Evaluating a coarse-grained IDP-model for structure and dynamics

Louise Kalander

Supervisors: Eric Fagerberg and Marie Skepö

Lund University | Division of Theoretical Chemistry

Bachelor's essay, 15 credits

RONNEBY, January 2021

Examinator: Pär Söderhjelm

Acknowledgements

This report is a Bachelor's thesis that has been done as part of the program "Master of Science in Biotechnology" at Lund University. The project has been carried out at the Division of Theoretical Chemistry at the Science Faculty, Lund University, and it corresponds to 15 credits. I would like to give a special thank you to my supervisors Eric Fagerberg and Marie Skepö for their support during the running of this project. They have given me useful comments and always tried their best to help me with my questions. I would also like to thank Kristoffer Modig who was very helpful by helping me get in touch with Marie Skepö and made this Bachelor's thesis possible. Furthermore, I would like to say thank you to my examinator Pär Söderhjelm for the task of examining my work.

Ronneby, January 2021.

Louise Kalander

Abstract

In this study a coarse-grained model of intrinsically disordered proteins (IDPs) is evaluated to determine whether the model is suitable to use for analyzing the structure and dynamics of IDPs. IDPs are proteins that lack a stable tertiary structure which gives them a flexible structure. The aim with this study is to find a model that is appropriate to use for analyzing dynamics events of IDPs. To evaluate if the model is appropriate to use, the structural characteristics of the protein are examined to verify that the generated results are consistent with previous studies. With a good model it will be possible to achieve a deeper understanding about these proteins, which is of considerable importance since IDPs are involved in several vital biological processes in eukaryotes.

In this study, a specific model, primarily described by Das et al [1], is evaluated whether it can be used as a suitable model. To evaluate the model, the results from different simulations, with the IDP Histatin 5 [2] used as the model protein, will be compared and analyzed. Running several simulations with different set-ups of parameters from the original model will give a deeper understanding of how these parameters affect the conformation and flexibility of the IDP, and thus give the ability to improve the model. The simulations were performed by using the molecular dynamics package GROMACS [3]. To evaluate if the model is appropriate to use for analyzing the structure and dynamics of IDPs, the results from the different simulations are compared with protein contact maps and experimental data from small angle X-ray scattering.

The results from a reference simulation, a simulation where there were as few changes as possible from the original model, indicates that the simulated Histatin 5 resembles a globular protein rather than an IDP that would have been the desired result. The conclusion can be drawn that the original model is not suitable to use for examining dynamic events of IDPs, but additional simulations have to be performed since some deviations made from the original model could have affected the results and thus the conclusion.

Four simulations with different set-ups of the parameters from the original model are performed in this study. The results from these simulations indicate that there are clear differences in the behavior of the simulated proteins compared to the protein simulated in the reference simulation. Most of the parameter changes made in this study gave a satisfactory result, where the properties of the simulated protein are more similar to an IDP than to a globular protein. This study is only a first step in the process of finding a suitable model to be able to examine both the structure and the dynamics of IDPs. The work has to be continued in order to find an appropriate model for this purpose and there are several simulations that could be performed to continue this study.

Table of contents

1. Introduction

1.1 Background

Around 30 years ago it was considered fiction that a protein that fails to form a stable tertiary structure could have an important biological function [4]. It was considered that only the proteins with a well-defined structure could exhibit any essential biological activity. It was not until the turn of the century, after decades of research with both experimental and computational approaches, that protein scientists finally changed their opinion [5]. IDPs are proteins that lack a stable tertiary structure which gives them a flexible structure [6]. It has been predicted that approximately 35% of the proteins in eukaryotic organisms contain significant intrinsically disordered regions and that about 25% of the proteins are likely to be completely disordered proteins [7]. The importance of these proteins is considerable since IDPs are involved in many biological processes and gives cause to many diseases within eukaryotes [6], such as Alzheimer's disease and diabetes [5]. The discoveries around IDPs have made IDP-related research to go from fiction to one of the most interesting and attractive fields in modern protein science [4], and the field is still expanding at a fast pace [5].

The characteristics of IDPs are that they lack the ability to fold under physiologic conditions [4]. Some of these proteins lose their flexibility upon binding to targets and folds into a stable configuration, others remain flexible. The IDPs that remain flexible constantly fluctuate between different structural states in physiological conditions, which results in a dynamic mixture of protein conformations [7]. The explanation to this behavior is explained in the encoding of the amino acid sequence of IDPs. A disordered region has many unique amino acid features compared to an ordered protein, such as high net charge and low hydrophobicity. IDPs also have the ability to go from disorder to order upon function, a fascinating ability that is absent in structured ordered proteins [4]. These variations make IDPs unique compared to structured proteins and makes them a prime target not only in protein science but also in other fields such as drug development and disease treatment [5].

Researchers have an interest in examining the structure and dynamics of IDPs to achieve a greater understanding about these proteins. Coarse-grained modeling and Monte Carlo simulations have previously been used to analyze the structure of IDPs in complex systems [8]. The disadvantage with Monte Carlo simulations is that this type of simulations is not possible to use for analyzing the dynamic events of a protein [3]. To be able to examine the dynamics of IDPs another approach has to be used. In this study, a specific model, primarily described by Das et al, is evaluated whether it can be used as a suitable model for simulations of dynamic events of IDPs. A prerequisite from our side is that the model will be coarse-grained because an atomistic model will become too demanding in terms of calculation at higher protein concentrations, which is the next step if the model proves to be suitable to use. This means that each amino acid in a protein is represented by one single particle, instead of multiple different atoms. These particles are from now on referred to as residues.

To evaluate the model, the results from five different simulations will be compared and analyzed. The simulations were performed by using the molecular dynamics package GROMACS [3]. Initially a reference simulation will be performed where an attempt is made to reproduce the original model, but some changes have been made to provide a more stable system and because some of the algorithms used in the original model are not possible to use in GROMACS. This is followed by a simulation where the Lennard-Jones (LJ) potential is scaled down to a third, which was also performed in the original model. Subsequently, simulations where the radius of the residues is decreased, the bond length is increased, and the protein concentration is increased, will be performed. In the last three simulations, the original value of the LJ potential is used. To evaluate if the model is appropriate to use for analyzing the dynamics of IDPs, the structural characteristics of the protein are examined to

verify that the generated results are consistent with previous studies. The results from the different simulations are compared with experimental data from small angle X-ray scattering (SAXS) performed by Cragnell et al [6] and protein contact maps generated by Fagerberg et al [9].

In this study, the model protein used is Histatin 5 (Hst5). Hst5 is a salivary cationic peptide occurring naturally in human saliva and provides a defense against oral candidiasis caused by *Candida albicans* [2]. There are 12 members of the Histatin family where Hst5 has the most potent antifungal activity. Hst5 also has an important role in the formation of a protective layer on the teeth surface that prevents microbial colonization. Furthermore, Hst5 can bind polyphenolic compounds such as tannin, which works as a bactericidal effect [6]. Hst5 is a short protein, consisting of 24 amino acids, and have a net charge of *+5* [10]. The protein is considered to belong to IDPs since nuclear magnetic resonance (NMR), SAXS, and circular dichroism (CD) measurements show that Hst5 has a flexible structure at physiological conditions [6].

1.2 Aim

There is a two-fold aim of this study:

i. To find a model that is appropriate to use for analyzing both structure and dynamics of IDPs, and thereby to achieve a deeper understanding about these proteins. The model primarily described by Das et al is reimplemented and evaluated, to be able to see if this is a suitable model to use to examine the dynamics events of IDPs.
ii. To find an answer to the following questions: How are the conformations and flexibility of an IDP affected if some parameters changed from the original model described by Das et al? If it turns out that the model is inappropriate to use for simulating the dynamic events of an IDP, what other simulations and tests could be carried out to continue this study to improve the model?

2. Materials and methods

In this section, an explanation of the evaluation approaches used in this study will initially be described. This is followed by a section describing exactly how the model is structured in the reference simulation, and which algorithms and equations that are applied in the simulation. Subsequently, there is a section with a detailed explanation of how all simulations were performed and what changes that have been made from the original model in the different simulations. At last, there is a section explaining the deviations made from the original model compared to the reference simulation.

2.1 Evaluation approaches

2.1.1 SAXS

SAXS is a method to study the structural characterizations of both ordered and disordered proteins in solution. The technique can provide low resolution information about the shape, conformation, and assembly state of proteins. From the SAXS measurements, the radius of gyration, the volume, the molecular mass, and the folding state can be determined to facilitate the examination of a protein. SAXS is one of few methods to quantitatively characterize the conformational ensemble in IDPs and other flexible, unfolded macromolecules [7].

When using SAXS in an experiment the solutions will be illuminated by a monochromatic beam of X-rays. Most of the X-ray beams go through the sample without interacting with it, while some of the X-rays scatter, see a schematic representation of a SAXS experiment in Figure 1 [7]. The scattering pattern of the sample and the pure solvent are collected using an X-ray detector. The scattering pattern from the solvent can then be subtracted from the sample solution pattern, what will remain is the scattering pattern from the particles of interest. This pattern can be used to examine the particles [7].

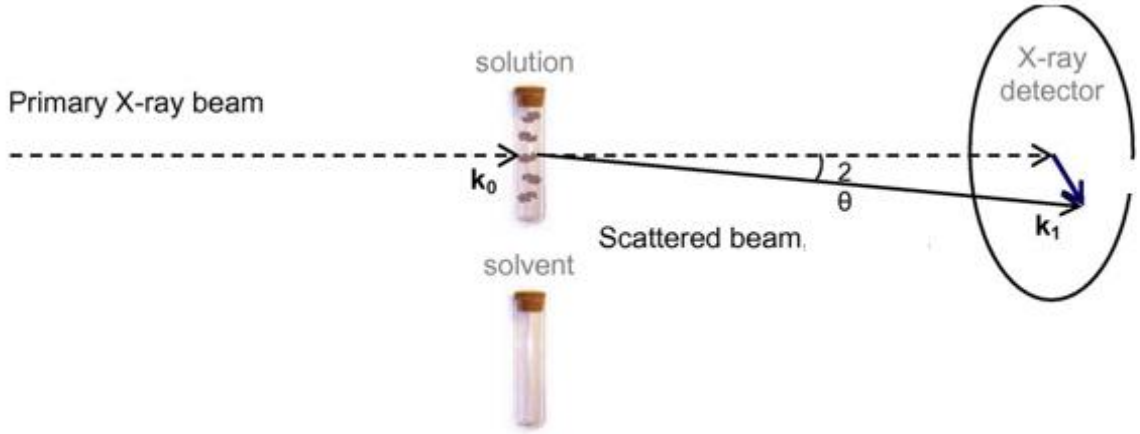

Figure 1. A schematic representation of a SAXS experiment [7].

