MULTISCALE MODELLING OF SQUID INSPIRED TANDEM REPEAT PROTEINS

by
OĞUZHAN ÇOLAK

Submitted to the Graduate School of Engineering and Natural Sciences
In partial fulfillment of
the requirements for the degree of Master of Science

Sabanci University
May 2020

MULTISCALE MODELLING OF SQUID INSPIRED TANDEM REPEAT PROTEINS

Approved by:

Prof. Canan ATILGAN ……..
(Thesis Supervisor)

Prof. Ali Rana ATILGAN ……..

Prof. Melik Cumhur DEMİREL ……..

Approval Date: March 9, 2020

ABSTRACT

MULTISCALE MODELLING OF SQUID INSPIRED TANDEM REPEAT PROTEINS

OĞUZHAN ÇOLAK

MATERIALS SCIENCE & NANO ENGINEERING M.S. THESIS, MAY 2020

Supervisor: Prof. Canan ATILGAN

Keywords: Molecular dynamics, Dissipative particle dynamics, Tandem repeat protein, Self-assembly

Squid ring teeth (SRT) proteins are structural proteins with repetitive amino acid sequences. They comprise two regions which are crystal forming and tie-chain regions. The mechanical properties of the protein known however the exact mechanism for the aggregation of the protein between the previously mentioned segments are still unknown. SRT proteins have unknown folding behavior and the size of those synthesized to date vary between 140 to 875 for a single chain. Considering these factors, we used Dissipative Particle Dynamics (DPD) simulations as our primary method of simulations rather than only using Molecular Dynamics (MD) simulations since MD simulations would be computationally expensive. So, in this study, we propose a method, which was previously used in polymers, of parameterizing the SRT proteins via multiscale simulations. To parameterize the system, we initially used binary MD simulations of each bead pair in the system at the atomistic detail. Then, we coarse-grained all the molecules into beads, and using the cohesive energy density values from the MD simulations, we constructed Flory-Huggins interaction parameters for all pairs in our system. We used four varying sizes of SRT proteins, n4, n7, n11, and n25 and two different solvents which were the good solvent HFIP and the hypothetical poor solvent P. Radial distribution function and structure factor calculations were used to characterize the structure of the SRT proteins in specified solvents. The results show that the SRT proteins swell in HFIP and they have no long-range order, but they cluster and form ordered structures in solvent P which show that the computational results agree with the experimental data.

ÖZET

AKDENİZ KALAMARINDAN ESİNLENİLMİŞ TEKRAR EDEN PROTEİNLERIN ÇOK ÖLÇEKLİ MODELLENMESİ

OĞUZHAN ÇOLAK

PROGRAM ADI YÜKSEK LİSANS TEZİ, MAYIS 2020

Tez Danışmanı: Prof. Canan ATILGAN

Anahtar Kelimeler: Moleküler dinamik, Dağılıcı parçacık dinamiği, Tekrar eden protein

Akdeniz kalamarı yüzük dişleri (AKYD) proteinleri tekrar eden amino asit dizilerinden oluşan yapı proteinleridir. Kristal yapı ve düzensiz-bağ zincirleri olmak üzere iki alt bölgeden oluşurlar. AKYD proteinlerinin mekanik özellikleri bilinse de alt bölgelerin topaklanma mekanizması hala bilinmemektedir. AKYD proteinlerinin katlanma davranışı da bilinmemekte ve bugüne dek sentezlenmiş protein zincirlerinin boyları 140 ile 875 amino asit uzunluğu arasında değişmektedir. Bu etmenler göz önüne alındığında, Moleküler Dinamik (MD) benzetimi, bilgisayar zamanı açısından pahalı bir yöntem olduğundan, benzetimlerde Dağılıcı Parçacık Dinamiği (DPD) metodu kullanıldı. Bu çalışmada, öncesinde polimerlerde kullanılan bir yöntemi, AKYD proteinlerinin benzetim değişkenlerini hesaplamak için kullanılmasını sunuyoruz. Sistem benzetim değişkenleri hesaplamak için, öncelikle her parçacık çifti için ikili MD benzetimlerini atomistik detayda koşturduk. Ardından, MD benzetimlerimdeki kohezif enerji yoğunluğu değerlerini kullanarak ve tüm molekülleri kürecikler halinde kaba modele dönüştürerek, sistemdeki tüm parçacık çiftleri için Flory-Huggins etkileşim değişkenlerini hesapladık. AKYD proteinlerinin n4, n7, n11 ve n25 olmak üzere dört farklı boyutunu, iyi çözücü HFIP ve yetersiz çözücü P olmak üzere iki ayrı çözücüde benzetimlerini koşturduk. AKYD proteinlerinin bu çözücülerdeki davranışını irdelemek için radyal dağılım fonksiyonu ve yapı çarpanı analizleri kullanıldı. Sonuçlar gösterdi ki, AKYD proteinleri HFIP içinde şişer ve uzun mesafe düzen olmamasına rağmen topaklanır; çözücü P'de ise, deneysel verilerle örtüşen düzenli yapılar oluştururlar.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Canan Atılgan for her patience and continuous support. She has been an impeccable source of wisdom and encouragement, and her door was always open to me since the time I met her in 2nd grade. She allowed me to solve problems by myself but kept me on the right track whenever I needed. Most importantly, she has been a role model for my work ethics, and working with her has solidified my desire to be in academia.

I would also like to thank Prof. Ali Rana Atılgan and Prof. Melik Cumhur Demirel for accepting to be in my thesis jury and the valuable discussions we had on my thesis topic.

I thank all MIDST Lab members, but I am particularly grateful for the assistance given by Tandaç Fürkan Güçlü. He is a quirky and fun person disguised as a cynical old man, but he was the person that kept me on course when I didn't know what to do and was the mentor I never knew I needed with his "tough love" attitude.

Finally, I must express my appreciation to my parents, Zafer Çolak and Arzu Çolak, and Ceren Özer for supporting and encouraging me to overcome the difficulties I encountered throughout my years of study and during this thesis. Additionally, I would like to thank Cansu Öztürk for the productive late-night discussions and the fun we have had over the years.

"Love all, trust a few, do wrong to none."

William Shakespeare

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Structural proteins have repetitive amino acid sequences that provide stability and mechanical strength to the proteins. The general name given to such proteins is tandem repeat (TR) proteins. For example, spider silk[1] is a TR protein that comprises repetitive [AAAAAAAA] sequence surrounded by glycine-rich regions which form crosslinked fibers. The alanine sequence increases the rigidity of the fiber whereas the glycine-rich regions increase the toughness of the spider silk. Another example would be elastin[2] which has repetitive [VGVPG] and [VGGVG] sequences that increase the elasticity of the protein. Similar to spider silk, squid ring teeth (SRT) proteins also have two distinct regions, one AVSTH-rich (crystal-forming) and one glycine-rich (amorphous). Native SRT proteins[3] were shown to change their various properties, e.g. mechanical, and thermal, depending on their chain lengths. However, for native SRT proteins, the chain length was non-uniform so the structural control of the properties was limited and the crystal-forming and tie-chain regions had inconsistent amino acid sequences among native SRT proteins. So, we used a synthetic form of the SRT proteins in this study which have been synthesized to have a predetermined sequence which is then repeated to create various chain lengths of the SRT protein. The repeat number of the SRT proteins used in this study are 4, 7, 11, and 25 to compare our results with the experimental findings [4] for the corresponding proteins.

The main assumption of this study is the similarities between polymer and protein structures. Both proteins and polymers have smaller building blocks and specifically, the structure of the SRT protein resembles a block copolymer with soft and hard segments. So, we used the Flory-Huggins (FH) theory of polymers combined with DPD simulations to parameterize and characterize the protein-solvent systems. Another reason for this choice is the computational cost of MD simulations compared to DPD simulations. MD

simulations are atomistic and the step size is on the order of femtoseconds, so a system with a solvent and multiple proteins would be computationally expensive to simulate. The method we use to parameterize the proteins in this study has been used in previous work[5, 6] to simulate polymer solutions. In one of the studies[5] (Avaz, 2017), poly(ethylene oxide) based poly(urethane-urea) copolymers were parameterized using similar methodology to understand the structure-property behavior of the polymer in binary solvent (THF and DMF) systems. The study concludes that the DPD simulations allowed for longer simulations with a higher number of chains of polymers at a lower computational cost compared to other methods while providing excellent agreement with experimentally observed morphologies.

In this study, our objective is to parameterize and characterize the SRT proteins in two different solvents, HFIP, and the hypothetical solvent P. HFIP is a known good solvent and the experimental results for the SRT proteins in HFIP is readily available to compare our computational results. Solvent P is based on the water which is a known poor solvent for the SRT proteins. We use solvent P to promote the clustering of the SRT proteins to observe their nanostructure. The parameterization starts by using MD simulations to calculate cohesive energy densities (CED) using binary MD simulations for all pairwise interactions in our system. After the MD simulations, we construct a forcefield for DPD simulations and use that to simulate our protein-solvent systems. Then, we use radial distribution function (RDF), structure factor, and contact map analysis to characterize the simulations.

2. THEORY AND METHODS

2.1. Molecular Dynamics (MD) Simulations

MD simulations solve Newton's 2nd law of motion over a specified period to simulate the trajectories of atoms. Definition of force in MD simulations can be written as,

$$F_i(t) = m_i a_i(t) \tag{1}$$

where F_i is the force exerted on the particle i, t is the time m_i is its mass and a_i is its acceleration. The second derivative of the position with respect to time provides the acceleration of a particle at a given point in time. Initial positions of the particles are assigned randomly which may not be at a realistic starting point. All non-bonded interactions between pairs of particles are treated by the Lennard-Jones potential energy function (Figure 2.1),

$$E_{LJ}(r_{ij}) = 4\varepsilon[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}] \tag{2}$$

In equation 2, r_{ij} is the distance between atoms i and j, σ_{ij} is the distance between atoms i and j at which the attractive and repulsive interactions cancel and ε is the energy minimum. Both σ_{ij} and ε depend on material properties of atoms i and j. In addition to the nonbonded interactions, depending on the geometry of the molecules to be investigated, additional terms are included in the force field. For example, bonded atoms are treated with a harmonic potential.

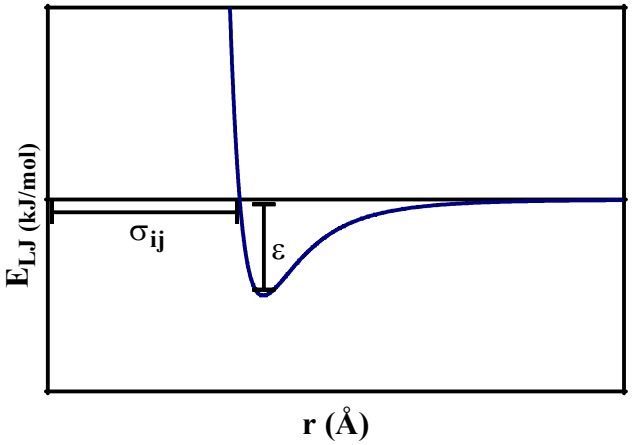


Figure 2.1 Representation of Lennard-Jones potential energy function

Before a simulation, minimization algorithms such as steepest descent or Newton-Raphson are used to eliminate possible clashes between atoms due to the random distribution of the particles; overlapping atoms cause unreasonably high energies in the system which in return lead to extremely large velocities of atoms in the MD simulations. During an MD simulation, the Verlet algorithm[7] is used to calculate velocity and acceleration from the initial position of the system for all the atoms in the system. Afterward, these initial values are iterated for number of times to calculate the positions and velocities in subsequent time points. Therefore, MD simulations are deterministic once the initial values and positions of the atoms are set following the minimization process. Forcefields determine the interactions of the particles in the system. In these forcefields, interactions due to quantum effects are approximated or ignored entirely depending on the forcefield. This is due to the difficulty of calculating these effects for every single atom and also the fact that MD simulations are used to analyze a system's bulk properties[8] (Density, Young's Modulus, Radial Distribution Function, etc.) which is determined using statistical mechanics whereby quantum effects are averaged out.

2.2. Binary MD Simulations for amino acid and solvent parametrization

Binary systems of amino acids and solvents were put into simulation boxes (Figure 3.1) using the Amorphous Cell module of Materials Studio 2018 (MS'18) [9] to ensure the data set is large enough to mimic realistic interaction parameters. The density of the

simulation boxes was fixed to 1.0 g/cm^3. Ewald summation technique was used for the computation of long-range interactions in the periodic system, and cut-off value for intermolecular interactions was set to 12.5 Å. As forcefield, widely used COMPASS[10] (Condensed-phase Optimized Molecular Potentials for Atomistic Simulation Studies), COMPASS27[5], and COMPASS II[11] forcefields were considered. Using several short simulations, we determined COMPASS II as the most suitable to use for our systems. COMPASS II builds on COMPASS and COMPASS27 forcefields which have been widely used to study conformational properties of polymers and were parameterized for a significant number of drugs in addition to the already existing chemicals. This results in COMPASS II being better suited for studying structures containing amino acids. The temperature of the simulation was regulated using Andersen[12] thermostat with a collision ratio of 1.0. To equilibrate the system, Geometry Optimization feature of the Forcite module was used, followed by 50 ps canonical ensemble (NVT) simulations. The calculation of cohesive energy density is done using the Forcite module. Cohesive energy density (CED) is the energy needed to separate a unit volume of a material to infinite distance from a solution. However, in theory, the equation used to calculate the cohesive energy is,

$$E_{coh,i} = \Delta H_{vap,i} - RT \tag{3}$$

where ΔH_{vap} is the enthalpy of vaporization of molecule i, E_{coh} is it's cohesive, R is the gas constant and T is the temperature of the system. CED is the cohesive energy of a system per unit volume area given by,

$$CED_i = \frac{E_{coh,i}}{V_{m,i}} \tag{4}$$

where $V_{m,i}$ is the molar volume of molecule i. To find the molar volume of each amino acid and solvent molecules, ACD/Labs Chem Sketch[13] freeware was used. We build all molecules using the software and optimize its geometry to calculate the molar volume of the molecule. Using the CED values obtained from MD simulations for all possible pairs of molecules in the study, we moved on to Dissipative Particle Dynamics (DPD) to construct and simulate the systems.

