TESTING THE ABILITY OF STANDARD MOLECULAR DYNAMIC SOFTWARE FORCE FIELDS TO ACCURATELY MODEL THE STRUCTURAL FEATURES OF INTRINSICALLY DISORDERED PROTEINS

Honors Thesis

Presented to the Honors Committee of
Texas State University
in Partial Fulfilment
of the Requirements

for Graduation in the Honors College

by

Micheal Jace Tarver

San Marcos, Texas
May 2015

TESTING THE ABILITY OF STANDARD MOLECULAR DYNAMIC SOFTWARE FORCE FIELDS TO ACCURATELY MODEL THE STRUCTURAL FEATURES OF INTRINSICALLY DISORDERED PROTEINS

Thesis Supervisor:

____

Steven T. Whitten, Ph.D.
Department of Chemistry and Biochemistry

Approved:

____

Heather C. Galloway, Ph.D.
Dean, Honors College

COPYRIGHT

FAIR USE AND AUTHOR’S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGMENTS

I would like to thank Dr. Steven Whitten for his continued support and guidance. I would also like to thank Dr. Albert Kwansa for lending his expertise with the MD simulations.

TABLE OF CONTENTS

Page

CHAPTER

List of Figures

Figure **Page**

ABSTRACT

TESTING THE ABILITY OF STANDARD MOLECULAR DYNAMIC SOFTWARE FORCE FIELDS TO ACCURATELY MODEL THE STRUCTURAL FEATURES OF INTRINSICALLY DISORDERED PROTEINS

A main tenet regarding the study of proteins states the structure of a protein defines its function. That is, the structure of the protein imbues characteristics that allow it to perform a specific function in the body. Intrinsically disordered proteins (IDPs) and intrinsically disordered regions of proteins (IDRPs) are common in all biological systems. It is estimated that up to forty percent of the human proteome is composed of IDPs or IDRPs. These proteins and regions of proteins, respectively, are responsible for gene regulation, cellular control, and molecular signaling pathways. Many human diseases are caused by mutations in IDPs and IDRPs, thus one could rationalize the mutated proteins fail to adopt appropriate structures to perform their intended function. It is estimated that up to fifty percent of human cancers are caused by mutations in the human tumor suppressor protein p53 (p53). p53 has an N-terminal IDRP mapping to residue positions 1 through 93. This region of p53 is crucial for transcription and apoptotic pathways. However, traditional techniques used to elucidate high resolution structural features (HRSF) of structured proteins, such as NMR and x-ray crystallography, are not amenable to the study of IDPs and IDRPs structural features, thus they remain unresolved. Previously, we used the intrinsically disordered N-terminal region of the human tumor

suppressor protein p53 (1-93) as a model system to study temperature and sequence effects on the structural features of IDPs using a combination of computational and laboratory experimental techniques. Using AMBER, a molecular dynamics suite software package, we now perform atomistic molecular dynamic simulations of the IDP p53(1-93) with the ff12SB force field and the ff99SB force field in explicit solvent in an attempt to recapitulate the experimental results from the temperature and sequence effects studies. Molecular dynamic (MD) simulations have the potential to help resolve the structural features of IDPs and IDRPs with atomic-scale resolution. Unfortunately, MD simulations are only as good as the force fields used. That is, if the force field parameters unrealistically express physical relationships between atoms in the simulation the results can end up being quite artificial and thus useless in elucidating molecular descriptions of protein interactions and consequently, protein structure. Here we analyze the structures produced from completed MD simulations and compare them to our experimental measurements. This analysis shows that the ff12SB protein force field did not accurately model the structural features of p53 (1-93). The ff99SB force field simulations have not yet completed.

CHAPTER ONE

INTRODUCTION

1.1 *Proteins Are Linear Polymers Composed of Amino Acids*

A protein is a biological polymer composed of monomeric units called amino acids. In humans there are twenty standard “alpha” amino acids. Excluding proline, the standard amino acids have the ubiquitous substituent arrangement of a carboxylic acid, primary amino group, and R group attached to the alpha carbon, hence the name alpha amino acids. The R group substituent, also known as the side chain, distinguishes amino acids from each other by conferring distinct physiochemical properties. Although the twenty standard amino acids vary, they can be assigned to three general categories based on side chain polarity at a physiological pH: non-polar, polar, and charged polar.

Non-polar amino acids have a wide range of side chain moieties ranging from the single hydrogen atom in glycine to the large aromatic and ring structures in tryptophan; all of which have varying degrees of hydrophobicity. The polar amino acids have thiol, hydroxyl, and amide side chains. Cysteine, a polar amino acid, is capable of forming a disulfide bond, a type of covalent bond, via its thiol side chain. The charged polar amino acids can be either positively or negatively charged. Glutamic acid and aspartic acid are negatively charged while arginine, lysine, and histidine are positively charged. Histidine is unique in that the pKa of its sidechain is near the physiological pH, thus histidine will have a positive charge or be neutral depending upon the microenvironment.

Amino acids polymerize through a condensation reaction that creates a peptide bond between the carboxylic acid moiety of the first amino acid and the free amine moiety of the second amino acid. This reaction always results in a linear chain of amino acids (1). A protein sequence is described by listing the amino acids in the order they’re connected, starting with the amino acid with the free amine, called the N-terminus, and ending with the C-terminus, the amino acid with the free carboxylic acid moiety.

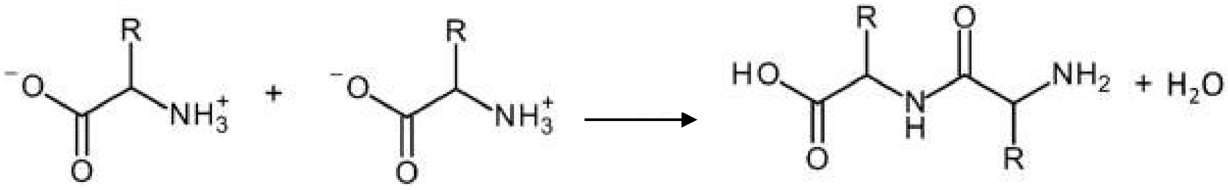


FIG. 1. **Protein Polymerization Reaction.** This reaction displays two amino acids forming a peptide bond. Proteins grow in a linear fashion through this reaction. The OH end is the C-terminus, while the NH_2 end is the N-terminus (1).

1.2 *Biological Significance and Classification of Protein Structure*

A main tenet regarding study of proteins is that the structure of the protein defines its function. That is, the structure of the protein imbues characteristics that allow it to perform a specific function in the body. Proteins can exhibit up to four levels of increasing structural complexity.

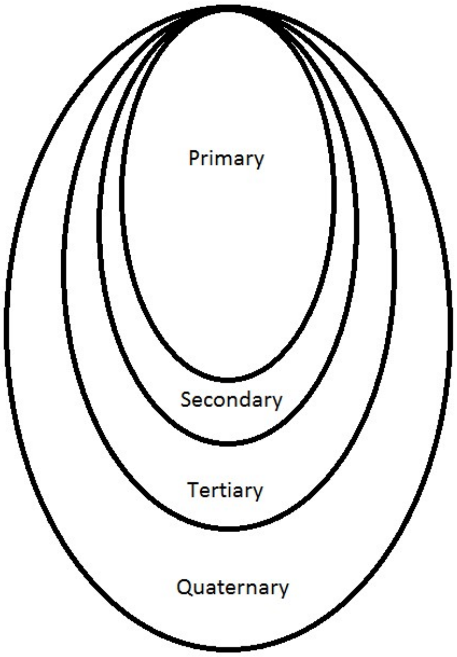


FIG. 2. **Classification of Protein Structure.** The diagram depicts how protein structure is classified in a manner that all levels of structural complexity incorporate the previous levels within them.

Primary Structure

The primary amino acid structure is the simplest of protein structures, consisting solely of the amino acid sequence starting at the N-terminus and ending at the C-terminus. The primary amino acid sequence only describes the order of amino acids that compose a protein.

