.
77. Vivcharuk, V.; Tomberli, B.; Tolokh, I. S.; Gray, C. G., Prediction of binding free energy for adsorption of antimicrobial peptide lactoferricin B on a POPC membrane. *Phys Rev E* **2008,** *77* (3).
78. Neale, C.; Bennett, W. F. D.; Tieleman, D. P.; Pomes, R., Statistical Convergence of Equilibrium Properties in Simulations of Molecular Solutes Embedded in Lipid Bilayers. *Journal of Chemical Theory and Computation* **2011,** *7* (12), 4175-4188.
79. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E., GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation* **2008,** *4* (3), 435-447.
80. Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H., The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* **2007,** *111* (27), 7812-24.

81. Senzel, L.; Gordon, M.; Blaustein, R. O.; Oh, K. J.; Collier, R. J.; Finkelstein, A., Topography of diphtheria Toxin's T domain in the open channel state. *J Gen Physiol* **2000,** *115* (4), 421-34.
82. Senzel, L.; Huynh, P. D.; Jakes, K. S.; Collier, R. J.; Finkelstein, A., The diphtheria toxin channel-forming T domain translocates its own NH2-terminal region across planar bilayers. *J Gen Physiol* **1998,** *112* (3), 317-24.
83. Oh, K. J.; Senzel, L.; Collier, R. J.; Finkelstein, A., Translocation of the catalytic domain of diphtheria toxin across planar phospholipid bilayers by its own T domain. *Proc Natl Acad Sci U S A* **1999,** *96* (15), 8467-70.
84. Flores-Canales, J. C.; Ladokhin, A. S.; Kurnikova, M., *Submitted*.
85. Fawzi, N. L.; Phillips, A. H.; Ruscio, J. Z.; Doucleff, M.; Wemmer, D. E.; Head-Gordon, T., Structure and Dynamics of the A beta(21-30) Peptide from the Interplay of NMR Experiments and Molecular Simulations (vol 130, pg 6145, 2008). *J Am Chem Soc* **2011,** *133* (30), 11816-11816.
86. Showalter, S. A.; Bruschweiler, R., Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *Journal of Chemical Theory and Computation* **2007,** *3* (3), 961-975.
87. Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E., Systematic Validation of Protein Force Fields against Experimental Data. *Plos One* **2012,** *7* (2).
88. Dickson, C. J.; Rosso, L.; Betz, R. M.; Walker, R. C.; Gould, I. R., GAFFlipid: a General Amber Force Field for the accurate molecular dynamics simulation of phospholipid. *Soft Matter* **2012,** *8* (37), 9617-9627.
89. Skjevik, A. A.; Madej, B. D.; Walker, R. C.; Teigen, K., LIPID11: A Modular Framework for Lipid Simulations Using Amber. *Journal of Physical Chemistry B* **2012,** *116* (36), 11124-11136.
90. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-1174.
91. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A., A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *J Phys Chem-Us* **1993,** *97* (40), 10269-10280.
92. Pan, J. J.; Heberle, F. A.; Tristram-Nagle, S.; Szymanski, M.; Koepfinger, M.; Katsaras, J.; Kucerka, N., Molecular structures of fluid phase phosphatidylglycerol bilayers as determined by small angle neutron and X-ray scattering. *Bba-Biomembranes* **2012,** *1818* (9), 2135-2148.
93. Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D.; Pastor, R. W., Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *Journal of Physical Chemistry B* **2010,** *114* (23), 7830-7843.
94. Dupradeau, F. Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P., The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys Chem Chem Phys* **2010,** *12* (28), 7821-39.
95. Frisch, M. J., et al., Gaussian 03; Gaussian Inc.: Wallingford, CT, 2003.

96. Henin, J.; Shinoda, W.; Klein, M. L., Models for Phosphatidylglycerol Lipids Put to a Structural Test. *Journal of Physical Chemistry B* **2009,** *113* (19), 6958-6963.
97. Tolokh, I. S.; Vivcharuk, V.; Tomberli, B.; Gray, C. G., Binding free energy and counterion release for adsorption of the antimicrobial peptide lactoferricin B on a POPG membrane. *Phys Rev E* **2009,** *80* (3).