2.3. DPD Simulations

DPD simulations are governed by Newton's laws of motion (1), similar to MD simulations. The force on bead *i* due to all interactions with neighboring beads *j* is defined by[14],

$$F_{DPD,i} = \sum (F_{ij}^{C} + F_{ij}^{D} + F_{ij}^{R} + F_{ij}^{S}) \tag{5}$$

where $F_{DPD,i}$ is the total force acting on bead *i* in a DPD simulation, F_{ij}^{C} is the conservative force acting on bead *i* due to the neighboring beads *j*, F_{ij}^{D} is the dissipative force acting on bead *i*, F_{ij}^{R} is the random force acting on bead *i* and F_{ij}^{R} is the spring force for covalently bonded beads, *i,* and *j* in a coarse-grained molecule/polymer. The conservative force, F_{ij}^{C}, is given by[14],

$$F_{ij}^{C}(r_{ij}) = \begin{cases} a_{ij}\left(1 - \dfrac{r_{ij}}{r_{c,ij}}\right)\hat{r}_{ij} & r_{ij} < r_{c,ij} \\ 0 & r_{ij} \geq r_{c,ij} \end{cases} \tag{6}$$

where $\hat{r}_{ij}$ is the unit vector indicating the direction of the conservative force, a_{ij} is the interaction between beads *i* and *j* in DPD units, r_{ij} is the distance between beads *i* and *j* and $r_{c,ij}$ is the cutoff distance. Conservative forces are the only form of force in DPD simulations to be related to material properties through the parameter a_{ij}. a_{ij} is in DPD units and is always positive. This property of a_{ij} leads to conservative forces being always positive, since $\left(1 - \frac{r_{ij}}{r_{c,ij}}\right)$ part equation 6 is always positive for $r_{ij} < r_{c,ij}$. Therefore, conservative forces are always positive and repulsive, so the only interaction between any given pair of beads is repulsive and there is no attraction in DPD simulations. The equation for $E_{DPD,ij}$ is given by,

$$E_{DPD,ij}(r_{ij}) = \begin{cases} \frac{a}{2}\left(1 - \frac{r_{ij}}{r_{c,ij}}\right)^2 & r_{ij} < r_{c,ij} \\ 0 & r_{ij} \geq r_{c,ij} \end{cases} \quad (7)$$

The intermolecular interaction energy, $E_{DPD,ij}$, in DPD simulations, is always positive similar to the conservative force. Comparing Figure 2.1 to Figure 2.2, it is clear that there are no energy minima in DPD simulations, unlike Lennard-Jones potentials.

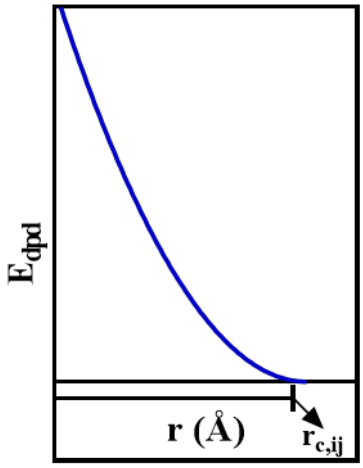


Figure 2.2 Representative graph of the intermolecular interaction energy in DPD simulations

As for the other types of forces, the equations for the dissipative force, F_{ij}^{D}, and the random force, $F_{ij}^{R}(r_{ij})$, are[14],

$$F_{ij}^{D}(r_{ij}) = -\mu\omega^{D}(r_{ij})(\hat{r}_{ij} \cdot v_{ij})\hat{r}_{ij} \quad (8)$$

$$F_{ij}^{R}(r_{ij}) = \sigma\omega^{R}(r_{ij})\xi\Delta t^{-1/2}\hat{r}_{ij} \quad (9)$$

where μ is the friction coefficient, σ is the amplitude of the random force, v_{ij} is the velocity of the beads *i* and *j* relative to each other, ξ is a randomly generated number with mean zero and range [-1,1], Δt is the time step of the simulation and $\omega^{D}(r_{ij})$ and $\omega^{R}(r_{ij})$ are the weight functions for the dissipative force and the random force, respectively; they depend on the distance between beads *i* and *j*. σ is given by[14],

$$\sigma = \sqrt{2\mu k_B T} \tag{10}$$

where k_B is the Boltzmann constant. We also relate two weight functions, $\omega^D(r_{ij})$ and $\omega^R(r_{ij})$, defined in the relations (8) and (9) by the expression[14],

$$\omega^D(r_{ij}) = [\omega^R(r_{ij})]^2 = \begin{cases} \left(1 - \frac{r_{ij}}{r_{c,ij}}\right)^2 & r_{ij} < r_{c,ij} \\ 0 & r_{ij} \geq r_{c,ij} \end{cases} \tag{8}$$

The role of the random and dissipative forces in DPD simulations is to correct the equilibrium of the system and keep the system in the canonical ensemble (NVT). The fourth component of the relation (5) is the spring force, $F_{ij}^S(r_{ij})$, defined by the expression[14],

$$F_{ij}^S(r_{ij}) = \sum C r_{ij} \tag{9}$$

where *C* is a spring constant defined in DPD units which is the same value for all the beads in the system. Other units of measurements are also taken in DPD units for simplicity and DPD units are unitless for calculation purposes. The values for $r_{c,ij}$ and mass *m* of a bead is taken as 1 DPD unit. Energy is in units of $k_B T$ which has a value of 1 DPD unit. Density ρ is a free variable in the DPD systems. Groot&Warren (1997) defined the relation of dimensionless compressibility, (κ^{-1}), to density and a_{ij} by[14],

$$\kappa^{-1} \approx 1 + \frac{0.2 a_{ii} \rho}{k_B T} \tag{10}$$

where a_{ii} is the interaction parameter of bead *i* with other beads of the same type. Note that this relation is a good approximation if ρ>2 holds[14]. To find the interaction parameter for the bead with itself, we evaluate the equation above with a known material and a suitable density value. The lowest possible density value that can be chosen while keeping the approximation good was found to be 3[14]. As for the κ^{-1} value, we chose water which has dimensionless compressibility of 16. Plugging the values chosen into

relation (13), we get [14],

$$a_{ii} = 25k_BT = 25 \tag{14}$$

So, for ρ=3, DPD systems have a self-interaction parameter of 25.

2.4. Flory-Huggins Theory of Polymers

Flory-Huggins (FH) theory[15] is a model that expresses the energy for a given polymer-solvent system and predicts the behavior of that mixture. The Helmholtz free energy is suitable for defining the free energy of the mixture since our simulations are at constant volume and temperature,

$$\Delta F_{FH} = \Delta U_{FH} - T\Delta S_{FH} \tag{15}$$

where ΔF_{FH} is the Helmholtz free energy of mixing for a system, ΔH_{FH} is the enthalpy of mixing and ΔS_{FH} is the entropy of mixing. There are several assumptions to consider for the FH Theory:

- The system is defined as a lattice with an equal volume of lattice units as seen in Figure 2.3.
- Each lattice point may only be occupied by a monomer or a solvent molecule.
- Monomers and solvent molecules are assumed to be hard spheres so there is no overlapping allowed between the lattice units.
- Only the configurational entropy is considered when calculating the entropy of the system.

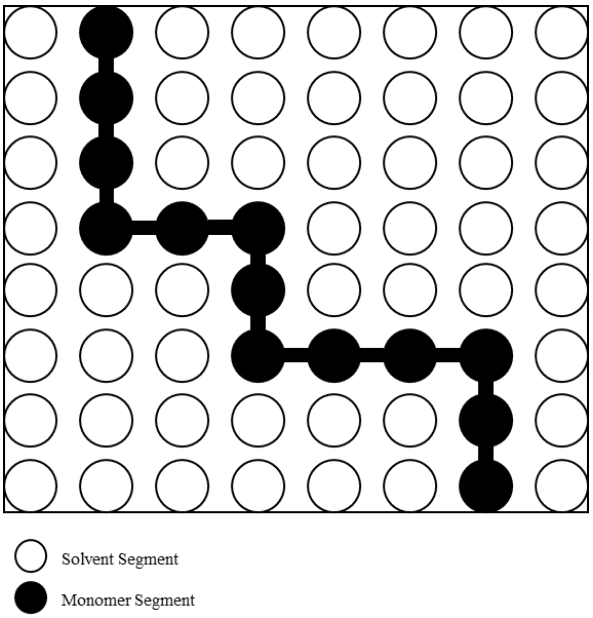


Figure 2.3 Visual representation of a polymer-solvent system defined by FH theory model

2.4.1. The entropy of Mixing for Polymer-Solvent systems

Entropy is related to the number of ways a system may be arranged,

$$S = k_B ln(W) \tag{16}$$

Here *S* is the absolute entropy and W is the number of ways monomers/molecules can be arranged on the lattice. For a given molecule, W is the same as the number of all lattice sites, n_l. If we consider a mixture of two types of species, solvent A and polymer P, then their change in entropy per lattice site is[16],

$$\begin{aligned} \Delta S_A &= k_B \ln(n_l) - k_B ln(n_l V_{f,A}) = -k_B ln(V_{f,A}) \\ \Delta S_P &= k_B \ln(n_l) - k_B ln(n_l V_{f,P}) = -k_B ln(V_{f,P}) \end{aligned} \tag{17}$$

where $V_{f,A}$ and $V_{f,P}$ are the volume fractions of the solvent molecule and the polymer,

respectively. We can construct the equation for the entropy of mixing per lattice site using the relations in (17)[16],

$$\Delta S_{FH} = -k_B(V_{f,A} ln(V_{f,A}) + \frac{V_{f,P}}{N_P} ln(V_{f,P})) \quad (18)$$

where N_P is the degree of polymerization for polymer P. Relation (18) shows that the entropy of mixing is directly dependent on the length of the polymer and the volume fractions of both the solvent and the polymer.

2.4.2. Free Energy of Mixing for Polymer-Solvent Systems

The enthalpy of mixing is the energy term which is the result of all the interactions in the mixture. When defining the free energy (equation 18), we did not include enthalpy directly into the equation. This is because enthalpy change is equivalent to the internal energy change for the Helmholtz free energy definition. We can prove it by writing the general expression for enthalpy change in a system,

$$\Delta H_{FT} = \Delta U_{FT} - P\Delta V \quad (11)$$

where ΔH_{FT} is the enthalpy change of the system, ΔV is the volume change of the system, and P is the pressure of the system. Since we assumed the lattice units would have no volume change, $P\Delta V$ term is 0. So, for our system, enthalpy and internal energy is equivalent. The internal energy of a polymer-solvent mixture per lattice site is defined as[16],

$$\Delta U_{FT} = \chi_{ij}\, V_{f,A} V_{f,P}\, k_B T \quad (20)$$

where χ_{ij} is the Flory-Huggins interaction parameter defined as[16],

$$\chi_{ij} = \frac{z}{2}(2\omega_{AP} - (\omega_{AA} + \omega_{PP})) \quad (21)$$

where z is the coordination number in the system (lattice), ω_{AA}, ω_{PP} and ω_{AP} are the interaction parameters for the solvent with itself, the polymer with itself and, between the solvent and the polymer, respectively. Combining equations (18) and (20) to construct the free energy of mixing per lattice for the polymer-solvent mixture, we arrive at the following expression[16],

$$\Delta F_{FH} = k_B T (V_{f,A} ln(V_{f,A}) + \frac{V_{f,P}}{N_P} ln(V_{f,P}) + \chi_{ij} V_{f,A} V_{f,P}) \quad (22)$$

2.4.3. Good, Poor, and Theta Solvent Behavior

The main attribute of a good solvent is the unfolding of the chain for our systems. In a good solvent, repulsive interactions are too weak to cause a phase separation between the chains and the solvent. So, the chains will swell and expand into the solution. FH theory quantifies the solvent behavior with χ_{ij} parameter. If the value of χ_{ij} is between 0 and 0.5[17] where i is the solvent and j is the bead of a chain, then that solvent is assumed to be a good solvent. As the value of χ_{ij} decreases, the interaction parameter between protein beads and the solvent beads become more alike since similar beads are assumed to have 0 as their χ_{ij} parameter[17]. In poor solvents, the solvent beads and the protein beads separate into two phases which are polymer-rich and solvent-rich. This is due to the strong repulsion between protein and solvent beads. The border between good and poor solvents are at χ_{ij}=0.5[17] which is defined as the theta solvent. For χ_{ij} greater than 0.5[17], the solvent is assumed to be poor and segregation of the protein will be observed in the system. However, in a theta solvent, the repulsion between the solvent beads and the protein beads are stronger than good solvent but not enough to cause a phase separation.

2.5. Coarse Graining

Coarse graining, in this thesis, is transforming a molecule into a hard sphere (bead). While we lose details at the atomic scale, the general behavior of the system does not change. We coarse-grained the amino acids according to a study[18] on the subject and depending on the sizes of the amino acids, we used either a single bead or two beads. We coarse-grained the solvent, HFIP, into a single bead since its relative size difference to the amino acids was negligible (see Figure 2.4 for an example).

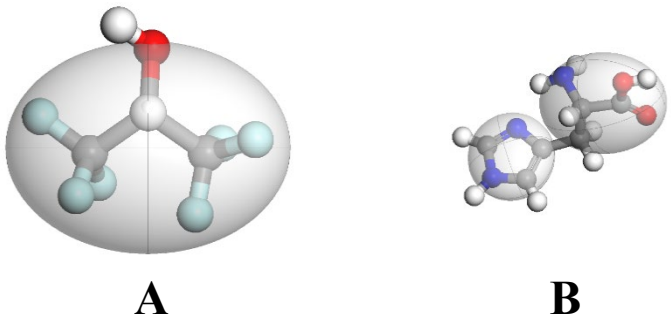


Figure 2.4 HFIP Molecule with its bead(A). Histidine with its beads(B).