Secondary Structure

The secondary structure of proteins is a property of hydrogen bonding patterns between amino acid “backbone” atoms. Proximity of hydrogen donors and acceptors is necessary in hydrogen bond formation and is facilitated by protein “backbone” flexibility (2). The protein “backbone” consists of the alpha-carbon, amine nitrogen, and the carboxylic acid’s carbon and oxygen atoms. The partial double bond character of the peptide bonds connecting adjoining amino acids does not allow free rotation about the bond. However, there are bonds that allow rotation of the “backbone” and thus increase flexibility.

Two common angles of rotation are referred to as Phi and Psi angles. Phi (φ) corresponds to the torsion angle, also known as a dihedral angle, between the nitrogen atom and the alpha-carbon, while Psi (ψ) refers to the dihedral angle between the alpha carbon and carboxylic acid’s carbon atom. There are only certain (φ) and (ψ) values allowed due to steric effects of the van der Waals radius of atoms (3). Van der Waals radii can be thought of as hard spheres that represent the volume or space that individual atoms occupy. Common folding patterns arise from repeating (φ) and (ψ) values (2).

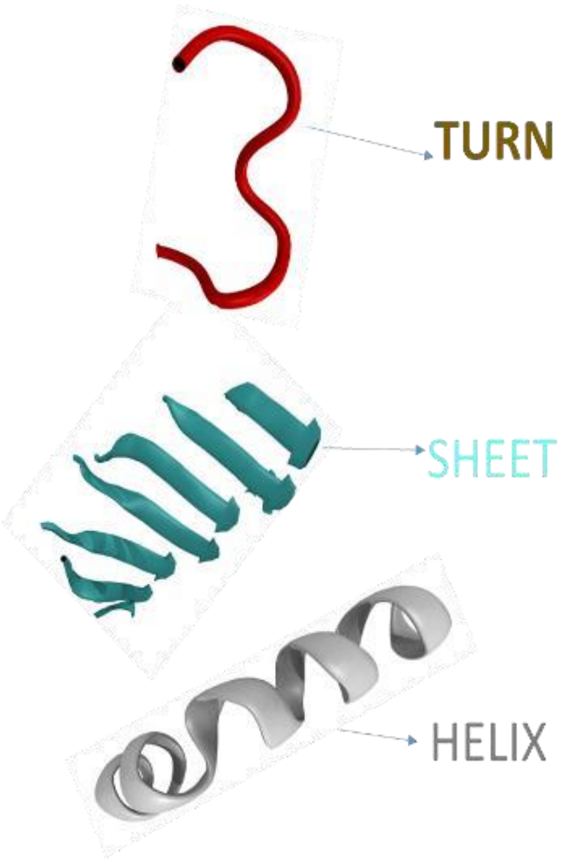


FIG. 3. **Common Secondary Structure Conformations.** Made with VMD 1.9.1, cartoon representations of common secondary structures arising from repeating (φ) and (ψ) values and their respective names (4).

Tertiary Structure

The tertiary structure is described by the three-dimensional arrangement of the protein. Tertiary structures consist of a single protein backbone and may have one or more secondary structural motifs. Large proteins may have several areas of secondary structures called domains. Each domain may possess unique physiochemical properties that allow a protein, specifically the domain of the protein, to perform a biological function.

Domains are often stable and their conformations conserved when studied independent of the protein sequence. For example, if a protein is normally 250 residues in length and a domain is present spanning residues 1-50, the domain consisting of residues 1-50 could be studied independently from the rest of the protein. The tertiary structures vary based upon the pH, ion concentration, and temperature of the microenvironment; therefore, the native structure of a protein is defined as the most common conformation in a physiological environment (5).

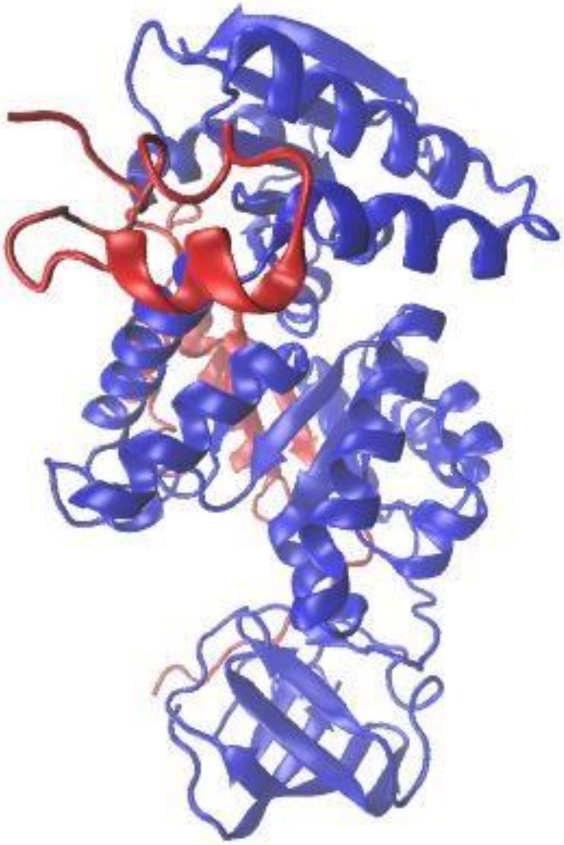


FIG. 4. **Tertiary Structure and Domains.** Made with VMD 1.9.1, this cartoon representation is of rabbit muscle pyruvate kinase complexed with $Mn2^+$, K^+, and pyruvate made from PDB structure 1PKN. There red piece highlights a helix and the bottom purple arrows are β-sheets .The 3-D arrangement is the tertiary structure of this protein (4,6).

Quaternary Structure

The highest level of structural complexity in proteins is quaternary structure. Quaternary structure consists of two or more protein chains bound together. Each protein chain comprising the quaternary structure is a subunit. In some proteins, changes in the tertiary structure of a single subunit can modulate the conformations of other subunits. Similar to the tertiary domains of proteins, subunits may have unique physiochemical properties that allow it to perform functions unique from other subunits in the protein complex.

Together the secondary, tertiary, and quaternary structures of proteins, all of which are resultant of unique primary structures, imbue characteristics essential for biological function. In light of this, there is a large class of proteins lacking in stable tertiary structures that possess uncorrelated residue dynamics and exhibit dynamic and flexible backbones (7,8). These dynamic proteins are said to have intrinsic disorder.

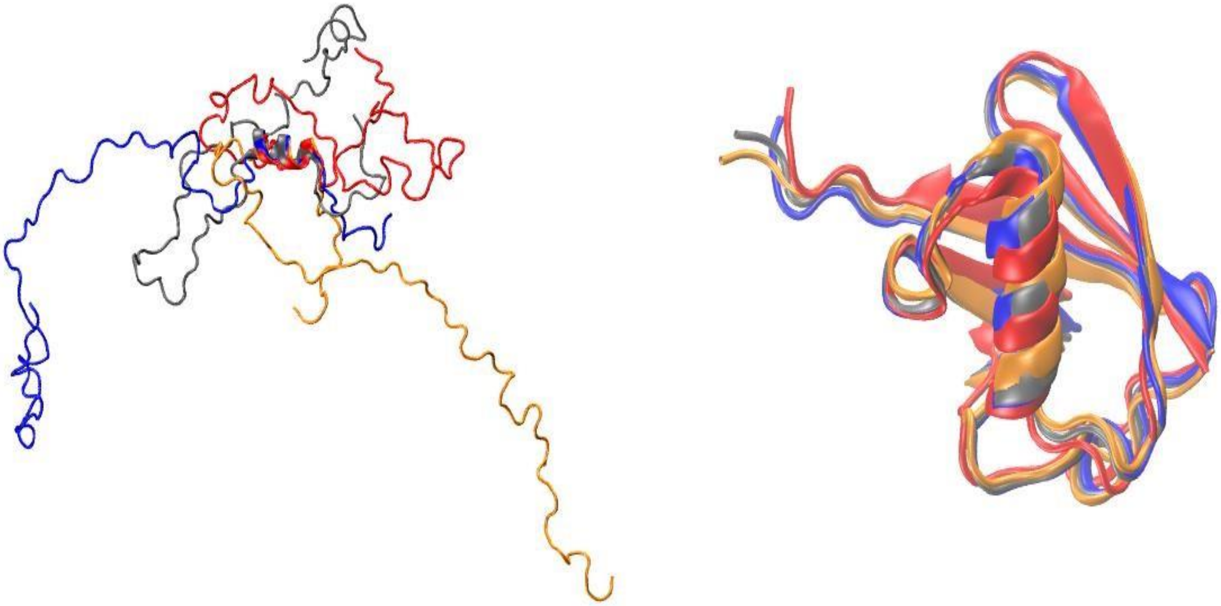


FIG. 5. **Disordered and Ordered Protein Dynamics.** Made with VMD 1.9.1, each color represents a snapshot of the protein at a different point in time. The left image represents a disordered protein created with PDB structure 2FFT (9). The right image is a well-folded protein using PDB structure 1D3Z (10). Although the well-folded protein exhibits a small degree of flexibility and movement, its structure is constrained to a specific conformation (4).