Theoretical scattering profiles from SAXS can be used to analyze the simulated Hst5 in this study to evaluate if the model is suitable for further studies or not. A scattering profile is a plot of *I(q)* as a function of *q*, where *I* is the scattering intensity and *q* is the momentum transfer. The momentum transfer *q* is defined as $q = \frac{4\pi sin\theta}{\lambda}$, where λ is the wavelength of the beam and 2θ is the scattering angle [6].

The radius of gyration, R_g, is a powerful tool to examine how the protein folds in the solution. R_g provides the overall size of the protein and is defined by:

$$R_g = \left(\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2\right)^{1/2} \quad (1)$$

where *n* equals the number of atoms, r_i correspond to the coordinates of the atom *i*, and r_{cm} correspond to the center of mass of the molecule [11]. If a protein is stably folded, R_g will remain at a steady value, unlike a single protein chain of an IDP that will generate a dynamic mixture of protein conformations [7], and therefore will have a value of R_g that changes over time [12]. For a stably folded protein with a compact shape, R_g is smaller compared to an unfolded protein with the same number of amino acids [7]. The simulated R_g value for Hst5 is then compared to an experimental reference value of R_g = 13.8 Å with a margin of error of 1-2 Å generated by Cragnell et al [6]. The approach and additional information about the SAXS experiments of Hst5 used to determine the experimental reference value can be found in reference [6].

To qualitatively measure the flexibility of the protein, a Kratky plot, which is a sort of scattering profile, can be used. A Kratky plot is a plot of $q^2I(q)$ as a function of *q.* Highly flexible proteins, as IDPs, should have a plateau or an ascending curve in the Krakty plot at a high *q,* while compact proteins will have a high peak before stabilizing at a low $q^2I(q)$ value. A partially flexible protein will have either a combination of the peak and the plateau, or a plateau that slowly decays to zero. Since Hst5 is a flexible IDP, the Kratky plot should have an increase in the $q^2I(q)$ value that stabilizes on a plateau at a high *q* [13]. Figure 2 show characteristic Kratky plots for different proteins [14]. The IDP in the figure is referred to as an unfolded protein.

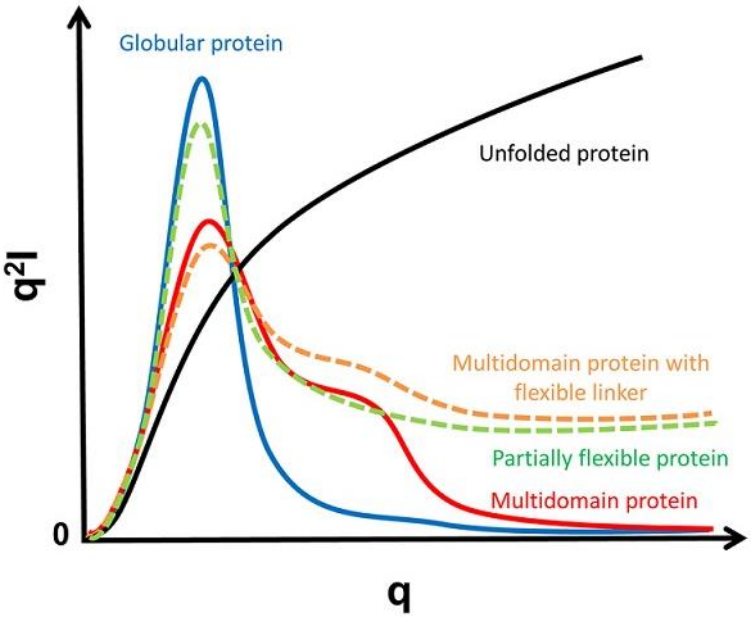

Figure 2. Typical Kratky plots for different proteins. The IDP is in the figure referred to as an unfolded protein [14].

The results from the simulations in this study are compared to experimental data generated by Cragnell et al [6]. In their study a dimensionless Kratky plot is used, where $qR_g{}^2\frac{I(q)}{I(0)}$ is plotted as a function of qR_g. The advantage of using a dimensionless Kratky plot is that it can provide semi-quantitative analyzes of both flexibility and disorder of a protein [13].

2.1.2 Contact map

The 3D structure of a protein can be represented in a more reduced way by using a protein contact map rather than studying the protein's full 3D atomic coordinates [15]. The advantage with the contact map is that it is invariant to rotations, which makes it easy to analyze the structure and conformation of a protein [16].

A protein contact map is generated by using a distance matrix that is a symmetric two-dimensional matrix containing the Euclidean distances between each pair of residues for a single snapshot [15]. The Euclidean distance is the length of a line between two points in Euclidean space, the space used in classical geometry [17]. To determine whether two residues are connected, the Euclidean distance between residues should be less than, or equal to, a specific cut-off value. The protein contact map is represented in a two-dimensional binary matrix and is produced by using the cut-off value and the distance matrix. If any two residues are in contact, the matrix cell value is set to one, and the matrix cell will get a black color. If there is no connection between two residues, the matrix cell value will be set to zero, and the matrix cell will get a white color. The protein contact map can be displayed and shows a black and white pattern that represents the structure of the protein [15].

In this study, a distance map will be used to examine the results from the simulations. That is a generalization of a protein contact map in which the distances are represented with a color gradient instead of black and white dots and is created directly from the distance matrix. Each of the numbers in the distance matrix corresponds to a distance between two residues, and the distances is then visualized as colors, where the darkest blue is the longest distance, and the darkest red is the shortest distance between two residues. The distance map makes it easier to distinguish the movements of the protein and to analyze how the proteins conformation changes during a simulation than by using a protein contact map [18].

The distance maps produced from Hst5 in this study are compared to a protein contact map of Hst5 generated by Fagerberg et al [9]. If the protein contact map visualizes a contact between two residues, the distance between these two residues is shorter than the cut-off value, and if there are no contact between the residues, the distance is greater than the cut-off value. In the study by Fagerberg et al, the cut-off value is set to 8 Å and their generated contact map shows that contacts between residues in Hst5 are very local except for in the end of the chain, and also to a minor degree in the middle of the chain.

2.2 Simulations

As described earlier, this study is based on the model primarily described by Das et al. The model is re-implemented in GROMACS to perform the molecular dynamic simulations. The program VMD, a molecular visualization program, is then used to be able to display the simulated proteins in 3D [19]. VMD is used to instantly examine the dynamics of the protein when changing various parameters from the original model to determine whether the change improved the model or not. In the original model reduced units are used throughout the report. Since GROMACS do not support reduced units as input, all units from the original model have been recalculated.

2.2.1 Model

To prevent the large computational load required in atomic simulations, a coarse-grained representation of the protein is used, similar to the original model. Figure 3 shows how the atomistic version of Hst5 has been transformed to a coarse-grained version that will be used for the simulations. The model of Hst5 in the Figure has been generated by the program VMD [19]. The amino acid sequence for Hst5 can also be seen in the Figure. The different colors in the Figure represents a specific amino acid: ASP = red, SER = green, HIS = dark blue, ALA = yellow, LYS = magenta, ARG = light blue, GLY = orange, TYR = purple, PHE = brown, GLU = silver.

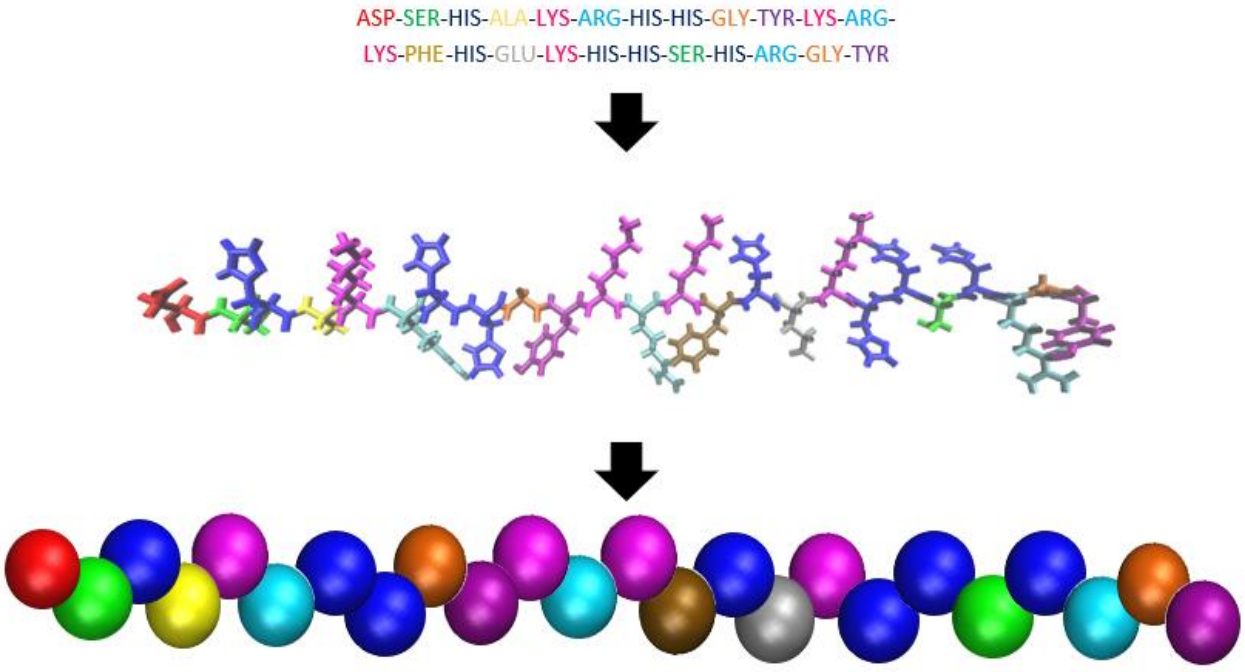

Figure 3. How the atomistic version of Hst5 is transformed to a coarse-grained version used for the simulations. The models of Hst5 in the figure have been generated by the program VMD [19]. The amino acid sequence for Hst5 is also displayed at the top of the figure. The different colors in the figure represents a specific amino acid: ASP = red, SER = green, HIS = dark blue, ALA = yellow, LYS = magenta, ARG = light blue, GLY = orange, TYR = purple, PHE = brown, GLU = silver.