2.6. Interaction Parameter Calculation

As defined in equation 14, the self-interaction parameter in DPD simulations is 25. Bead-bead interactions are defined as deviations from this self-interaction parameter and given by the following expression,

$$a_{ij} = a_{ii} + \Delta a_{ij} \tag{23}$$

where a_{ij} is the interaction between beads *i* and *j,* and Δa_{ij} is the deviation from the self-interaction due to bead-bead interactions. The studies of Groot&Warren (1997) has shown that Δa_{ij} parameter is directly related to the FH interaction parameter with the following relation[14],

$$\Delta a_{ij} = \frac{\chi_{ij} k_B T}{0.306} \tag{24}$$

Substituting equation 24 into equation 23, we arrive at the relation between DPD interaction parameter and the FH interaction parameter,

$$a_{ij} = 25 + 3.27\chi_{ij} \tag{25}$$

In equation 25, a_{ij} is the interaction parameter between the beads *i* and *j.* To find the interaction parameters between the beads, we needed to define the FH interaction parameter for our system. To find the energy of mixing of the system, we used CED values obtained from the MD simulations discussed in section 2.2. CED is also the square of the Hildebrand solubility parameter, δ,[19] which is used to quantify the solubility of systems containing polymers. The energy of mixing, $\Delta E_{mix,ij}$, is related to the CED of the beads, for both pure and mixed state, with the following relations [6],

$$\Delta E_{mix,ij} = \Phi_i CED_i + \Phi_j CED_j - CED_{ij} \tag{26}$$

where Φ_i is the volume fraction of amino acids or solvent molecules, i, in the MD simulations. The relation between the FH interaction parameter, χ_{ij}, and the energy of mixing as follows[6, 20],

$$\chi_{ij} = \left(\frac{\Delta E_{mix,ij}}{RT}\right) V_{bead} \tag{27}$$

where V_{bead} is the average volume of beads i and j. Flory-Huggins interaction parameters may be obtained experimentally, or they may be related to a material's solubility parameter. $\Delta E_{mix,ij}$ is calculated using solubility parameter $\delta = \sqrt{(CED)}$. By combining equations 26 and 27, we find the FH parameter for all pairs in the system. Then, using equation 25, we construct the DPD interaction parameters for all the beads in the system.

2.7. Radial Distribution Function (RDF)

RDF measures how atoms are distributed in a system with reference to a particle. To calculate RDF for the whole system, shells with a fixed width are used. RDF is given by the following relation[21],

$$g(r) = \frac{dn_r}{\rho(r) dV_{shell}} \tag{28}$$

where $\rho(r)$ is the density of the material within the specified radius, $dV_{shell} = 4\pi r^2 dr$ is the volume of the spherical shell and dn_r is the number of atoms in the spherical shell [22].

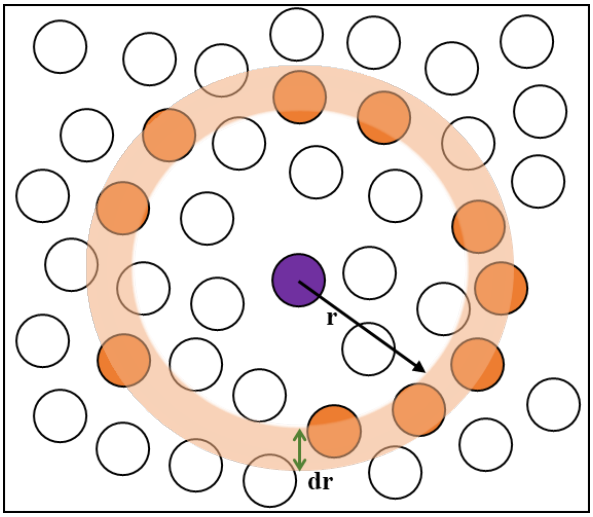


Figure 2.5 Visualization of RDF shells

As schematically shown in Figure 2.5, the labeled atoms are the only ones that are in the shell, so they are counted as being in the shell with the thickness of *dr*. The value for *dr* is arbitrary but it should be chosen according to the system size to account for the noise and the desired level of details. The long-range value of the RDF converges to one in amorphous systems. Thus, the main benefit of the RDF analysis is the local structural information it can provide for a system. Especially, the crystal structure of a material has considerably higher intensity in its RDF values compared to amorphous regions.

2.8. Structure Factor Analysis

Structure factor, $I(s)$, is experimentally obtained using methods such as x-ray scattering. In the case of small-angle x-ray scattering (SAXS), it is used to characterize the overall structure of materials. In computational studies, the structure factor may be related to the coordinates of the atoms in a simulation since it is the Fourier transform of the RDF. The relation between RDF and structure factor is given by the following relation[23],

$$g(r) = \frac{r^2}{2\pi^2}\int_0^\infty s^2 I(s)\frac{\sin(sr)}{sr}ds \tag{29}$$

where s is related to the diffraction angle θ and the wavelength of the incoming light λ through,

$$s = \frac{4\pi \sin\theta}{\lambda}\text{Å}^{-1} \tag{29}$$

While r in RDF analysis gives information on the short-range structure of the material, since s has the inverse units of r, structure factor analysis gives information about the structure of the overall system rather than the local details. Smaller s values such as 0.01 Å^{-1} corresponds to distances on the order of 100 Å and can describe the interaction between large nanoscale assemblies at such distances, as well as the size and shape of those nanostructıres.

2.9. Contact Map Analysis

Contact Maps[24, 25] are used to visualize the close contacts in a structure, for pairs of selected particles residing at distances below a preselected threshold. This information is arranged into a symmetric matrix; e.g. to disclose local structural organizations such as alpha helices or beta sheets formed by amino acid arrangements in folded proteins. For a system of N amino acids, this information is arranged in a $N \times N$ matrix. If the distance between amino acids i and j is lower than the threshold value, the value of (i,j) is 1 or otherwise, it is 0. Using this simple method, it is possible to identify secondary structures of proteins and other trends emerging from the packing of amino acids in specific orders.

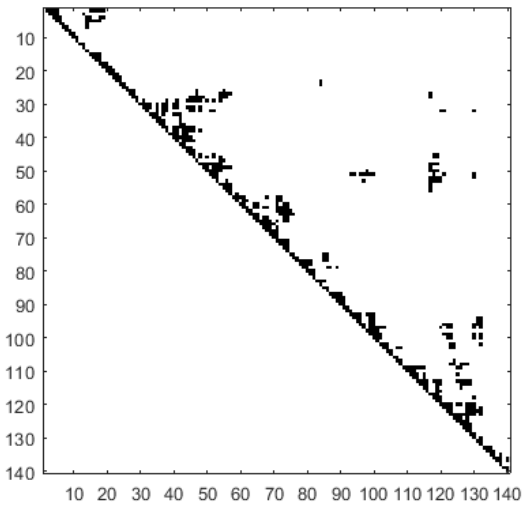


Figure 2.6 Contact Map analysis of a single chain Squid Ring Teeth protein in Poor Solvent

An example contact map is shown in Figure 2.6. To quantify our findings with the contact map analysis, we counted and classified how many beads are clustered together for a given system. The exact method of finding and counting clusters is given by the following pseudo codes:

Cluster_Finder.m (Appendix A-1)

1. Input: Contact Map as an upper-diagonal binary matrix
2. Check two elements at a time until both are 1 and record the pointer position
3. If found, put them into an array, check other elements along the diagonal
4. Repeat until all the neighboring elements are checked and recorded into the array at Step 3
5. When no other elements are found in the group, record all their indices (vertical and horizontal separately) as arrays and erase them from the original array
6. Move the pointer one element forward
7. Return to Step 2
8. Run the script until all the matrix has been checked
9. Output: Indices as arrays from Step 5

Cluster_Counter.m (Appendix A-2)

1. Input: Indices as arrays (V-H Arrays) from the first script
2. Two empty (Temp) arrays are used to store elements from the arrays of Step 1
3. If any two elements have consecutive indices, they are added to Temp arrays created in Step 2
4. All the elements that are added to Temp arrays in Step 3 are deleted from V-H arrays from Step 1
5. Repeat steps 2-4 until all the consecutive indices are found for a single group
6. After all the elements for a group is found, the size of Temp arrays is recorded in an array (Out) and then Temp arrays are emptied
7. Repeat from Step 1 using the updated V-H arrays until they are empty
8. Output: Out array with the cluster sizes of every group counted

To generate the contact map matrices, R[26] freeware was used with **Bio3D**[27] package. Then, using MATLAB R2020a[28] Software, the scripts mentioned above were used to process the contact map matrices.

3. RESULTS AND DISCUSSION

Two different solvents were used in the study which is HFIP and a hypothetical poor solvent denoted P. The decision to use these HFIP came from the experimental results[29]. HFIP is known to be a good solvent for the SRT proteins regardless of their sizes. This enabled us to compare our computational results to the experimental results. However, to examine the secondary structure formation of the system, a poor solvent was needed for the SRT proteins. To that end, solvent P was arbitrarily parametrized to have high FH interaction parameters with all the amino acids in the SRT proteins.

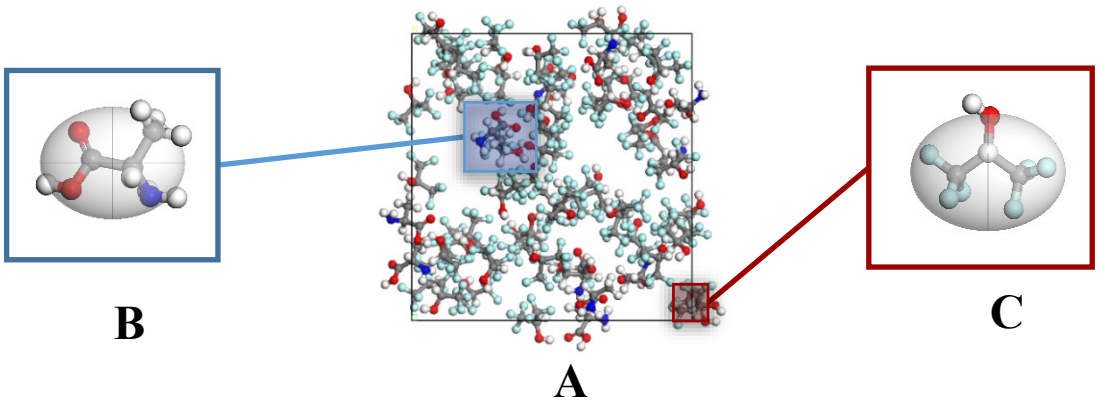


Figure 3.1 Alanine/HFIP Molecular Dynamics Box(A). Alanine as a molecule and a bead(B). Hexafluoro-2-propanol as a molecule and a bead(C)

We examined four SRT proteins in this study which were named n4, n7, n11, and n25. The difference between the SRT proteins is the repeat number of the same sequence of amino acids. Repeat units represent the number of times the sequence of the SRT proteins is repeated for a given chain. Thus, the repeat units for the SRT proteins are 4, 7, 11, and 25 for SRT proteins n4, n7, n11, and n25, respectively. These segments are made up of (Ala), glycine (Gly), histidine (His), leucine (Leu), proline (Pro), serine (Ser), valine

(Val), and tyrosine (Tyr) amino acids. The amino acid sequence for the segments is as follows[30],

$$(STGTLSGYGGYGGLGYGGLGYGGLGGYG)_1P(AAASVSTVGHGH)_2P$$

where the parentheses indicate regions numbered one and two which are the tie-chain, and the crystal forming regions, respectively (Figure 3.2). The tie-chain region is shown to be amorphous[30] and the crystal forming region is rigid. However, the exact secondary structure of the protein is unknown.

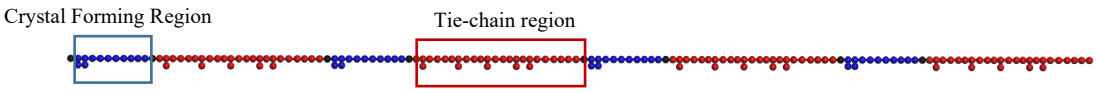


Figure 3.2 Coarse grained SRT protein with four repeat units (n4). Blue labeled region is the crystal forming region and red labeled region is the tie-chain region

To parametrize amino acids, we assigned a single bead to all of them except tyrosine and histidine. Histidine and tyrosine were divided into two beads[18], one for their backbone and another for their side chain. Their backbone beads were assumed to interact similarly to glycine so the interaction parameter for glycine is assigned to the backbone beads. The side chain beads were parameterized separately. Each distinct amino acid and solvent structure were constructed using MS'18 **sketch atom** feature. To equilibrate the structures, we used the **geometry optimization** feature of the Forcite module of MS'18 with a maximum of 50000 iterations each. These are followed by MD simulations as described in section 2.2

3.1. Comparison of Two Solvents and DPD Interaction Parameters

To compare HFIP and solvent P systems, we kept the concentration of the systems at 2 wt.% of the solvent. The last snapshot of the simulation in Figure 3.3 shows the SRT-n4 proteins as swollen in HFIP which indicates that the solvent beads have attractive or weak repulsive interactions with the amino acid beads. Similarly, experimental results indicate that HFIP is a good solvent for SRT protein systems since it is a hydrogen bond-forming polar solvent[29]. However, the SRT-n4 proteins cluster together in solvent P. The repulsion of the solvent P beads and the amino acid beads are strong enough to separate the system into two distinct phases so solvent P is a poor solvent for SRT-n4 proteins. Additionally, we simulated identical systems with SRT-n7, SRT-n11, and SRT-n25 proteins with HFIP and solvent P. The results of those simulations mirror the findings from the SRT-n4 proteins and solvent systems, and their last snapshots are displayed in Appendix B-1 and B-2.