1.3 *Intrinsically Disordered Proteins and Their Biological Significance*

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions of proteins (IDRPs) are responsible for gene regulation, cellular control, and signaling pathways of biological systems (11,12). It's estimated that up to forty percent of the proteome is composed of IDPs or IDRPs (7). Furthermore, greater than 70 percent of transcription factors are IDPs or contain extended regions of disorder (12). IDPs and IDRPs generally have a high net charge and adopt more extended structures, relative to a statistical coil, than structured proteins of the same size, which adopt compact structures relative to a statistical coil (7,13,14). A random coil size is calculated from a statistical distribution of protein conformations where residues are randomly oriented. As a protein travels through solution it tumbles and because the protein has an average length it creates a spherical volume defined by this length. A tumbling extended structure would circumscribe a larger volume than a relatively more compact structure. Using the Stokes-Einstein equation a measurement of extendedness or compaction, called the hydrodynamic radius (Rh), can be calculated experimentally with techniques such as dynamic light scattering. Rh has become an important global feature when structurally characterizing IDPs and IDRPs due to its ease of measurement and difficulty in measurement of local structural features (7).

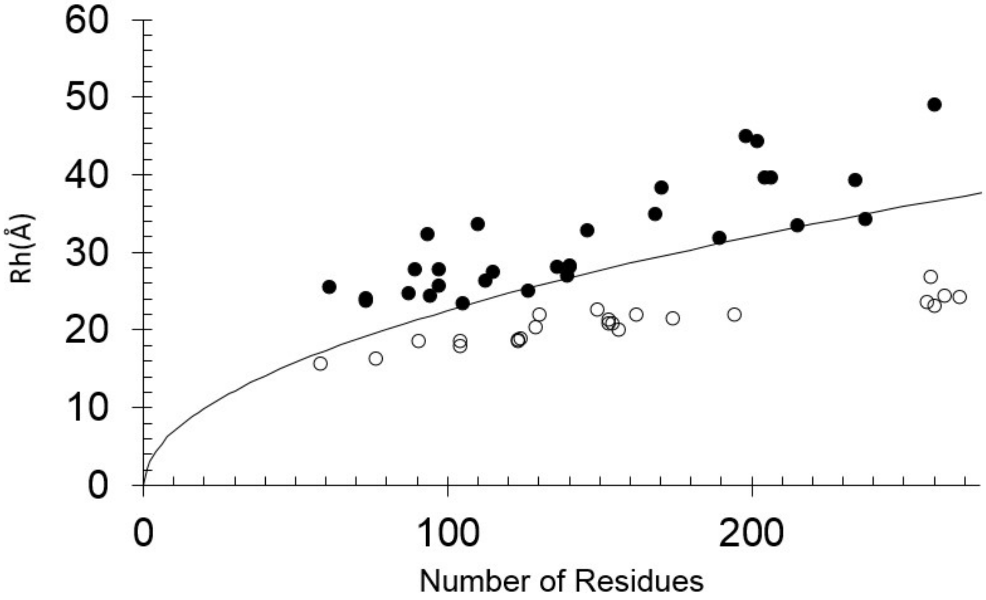


FIG. 6. **Extended Structures of IDPs and IDRPs.** Rh vs number of residues present in the protein demonstrates that IDPs, closed circles, adopt more extended structures than folded proteins, open circles, relative to a random coil, the solid line. Rh values and residue numbers were taken from published data (13,15-35).

While structural characterization techniques, such as NMR, currently describe the tertiary structures of IDPs and IDRPs as transient and unstable, many human diseases are caused by mutations in IDPs and IDRPs, thus one could rationalize the mutated proteins fail to adopt the appropriate structures to perform their intended functions.

Human Tumor Suppressor Protein p53

Up to fifty percent of human cancers are caused by mutations in the Human Tumor Suppressor Protein p53 (p53) (36). Sometimes referred to as the guardian of the genome, p53 plays a vital role in tumor suppression (37). The N-terminal region of p53, p53 (1-93), is responsible for activation of both transcription and apoptotic pathways.

Residues 1-42 compose the major acidic transcription-activation domain (TAD) known as AD1. Complimentary to AD1 is the minor TAD, AD2, between residues 43 and 63. Furthermore, the proline rich domain between residues 64-92 is essential for apoptotic activity (38).

Difficulty Studying IDPs and IDRPs

Rapid fluctuations of IDPs and IDRPs make it difficult to elucidate high-resolution structural features (HRSF) through traditional techniques such as NMR and X-ray crystallography.

1.4 *Molecular Dynamic Simulations*

Molecular dynamics (MD) simulations have the potential to help resolve IDPs and IDRPs with atomic-scale resolution. Unfortunately, MD simulations are only as good as the force fields used. That is, if the force field parameters unrealistically express physical relationships between atoms in a simulation, the results can end up being quite artificial and thus useless in elucidating molecular descriptions of protein interactions and consequently protein structure.

Atomistic MD Simulations with AMBER Force Fields

AMBER (Assisted Model Building with Energy Refinement) is a software package that can be used to run atomistic MD simulations (39). Atomistic MD is one of the various computational techniques used to study proteins. In atomistic MD, each atom of the protein is represented. Forces that the atoms within the simulation experience and exert upon one another are dictated by a mathematical expression consisting of several terms and representing molecular potential energy (*$V(r^N)$*); this mathematical expression and its constituent parameters are collectively referred to as a force field (40). As progress is made in understanding these interactions through experiment and quantum mechanical calculations, the force field parameters are revised to represent atomic interactions in the real world more accurately.

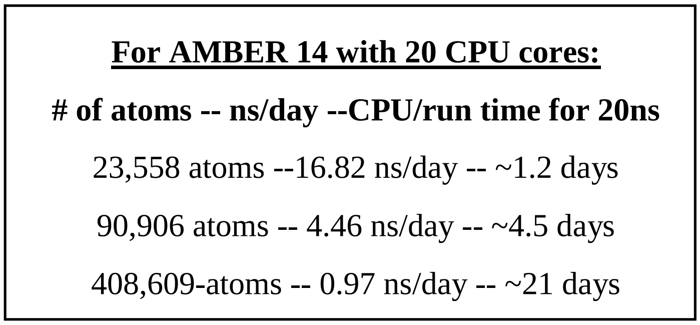
$$V(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2$$

$$+ \sum_{\text{torsions}} \sum_{n} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

FIG. 7. **Potential Energy Equations.** These four terms calculate the potential energy of each atom in a MD simulation. Labels below terms 1-3 describe the type of potential energy calculated. The fourth term consist of two long range potential energy calculations, the van der Waals and electrostatic interactions (40).