In the model, water is treated implicitly. The implicit solvent is simulated by using the dielectric constant of water for different temperatures. The values for the dielectric constants used in this study is 122.2 for a temperature of 200 K, 97.4 for a temperature of 250 K and 77.6 for a temperature of 300 K [20].

The long-spatial-ranged electrostatic interactions among the residues are treated by the Particle-Particle Particle-Mesh (PPPM) algorithm. The PPPM method separates the total interaction between residues into a sum of short-ranged and long-ranged interactions. The former are then computed by direct particle-particle summation, while the latter are calculated by solving Poisson's equation using periodic boundary conditions [21].

The electrostatic interactions are modeled using the following equation:

$$(U_{el})_{\mu i,\nu j} = \frac{\sigma_{\mu i}\sigma_{\nu j}e^2}{4\pi\varepsilon_0\varepsilon_r r_{\mu i,\nu j}} \quad (2)$$

where μ,ν = 1,2,…,n label the IDP molecules where n is the total number of chains in the simulation and i,j = 1,2,…, N label the N residues in each chain. σ is the residues charges in units of elementary electronic charge e, ε_0 is the vacuum permittivity, ε_r is the dielectric constant and $r_{\mu i,\nu j}$ is the distance between two residues.

All non-bonded residue pairs also interact by the LJ potential:

$$(U_{el})_{\mu i,\nu j} = 4\varepsilon\left[\left(\frac{a}{r_{\mu i,\nu j}}\right)^{12} - \left(\frac{a}{r_{\mu i,\nu j}}\right)^{6}\right] \quad (3)$$

where r is the distance between two residues, ε is the well depth, and a is the LJ interaction range. Similar to the original model, the well depth is $\varepsilon = 1.67\ k_B T$, where T = 300 K, and the same a is used for all residues, see equation 4.

The LJ interaction range, a, is calculated by the same equation as in the original model:

$$a = \frac{e^2}{4\pi\varepsilon_0\varepsilon_r\varepsilon} \quad (4)$$

where e is the elementary charge, ε_0 is the vacuum permittivity, ε_r is the dielectric constant for a specific temperature. In the original model is ε_r = 80, so a value of a can be calculated to 4.18 Å.

In the simulations, a distance of $6a$ is used as the LJ cutoff and for simplicity the same cutoff is applied for the electrostatic interactions. In this study, similar to the original model, the mass is assumed to be the same for all residues. The value of the mass that has been used in the simulations is 126.5 Da since that is the average mass of the amino acids that Hst5 consists of. The time-step used in the simulations is 0.001 ns.

The bonded energy is described by a harmonic potential:

$$U_{bond}(r_{\mu i,\nu j}) = \frac{K_{bond}(r_{\mu i,\nu j}-a)^2}{2} \quad (5)$$

where $r_{\mu i,\nu j}$ refers to the distance between two beads, a refers to the equilibrium distance and K_{bond} is the force constant. The force constant is set to $K_{bond} = 1.8 \cdot 10^6$ kJ/(mol · nm^2). The length of the bonds is set to the same value as a, that is 4.18 Å. Two residue pairs that are sequential neighbors along the chain only interact by the bonded energy so consequently, the electrostatics- and LJ-interactions are neglected for these residues.

Langevin dynamics is used to simulate the kinetic properties in the system by using the leap-frog algorithm and is used in the simulations to extend the molecular dynamics to also account for effects that the solvent has had on the system [22]. The leap-frog algorithm is the default algorithm used in GROMACS and requires the start coordinates of the molecules as well as the initial velocities. The initial velocities are generated with a Maxwell-Boltzmann distribution at a given temperature, T**.** The leap-frog algorithm then updates the positions and the velocities using the forces at every time-step by the following relations [3]:

$$\mathbf{v}\left(t+\tfrac{1}{2}\Delta t\right) = \mathbf{v}\left(t-\tfrac{1}{2}\Delta t\right) + \frac{\Delta t}{m}\mathbf{F}(\mathbf{t}) \quad (6)$$

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v}\left(t+\tfrac{1}{2}\Delta t\right) \quad (7)$$

where **v** is the velocity, *t* the time, *m* the real mass of the residue, **r** the positions, and **F(t)** is the force determined by the position at a specific time *t*.

For temperature coupling, a weak Langevin thermostat is used, with a friction factor of 5.5 Da/ps, and a time-step of 0.001 ns. A Langevin thermostat is a stochastic thermostat used to add or remove energy from the boundaries of the system in a realistic way to approximate a canonical ensemble, which is an ensemble (*NVT)* where the number of particles, *N*, the volume, *V*, and the temperature, *T*, are held constant [23].

Before the simulation, an energy-minimization of the system is performed by applying a steepest descent algorithm. This algorithm is the recommended algorithm to use in GROMACS since it is robust and easy to implement, although there are other algorithms more efficient for searching. The algorithm requires an initially maximum displacement before the calculations can be executed, this is referred to as the energy step size later in the report [3].

2.2.2 Simulation aspects

Before the simulation starts the amino acid sequence in Hst5 is defined as a chain with coarse-grained residues. This is accomplished by writing a topology-file where the amino acids are defined as single particles. All residues are given the same mass and van der Waals parameters. The only difference between the residues is the elementary charge, it is either positive, negative, or neutral. The charge of the residues corresponds to the charge of the amino acid that the residue represents. In the topology-file, the counterions are included as well. Here chloride ions are used to neutralize the net charge of the system. The counterions are defined with the same parameters as the amino acid residues, that is, the same mass and the same parameters for van der Waals interactions.

A molecular structure file is composed in a similar way as the topology-file and describes the molecules position before the simulation begins. For simplicity, the residues are positioned with the same coordinates in two directions (*y* and *z*), and only vary in positions in one direction (*x*). This means that the residues of the protein are placed in a straight line as a start position for the simulation. The distance between the residues is equal to the bond length, which in this study is equal to the LJ interaction range, *a.*

The simulation begins with the chain of residues being centered in a periodic cubic box with a box length of 70*a*. After the protein is placed in the box the counterions are added. Since Hst5 has a net charge of *+5*, five chlorine ions are placed randomly inside the box to charge neutralize the system. This is followed by an energy-minimization of the chain configuration. An energy step size of 0.001 nm is used, and the energy-minimization is carried out until there is no difference in energy between two consecutive steps.

The simulation box is then heated to 300 K in three steps. First the system is heated to 200 K and equilibrated for five ns. Thereafter, the system is heated to 250 K for five ns. Finally, the system is heated to 300 K and equilibrated for 20 ns to be sure that the system has had enough time to reach equilibrium. The heating is carried out in three steps to minimize the fluctuation of the temperature by preventing that the system is heated to fast. Important to remember during the heating is that the dielectric constant changes with the temperature [20].

After this initial preparation, the production run is carried out for 40 ns, which is then followed by analysis of the results, where the trajectory from the production run is corrected by center the chain in the simulation box to avoid that the protein may appear broken or may jump across to the other side of the box [12]. This corrected trajectory will underlie all further analyzes in this study.

Before the results from the simulations are analyzed, the program VMD was used to display the simulated proteins in 3D to be able to examine the conformations of the proteins [19]. The conformations give a first sign if the simulations can simulate Hist5 correctly. R_g is computed by using the "gmx gyrate" method implemented in GROMACS that calculates R_g directly from the trajectories. To calculate the standard deviation the "gmx analyze" method implemented in GROMACS is used to calculate the standard deviation of R_g. The calculated R_g values from the simulations were directly compared to the experimental reference value generated by Cragnell et al [6] to be able to determine as a first step whether the different simulations could simulate Hist5 correctly or not.

The first step in the process of analyzing the results is to generate the SAXS data by using FoXS [24] to produce scattering profiles from the simulations. FoXS compute the theoretical scattering profiles by using the Debye formula:

$$I_m(q) = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} f_i(q) f_j(q) \frac{\sin(qd_{ij})}{qd_{ij}} \quad (8)$$

where d_{ij} is the Euclidean distance between residue *i* and *j*, $f_i(q)$ and $f_j(q)$ represents the atomic form factors of the residues [24].

FoXS requires a PDB format file as input to determine the SAXS scattering profile. To achieve this, the trajectory from the simulation is divided into multiple PDB-files. Every 0.5 ns a PDB-file is saved from the trajectory and stored in a PDB-file library. This library is then used as an input to FoXS that calculates a SAXS scattering profile for each PDB-file. An average of these profiles is used as the final SAXS scattering profile to generate a Kratky plot to be able to analyze the flexibility of Hst5.

To generate the distance maps, the "gmx mdmat" method implemented in the GROMACS package is used. The distance matrix is generated directly from the corrected trajectory, and a file containing a colored distance map is produced, where the colors represent the average distance between any two residues.

As described earlier, five different simulations are performed with different changes from the original model to give a deeper understanding how these parameters affect the IDP. The simulations are described in the list below:

I. A reference simulation with as few changes from the original model as possible. Some changes have been made to provide a more stable system and because some of the algorithms used in the original model are not possible to use in GROMACS.
II. A simulation where the LJ potential is scaled down to a third.
III. A simulation where the radius of the residues is decreased from 4.18 Å to 3.78 Å.
IV. A simulation where the bond length is increased from 4.18 Å to 5.00 Å.
V. A simulation where the protein concentration is increased to 200 IDP chains.

In the last three simulations, the original value for the LJ potential is used. The simulation aspects are exactly the same for all different simulation, except for the simulation with a higher protein concentration. The difference for this simulation is that 200 protein chains and 1000 counterions, five ions for each protein chain, are randomly added to the simulation box, instead of simulating one single protein chain located in the center of the box with five counterions added randomly. In this simulation is the simulation box the same size of 70*a* as in the previous simulations.

Files discussed and used in this section for all simulations can be found in the appendices.