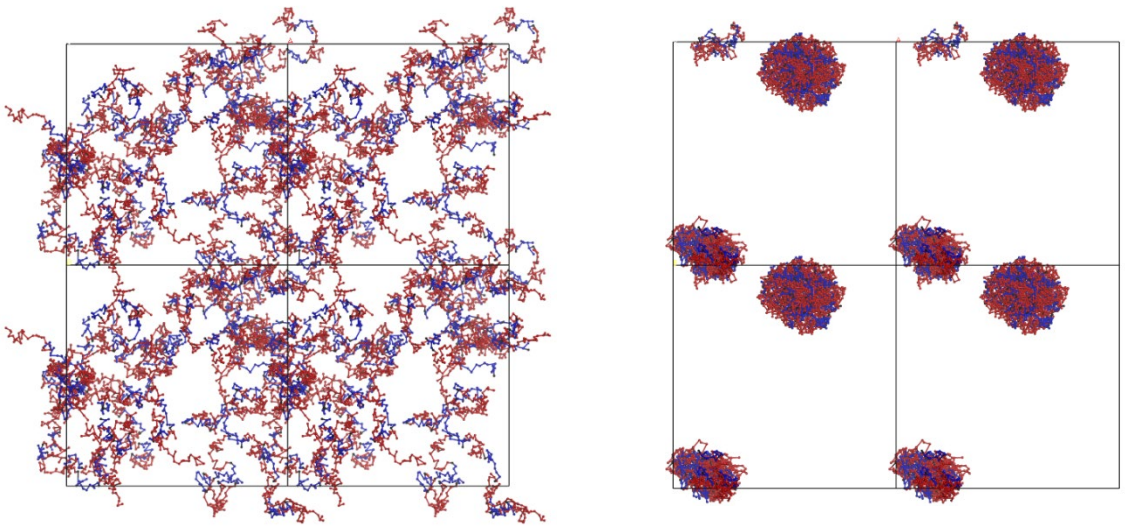


Figure 3.3 Last Snapshots of the HFIP-SRT-n4 (A) and Dilute Solvent P-SRT-n4 (B) DPD simulations

All pairwise interactions between the amino acid and solvent molecules are displayed in Table 3.1. The most notable interactions between amino acids include the side chain of histidine. The most attractive interactions in the system are between leucine and histidine,

followed by tyrosine and histidine. The common property of tyrosine and leucine is their hydrophobicity. Adding to this, the reactivity of histidine and tyrosine might have increased by the removal of their backbone. Other amino acid-amino acid interactions are either weak repulsions or attractions. The interaction parameters between HFIP and amino acids clearly show that HFIP is a good solvent for the system. Aside from leucine-HFIP pair, all other FH interaction parameters are below 0.5. These values also agree with the observations made in DPD simulations from Figure 3.3. Solvent P was parameterized using water molecules and amino acids in binary MD simulations. However, when calculating the FH interaction parameter, we did not account for the volume difference between water molecules and amino acids molecule to have a solvent which repels all amino acids strongly. So, the final FH parameters for solvent P are based on water, which is a known poor solvent for SRT proteins, but the values of the parameters are not representative of water molecules with the amino acids in the study. The value for the FH interaction parameter is over 5 for all interactions between the amino acids and solvent P which is an order of magnitude higher than the poor solvent interaction parameter limit.

Table 3.1 The FH Interaction Parameters, χ_{ij}, are in bold and in the upper diagonal part of the table whereas DPD Interaction Parameters, a_{ij}, are in the lower diagonal part of the table. Colored borders indicate those beads that are only found in a specific region in the SRT, red for tie-chain region and blue for crystal forming region

	Ala	Ser	Thr	Gly	Leu	Tyr*	Pro	Val	His*	HFIP	Solvent P
Ala	25/**0**	**0.25**	**0.05**	**0.19**	**-0.03**	**-0.31**	**0.11**	**0.11**	**-0.07**	**0.02**	**5.13**
Ser	25.82	25/**0**	**-0.17**	**0.08**	**-0.03**	**0.02**	**0.15**	**0.26**	**0.03**	**0.39**	**5.21**
Thr	25.17	24.44	25/**0**	**-0.06**	**-0.17**	**-0.24**	**0.37**	**0.19**	**-0.26**	**0.34**	**6.19**
Gly	25.62	25.25	24.79	25/**0**	**-0.37**	**-0.27**	**0.05**	**0.05**	**0.00**	**0.19**	**4.90**
Leu	24.91	24.89	24.44	23.80	25/**0**	**-0.07**	**0.35**	**0.37**	**-1.27**	**0.85**	**7.82**
Tyr*	24.00	25.08	24.21	24.11	24.77	25/**0**	**-0.06**	**-0.03**	**-0.52**	**-0.01**	**6.62**
Pro	25.35	25.48	26.21	25.17	26.15	24.79	25/**0**	**0.38**	**-0.04**	**0.22**	**6.22**
Val	25.34	25.84	25.63	25.17	26.20	24.89	26.24	25/**0**	**-0.09**	**0.45**	**6.59**
His*	24.76	25.09	24.15	24.99	20.86	23.30	24.86	24.71	25/**0**	**-0.49**	**5.32**
HFIP	25.08	26.26	26.11	25.62	27.76	24.97	25.72	26.48	23.39	25/**0**	-
Solvent P	41.79	42.04	45.24	41.04	50.58	46.65	45.34	46.56	42.41	-	25/**0**

** The values for Histidine and Tyrosine are for the side chain beads, not the whole amino acid.*

3.2. Effect of Solvent Type on the Organization of the TR proteins

Radial distribution function gives valuable information on the short-range order in a molecular system. For our protein, the interactions between crystal and tie-chain parts are essential in understanding the nanostructure of the protein. To analyze our systems, we used the Mesocite Analysis module of MS'18 Software. Our cutoff distance was 20 Å and our interval dr (Section 2.6) for all RDF results was 0.05 Å.

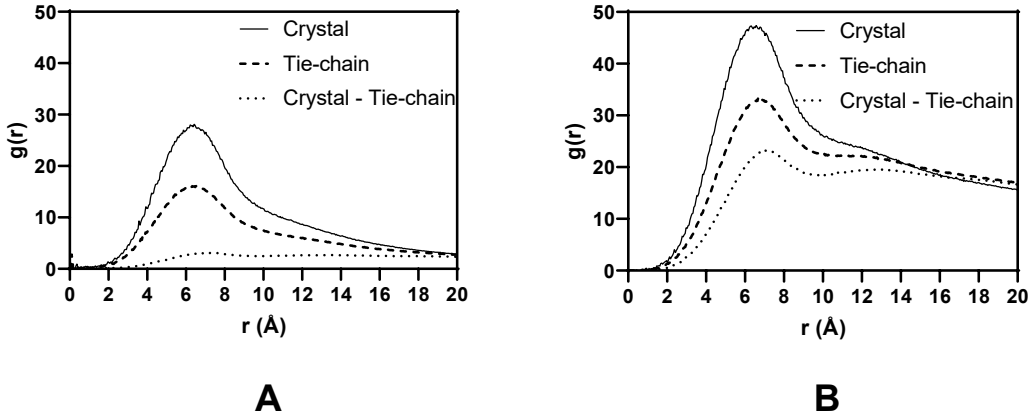


Figure 3.4 RDF Analysis of SRT-n4 Protein in Good (A) and Dilute Poor (B) Solvent.

In Figure 3.4, we display the results for the SRT-n4 as a representative system. The RDF results are almost identical for SRT-n7, SRT-n11, and SRT-n25 proteins in both solvents, and are not displayed separately. Good solvent and dilute poor solvent refer to HFIP and solvent P, respectively. The main RDF peak for both systems is around 7 Å which is almost twice the length of virtual bonds connecting the beads that are 4 Å. The intensities of the curves show that the beads in the crystal-forming region and tie-chain region prefer to interact with similar beads since the least interaction is between the tie-chain regions and crystal-forming regions. Crystal forming region and tie-chain region peaks are at the same distance; however, the intensity for the crystal forming region's peak is higher than that of the tie-chain. Experimentally[4], the crystal-forming region is denser compared to the amorphous tie-chain region. Therefore, the computational results agree with the

experimental observations.

Comparing the solvents, solvent P has higher peaks compared to HFIP. Since the RDF results are indicative of short-range interactions, they show that solvent P has more short-range interactions. Solvent P is a poor solvent and the SRT proteins cluster in this solvent as shown in Figure 3.3. Therefore, the higher number of short-range interactions are consistent with the properties of the solvent P. Short-range interactions are lower in HFIP because HFIP is a good solvent and the SRT proteins swell in good solvents. The overall distance between the amino acids increases as the protein swells and, as a result, the number of interactions between regions in the SRT proteins decreases.

Figure 3.5 shows the total system, intrachain, and interchain RDF results for the SRT proteins in a good solvent. The total RDF results show that the intensity of the crystal forming region's RDF peak is higher than the tie-chain region's RDF peak for all SRT protein systems. Intrachain RDF results are identical to the total RDF results for both crystal-forming region and the tie-chain region which indicate that the system is dominated by the intrachain interactions. Interchain RDF results also support this claim since the intensity of the RDF results is more than an order of magnitude lower than the intrachain RDF results. These results are consistent with the previous findings since in a good solvent, the protein chains are swollen and the interaction between the chains is minimal. Figure 3.6 shows the total system, intrachain, and interchain RDF results for the SRT proteins in a dilute poor solvent. The total RDF, the intrachain RDF, and the interchain RDF results show that the crystal-forming regions have a higher intensity peak compared to the tie-chain regions. The SRT protein size difference between the systems only affects intrachain and interchain RDF results, not total RDF results. Both intramolecular RDF results, crystal-forming region, and tie-chain region, has higher peaks for longer chains of the SRT proteins in the systems. So, the longer chains interact with themselves more than the shorter chains interact with themselves. The interchain RDF results show the opposite trend compared to the intrachain RDF results. Longer chains have lower intensities in the interchain RDF results, so the shorter chains interact with other chains more than longer chains interact with others. Since the total RDF results of the systems is a combination of the interchain and the intrachain RDF results, the size dependency of the interchain and the intrachain RDF results cancel each other in the total RDF results.

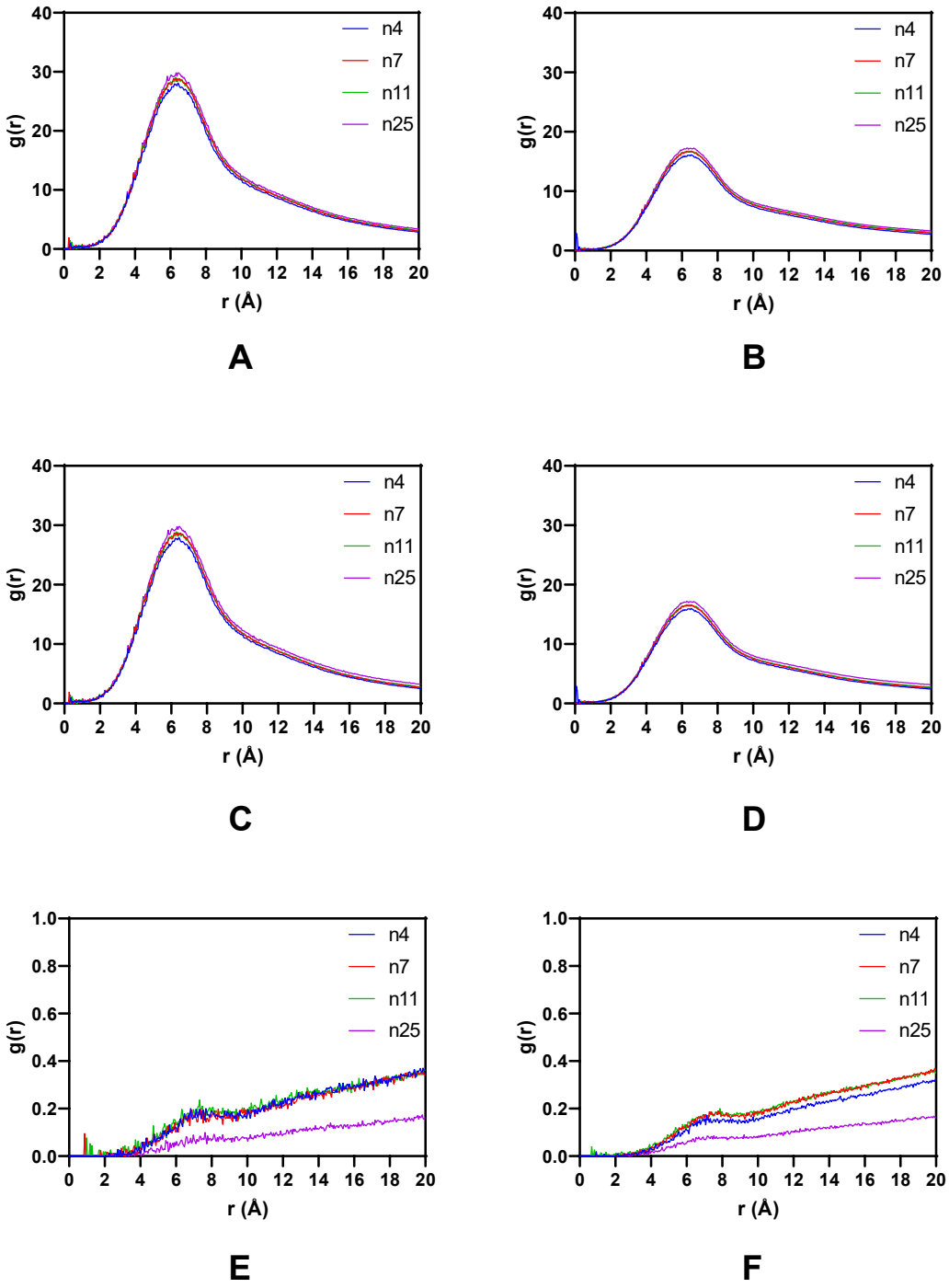


Figure 3.5 SRT Proteins in HFIP. Total System RDF Analysis of Crystal region (A) and Tie-chain region (B); Intrachain RDF Analysis of Crystal region (C) and Tie-chain region (D); Interchain RDF Analysis of Crystal region (E) and Tie-chain region (F).