How Energies are Recorded and Simulation Progression

During MD simulations, the protein samples many conformations before it eventually finds an energetically favorable structure. The MD simulation progresses in increments called timesteps, which are small units of time in the simulation; 2 femtoseconds (fs) is commonly used for atomistic MD. The amount of actual time it takes for 2 fs of time to pass in the simulation depends largely on the number of atoms present in the simulation. As the number of atoms increases, the amount of time it takes to reach a certain simulation time increases (39).



For AMBER 14 with 20 CPU cores:

# of atoms -- ns/day --CPU/run time for 20ns

23,558 atoms --16.82 ns/day -- ~1.2 days

90,906 atoms -- 4.46 ns/day -- ~4.5 days

408,609-atoms -- 0.97 ns/day -- ~21 days

FIG. 8. **Simulation Progress Dependency on Atom Number.** This diagram shows the simulation time for a 20ns simulation with 20 computer processors for AMBER 14 (39).

This is a result of the MD simulations having to solve for Newtonian equations of motion while using the above $V(r^N)$ equation to determine the energies of each atomic interaction. Every X number of steps, as defined by the input file, the computer records the energies, coordinates, and velocities of all of the atoms, to files for subsequent processing and analysis. Each recording is called a frame.

Implicit vs Explicit Solvent Simulations

Atomistic MD simulations can be generally broken into two categories, implicit and explicit solvent simulations. Implicit solvent simulations approximate solvent-ion-solute interactions and exclude actual solvent atoms and ions. Implicit solvent simulations are thus less computationally expensive. However, due to the use of approximations, and in some cases the exclusion of solvent-solute interactions, they are generally less accurate than explicit solvent simulations (41,42).

Explicit solvent simulations incorporate solvent atoms and ions, such as Na^+ and Cl^-, into the simulation. This drastically increases the computation time. For example, you could only have around one or two thousand atoms in an implicit solvent simulation depending upon the size of the protein, but once a 10-angstrom water box and ions are added, the number of atoms present in the simulation can easily exceed 100,000.

Preparing the System and Running Production MD Simulation

There are three basic steps for setting up and running MD simulations: Building the system, setting conditions, and starting the simulation.

Building the System

Some aspects of building the system are the same for both implicit and explicit solvent simulation. Initially a starting structure is created and saved in the Protein Data Bank (PDB) file format. A program within the AMBER software package called xleap can be used to build the system. First, force field parameters are loaded followed by the PDB structure. Addition of ions and solvent is an extra step that is required for building the system of an explicit solvent simulation; this step is also carried out with xleap.

Once the appropriate force fields have been set a parameter-topology file (.prmtop) and a coordinate file (.inpcrd) are created. The .prmtop file expresses the connectivity of the atoms in the simulation and various parameters (e.g., those in the $V(r^N)$ equation shown above), while the .inpcrd file expresses the locations of the atoms in three-dimensional space.

Setting Conditions

Multiple input files are needed to set the conditions for consecutive stages of MD simulations; typically, four stages are used: Minimization, heating, equilibration, and production. There are various settings such as the timestep, total number of steps, cutoff distance (used to exclude the direct calculation of long-range interactions), and settings related to pressure and temperature control that can be set for each of these MD stages. The minimization stage optimizes the bond lengths and angles of all the atoms in the simulation (i.e., to locally minimize the value of the $V(r^N)$ equation mentioned earlier). Once the system energy has been minimized, the heating stage can begin, during which the system is heated to an appropriate temperature. Then the equilibration stage is used to insure that the temperature has stabilized and to introduce and stabilize the pressure. After the equilibration stage, the production MD stage can begin. The minimization, heating, and equilibration stages serve to prepare the system, while the production MD stage provides the data to be subsequently analyzed.

Starting the Simulation

MD software can be configured to spread the atoms of a simulation over multiple computer processors and across multiple interconnected computers in a process known as parallelization. AMBER has two “engines” that are used to run MD simulations in parallel, namely, sander.MPI and pmemd.MPI. Pmemd.MPI is a newer MD engine optimized for better parallel performance (i.e., can complete the same simulation in less time relative to sander.MPI); however, pmemd.MPI does not yet support all of the features built into sander.MPI. The three input files created in the above “setting conditions” section are input into one of the two “engines” along with the .prmtop and .inpcrd files from the “building the system” section to begin the simulation stages.

1.5 Inconsistencies with MD Simulations of IDPs and IDRPs and Previous Studies of p53 (1-93)

Previously, p53 (1-93) has been used as a model to study the sequence and temperature effects on IDPs and IDRPs (13,43). Sequence mutations have been used to study contributions of specific amino acids to the transient tertiary conformations. Specifically, 3 mutants were made, Pro-, Ala-, and Ala-_Pro-, by substituting the proline, alanine, and both proline and alanine residues, respectively, with glycine in the background of the wild-type (WT).

WT :

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIE
QWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPL

Ala- (12 amino acids) :

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIE
QWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPL

Pro- (22 amino acids):

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIE
QWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPL

Ala- Pro- (34 amino acids):

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIE
QWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPL

FIG. 9. **p53 (1-93) Variants Used in Previous Studies.** Labeled in blue are the sequences of the four variants used in our previous studies. In parentheses are the number of amino acids replaced with glycine, colored in red. For example, the 12 alanine residues in the Ala- variant have all been substituted with glycine. Glycine was used because it has similar charge and polarity of the amino acids it substitutes.

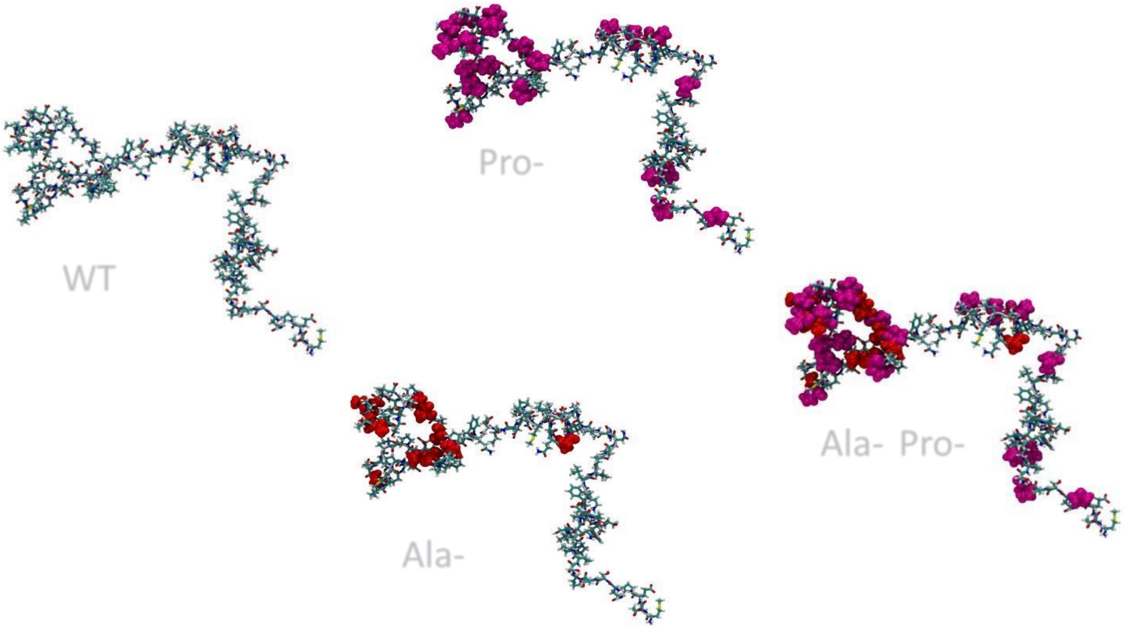


FIG. 10. **Atomistic Rendering of p53 (1-93) Variants.** Made with VMD 1.9.1, this atomistic representation of the four variants used in previous studies. Red dots are amino acids substituted with glycine (4).

Temperature effects were assessed with various techniques such as dynamic light scattering, SDS-PAGE, CD-spectroscopy, and SEC. It was shown that p53 (1-93) had a large hydrodynamic radius at low temperatures and an increasingly smaller hydrodynamic radius as the temperature increased (13,43).