2.2.3 Deviations from original model

In the original model it is studied how liquid-liquid phase separation (LLPS) of IDPs depends on their sequence charge patterns [1]. When a solution of proteins undergoes LLPS they condense into a dense

phase that coexists with a dilute phase. Whether a solution undergoes LLPS depends on the concentration and the characteristics of the protein in the solution, as well as the environmental conditions such as temperature, pH, and the volume [25]. In this study the protein is not undergoing LLPS since mainly single chain simulations are performed, then is LLPS not possible.

Other deviations made is that in the original model, 500 IDP chains are placed in a periodic cubic box at the start of the simulations. In this study is only 200 IDP chains simulated, thus a lower protein concentration than in the original model. The amino acid sequence and the length of the protein chain is also a deviation since Hst5 is not used as a model protein in the original model. In the original model are proteins containing only lysine and glutamic acid, with a length of 50 amino acids, used in the simulations. Counterions are not mentioned in the original model, which makes it unclear if counterions have been present during their simulations or not. It has been decided in this study to include counterions since their presence will provide a more stable system. Another thing not mentioned in the original model is the value of the mass. In this study a value of 126.5 Da is used since that is the average mass of the amino acids in Hst5.

How the energy minimization was carried out in the original model is not explained in the article by Das et al. In this study a steepest descent algorithm is used since that is the recommended algorithm to use in GROMACS for energy-minimization. The algorithm gives a satisfactory result in terms of minimizing the potential in the system, and is therefore suitable to use for this study.

During the simulations in this study, a time-step of 0.001 ns has been used, unlike the original model where a time-step of 0.0023 ns was used. The reason for not using 0.0023 ns in this study is because that gives too high fluctuations of the temperature. To minimize the fluctuations, a decision was made to make a change from the original model and decrease the time-step to 0.001 ns.

The kinetic properties of the system are modeled by Langevin dynamics using the leap-frog algorithm. In the original model, the velocity-Verlet algorithm is used instead. The two algorithms will generate the same trajectories if there is no pressure coupling and with corresponding starting points, which can be determined by comparing the equations for velocity-Verlet, equation 9 and 10, with the equations for the leap-frog algorithm, equation 6 and 7 [3]:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v} + \frac{\Delta t^2}{2m}\mathbf{F}(t) \quad (9)$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\Delta t}{2m}[\mathbf{F}(t) + \mathbf{F}\,(t + \Delta t)] \quad (10)$$

where **v** is the velocity, *t* the time, *m* the mass of the residue, **r** the positions and **F(t)** is the force determined by the position at a specific time *t*.

However, if a starting file with the same starting points is given, leap-frog and velocity-Verlet will give different trajectories as leap-frog interprets the velocities according to $t = -\frac{1}{2}\Delta t$ (see equation 6 and 7) and velocity-Verlet interpret the velocities corresponding to the timepoint $t = 0$ (see equations 9 and 10). This is not a problem in this study since no pressure coupling is used and the two different algorithms will give a negligible difference in the trajectories they generate. Why the leap-frog algorithm is used in this study instead of the velocity-Verlet algorithm is because the leap-frog algorithm is the only algorithm integrated with Langevin dynamics and the PPPM method in GROMACS [3].

3. Results

3.1 Radius of gyration

As a first step in the evaluation process of the model, R_g is determined as a function of time. The R_g values generated from the simulations can be seen in Figure 4, where each graph represents the result from a specific simulation. From the graphs the fluctuation of R_g over time can be examined. The calculated average R_g value with a standard deviation due to fluctuation is also displayed in the Figure.

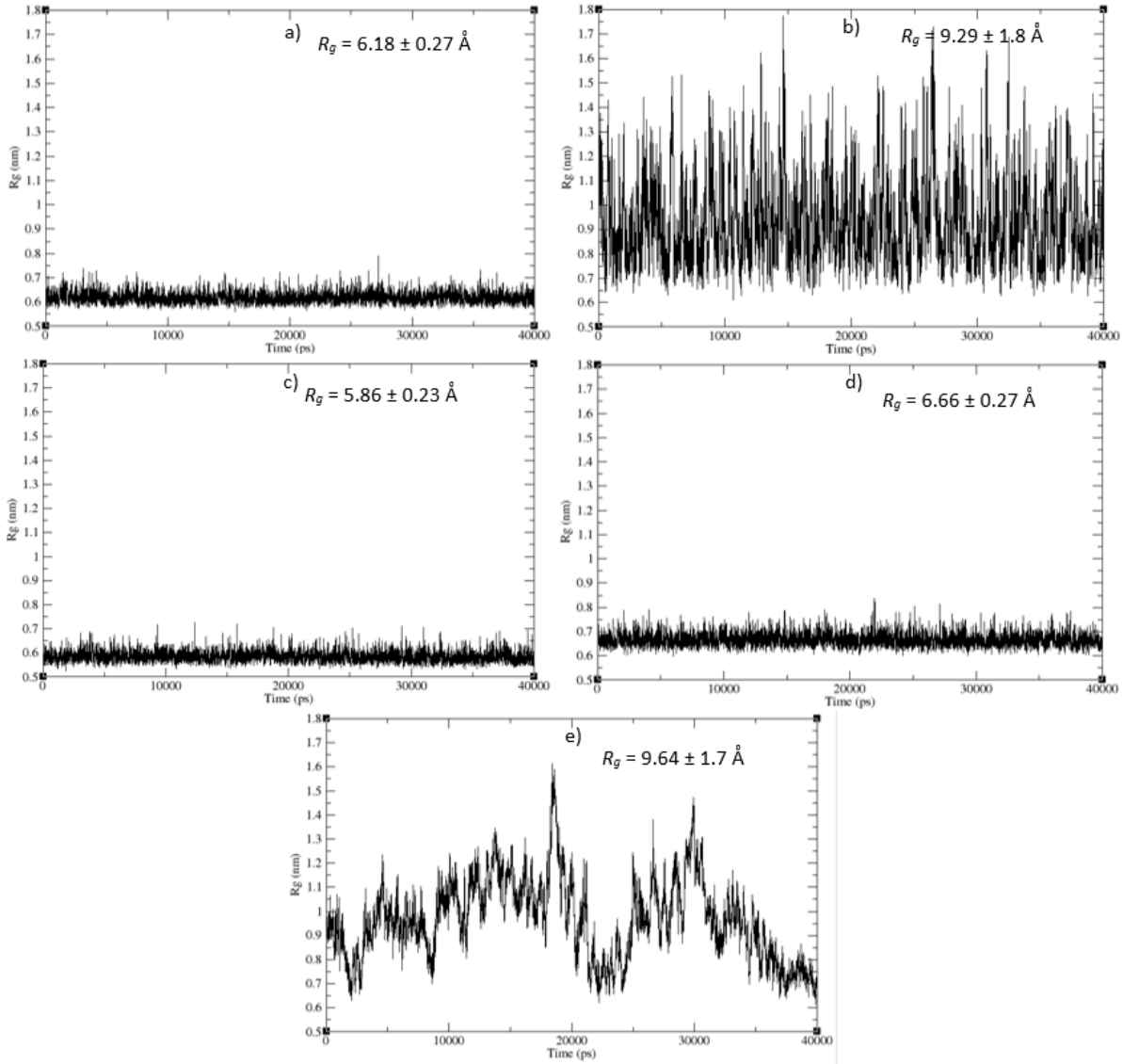

Figure 4. Data of R_g simulated from Hst5 using the implemented method in GROMACS. R_g (in nm) as a function of time (in ps). The calculated average R_g value with a margin of error is displayed in the Figure with the unit Å. The results are from: (a) the reference simulation, (b) the simulation with the LJ potential scaled down to a third, (c) the simulation with a decreased residue radius, (d) the simulation with an increased bond length, and (e) the simulation with high protein concentration.

The results in Figure 4 are compared with the experimental reference data of R_g = 13.8 Å with a margin of error of 1-2 Å generated by Cragnell et al [6]. The R_g generated for the reference simulation, Figure 4(a), indicates that Hst5 is stably folded and does not resemble an IDP. This conclusion can be drawn since R_g remains at a steady value when a protein is stably folded. The R_g value is determined to 6.18 ± 0.27 Å, which is about half of the experimental data for Hst5 according to the experimental value. Figure 4(b) shows the result from the simulation with a LJ potential scaled down to a third. The Figure

shows that R_g in this simulation has large fluctuations, and this indicates that the simulated protein resembles an IDP, since a flexible protein has a R_g that fluctuates significantly. The simulated Hst5 in this simulation has a R_g value of 9.29 ± 1.8 Å, which is closer to the experimental reference than the reference simulation could generate, see Figure 4(a).

The simulation where the radius of the residues is decreased, see Figure 4(c), generates a R_g value of 5.86 ± 0.23 Å. This value is not close to the experimental reference data and indicates that this simulation cannot simulate Hst5 correctly. From Figure 4(d), which represent the result from the simulation with an increased bond length, an R_g value of 6.66 ± 0.27 Å can be determined. This value is somewhat higher than the R_g generated from the reference simulation and could indicate that the simulated protein is slightly more unfolded than the simulated protein in the reference simulation. However, the R_g has no fluctuations which indicates that the simulated protein does not resemble an IDP. Figure 4(e) shows that R_g has large fluctuations in the simulation with a high protein concentration, and is determined to 9.64 ± 1.7 Å, which is significantly closer to the experimental reference data. These two observations indicate that the simulated protein in this simulation is more flexible than the simulated protein in the reference simulation.

To draw a conclusion if the simulation time of 40 ns were sufficient for the simulations the time autocorrelation function of R_g is calculated:

$$C(t) = \frac{\langle (R_g(t) - \langle R_g \rangle)(R_g(0) - \langle R_g \rangle) \rangle}{\left(\langle R_g^2 \rangle - \langle R_g \rangle\right)^2} \tag{11}$$

where $R_g(t)$ is the R_g value at time t, $\langle R_g \rangle$ is the average value and $R_g(0)$ is the R_g value at time t=0 [26]. Autocorrelation refers to the degree of correlation of the same variables between successive time intervals. It is used to determine if the simulation time used in this study is sufficient by measure how a lagged version of the simulated values is related to the original version [27]. The autocorrelation function was calculated by the “gmx analyze” method implemented in the GROMACS package [3]. The results of the calculations are displayed in Figure 5 where C(t) is a function of time (ps).