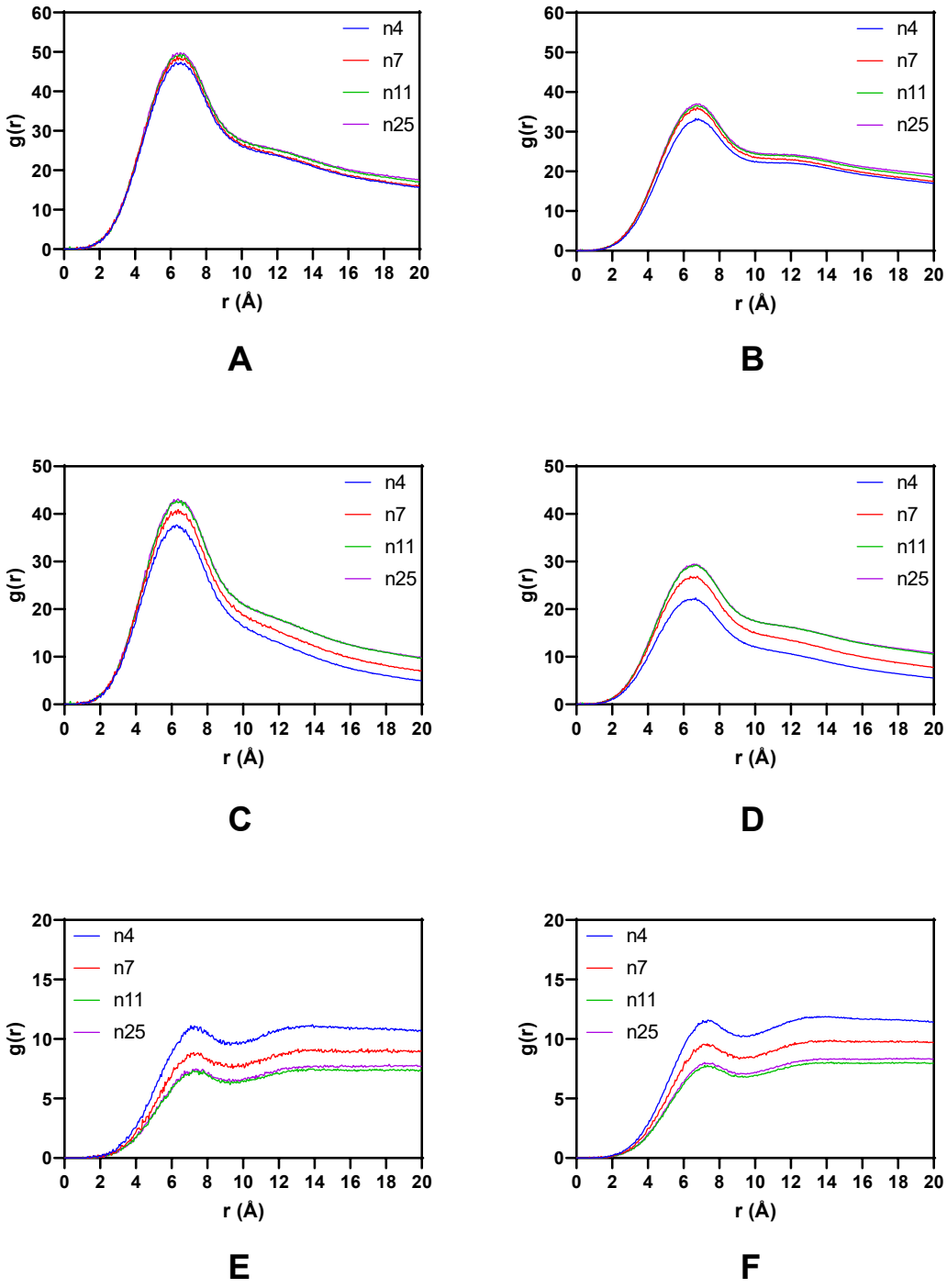


Figure 3.6 SRT Proteins in Dilute Poor Solvent Total System RDF Analysis of Crystal region (A) and Tie-chain region (B); Intrachain RDF Analysis of Crystal region (C) and Tie-chain region (D); Interchain RDF Analysis of Crystal region (E) and Tie-chain region (F).

The RDF results for both the good solvent and the dilute poor solvent show that the crystal forming region's beads are more clustered than the tie-chain region's beads. The interaction within a chain is high in both solvents, however, interchain interactions are only present in the dilute poor solvent. Also, the size of the SRT proteins change the interaction amount in the dilute poor solvent, but do not have any effect in the good solvent.

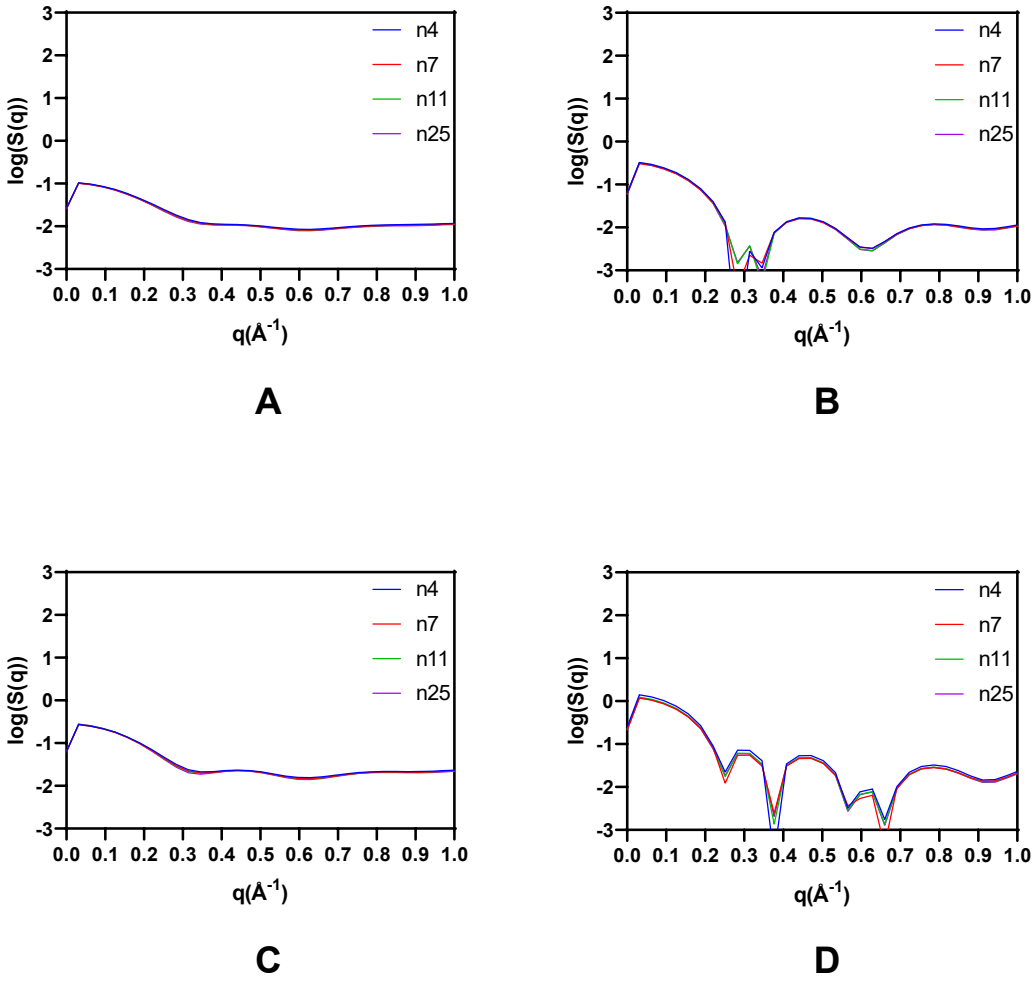


Figure 3.7 Structure Factor calculation of SRT Proteins of 1. Crystal region in Good (A) and Dilute Poor (B) Solvent 2. Tie-chain region Good (A) and Dilute Poor (B) Solvent.

Figure 3.7 shows the structure factor calculation results of the SRT proteins in the good solvent and the dilute poor solvent. The structure factor graphs for the crystal-forming region and the tie-chain region are identical for the SRT proteins in both solvents. The structure factor calculations of the systems containing SRT proteins and the good solvent have a limited amount of detail. The lack of long-range order is consistent with the established characteristics of SRT proteins in a good solvent. In contrast, the structure

factor calculations for the SRT proteins in dilute poor solvent point to a long-range order in the systems. However, the number of chains in the simulations were too few to understand the nanostructure of the systems.

3.3. Effect of Solvent Concentration on the Morphology of the TR Proteins in a Poor Solvent

We simulated the concentrated poor solvent systems with the same parameters as the good solvent and the dilute poor solvent parameter, except for the concentration of the solvent in the simulations. To find the optimal concentration, we simulated six systems that had 5wt.%, 10wt.%, 15wt.%, 20wt.%, 25wt.%, and 30wt.% poor solvent in SRT-n4 protein. The last snapshots of the simulations are in Appendix C. We observed that the morphology of the systems significantly changed after 25wt.% and the SRT proteins formed cylindrical continuous structures rather than forming clusters due to oversaturation. So, we increased the solvent concentration from 2wt.% to 20wt.% for the concentrated poor solvent simulations since the main purpose of these simulations over the dilute poor solvent simulations was to provide additional information on the nanostructure of the SRT protein systems.

Figure 3.8 shows the final morphologies of the SRT proteins in the concentrated poor solvent. All the SRT systems have simple cubic structures, but the connectivity of the spherical clusters changes as the number of repeats is increased from SRT-n4 to SRT-n25. The SRT-n4 system has a 1-D necklace-type morphology. The proteins form many spheres in the simulation box and the spheres connect in one direction only. The SRT-n7 system adds one more dimension to its morphology compared to the SRT-n4 system and forms a grid with the spheres and their connections. The SRT-n11 adds another dimension to the SRT-n7 system and creates a 3-D cubic structure. The chains stack around the edges of the cube to form spherical clusters and the clusters are connected via swollen chains. The SRT-n25 system's morphology is similar to the SRT-n11 system since they are both 3-D cubic structures. However, the SRT-n25 has additional connections along the main diagonal of the cube.

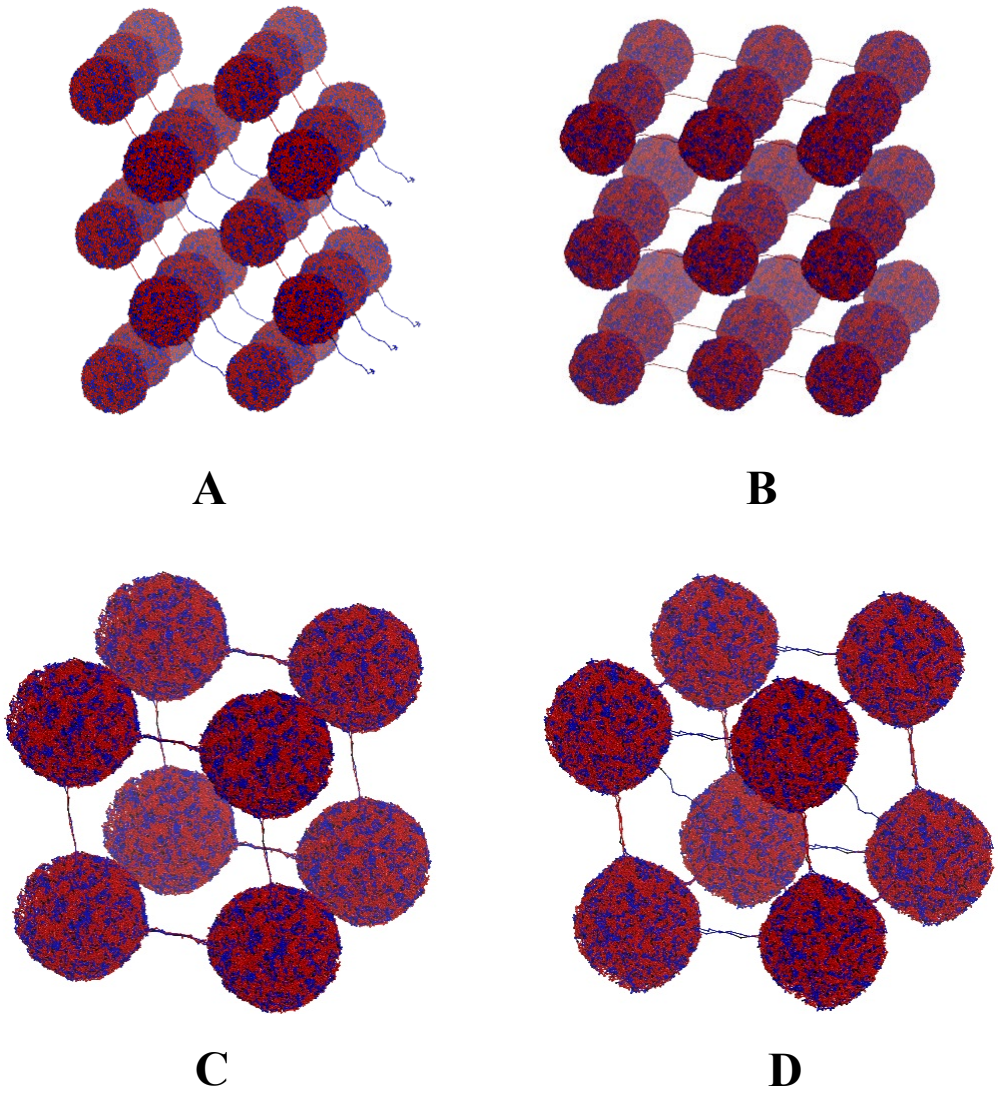


Figure 3.8 Last Snapshots of the HFIP-SRT n4 (A), n7 (B), n11 (C), and n25 (D) in concentrated poor solvent DPD simulations

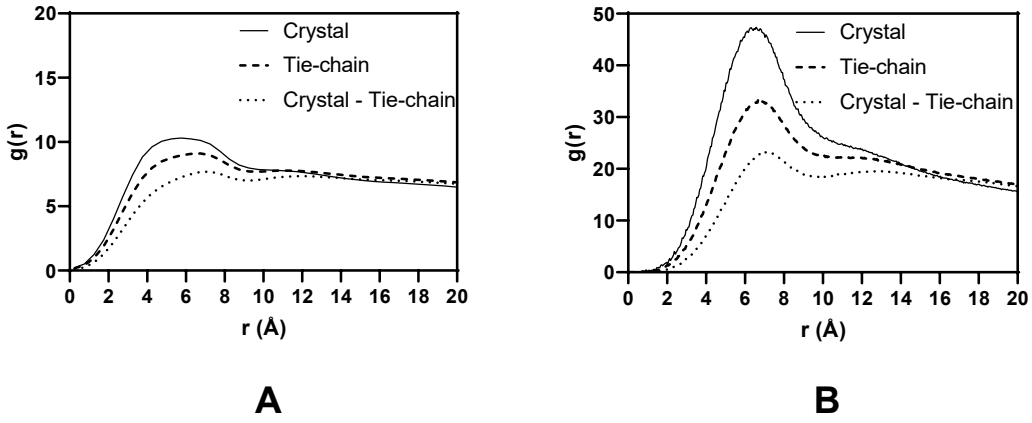


Figure 3.9 RDF Analysis of SRT-n4 Protein in Concentrated (A) and Dilute Poor (B) Solvent.