98. Cojocaru, V.; Balali-Mood, K.; Sansom, M. S.; Wade, R. C., Structure and dynamics of the membrane-bound cytochrome P450 2C9. *PLoS Comput Biol* **2011,** *7* (8), e1002152.
99. Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation* **2012,** *8* (5), 1542-1555.
100. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Function and Bioinformatics* **2010,** *78* (8), 1950-1958.
101. Shaw, D. E., Millisecond-long molecular dynamics simulations of proteins on a special-purpose machine. *Abstr Pap Am Chem S* **2010,** *240*.
102. Martyna, G. J.; Klein, M. L.; Tuckerman, M., Nose-Hoover Chains - the Canonical Ensemble Via Continuous Dynamics. *J Chem Phys* **1992,** *97* (4), 2635-2643.
103. Martyna, G. J.; Tobias, D. J.; Klein, M. L., Constant-Pressure Molecular-Dynamics Algorithms. *J Chem Phys* **1994,** *101* (5), 4177-4189.
104. Fendos, J.; Barrera, F. N.; Engelman, D. M., Aspartate Embedding Depth Affects pHLIP's Insertion pK(a). *Biochemistry* **2013,** *52* (27), 4595-4604.
105. White, S. http://blanco.biomol.uci.edu/mpstruc/. (accessed 08/23/2014).
106. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000,** *28* (1), 235-242.
107. Simms, J.; Booth, P. J., Membrane proteins by accident or design. *Curr Opin Chem Biol* **2013,** *17* (6), 976-81.
108. Bill, R. M., et al., Overcoming barriers to membrane protein structure determination. *Nat Biotechnol* **2011,** *29* (4), 335-340.
109. Walther, T. H.; Ulrich, A. S., Transmembrane helix assembly and the role of salt bridges. *Curr Opin Struct Biol* **2014,** *27C*, 63-68.
110. Bowie, J. U., Membrane protein folding: how important are hydrogen bonds? *Curr Opin Struc Biol* **2011,** *21* (1), 42-49.
111. Ozdirekcan, S.; Rijkers, D. T. S.; Liskamp, R. M. J.; Killian, J. A., Influence of flanking residues on tilt and rotation angles of transmembrane peptides in lipid bilayers. A solid-state H-2 NMR study. *Biochemistry* **2005,** *44* (3), 1004-1012.
112. Nyholm, T. K. M.; Ozdirekcan, S.; Killian, J. A., How protein transmembrane segments sense the lipid environment. *Biochemistry* **2007,** *46* (6), 1457-1465.
113. Benjamini, A.; Smit, B., Robust Driving Forces for Transmembrane Helix Packing. *Biophysical Journal* **2012,** *103* (6), 1227-1235.
114. Weiner, B. E.; Woetzel, N.; Karakas, M.; Alexander, N.; Meiler, J., BCL::MP-Fold: Folding Membrane Proteins through Assembly of Transmembrane Helices. *Structure* **2013,** *21* (7), 1107-1117.

115. Barth, P.; Schonbrun, J.; Baker, D., Toward high-resolution prediction and design of transmembrane helical protein structures. *P Natl Acad Sci USA* **2007,** *104* (40), 15682-15687.
116. Hopf, T. A.; Colwell, L. J.; Sheridan, R.; Rost, B.; Sander, C.; Marks, D. S., Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **2012,** *149* (7), 1607-1621.
117. Kokubo, H.; Okamoto, Y., Analysis of Helix-Helix Interactions of Bacteriorhodopsin by Replica-Exchange Simulations. *Biophysical Journal* **2009,** *96* (3), 765-776.
118. Michino, M.; Chen, J. H.; Stevens, R. C.; Brooks, C. L., FoldGPCR: Structure prediction protocol for the transmembrane domain of G protein-coupled receptors from class A. *Proteins-Structure Function and Bioinformatics* **2010,** *78* (10), 2189-2201.