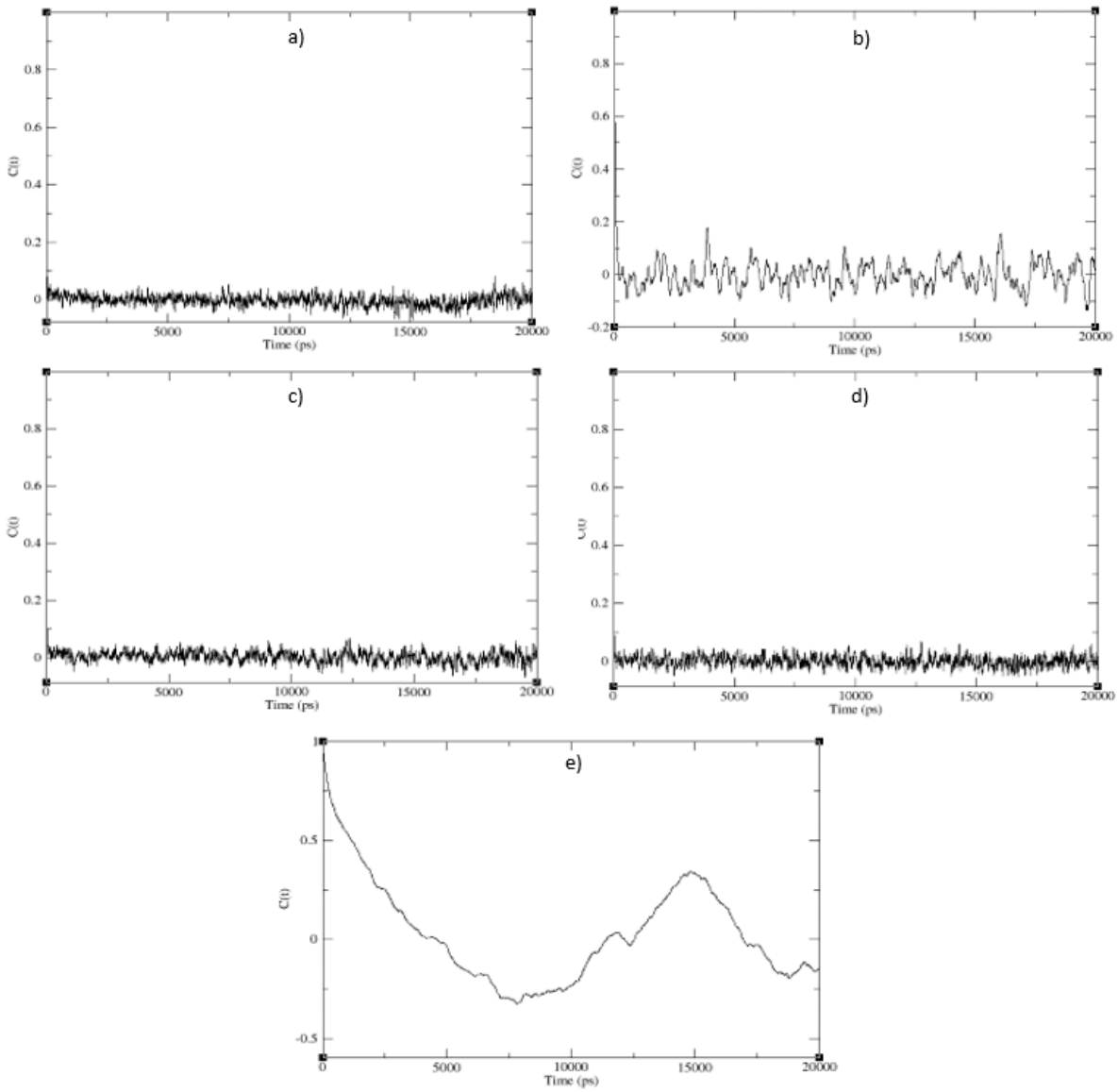

Figure 5. Time autocorrelation function of R_g, where C(t) is a function of time (ps).

From Figure 5a-5d it can be determined that the simulation time of 40 ns is sufficient for these simulations. This can be determined since there is no consistent change in autocorrelation, which indicates that these systems have converged and does not need more time to stabilize. From Figure 5(e) it can be determined that the simulation time of 40 ns may not have been sufficient for the simulation. The consistent change in autocorrelation indicates that the system has not converged and may need more time to stabilize. From Figure 4(e) it can at least be determined that the average R_g value is not close to either the R_g value generated by the reference simulation (6.18 ± 0.27 Å) or the experimental data of 13.8 Å.

3.2 Snapshot of the protein conformation

The snapshots are visualizing the conformation of the protein and are used to determine the structure of the protein. The snapshots in Figure 6 are generated with the program VMD [19], the black dots represent the residues, and the pink dots represent the counterions. In the snapshot from the simulation with a high protein concentration, see Figure 6(e), one single protein chain is represented with bonds between the residues for clarity.

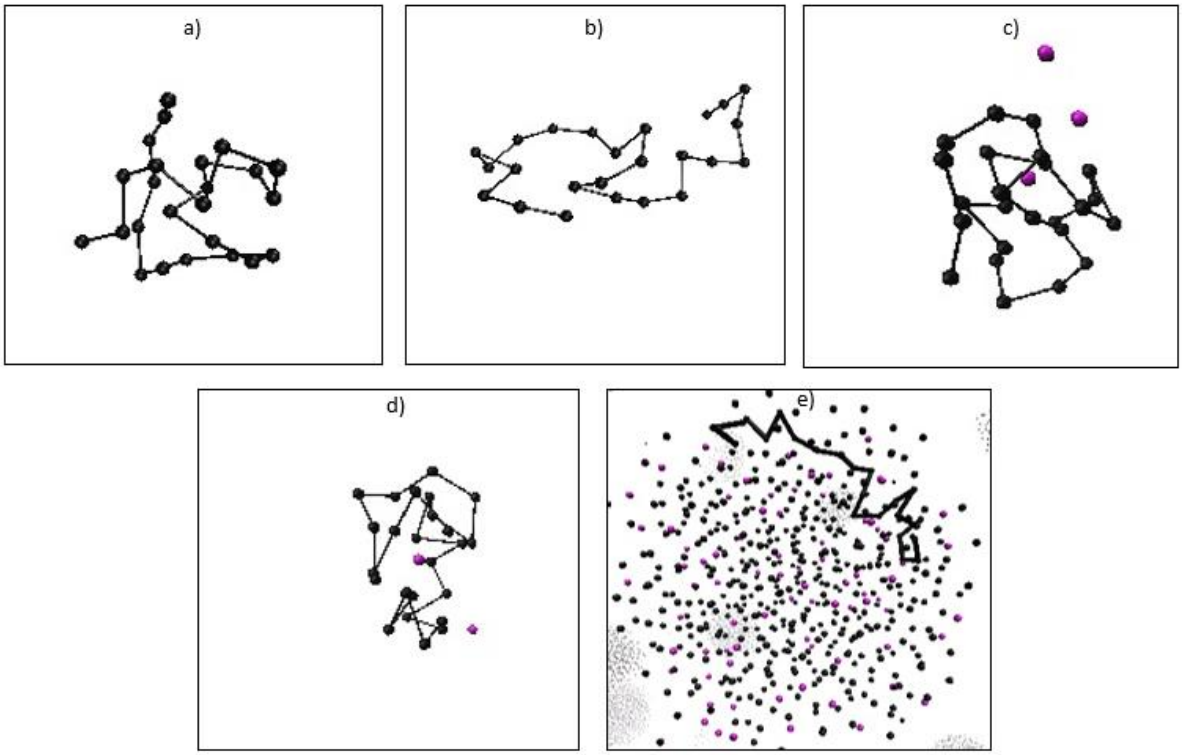

Figure 6. Snapshots of the simulated protein generated with the program VMD [19]. The black dots represent the residues, and the pink dots represent the counterions. The results are from: (a) the reference simulation, (b) the simulation with the LJ potential scaled down to a third, (c) the simulation with a decreased residue radius, (d) the simulation with an increased bond length, and (e) the simulation with high protein concentration.

The characteristic of IDPs is that they lack the ability to fold under physiological conditions [4]. Since the simulations in this study are performed under such conditions, the simulated protein should be unable to fold. The snapshots visualizing the average conformation of the protein, see Figure 6, should visualize unfolded proteins if the simulated protein resembles an IDP. Figure 6(a), representing the reference simulation, and Figure 6(c) and 6(d), representing the simulations with a decreased radius and an increased bond length, indicates that the simulated proteins in these simulations are folded and resembles globular proteins rather than IDPs. Figure 6(b) and 6(e), representing the simulation with a LJ potential scaled down to a third and the simulation with a high protein concentration, respectively, show that the simulated proteins in these simulations are unfolded and could behave like IDPs.

3.3 Scattering curves

The theoretical scattering curves, a logarithmic plot of the scattering intensity I(q) as a function of q, obtained from the simulations of Hst5 are presented in Figure 7. All theoretical scattering profiles obtained by using FoXS [24] are represented in the Figure as blue lines, whereas the thicker red line represents the average of these scattering profiles.