In Figure 3.9, we used SRT-n4 in concentrated and dilute solvents to observe the difference between the systems. The crystal forming and tie-chain region trends are similar between the systems, however, the RDF peaks are sharper for the dilute poor solvent system. Also, the peak of the RDF results shifts from 7 Å to 6 Å between the dilute poor solvent and the concentrated poor solvent. The concentrated poor solvent system has considerably more SRT protein chains than the dilute solvent system and the SRT proteins cluster in the poor solvent so the shift and broadening of the peaks in RDF results is the formation of bigger clusters in a concentrated poor solvent. As the cluster size increases, the number of interacting beads in proximity also increases. This increase forces both crystal forming and tie-chain region to interact within themselves and with one another more and it is the reason for the decrease in the difference between the RDF results of separate regions in SRT proteins.

Figure 3.10 shows the RDF results of the SRT proteins in concentrated solvent for the total system, intrachain interactions, and interchain interactions. The interactions of the crystal forming region and the tie-chain region are identical for the intrachain and the interchain RDF results. The chains interact with themselves until around 10 Å; however, after that point, the interaction within the chain starts to converge to zero. The interchain interactions peak around 6 Å, similarly to intrachain interactions, but as we move away from 6 Å, the RDF value stays constant. So, the interchain interactions are more

pronounced at higher distances of the RDF graphs. The size of the chains does not affect the RDF results of the systems.

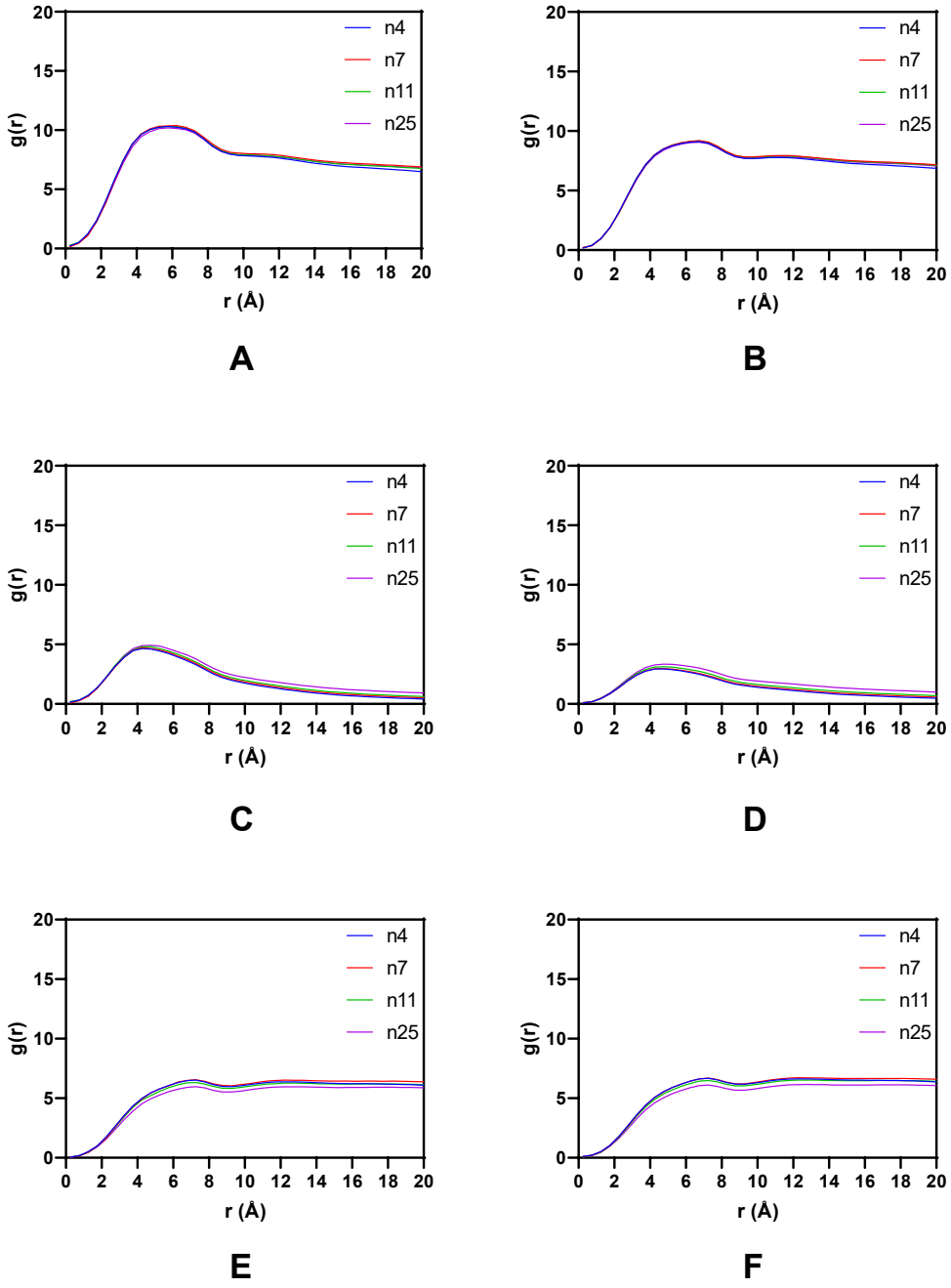


Figure 3.10 SRT Proteins in Concentrated Poor Solvent. Total System RDF Analysis of Crystal region (A) and Tie-chain region (B); Intrachain RDF Analysis of Crystal region (C) and Tie-chain region (D); Interchain RDF Analysis of Crystal region (E) and Tie-chain region (F)

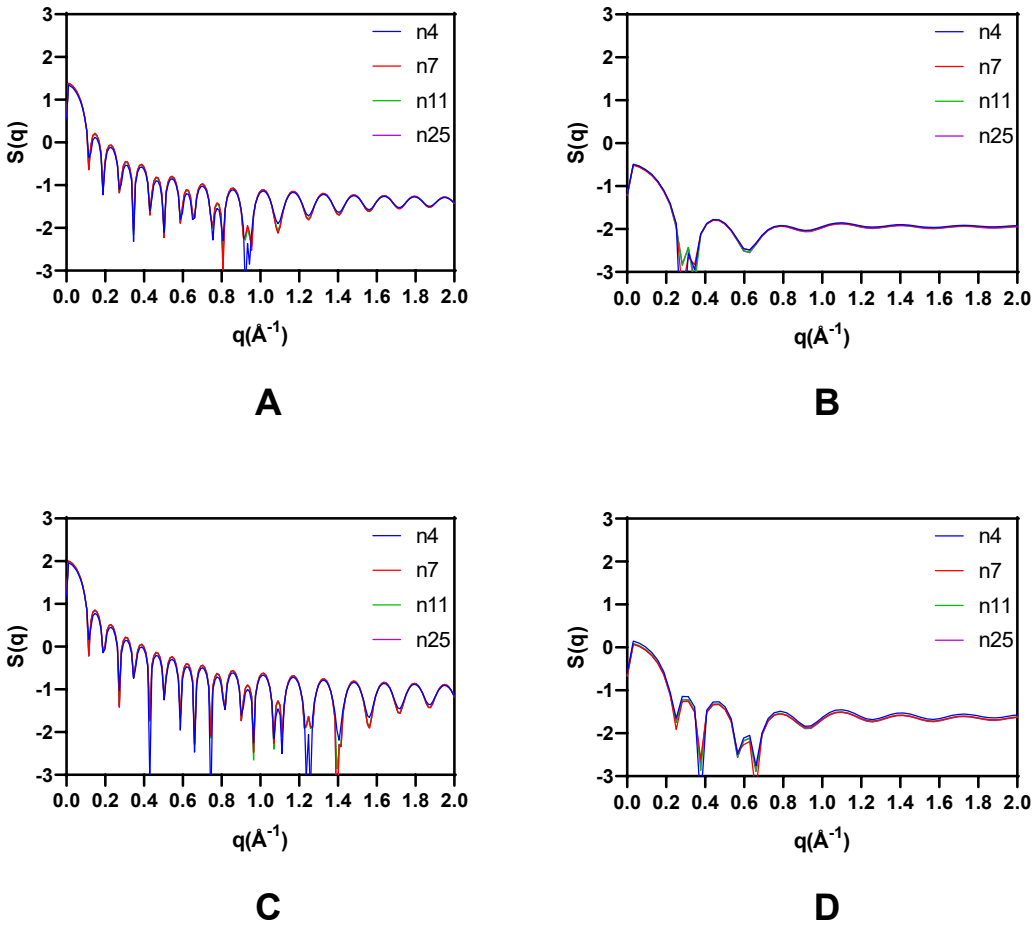


Figure 3.11 Structure Factor calculation of SRT Proteins of 1. Crystal region in Concentrated (A) and Dilute (B) Poor Solvent 2. Tie-chain region Concentrated (A) and Dilute (B) Poor Solvent.

In Figure 3.11, we compare the structure factors of the dilute and the concentrated poor solvent systems. The structure factor graphs indicate that the crystal forming region and the tie-chain region are identical for both poor solvent systems. The dilute poor solvent system's structure factor graphs have similar features to the corresponding graphs from the concentrated poor solvent; however, many of the details are missing. The concentration increase enabled us to study the system in a more detailed manner. The structure factor calculation results of the concentrated poor solvent show that there is an order in the system[23].

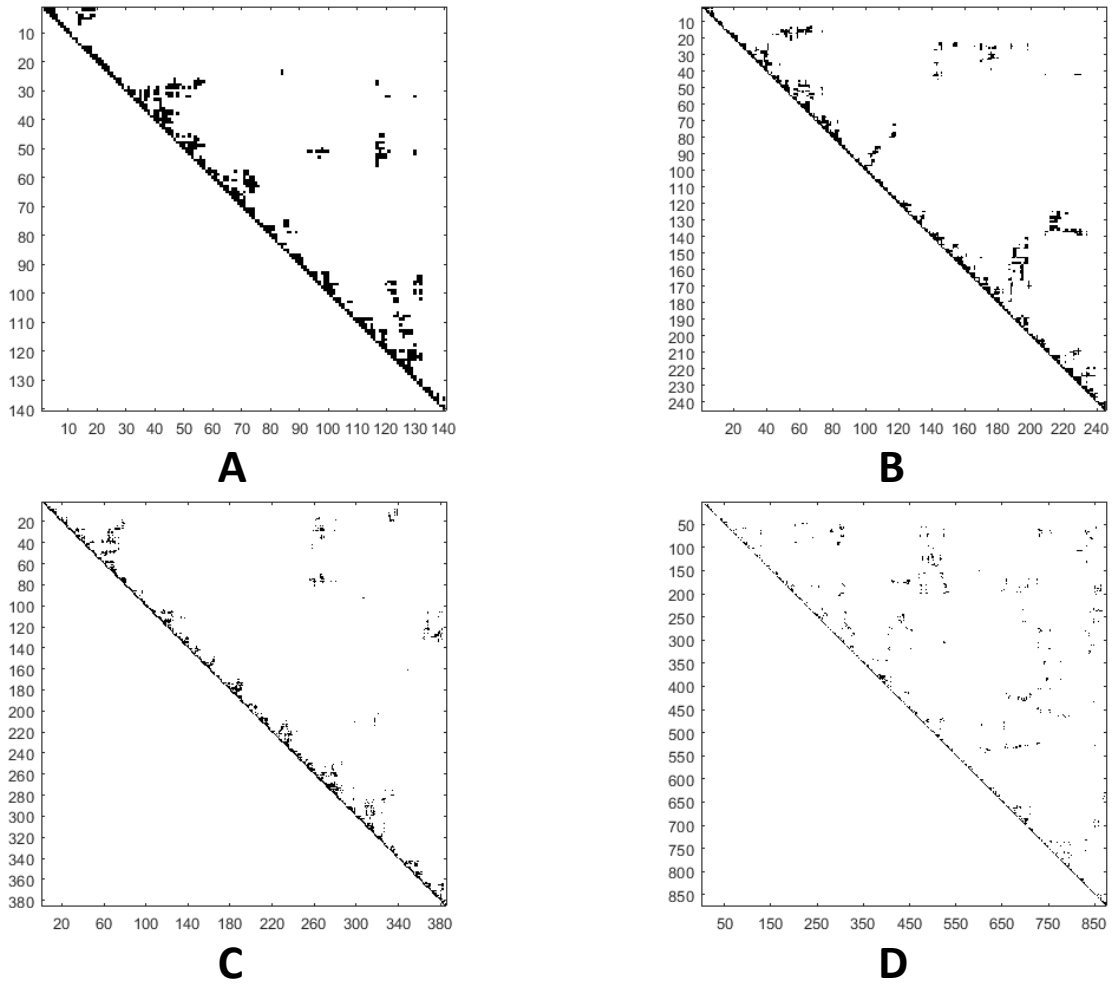


Figure 3.12 Contact Map Analysis of Single Chain SRT n4 (A), n7 (B), n11 (C), n25 (D) in concentrated poor solvent

Figure 3.12 shows the contact maps of single-chain SRT proteins in a concentrated poor solvent. Since the entire system contact map matrices are too large to plot and visually examine, the single chain contact maps are displayed. All the chains have multiple interaction regions between their amino acid beads as indicated by the clustered points in the plots. To quantify how the clusters form, we have resorted to counting the number and type of residues in the clusters as outlined in section 2.8.