Correlations between changes in molar residue ellipticity and magnitude of Rh with temperature fluctuations and sequence variation, along with a local maximum at 221nm via CD spectroscopy, made a case that changes in Rh were related to conformational changes in the backbone polyproline II propensities (ppII); a common secondary structure present in IDPs (13).

Fluorescent studies assessed whether compaction as a function of temperature and sequence variation was due to folding of the hydrophobic residues tryptophan, at residue positions 23, 53, and 91, into a hydrophobic core or a result of other phenomena (43). The fluorescent studies results were consistent with the tryptophan residues being fully exposed to solvent in all variants across the experimental temperature range (43). Using the Rh data and a quantitative model that correlates Rh as a function of ppII propensities we calculated ppII propensities for the proline and alanine.

1.6 *Goals*

Using AMBER, a MD suite software package, we now perform atomistic MD simulations of p53 (1-93) with the ff12SB force field and the ff99SB force field in explicit solvent in an attempt to recapitulate the previous experimental results from the temperature and sequence effects studies.

CHAPTER TWO

METHODS

2.1 *Model Protein and Variants*

We used the same four variants of p53 (1-93) used in the studies of temperature and sequence effects.

2.2 *Hardware and Software*

The AMBER suite software package was used to run the MD simulations using pmemd.MPI on the Texas State University STAR Cluster and North Carolina State University High Performance Computing Center. We created the protein structures with Discover Studio Viewer 4.0. Once all of the preliminary software and hardware was setup we began with implicit solvent simulations of the p53 (1-93) variants followed by explicit simulations.

2.3 *Simulation Force Fields and Parameters*

Implicit Solvent Simulation

Implicit solvent simulations were performed first to create a semi-folded protein structure. This decreased the overall computation time. For the implicit solvent simulations, we used the Amber force field ff12SB. The semi-folded protein structures obtained from the implicit solvent simulations were used as the starting structures for the explicit solvent simulations.

Explicit Solvent Simulations

Two sets of explicit solvent simulations were ran using two different force fields, ff99SB and ff12SB. The ff12SB force field is biased toward secondary structure formation while the ff99SB force field is not. Both sets of simulations used the AMBER gaff force field and frcmod.ionsJC_tip3p. For both simulations a water box that covered the entirety of the protein was created using TIP3CP water, then using the known volume of the water box, a 100mM concentration of NaCl was added to mimic experimental conditions. Counter-ions were also added to balance the negative 15 net charge of p53 (1-93).

$$\frac{([M] * 6.023 * \text{Å}^3)}{10^4}$$

FIG. 11. **Ion Calculation Equation.** Using this equation the number of ions needed in a simulation to replicated experimental conditions can be calculated in molarity ([M]). The volume of the water box, in angstroms, is represented in the equation as Å^3.

Temperature, Duration, and Total Number of Simulations

Both the ff99SB and ff12SB simulations used the four p53 (1-93) variants described above. Simulation were ran at two different temperatures, 300K and 350K, for a total of sixteen simulations. Each simulation was scheduled to run for a total of 100ns in hopes the simulations would reach an equilibrium.

Analysis

Analysis of RMSD was done using the AMBER analysis software CPPTRAJ and PTRAJ. RMSD was used to monitor the evolution of the protein structure and ascertain when it had reached equilibrium.

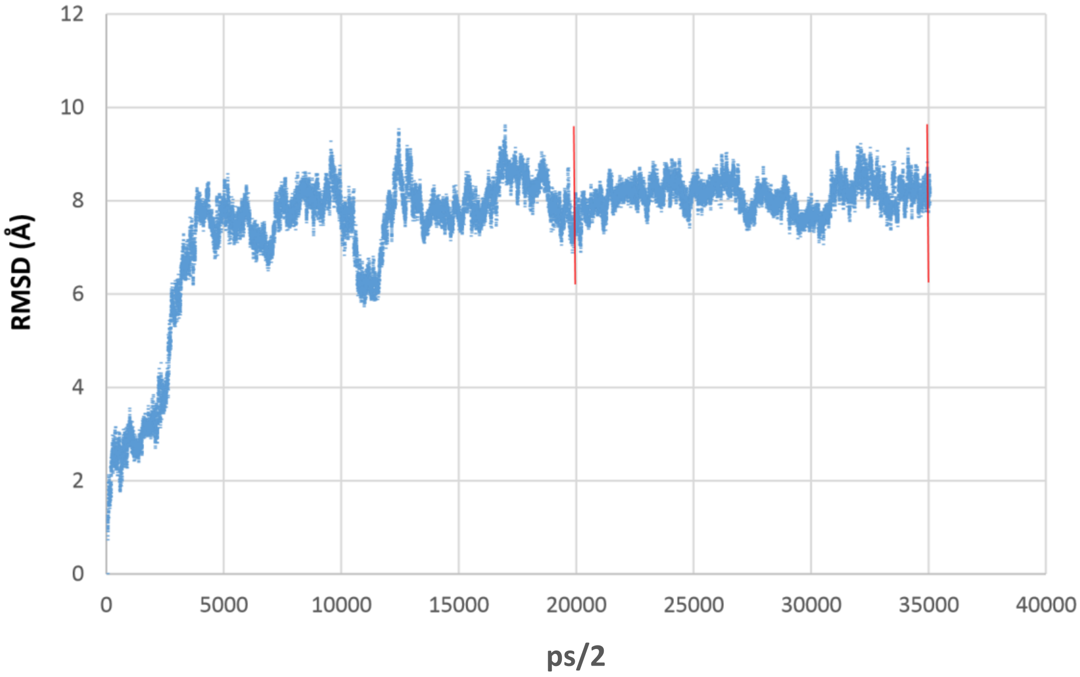


FIG. 12. **RMSD Plot Describing Equilibrium.** Root mean square deviation (RMSD) is used to ascertain when a simulation has reached an equilibrium. It is a measure of the displacement of a structure, in angstroms, from the starting structure. This graph is from the simulation of Pro- ff12SB at 350K. The RMSD of the simulation undulates around 8 in the frames between the red bars, thus it is considered to have converged to an energetically favorable structure and be in equilibrium.

The Rh for each structure was calculated by measuring the distance between the two farthest alpha-carbons of the structure. This would represent the diameter of the sphere circumscribed by the protein tumbling through solution. Then, this value was halved to calculate the radius of the sphere, also known as Rh.

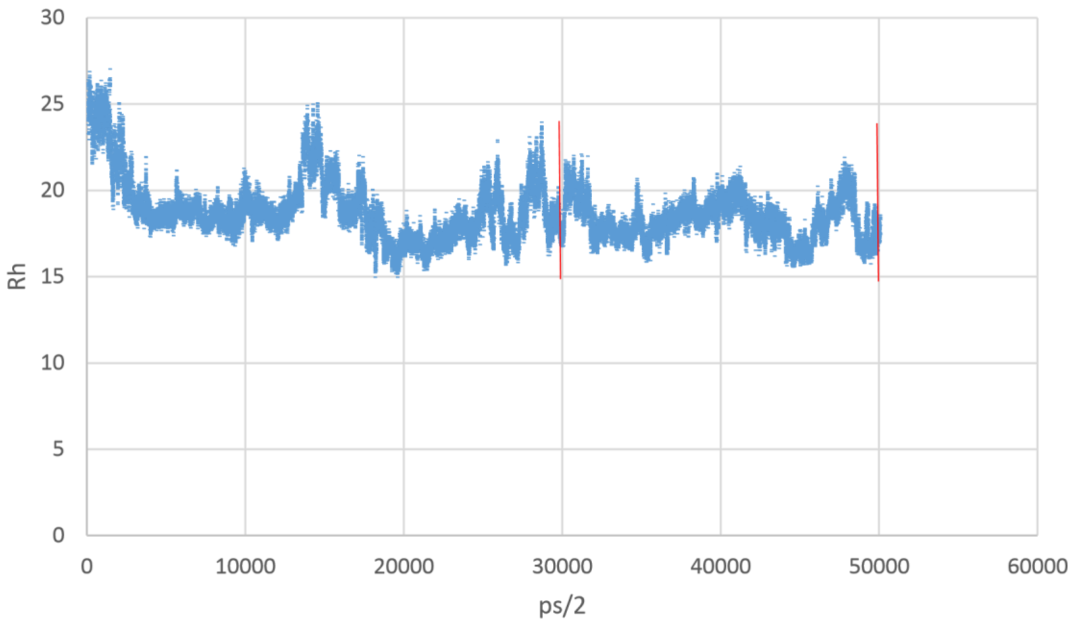


FIG. 13. **Rh vs Time.** As the simulation reaches an equilibrium the Rh of the protein stabilizes. This graph is from the simulation of Pro- ff12SB at 350K. The Rh of the simulation undulates around 17 in the frames between the red bars, thus it is considered to have converged to an energetically favorable structure and be in equilibrium.