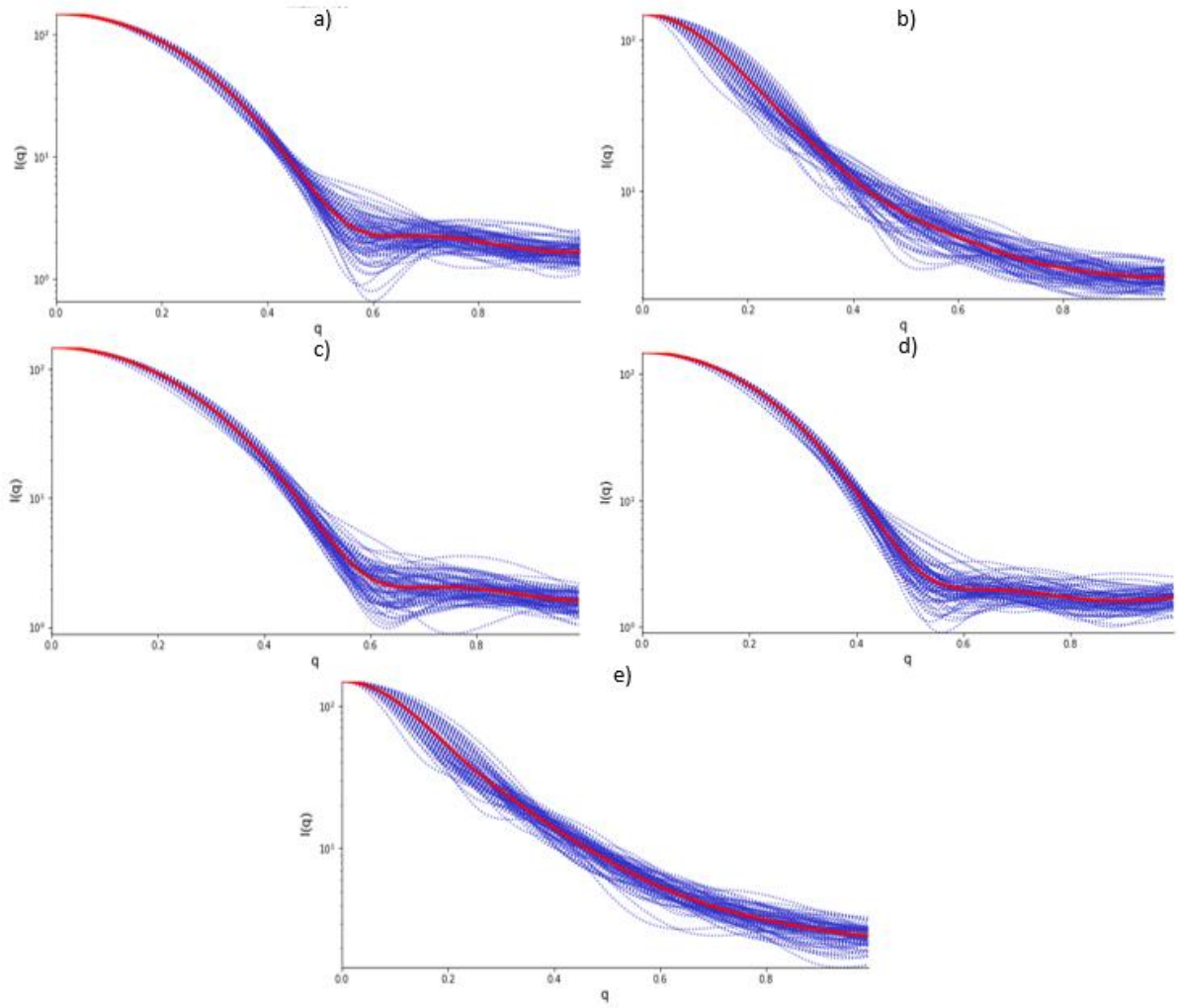

Figure 7. Logarithmic plot of the scattering intensity I(q) as a function of q. Each blue line represents one theoretical scattering profile obtained from the PDB-library using FoXS [24]. The thicker red line represents the average of these scattering profiles. The results are from: (a) the reference simulation, (b) the simulation with the LJ potential scaled down to a third, (c) the simulation with a decreased residue radius, (d) the simulation with an increased bond length, and (e) the simulation with high protein concentration.

A scattering profile for a globular protein has a specific feature, unlike the scattering profile for an IDP. Since an IDP have many different conformations, which all display a different scattering profile, the resulting average scattering curve is considerably smoother compared to the scattering curve for a globular protein [28]. Figure 8 show a characteristic scattering curve, a logarithmic plot of the scattering intensity I(q) as a function of q, for a globular protein, a partially folded protein and an IDP [7]. The IDP is in the figure referred to as a natively unfolded protein.

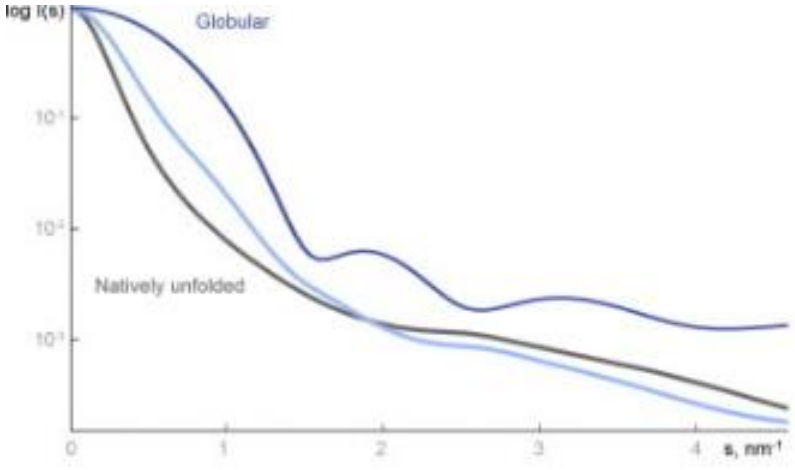

Figure 8. A characteristic scattering curve, a logarithmic plot of the scattering intensity I(q) as a function of q, for a globular protein, dark blue line, a partially folded protein, light blue line and a natively unfolded protein, black line. The momentum transfer, q, is in the Figure referred to as s [7].

The scattering profile generated in the reference simulation, see Figure 7(a), is not as smooth as the scattering profile should be for an IDP according to Figure 8, nor have the generated scattering profile from the simulation the specific features that a globular protein provides. This indicates that the simulated protein in the reference simulation behaves like a partially folded protein. Figure 7(c) and 7(d), representing the simulations with a decreased radius and an increased bond length, respectively, are similar to Figure 7 (a). From this, it can be assumed that the simulated protein in these two simulations also have the characteristics of a partially folded protein.

If the scattering profiles from the simulations with reduced LJ potential and high protein concentration, respectively, see Figure 7(b) and 7(e), are compared to the scattering profile for the natively unfolded protein in Figure 8, it can be determined that the simulated proteins in these two simulations resemble natively unfolded proteins or IDPs.

3.4 Kratky plot

The Kratky plots, $q^2I(q)$ as a function of q, generated from the simulations are shown in Figure 9. Similar to the scattering curves are the results from all theoretical scattering profiles obtained from the PDB-library displayed as a blue line in the Figure. The thicker red line represents the average of these scattering profiles.

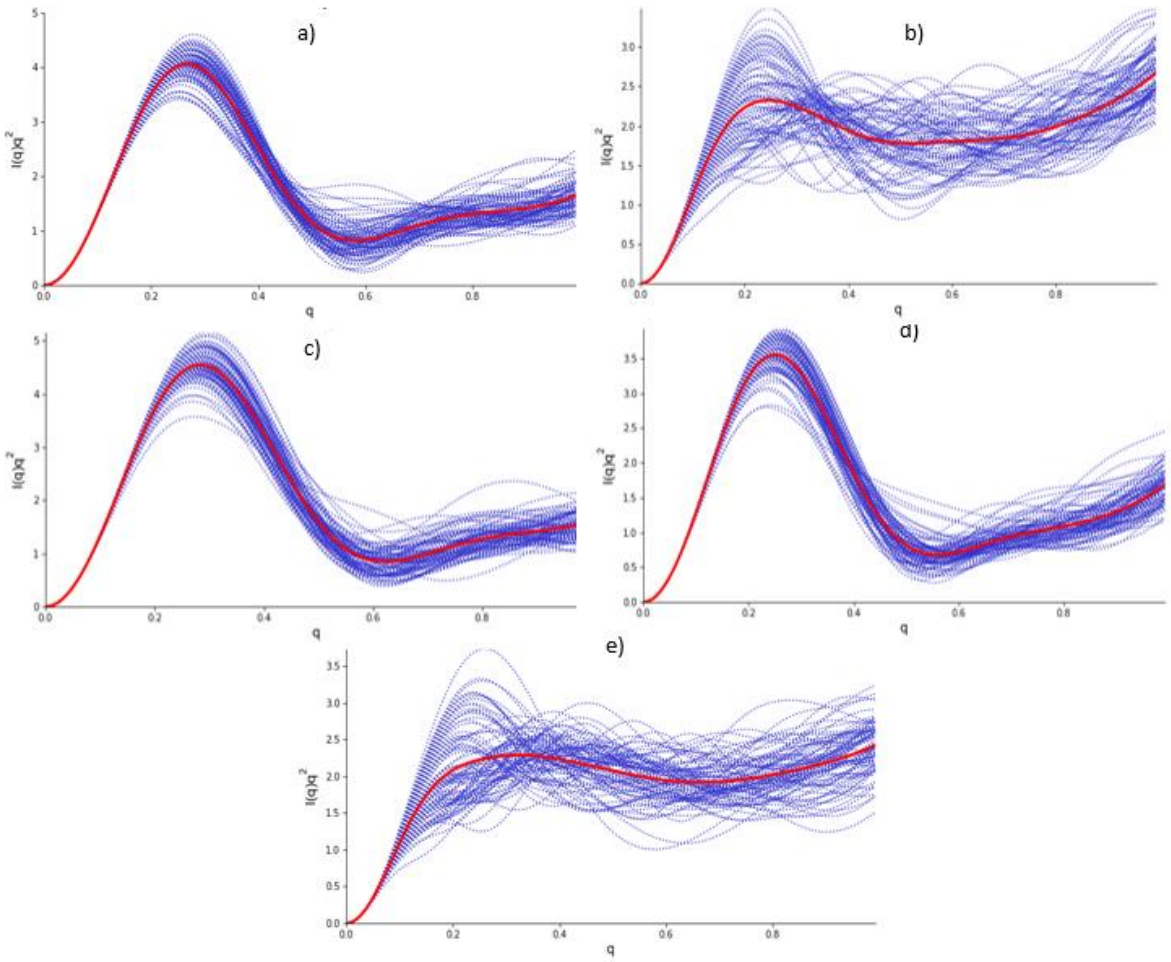

Figure 9. Kratky plot, $q^2I(q)$ as a function of *q. The blue lines represent all theoretical scattering profiles obtained from the PDB-library and the thicker red line represent the average of these scattering profiles. The results are from: (a) the reference simulation, (b) the simulation with the LJ potential scaled down to a third, (c) the simulation with a decreased residue radius, (d) the simulation with an increased bond length, and (e) the simulation with high protein concentration.*

The results from the simulations in this study are compared to an experimental Kratky plot for Hst5 generated by Cragnell et al, see Figure 10 [6] where a dimensionless Kratky plot used, i.e. $qR_g{}^2\frac{I(q)}{I(0)}$ is plotted as a function of qR_g. The black curve is the experimentally measured form factor. The white, blue and red curves show the results from simulations performed by Cragnell et al.