119. Kuhlman, B.; Baker, D., Native protein sequences are close to optimal for their structures. *P Natl Acad Sci USA* **2000,** *97* (19), 10383-10388.
120. Zhang, Y., I-TASSER server for protein 3D structure prediction. *Bmc Bioinformatics* **2008,** *9*.
121. Sippl, M. J., Calculation of Conformational Ensembles from Potentials of Mean Force - an Approach to the Knowledge-Based Prediction of Local Structures in Globular-Proteins. *Journal of Molecular Biology* **1990,** *213* (4), 859-883.
122. Yarov-Yarovoy, V.; Schonbrun, J.; Baker, D., Multipass membrane protein structure prediction using Rosetta. *Proteins-Structure Function and Bioinformatics* **2006,** *62* (4), 1010-1025.
123. Wendel, C.; Gohlke, H., Predicting transmembrane helix pair configurations with knowledge-based distance-dependent pair potentials. *Proteins* **2008,** *70* (3), 984-99.
124. Hessa, T.; Meindl-Beinker, N. M.; Bernsel, A.; Kim, H.; Sato, Y.; Lerch-Bader, M.; Nilsson, I.; White, S. H.; von Heijne, G., Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **2007,** *450* (7172), 1026-U2.
125. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of State Calculations by Fast Computing Machines. *J Chem Phys* **1953,** *21* (6), 1087-1092.
126. Kurnikov, I. V. *HARLEM - Biomolecular simulation program*, 2003.
127. Zhou, H. Y.; Zhou, Y. Q., Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science* **2002,** *11* (11), 2714-2726.
128. Zhang, C.; Liu, S.; Zhou, H. Y.; Zhou, Y. Q., An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Science* **2004,** *13* (2), 400-411.
129. Larriva, M.; de Sancho, D.; Rey, A., Evaluation of a mean field potential for protein folding with different interaction centers. *Physica A* **2006,** *371* (2), 449-462.
130. Kocher, J. P. A.; Rooman, M. J.; Wodak, S. J., Factors Influencing the Ability of Knowledge-Based Potentials to Identify Native Sequence-Structure Matches. *Journal of Molecular Biology* **1994,** *235* (5), 1598-1613.
131. MacKenzie, K. R.; Prestegard, J. H.; Engelman, D. M., A transmembrane helix dimer: structure and implications. *Science* **1997,** *276* (5309), 131-3.

132. Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G., Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **2005,** *433* (7024), 377-381.
133. Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I., OPM: Orientations of proteins in membranes database. *Bioinformatics* **2006,** *22* (5), 623-625.
134. Lemmon, M. A.; Flanagan, J. M.; Hunt, J. F.; Adair, B. D.; Bormann, B. J.; Dempsey, C. E.; Engelman, D. M., Glycophorin-a Dimerization Is Driven by Specific Interactions between Transmembrane Alpha-Helices. *Journal of Biological Chemistry* **1992,** *267* (11), 7683-7689.
135. Psachoulia, E.; Marshall, D. P.; Sansom, M. S., Molecular dynamics simulations of the dimerization of transmembrane alpha-helices. *Acc Chem Res* **2010,** *43* (3), 388-96.
136. Sengupta, D.; Marrink, S. J., Lipid-mediated interactions tune the association of glycophorin A helix and its disruptive mutants in membranes. *Phys Chem Chem Phys* **2010,** *12* (40), 12987-96.
137. Bocharov, E. V., et al., Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger. *J Biol Chem* **2007,** *282* (22), 16256-66.
138. Sulistijo, E. S.; Mackenzie, K. R., Structural basis for dimerization of the BNIP3 transmembrane domain. *Biochemistry* **2009,** *48* (23), 5106-20.
139. Bocharov, E. V.; Mayzel, M. L.; Volynsky, P. E.; Goncharuk, M. V.; Ermolyuk, Y. S.; Schulga, A. A.; Artemenko, E. O.; Efremov, R. G.; Arseniev, A. S., Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1. *J Biol Chem* **2008,** *283* (43), 29385-95.
140. Bocharov, E. V., et al., Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. *J Biol Chem* **2008,** *283* (11), 6950-6.