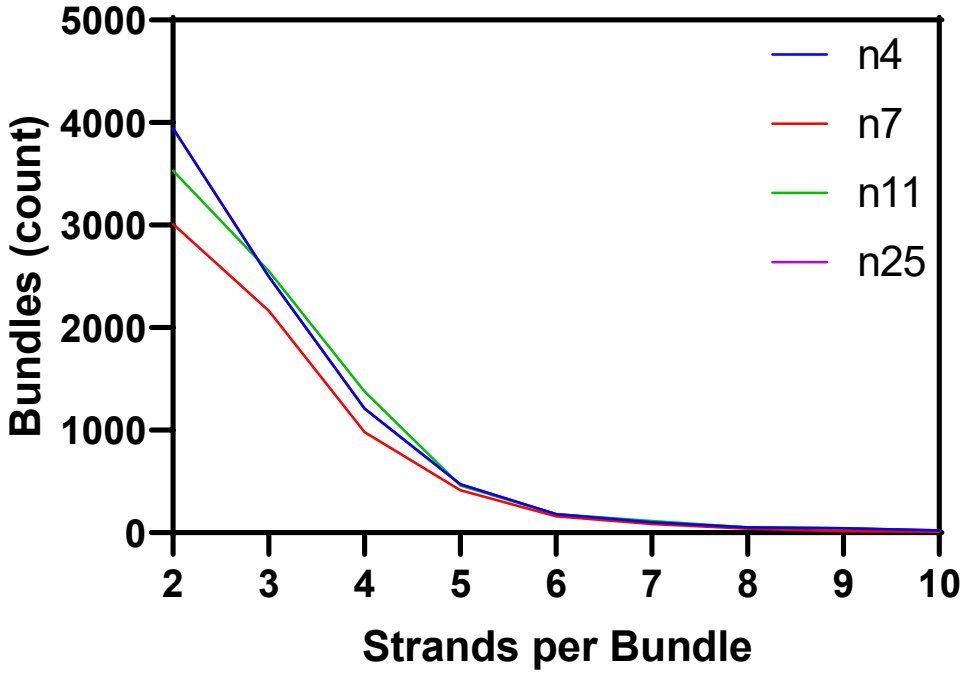


Figure 3.13 Analysis of the Contact Map of the SRT proteins in a concentrated poor solvent. Average strands per bundle in SRT n4, n7, n11, and n25 are 3.01, 3.01, 3.06, and 3.05, respectively.

Based on the experimental findings, the proposed nanostructure of the SRT proteins is ordered clusters that are on the order of 3-4 nm which are formed by the crystal forming regions and connected via tie-chain regions of the SRT proteins. [3] Figure 3.13 displays the number of strands per bundle residing in the computational morphologies. Experimentally, the number of strands per bundle is unknown, but it is known that the number is constant between all SRT proteins. So, the computational analysis of the bundles partially agrees with the experimental results, but the answer is not yet conclusive.

4. CONCLUSIONS

In this study, we used multiscale simulations to parameterize and model the SRT protein systems. The initial part of the study the parametrization of the amino acid and solvent molecules. We used HFIP and SRT protein system as a reference to check our simulations against the experimental data[3]. The swelling of the proteins, RDF results given in Figure 3.5, and the lack of long-range order shown by the structure factor results in Figure 3.7 clearly show that HFIP parameters that we calculated imitate the real-life interactions between HFIP and SRT proteins.

We used solvent P to characterize SRT proteins in a highly repulsive solvent. The first set of simulations we did with solvent P was the 2wt.% concentration solvent simulations in which we observed the clustering of the SRT proteins. However, the concentration was too low to get detailed information on the possible secondary structure formation of the SRT proteins. So, we increased the concentration of solvent P in our simulations from 2 wt.% to 20 wt.% after a concentration sweep study detailed in section 3.3. The second set of simulations we did with solvent P was the 20 wt.% solvent simulations. The SRT proteins clustered in these simulations and the long-range order of the system was more apparent compared to the 2wt.% solvent simulations as seen in Figure 3.3. Later stages of the study focused on the characterization of the clusters in a concentrated poor solvent. We used two MATLAB scripts on the contact map analysis results of the SRT protein systems to count all the β-sheet-like structures inside the clusters and classify them according to their sizes which are shown in Figure 3.13. These results point to ordered structures inside the clusters. However, the exact properties of these ordered structures are unknown currently.

Overall, the novel multiscale modeling process in this study is shown to be a good

approximation in the parameterization and prediction of the behavior of the SRT proteins in different solvents. This process is valuable for systems without proper databases in the literature which was the case for the SRT proteins. The forcefield parameterization and the MD simulations are computationally expensive compared to the parameterization process in this study. One advantage of the SRT proteins used in this study is that it contains 9 of the 20 naturally occurring amino acids, and that they do not contain charged residues, as modeling of charges in DPD simulations have additional difficulties that we did not need to address in this thesis.

This thesis is a first step towards building an efficient scheme for modeling SRT proteins in various solvents. As future work, modeling the proteins in a solvent that realistically represents water and DMSO will be the first goal. This will enable comparing the morphologies obtained in HFIP, DMSO and water. While the mechanical properties of proteins obtained in these different solvents differ, it has not yet been possible to delineate the origin of these differences[31]. To gain information at the atomistic detail, in particular the identification of the secondary structure of SRT proteins, the morphologies of the coarse-grained equilibrated structures may be reverse-mapped as exemplified in the literature for obtaining helical structures of PIPOX chains[6]. Our ultimate goal is to relate SRT morphologies to various material properties to develop SRT-based advanced technology materials.

BIBLIOGRAPHY

1. Römer, L. and T. Scheibel, *The elaborate structure of spider silk: structure and function of a natural high performance fiber.* Prion, 2008. **2**(4): p. 154-161.

2. Wise, S.G., S.M. Mithieux, and A.S. Weiss, *Engineered Tropoelastin and Elastin-Based Biomaterials*, in *Advances in Protein Chemistry and Structural Biology*, A. McPherson, Editor. 2009, Academic Press. p. 1-24.

3. Jung, H., et al., *Molecular tandem repeat strategy for elucidating mechanical properties of high-strength proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**: p. 6478-6483.

4. Pena-Francesch, A. and M.C. Demirel, *Squid-Inspired Tandem Repeat Proteins: Functional Fibers and Films.* Frontiers in Chemistry, 2019. **7**(69).

5. Avaz, S., et al., *Soft segment length controls morphology of poly(ethylene oxide) based segmented poly(urethane-urea) copolymers in a binary solvent.* Computational Materials Science, 2017. **138**: p. 58-69.

6. Furuncuoğlu Özaltın, T., et al., *Multiscale modeling of poly(2-isopropyl-2-oxazoline) chains in aqueous solution.* European Polymer Journal, 2017. **88**: p. 594-604.

7. Verlet, L., *Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules.* Physical Review, 1967. **159**(1): p. 98-103.

8. Hansson, T., C. Oostenbrink, and W. van Gunsteren, *Molecular dynamics simulations.* Current Opinion in Structural Biology, 2002. **12**(2): p. 190-196.

9. BIOVIA, D.S., *MATERIALS STUDIO 2018*. 2006, San Diego: Dassault Systèmes.

10. Sun, H., *COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds.* The Journal of Physical Chemistry B, 1998. **102**: p. 7338-7364.

11. Sun, H., et al., *COMPASS II: extended coverage for polymer and drug-like molecule databases.* Journal of Molecular Modeling, 2016. **22**: p. 1-10.

12. Andersen, H.C., *Molecular dynamics simulations at constant pressure and/or temperature.* The Journal of Chemical Physics, 1980. **72**: p. 2384-2393.

13. ACD/Chemsketch, v. 2019: Advanced Chemistry Development, Inc., Toronto, On, Canada, www.acdlabs.com.

14. Groot, R.D. and P.B. Warren, *Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation.* Journal of Chemical Physics, 1997. **107**: p. 4423-4435.

15. Flory, P.J., *Principles of polymer chemistry*. 1953: Cornell University Press.

16. Rubinstein, M. and R.H. Colby, *Polymer Physics*. 2003, Oxford: Oxford University Press. 137-143.

17. Gennes, P.-G., *Scaling concepts in polymer physics*. 1979, Ithaca, N.Y: Cornell University Press.

18. Gu, J., et al., *A Generic Force Field for Protein Coarse-Grained Molecular Dynamics Simulation.* International Journal of Molecular Sciences, 2012. **13**: p. 14451-14469.

19. Hildebrand, J.H., *Solubility of Non-Electrolytes*. 1936, New York: Reinhold.

20. Maiti, A. and S. Mcgrother, *Perspective: Dissipative particle dynamics.* The Journal of Chemical Physics, 2004. **120**: p. 150901.

21. Levine, B.G., J.E. Stone, and A. Kohlmeyer, *Fast analysis of molecular dynamics trajectories with graphics processing units—Radial distribution function histogramming.* Journal of Computational Physics, 2011. **230**(9): p. 3556-3569.

22. Atkins, P. and J. De Paula, *Physical chemistry*. Ninth Edition ed. 2006, Oxford: Oxford University Press.

23. Mertens, H.D. and D.I. Svergun, *Structural characterization of proteins and complexes using small-angle X-ray solution scattering.* J Struct Biol, 2010. **172**(1): p. 128-41.

24. Göbel, U., et al., *Correlated mutations and residue contacts in proteins.* Proteins: Structure, Function, and Bioinformatics, 1994. **18**(4): p. 309-317.

25. Vassura, M., et al., *Reconstruction of 3D Structures From Protein Contact Maps.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2008. **5**(3): p. 357-367.

26. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2017.

27. B.J., G., et al., *Bio3D: An R package for the comparative analysis of protein structures.* Bioinformatics, 2006. **22**: p. 2695--2696.

28. MATLAB, *9.8.0.1380330 (R2020a) Update 2*. 2020: The MathWorks Inc.

29. Sariola, V., et al., *Segmented molecular design of self-healing proteinaceous materials.* Scientific Reports, 2015. **5**(1): p. 13482.

30. Pena-Francesch, A., et al., *Research Update: Programmable tandem repeat proteins inspired by squid ring teeth.* APL Materials, 2018. **6**(1): p. 010701.

31. Pena-Francesch, A., et al., *Mechanical Properties of Tandem-Repeat Proteins Are Governed by Network Defects.* ACS Biomaterials Science and Engineering, 2018. **4**: p. 884-891.

APPENDIX A

Cluster_Finder.m

```
n = length(m);

data1 = [];
data2 = [];

countloopP_data = [];
countloopPbreadata = [];
count = 0;
countLoop = 0;
countLoopDummy = 0;
countLoopBreak = 0;

for k = 1:n
   for i = 1:n

      if i > n || i + k + 3 > n


         break
      end


      if m(i, i+1+k) == 1 && m(i+1, i+k+2) == 1 && i ~= i + 1 + k


         if countLoop == 0
            countLoop = 2;
            if m(i+2, i+3+k) == 1
               data1 = [data1; i];
               data2 = [data2; i + 1 + k];

            end

         else


            data1 = [data1; i];
            data2 = [data2; i + 1 + k];
            countLoop = countLoop + 1;

         end

      else
```

```
        if (countLoop ~= 0)
            countLoopDummy = countLoop;
            data1 = [data1; i];
            data2 = [data2; i + 1 + k];
            data1 = [data1; 0];
            data2 = [data2; 0];
            countloopP_data = [countloopP_data; countLoopDummy];
        end
        countLoop = 0;

    end


  end

end


for i = 1:length(data1) - 1
  if data1(i+1) - data1(i) ~= 1 && data1(i+1) ~= 0 && data1(i) ~= 0
    data1(i+1) = data1(i) + 1;
    data2(i+1) = data2(i) + 1;
  end


end


for i = 1:length(data1) - 2

  if data1(i) == 0 && data1(i+2) == 0
    data1(i+1) = 0;
    data2(i+1) = 0;

  end

end

i = 1;
while true

  if data1(i) == 0 && data1(i+1) == 0
    data1(i) = [];
    data2(i) = [];
    i = i - 2;

  end
  i = i + 1;
end
```