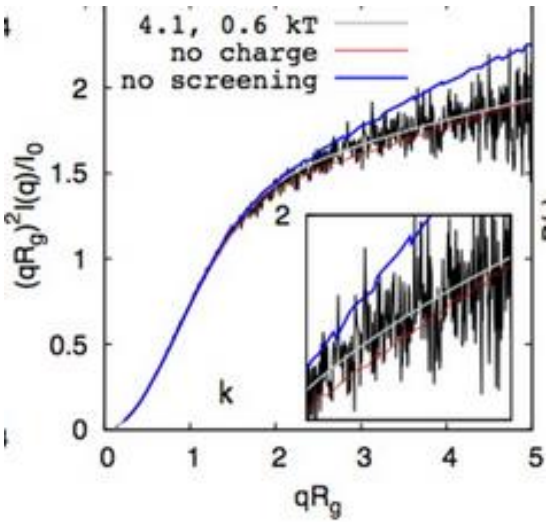

Figure 10. Experimental Kratky plot for Hst5 generated by Cragnell et al [6], where a dimensionless Kratky plot used, i.e., $qR_g{}^2\frac{I(q)}{I(0)}$ is plotted as a function of qR_g. The black curve is the experimentally measured form factor. The white, blue and red curves show the results from simulations performed by Cragnell et al.

A Kratky plot should have a plateau or an ascending curve at a high *q* for an IDP, as shown in Figure 10. The Kratky plot from the reference simulation, see Figure 9(a), has a high peak before the curve decreases to a low $q^2I(q)$-value. This is not comparable with Figure 10 and indicates that the simulated protein is not flexible. The generated Kratky plots for the simulation with a decreased radius, and the simulation with an increased bond length, see Figure 9(c) and 9(d), are almost identical with the Kratky plot generated by the reference simulation. This indicates that the simulated proteins in these two simulations are less flexible under physiological conditions than expected for an IDP. In Figure 9(b), representing the simulation with a reduced LJ potential, it can be seen that the Kratky plot has a plateau at a high *q* value, which indicates a flexible protein. However, the simulated protein is not a perfect IDP since there is a small peak at $q = 0.2$ which is not present in the experimental Kratky plot, see Figure 10. Figure 9(e), representing the simulation with a high protein concentration, shows that the generated Kratky plot gives a plateau at high *q*, but the generated Kratky plot is not identical to Figure 10, since there is still a small peak at $q = 0.2$ and the ascending curve is not as steep as in the experimental Kratky plot. This can be due to the simulated protein is not a perfect IDP.

3.5 Distance map

The distance maps obtained from each simulation are displayed in Figure 11. The colors represent the distance in nm between any two residues. What distance each color symbolizes can be obtained by using the color bar at the bottom of each distance map. Both the x-axis and the y-axis represent the residue index number, from 1 to 24.

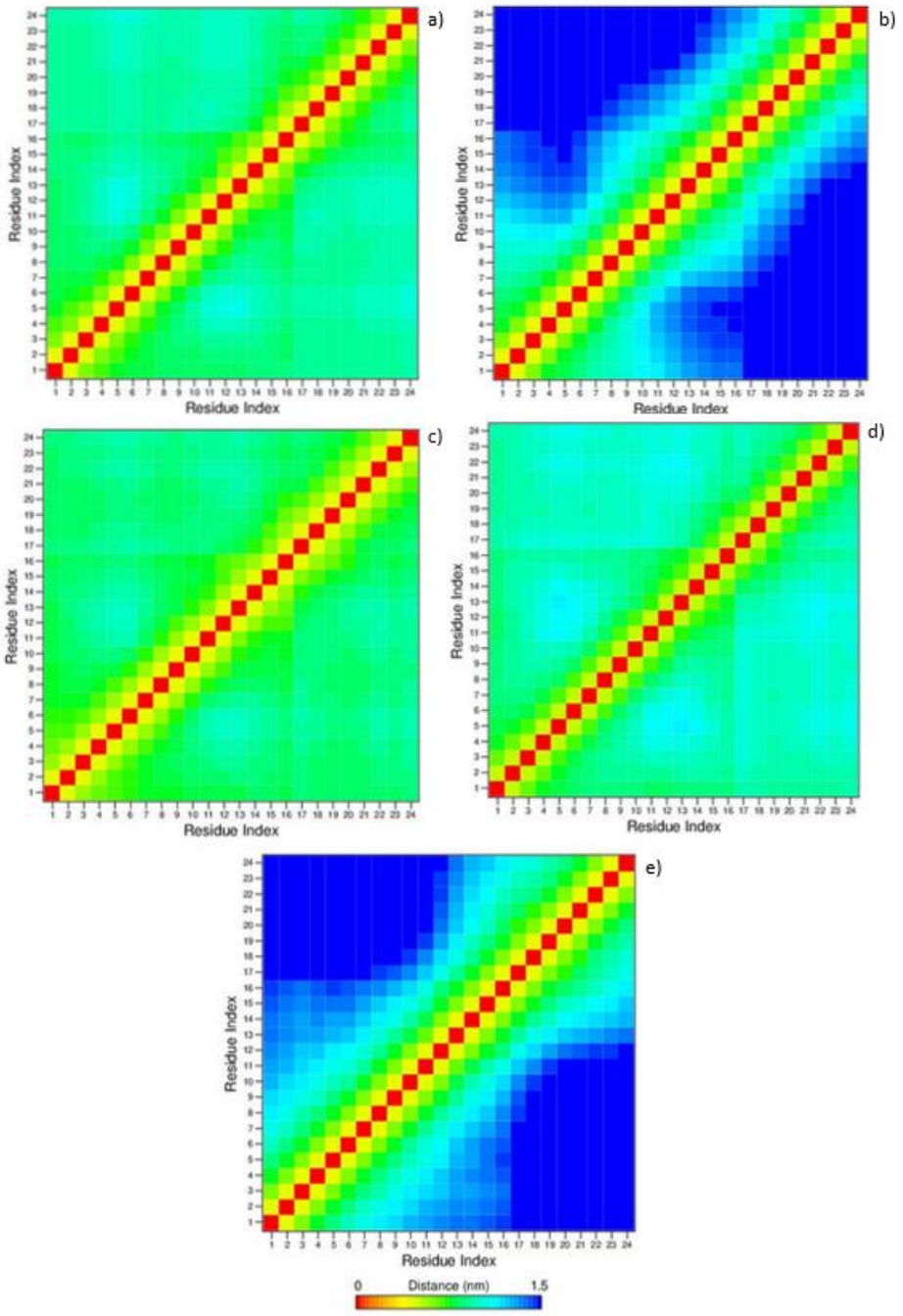

Figure 11. Distance map obtained from each simulation of Hst5. The colors represent the distance between any two residues in nm. The color bar at the bottom of the distance map can be used to determine what distance each color represents. The residue index number, from 1 to 24, is presented both on the x-axis and the y-axis. The results are from: (a) the reference simulation, (b) the simulation with the LJ potential scaled down to a third, (c) the simulation with a decreased residue radius, (d) the simulation with an increased bond length, and (e) the simulation with high protein concentration.

The distance maps produced from Hst5 in this study are compared to a protein contact map of Hst5 generated by Fagerberg et al, see Figure 12 [9]. The protein contact map shows that the contacts between residues are very local except for at the end of the chain, and also to a minor degree in the middle of the chain. The cut-off value in the study is set to 8 Å. The color bar at the side of the contact map is used to determine the contact between any two residues.

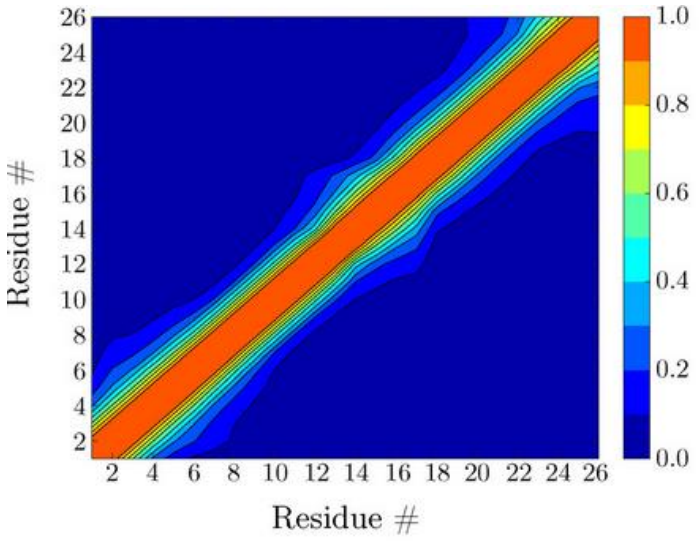

Figure 12. A protein contact map generated by Fagerberg et al [9]. The cut-off value in the study is set to 8 Å, and the color bar at the side of the contact map is used to determine the contact between any two residues.

From the generated distance map of the reference simulation, see Figure 11(a), it can be determined that the simulated Hst5 has some contact between all residues which can indicate that the protein has some flexibility but on the whole the protein is very stable. The result from the simulation indicates that Hst5 appears more as a globular protein than an IDP. This is not comparable to the contact map in Figure 12. The distance map for the simulation where the LJ potential is reduced, see Figure 11(b), show that the residues have least contact between residue 1-10, and most contact in the region from residue 10-24. This could indicate that the simulated protein is more flexible than the protein simulated in the reference simulation, but there are still some differences between the generated distance map and the protein contact map generated by Fagerberg et al.

The distance map generated from the simulation with a decrease in the radius of the residues, see Figure 11(c), is quite similar to the contact map generated from the reference simulation. However, some small differences could indicate that the simulated protein from the simulation with a decrease in the radius is even less flexible than the simulated protein in the reference simulation. The simulation with an increased bond length, see Figure 11(d), has greater distances between residues, especially at the ends of the chain, compared to the reference simulation. This means that the simulated Hst5 could have a higher flexibility and to some extent behaves more like an IDP than the Hst5 simulated in the reference simulation, although none of these two simulations are close to generate a distance map similar to the contact map generated by Fagerberg et al, see Figure 12.