We calculated ppII propensities for each structure by measuring the (φ) and (ψ) values for the protein. Then (φ) and (ψ) values within the canonical (φ) and (ψ) values range of -75 and 145 -/+ 10 degrees were designated as being ppII conformation. Once the number of alanine and proline residues in ppII conformations was known, the ppII propensities of alanine and proline could be calculated by dividing the number of residues in ppII conformations by the total number of respective residues present in the protein.

CHAPTER THREE

RESULTS

3.1 *Progress of Simulations*

Out of the sixteen total simulations started, the five that completed were analyzed. The completed simulations consisted of the four variants using the AMBER protein force field ff12SB at 300K and the Pro- variant at 350K using the ff12SB AMBER protein force field.

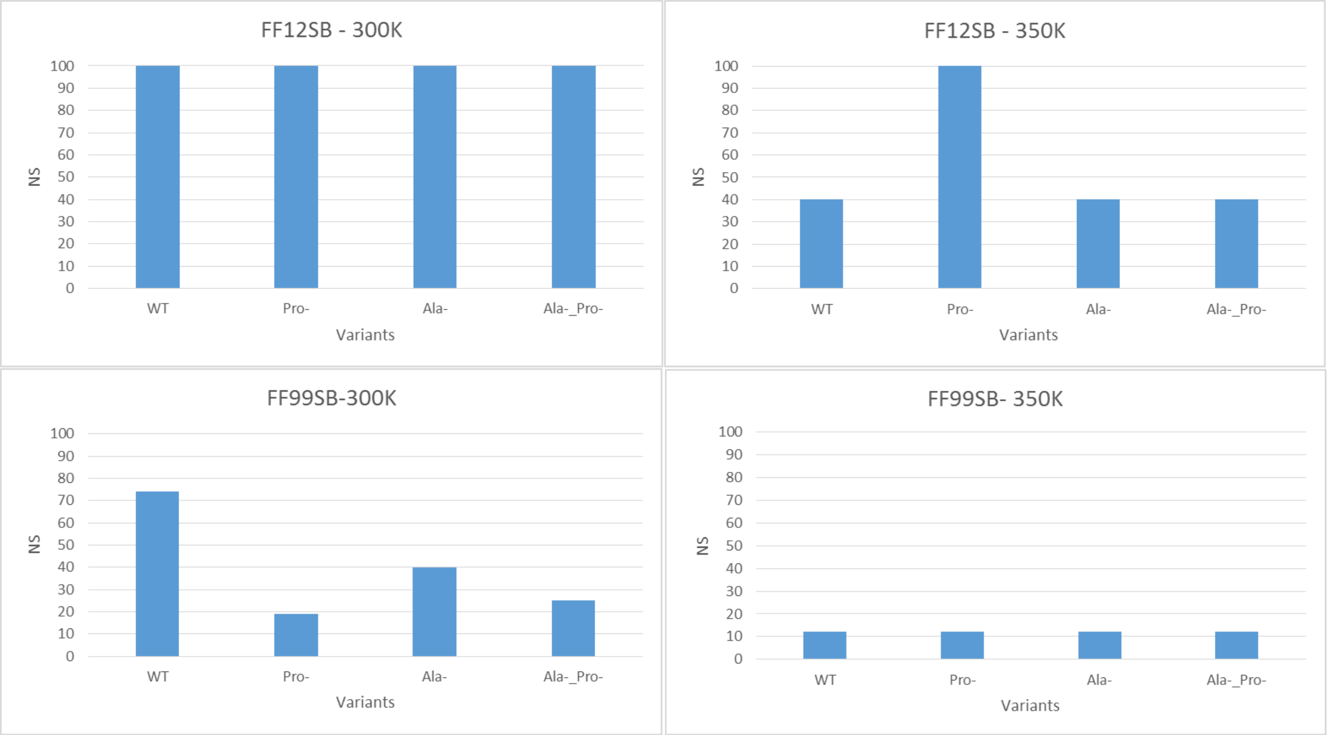
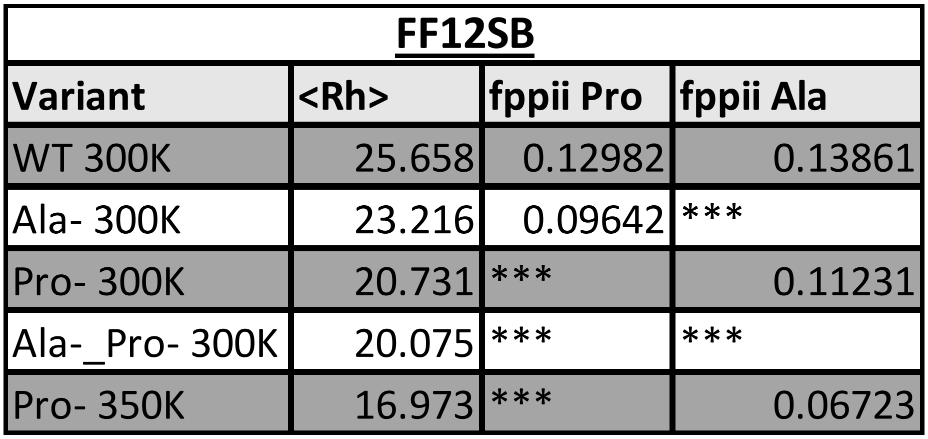


FIG. 14. **Progress of MD Simulations.** Progress of each simulation is displayed above in the four graphs. The titles display the force field used and temperature of each simulation. The four bars represent the four variants. Their respective names are listed below each bar.

3.2 *Polyproline II Propensities and Hydrodynamic Radius*



FF12SB			
Variant	**<Rh>**	**fppii Pro**	**fppii Ala**
WT 300K	25.658	0.12982	0.13861
Ala- 300K	23.216	0.09642	***
Pro- 300K	20.731	***	0.11231
Ala-_Pro- 300K	20.075	***	***
Pro- 350K	16.973	***	0.06723

FIG. 15. **ppII Propensities and Rh Values for Completed Simulations.** The title displays the force field used. From left to right, the columns display the variant and temperature, the Rh, the ppII propensity of proline, and the ppII propensity of alanine. Asterisk demark values not calculated.