Cluster_Counter.m

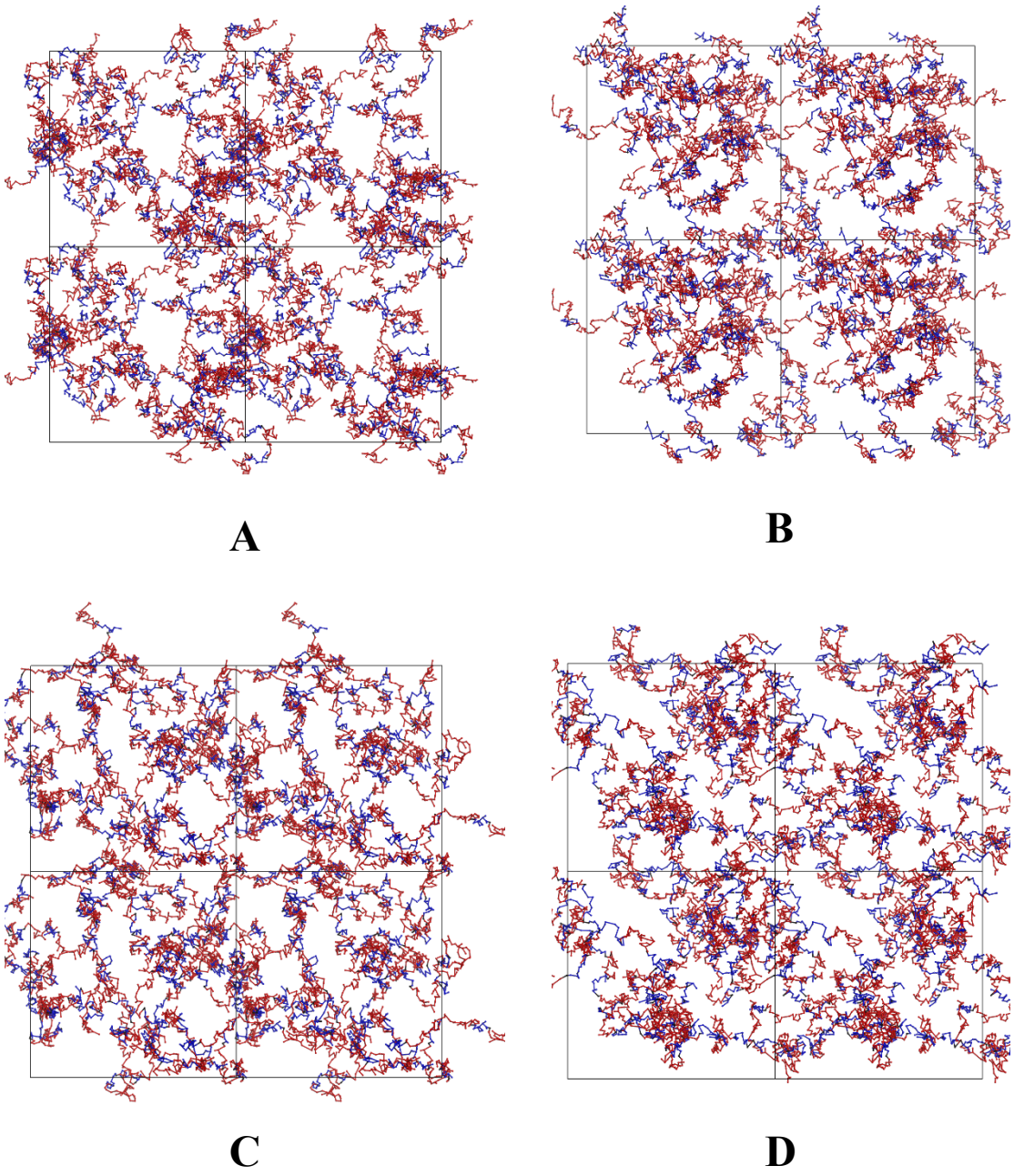
```
i = 1;
dataB1 = [];
dataB1i = [];
dataB1j = [];
matDati = {};
matDatj = {};
dataB2i = [];
dataB2j = [];
countBA = [];


while true
   i = 1;
   while data1(i) ~= 0
      a = data1(i);
      b = data2(i);

      dataB1i = [dataB1i; a];
      dataB1j = [dataB1j; b];


      i = i + 1;

   end

   temp = (false);
   matDati = {dataB1i};
   matDatj = {dataB1j};
   data1(1:i) = [];
   data2(1:i) = [];
   i = 1;
   while temp == 0

      while data1(i) ~= 0
         a = data1(i);
         b = data2(i);

         dataB2i = [dataB2i; a];
         dataB2j = [dataB2j; b];


         i = i + 1;

      end


      if length(dataB1i) + length(dataB2i) ~= length(union(dataB1i, dataB2i))
         i = i + 1;
```

```
    if length(dataB1j) + length(dataB2j) == length(union(dataB1j, dataB2j))

        matDati{end+1} = dataB2j;
        g = 1;
        while data1(g) ~= 0
            g = g + 1;
        end
        data1(1:g) = [];
        data2(1:g) = [];
        dataB1i = dataB2i;
        dataB1j = dataB2j;
        dataB2i = [];
        dataB2j = [];
        i = 1;


    end


elseif length(dataB1i) + length(dataB2j) ~= length(union(dataB1i, dataB2j))
    i = i + 1;

    if length(dataB1j) + length(dataB2i) == length(union(dataB1j, dataB2i))


        matDati{end+1} = dataB2i;
        g = 1;
        while data1(g) ~= 0
            g = g + 1;
        end
        data1(1:g) = [];
        data2(1:g) = [];
        dataB1i = dataB2i;
        dataB1j = dataB2j;
        dataB2i = [];
        dataB2j = [];
        i = 1;

    end


elseif length(dataB1j) + length(dataB2i) ~= length(union(dataB1j, dataB2i))
    i = i + 1;

    if length(dataB1i) + length(dataB2j) == length(union(dataB1i, dataB2j))


        matDatj{end+1} = dataB2j;
        g = 1;
```

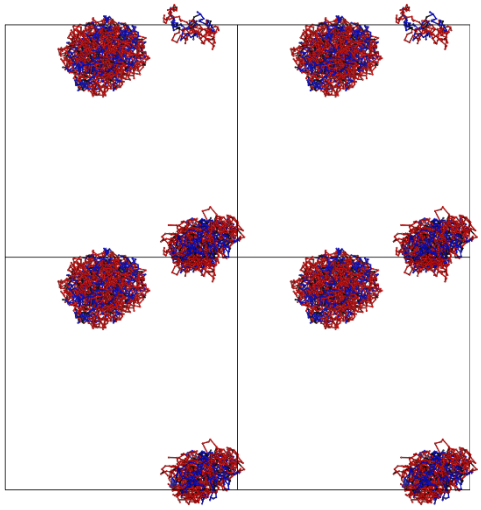
```
            while data1(g) ~= 0
                g = g + 1;
            end
            data1(1:g) = [];
            data2(1:g) = [];
            dataB1i = dataB2i;
            dataB1j = dataB2j;
            dataB2i = [];
            dataB2j = [];
            i = 1;

        end


    elseif length(dataB1j) + length(dataB2j) ~= length(union(dataB1j, dataB2j))
        i = i + 1;

        if length(dataB1i) + length(dataB2i) == length(union(dataB1i, dataB2i))


            matDatj{end+1} = dataB2i;
            g = 1;
            while data1(g) ~= 0
                g = g + 1;
            end
            data1(1:g) = [];
            data2(1:g) = [];
            dataB1i = dataB2i;
            dataB1j = dataB2j;
            dataB2i = [];
            dataB2j = [];
            i = 1;

        end

    else

        i = i + 1;
        dataB2i = [];
        dataB2j = [];

    end


    if isempty(data1) == 1 || i > length(data1)
        temp = (true);
    end

end
[a, b] = size(matDati);
```

```
    [x, y] = size(matDatj);
    countBundle = b + y;
    countBA = [countBA; countBundle];
    matDati = [];
    matDatj = [];
    dataB1i = [];
    dataB1j = [];
    dataB2i = [];
    dataB2j = [];


    if isempty(data1) == 1
       break
    end


end
```

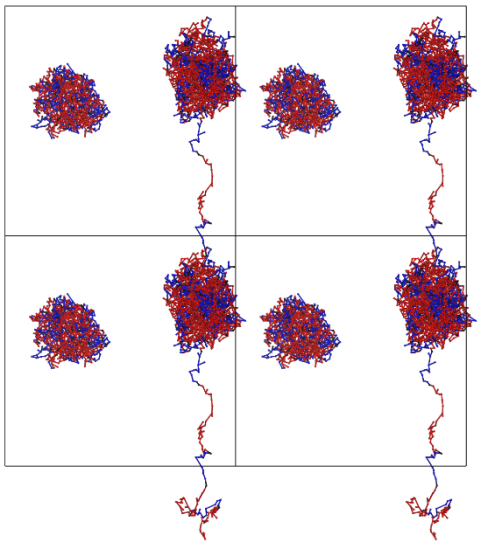
APPENDIX B



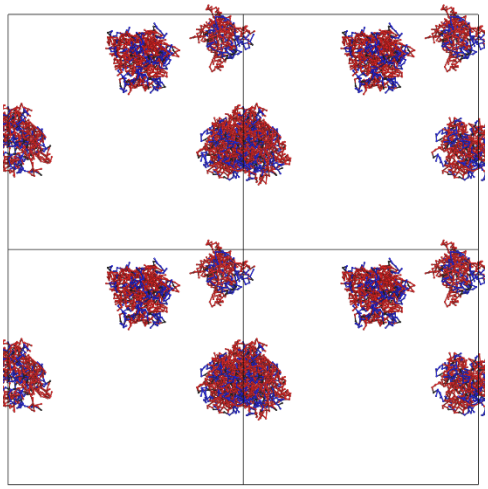
Appendix B-1 Last Snapshots of the HFIP-SRT n4 (A), n7 (B), n11 (C), and n25 (D) DPD simulations



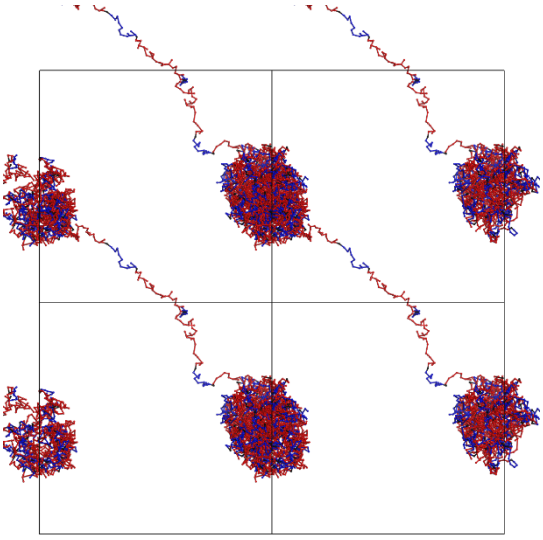
A



B



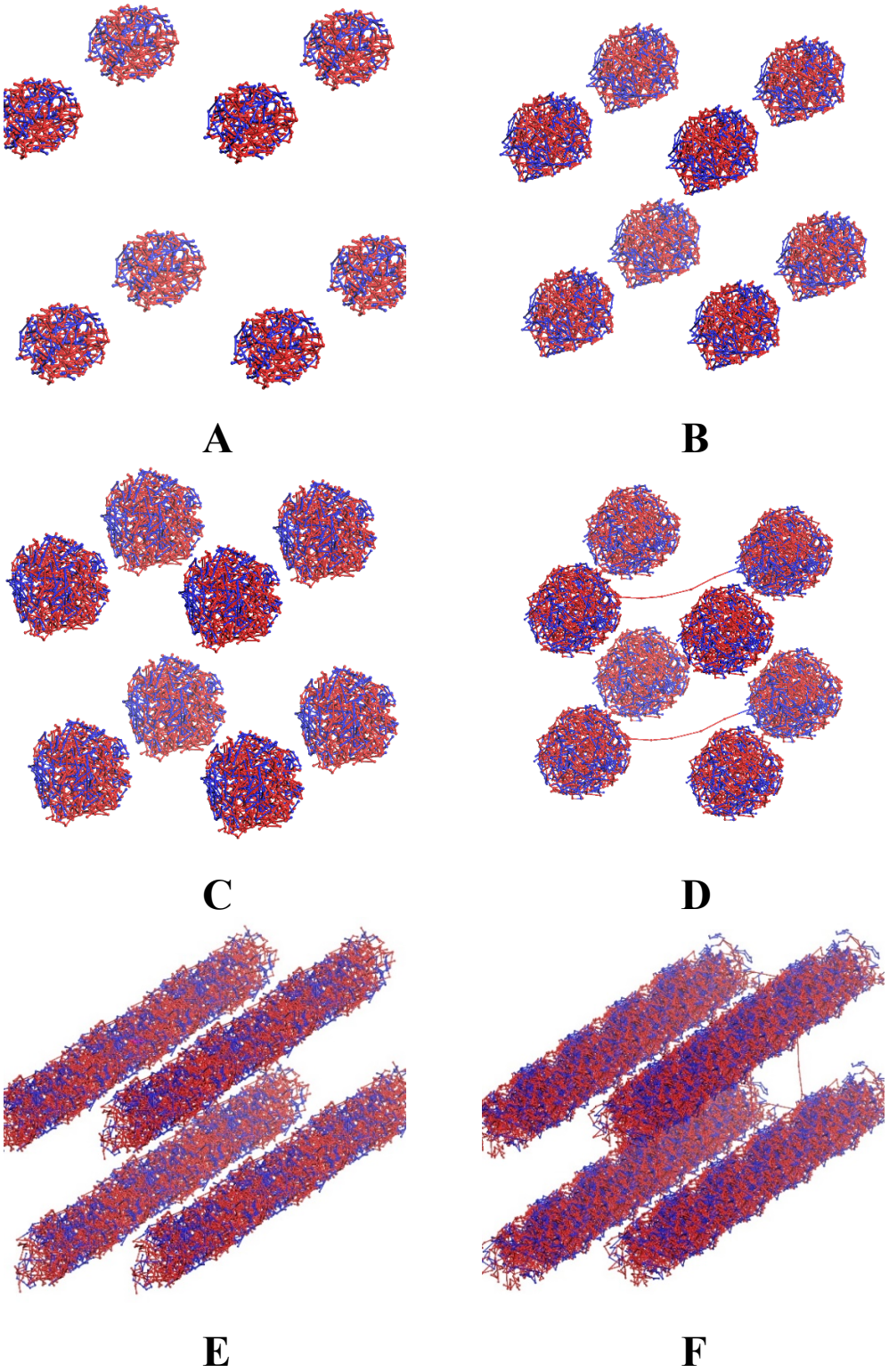
C



D

Appendix B-2 Last Snapshots of the Dilute Solvent-SRT n4 (A), n7 (B), n11 (C), and n25 (D) DPD simulations

APPENDIX C



Appendix C Last snapshots of the concentration sweep of SRT-n4 protein in 5wt.% (A), 10wt.% (B), 15wt.% (C), 20wt.% (D), 25wt.% (E), 30wt.% (F) solvent P