CHAPTER FOUR

DISCUSSION

4.1 *Conclusions*

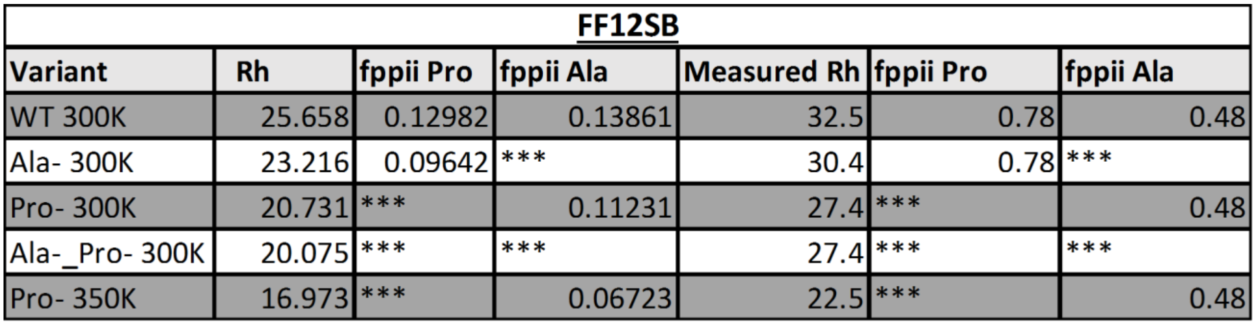
The Hydrodynamic Radius

Rh values from the simulation are both consistent and inconsistent with experimental measurements. Consistencies with experimental measurements are seen with the sequence mutation trends with the order Rh being, from largest to smallest, WT, Ala-, and Pro- & Ala-_Pro- (43). However, all of the variants respective Rh values are significantly smaller in magnitude compared to experimentally measured Rh values. The MD simulations also seem to have captured the temperature effects. The Pro- variant becomes more compact as the simulation temperature increases from 300K to 350K.

Polyproline II Propensities

Calculated ppII propensities for alanine and proline residues from the WT simulation at 300K with the ff12SB force field are .13 and .12 respectively. They are significantly smaller than previously calculated ppII propensities, .48 and .78 respectively (43). Decreases in calculated ppII propensities from the simulations are observed when temperature is increased from 300K to 350K with the Pro- variant. The ppII propensities from the simulation also decrease as Rh values decrease due to sequence effects.

However, ppII propensities from the simulation are so small that variation in Rh between variants and temperatures with the Pro- variant are likely not a result of backbone ppII conformations propensity fluctuation. This, if true, is contradictory to our previous computational and experimental studies. Also, inconsistent with our previous studies is the order of alanine and proline ppII propensities. Alanine and proline ppII propensities previously calculated have proline with a greater ppII propensity than alanine, .78 and .48, respectively, while the MD simulations have alanine with greater ppII propensities than proline.

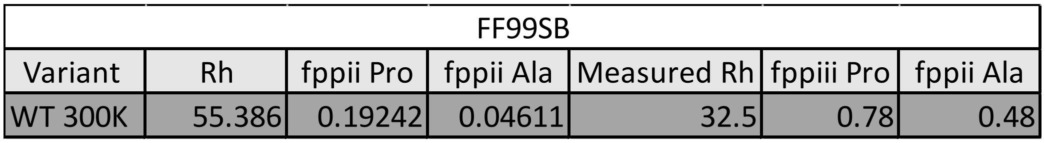


FF12SB						
Variant	Rh	fppii Pro	fppii Ala	Measured Rh	fppii Pro	fppii Ala
WT 300K	25.658	0.12982	0.13861	32.5	0.78	0.48
Ala- 300K	23.216	0.09642	***	30.4	0.78	***
Pro- 300K	20.731	***	0.11231	27.4	***	0.48
Ala-_Pro- 300K	20.075	***	***	27.4	***	***
Pro- 350K	16.973	***	0.06723	22.5	***	0.48

FIG. 16. **Comparison of Completed ff12B Simulations to Previous Experiments.** This table compares the simulation results of the completed simulations with the ff12SB force field, columns 2-4, to the results of our previous studies, columns 5-7.

4.2 *Preliminary Results of ff99SB Fore Field Simulation*

Preliminary results of the ff99SB force field simulations give credence to the dismissal of correlations between Rh fluctuations and ppII propensities in the MD simulations. Interestingly, the preliminary results of Rh on the unfinished WT simulation with the ff99SB force field at 300k are significantly larger than experimentally measure Rh values. This sharply contrast with the results of the ff12SB WT p53 (1-93) at 300K with a calculated Rh of 25.6 angstroms.



FF99SB						
Variant	Rh	fppii Pro	fppii Ala	Measured Rh	fppiii Pro	fppii Ala
WT 300K	55.386	0.19242	0.04611	32.5	0.78	0.48

FIG. 16. **Preliminary Results of ff99SB Simulation to Results of Previous Studies.** This table show the comparison of the preliminary results of the ff99SB simulation with the WT variant at 300K, columns 2-4, with the results of our previous studies, columns 5-7.

The ppII propensities for a 93 residue protein with an Rh of 55.3 angstroms would be significantly greater according to our previous model (15). The extent of protein expansion is also unrealistic when compared to the experimentally measured Rh proteins that are more than 2.5 greater in residue number (FIG. 6.)

4.3 *Future Experimental Goals*

In the future we may expand the definition of ppII conformations from -/+ 10 degrees of the canonical (φ) and (ψ) values to -/+ 20 degrees in order to reaffirm that the simulations are not producing structures with similar ppII propensities to our previous calculations.

Using AMBER analysis software CPPTRAJ, we want to calculate the solvent accessible surface area (SASA) of the tryptophan sidechains in our MD simulation to compare to our previous experiments. To do this we will need to first calculate the SASA of a tryptophan sidechain in a 5-mer peptide with the sequence: Alanine-Alanine-Tryptophan-Alanine-Alanine. This will be used as a benchmark for the SASA of a fully exposed tryptophan sidechain. The methyl sidechain of alanine will give a benchmark that is not unrealistically exposed as would be the case if a glycine or NATA molecule was used.

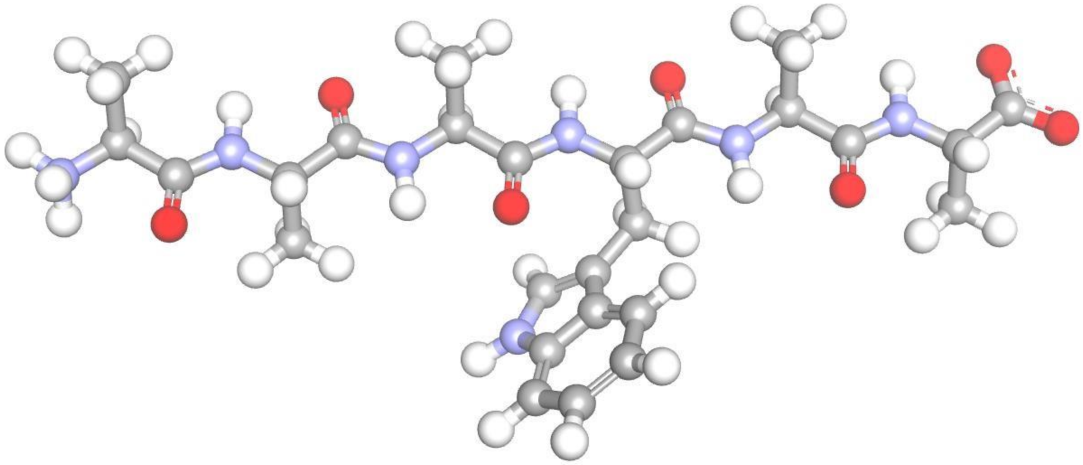


FIG. 18. **5-mer Peptide Used to Calculate Benchmark Tryptophan SASA.** Made with VMD 1.9.1, this molecule is the 5-mer peptide with sequence, Alanine-Alanine-Tryptophan-Alanine-Alanine, used to calculate the SASA for a solvent exposed tryptophan (4).

It appears the MD simulations are not capturing important structural features of IDPs and IDRPs when compared to our previous studies. Once all of the simulations have completed we will have a better understanding of how standard MD software force fields, such as the AMBER ff99SB and ff12SB force fields, affect various properties of the p53 (1-93) specifically, and in general, intrinsically disordered proteins. If the trends seen in the preliminary data continue, the completed simulation results will highlight the need for development of more robust MD force fields that can elucidate high-resolution molecular descriptions of IDPs and IDRPs structural features; thus, ultimately offering a deeper understanding of their role in normal cellular functions and disease pathologies.

REFERENCES

1. Griffith, C. E., and VaidaIn, V. (2012) Situ observation of peptide bond formation at the water–air interface. *PNAS* 109, 15697–15701.
2. Branson, H. R., Corey, H.R., Pauling, L. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *PNAS* 37, 205–211.
3. Flory, P.J., (1969) Statistical mechanics of chain molecules. *Interscience*, 253.
4. Humphrey, W., Dalke, A. and Schulten, K., (1996) VMD- Visual Molecular Dynamics. *J. Molec.. Graphics* 14, 33-38.
5. Richardson J. S. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34, 167–339.
6. Holden, H.M., Larsen, T.M., Laughlin, L.T., Rayment, I., Reed, G.H. (1994). Structure of rabbit muscle pyruvate kinase complexed with Mn2+, K+, and pyruvate. *Biochemistry* 33, 6301-6309
7. Tompa, P. (2012). Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* 37, 509-516
8. Eliezer, D. (2009). Biophysical characterization of intrinsically disordered proteins. *Curr Opin Struct Biol* 19, 23-30.
9. Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications., Song, J., Lee, M.S., Carlberg, I., Markley, J.L., Vener, A.V. (2006) *Biochemistry* 45, 15633-15643
10. Bax, A., Cornilescu, G., Marquardt, J.L., Ottiger, M. (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J.Am.Chem.Soc* 120, 6836-6837
11. Ausio, J., Bailer, R.W., Brown, Campen, A.M., Chiu, W., C.J., Dunker, A.K., Garner, E.E, Griswold, M.D., Hipps, K.W., Kang, C., Kissinger, C.R., Lawson, J.D., Nissen, M.S., Obradovic, Z., Oh, J.S., Oldfield, C.J., Ratliff, C.M., Reeves, R., Romero, P., Williams, R.M. (2001). Intrinsically disordered protein. *J Mol Graph Model* 19, 26-59
12. Liu, J., Oldfield, C.J., Perumal, N.B., Su, E.W., Uversky, V.N., Dunker, A.K., (2006). Intrinsic disorder in transcription factors. *Biochemistry* 45, 6873-6888.
13. Langridge, T.D., Tarver, M.J., Whitten, S.T. (2014). Temperature Effects on the hydrodynamic radius of the intrinsically disordered N-terminal region of the p53 protein. *Protein: Structure, Function, Bioinformatics* 82, 668-678.

14. Marsh, J.A., Forman-Kay, J.D. (2010). Sequence determinants of compaction in intrinsically disordered proteins. *Biophys J* 98, 2383-2390.
15. Choi UB, McCann JJ, Weninger KR, Bowen ME. Beyond the random coil: stochastic conformational switching in intrinsically disordered proteins. Structure 2011;19:566–576.
16. Wilkins DK, Grimshaw SB, Receveur V, Dobson CM, Jones JA, Smith LJ. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. Biochemistry 1999;38:16424–16431.
17. Chong PA, Ozdamar B, Wrana JL, Forman-Kay JD. Disorder in a target for the smad2 mad homology 2 domain and its implications for binding and specificity. J Biol Chem 2004;279:40707–40714.
18. Sivakolundu SG, Nourse A, Moshiach S, Bothner B, Ashley C, Satumba J, Lahti J, Kriwacki RW. Intrinsically unstructured domains of Arf and Hdm2 form bimolecular oligomeric structures in vitro and in vivo. J Mol Biol 2008;384:240–254.
19. Danielsson J, Liljedahl L, Bárány-Wallje E, Sønderby P, Kristensen LH, MartinezYamout MA, Dyson HJ, Wright PE, Poulsen FM, Mäler L, Gräslund A, Kragelund BB. The intrinsically disordered RNR inhibitor Sml1 is a dynamic dimer. Biochemistry 2008;47:13428–13437.
20. Yi S, Boys BL, Brickenden A, Konermann L, Choy WY. Effects of zinc binding on the structure and dynamics of the intrinsically disordered protein prothymosin alpha: evidence for metalation as an entropic switch. Biochemistry 2007;46:13120–13130.
21. Paleologou KE, Schmid AW, Rospigliosi CC, Kim HY, Lamberto GR, Fredenburg RA, Lansbury PT Jr, Fernandez CO, Eliezer D, Zweckstetter M, Lashuel HA. J Biol Chem 2008;283:16895–16905.
22. Baker JMR. Structural characterization and interactions of the CFTR regulatory region (PhD Thesis). Department of Biochemistry, University of Toronto, Toronto; 2009.
23. Soragni A, Zambelli B, Mukrasch MD, Biernat J, Jeganathan S, Griesinger C, Ciurli S, Mandelkow E, Zweckstetter M. Structural characterization of binding of Cu(II) to tau protein. Biochemistry 2008;47:10841–10851.
24. Lowry DF, Stancik A, Shrestha RM, Daughdrill GW. Modeling the accessible conformations of the intrinsically unstructured transactivation domain of p53. Proteins 2008;71:587–598.
25. Adkins JN, Lumb KJ. Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57Kip2. Proteins 2002;46:1–7.
26. Uversky VN, Permyakov SE, Zagranichny VE, Rodionov IL, Fink AL, Cherskaya AM, Wasserman LA, Permyakov EA. Effect of zinc and temperature on the conformation of the gamma subunit of retinal phosphodiesterase: a natively unfolded protein. J Proteome Res 2002;1:149–159.

27. Donaldson L, Capone JP. Purification and characterization of the carboxyl-terminal transactivation domain of Vmw65 from herpes simplex virus type 1. J Biol Chem 1992;267:1411–1414.
28. Haaning S, Radutoiu S, Hoffmann SV, Dittmer J, Giehm L, Otzen DE, Stougaard J. An unusual intrinsically disordered protein from the model legume Lotus japonicus stabilizes proteins in vitro. J Biol Chem 2008;283:31142–31152.
29. Geething NC1, Spudich JA. Identification of a minimal myosin Va binding site within an intrinsically unstructured domain of melanophilin. J Biol Chem 2007;282:21518–21528.
30. Permyakov SE, Millett IS, Doniach S, Permyakov EA, Uversky VN. Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. Proteins 2003;53:855–862.
31. Magidovich E, Orr I, Fass D, Abdu U, Yifrach O. Intrinsic disorder in the Cterminal domain of the Shaker voltage-activated K+ channel modulates its interaction with scaffold proteins. Proc Natl Acad Sci USA 2007;104:13022– 13027.
32. Campbell KM1, Terrell AR, Laybourn PJ, Lumb KJ. Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. Biochemistry 2000;39:2708–2713.
33. Sánchez-Puig N, Veprintsev DB, Fersht AR. Binding of natively unfolded HIF1alpha ODD domain to p53. Mol Cell 2005;17:11–21.
34. Sánchez-Puig N, Veprintsev DB, Fersht AR. Human full-length securin is a natively unfolded protein. Protein Sci 2005;14:1410–1418.
35. Tcherkasskaya O, Davidson EA, Uversky VN. Biophysical constraints for protein structure prediction. J Proteome Res 2003;2:37–42.
36. Read, A. P.; Strachan, T. (1999). *Human molecular genetics.2*. New York: Wiley. Chapter 18: Cancer Genetics
37. Bourdon J.C., Khoury M.P., Surget, S. (2013). Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *OncoTargets and Therapy* 7, 57–68.
38. Venot, C., Maratrat, M., Dureuil, C., Conseiller, E., Bracco, L., Debussche, L. (1998). The requirement for the p53 proline-rich functional domain for mediation of apoptosis is correlated with specific PIG3 gene transactivation and with transcriptional repression. *EMBO J.* 17, 4668–4679
39. D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman (2014), AMBER 14, University of California, San Francisco.
40. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995). A second generation

force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117, 5179–5197

41. Cino EA, Wong-Ekkabut J, Karttunen M, Choy WY (2011) *PLoS One* 6
42. Lomize, A.L., Pogozheva, I.D., Mosberg, H.I (2011). Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides and proteins in membranes. *J Chem Inf Model* 51, 930–946
43. Perez, R.B., Tischer, A., Auton, M., Whitten, S.T. (2014). Alanine and proline content modulate global sensitivity to discrete perurbations in disordered proteins. *Proteins: Structure, Function, Bioinformatics* 82, 3373-3384.