Figure 11(e), represent the simulation with a high protein concentration, suggests that the simulated Hst5 is most flexible in the middle of the chain and least flexible at the end of the chain. This is similar to the contact map generated by Fagerberg et al.

4. Discussion

4.1 Improvement of the model

From the comparisons made of the results it can be determined which of the changes that are favorable, and how the model can be changed to be suitable for further studies of dynamics of IDPs. A reduction of the LJ potential generated results that better resemble an IDP than the simulated protein in the reference simulation. This is due to a lower LJ potential giving a decrease in the non-bonded interactions between residues, which will give the residues more freedom and flexibility, and thereby generate a simulation of a highly flexible protein that behaves similar to an IDP.

From the results it can be determined that an increase in bond length affects the protein to behave slightly more like an IDP than the simulated protein in the reference simulation did. The results also indicate that the protein changed less with an increased bond length than with a reduced LJ potential, and this may be due to the change of the LJ potential is greater than the change of the bond length. The LJ potential is scaled down to a third, while the bond length is increased from 4 Å to 5 Å, which is a lesser percentage change. However, increasing the bond length is a favorable change in the process of modifying the model to be suitable for further studies of IDPs.

It is difficult to draw a conclusion how the decrease of the radius of the residues affects the protein, because of the minor change made, only 0.4 Å. The results show that a decrease of the radius give a simulated protein that resemble a globular protein rather than an IDP. From the results it could be assumed that an increase of the size of the amino acids in a protein could be a favorable change to improve the model. Since a protein with larger residues does not necessarily need to become more flexible, is an increase of the radius not certain to be a positive change. Different residues contribute differently to the flexibility of the protein, some residues contribute a lot to the disorder while some residues do not contribute at all. How much a residue contribute to the disorder is not connected with the radius of the residue, but rather to the properties of the residue [29].

An increase of the protein concentration, from one single protein to 200 chains, was a positive change, and the results showed that the simulated Hst5 resembled an IDP. A higher protein concentration allows the proteins in the simulation to interact with each other. It can be hard to imitate reality when one single chain is simulated since then the protein is forced to only interact with itself, the counterions, and the implicit solvent. The result of such a simulation can either be that the protein becomes more compact or more flexible, depending on the characteristics of the simulated protein.

4.2 The effects of the deviations

The reference simulation was an attempt to reproduce the original model, but some changes have been made to provide a more stable system and because some of the algorithms used in the original model is not possible to use in GROMACS. Some of the deviations made can have affected the results, and gives an inaccurate view of the model's ability to be used for simulating dynamic events of IDPs. In the original model there are 500 IDP chains simulated in one simulation box, unlike the simulations in this study where a maximum of 200 IDP chains were simulated at the same time. From the results obtained in this study, it can be determined that a high concentration gives proteins that resembles IDPs. An even higher concentration, such as 500 IDP chains, would likely give even better results than a simulation of 200 IDP chains. Since a concentration that high has not been simulated in this study, it is hard to draw a conclusion how well the original model is suitable for simulating dynamics events of IDPs when studying systems with a high protein concentration.

Another thing that could have affected the results is the value of the mass. The value of mass is not mentioned in the original model so in this study a value of 126.5 Da is used since that is the average

mass of the amino acids in Hst5. In the original model is the friction factor calculated by the following equation:

$$Friction\ factor = \frac{0.1m}{\sqrt{\frac{ma^2}{\varepsilon}}} \quad (12)$$

where *m* is the mass, *ε* is the well depth, and *a* is the LJ interaction range [1]. In this study the same value for the mass was used for all residues in all simulations, and therefore is a fixed value for the friction factor of 5.5 Da/ps is used throughout the study. This value is calculated from equation 12 with a mass of 126.5 Da. The friction factor is affecting the kinetic properties of the protein, a high friction factor means that the protein gets a reduced kinetic energy and a decrease in the velocity for each residue. With a decreased velocity the movement of the protein will be restricted, and the simulated protein will be less flexible. It is possible that another value of the mass has been used in the original model than in this study, and therefore also a different friction factor.

5. Conclusions

The first aim of this study was to investigate if the coarse-grained model primary described by Das et al is suitable to use for analyzing the structure and dynamics of IDPs. The model protein used in this study is Hst5. The original model was re-implemented using GROMACS to run the simulations, and evaluated by comparing theoretical produced SAXS data and a distance map, with experimental data and a contact map produced for Hst5 in previous studies. The results from the reference simulation, indicates that the simulated Hst5 resembles a globular protein rather than an IDP that would have been the desired result. R_g has a value of 6.18 ± 0.27 Å and this result is approximately half of the R_g produced in previous studies, which indicates that the simulated Hst5 in this study is more compact and less flexible than the protein is in reality. The Kratky plot as well as the distance map generated in this study also shows that the simulated protein behaves like a globular protein rather than an IDP. From these results the conclusion can be drawn that the original model is not suitable to use for examining dynamic events of an IDP. Worth mentioning is that in this study some deviations have been made from the original model that could have affected the results and thus the conclusion. For instance, the high protein concentration used in the original model was not simulated in this study. In order to be able to draw a conclusion about the suitability of the model with certainty, additional simulations with higher protein concentrations have to be performed.

The second aim of this study was to respond to the questions: How is the conformation and flexibility of an IDP affected if some parameters changed from the original model? If it turns out that the model is inappropriate to use for simulating the dynamic events of an IDP, what other simulations and tests could be carried out to continue this study to improve the model?

To begin with the first question: the changes that have been performed include a reduction in the LJ potential, a decrease of the radius of the residues, an increase of the bond length, and finally a simulation with a higher protein concentration. Each change has led to an improvement of the results, and the simulated protein has behaved more like an IDP than a globular protein, except the simulation where the radius of the residues was decreased. In the simulation with a decreased radius, the results show that the simulated Hst5 became even less flexible than in the reference simulation.

In response to the second question, there are several tests and simulations that could be performed to continue this study. For instance, simulations could be performed where counterions are excluded, where angular potential has been included, or a simulation were an N- and C-terminus have been added to the protein. This study is only a first step in the process of finding a suitable model to be able to examine both the structure and the dynamics of IDPs. The work in this study has to be continued in order to find an appropriate model for this purpose.

6. Future aspects

There are several different simulations that could be performed to continue this study and the evaluation of the model. For a start it could be interesting to perform a simulation where counterions are excluded to examine if there will be any differences in the properties of the protein. With this simulation it can be determined whether the counterions affect the protein's characteristics or not. Another simulation based on this study could be to reduce the LJ potential and in the same simulation increase the protein concentration. The results from this study indicates that these two changes had the greatest effect on the flexibility of the protein. If these two changes were combined, the model may be able to simulate the dynamics events of IDPs. Another sensible change could be to perform a simulation where angular potentials between residues are included. The angular potential would prevent the residues to have excessive movements around their own bonds. This would prevent the protein to fold too compactly and the protein would likely become more flexible. Finally, a simulation where a N- and C-terminus is inserted at the end of the protein chain could be performed. Both the N- and C-terminus have an elementary charge, and it could be interesting to perform this change to analyze how this will affect the simulated protein.

References

[1] Das S, Amin AN, Lin YH, Chan HS. Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. *Phys. Chem. Chem. Phys.* 2018; 20: 28558-28574.
DOI: https://doi.org/10.1039/C8CP05095C

[2] Puri S, Edgerton M. How Does It Kill?: Understanding the Candidacidal Mechanism of Salivary Histatin 5. *Eukaryotic Cell*. 2014 Jul; 13 (8): 958-964.
DOI: 10.1128/EC.00095-14

[3] M.J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, GROMACS User Manual version 5.0.7, www.gromacs.org (2015)

[4] Uversky VN. Intrinsically disordered proteins from A to Z. *Int J. Biochem. Cell Biol*. 2011; 43: 1090-1103.
DOI: https://doi.org/10.1016/j.biocel.2011.04.001

[5] Tompa P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci.* 2012; 37: 509-516.
DOI: https://doi.org/10.1016/j.tibs.2012.08.004

[6] Cragnell C, Durand D, Cabane B, Skepö M. Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS. *Proteins.* 2016 Jun; 84(6): 777-91.
DOI: 10.1002/prot.25025.

[7] Kikhney A, Svergun D. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters.* 2015; 589 (19A): 2570-2577.
DOI: https://doi.org/10.1016/j.febslet.2015.08.027.

[8] Cragnell C, Rieloff E, Skepö M. Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensamble of Intrinsically Disordered Proteins and Regions. *Journal of Molecular Biology*. 2018; 430(16): 2478-2492.
DOI: https://doi.org/10.1016/j.jmb.2018.03.006.

[9] Fagerberg E, Lenton S, Skepö M. Evaluating Models of Varying Complexity of Crowded Intrinsically Disordered Protein Solutions Against SAXS. *Journal of chemical Theory and Computation.* 2019; 15 (12): 6968-6983.
DOI: 10.1021/acs.jctc.9b00723

[10] Fagerberg E, Lenton S, Skepö M. Evaluating Models of Varying Complexity of Crowded Intrinsically Disordered Protein Solutions Against SAXS. *Journal of Chemical Theory and Computation.* 2019; 15(12): 6968-6983.
DOI: 10.1021/acs.jctc.9b00723

[11] Rieloff E, Skepö M. Determining R_g of IDPs from SAXS Data. *Methods in Molecular Biology*. 2020; 2141: 271-283.
DOI: https://doi.org/10.1007/978-1-0716-0524-0_13

[12] Lemkul JA. From proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package, v1.0. *Living J. Comp. Mol. Sci.* 2018. In Press.

[13] Nielsen S, Toft K, Snakenborg D, Jeppesen MG, Jacobsen J, Vestergaard B, et al. BioXTAS RAW, a software program for high-throughput automated small-angle X-ray scattering data reduction and preliminary analysis. *J. Appl. Cryst.* 2009; 42 (5): 959-964.
DOI: 10.1107/S0021889809023863

[14] Structural Molecular Biology. Initial Analysis and Quality Assessment of Solution Scattering Data [Internet]. Stanford: Stanford University; 2017 [updated 2017 Feb 01; cited 2020 Jan 21]. Available from: https://www-ssrl.slac.stanford.edu/~saxs/analysis/assessment.htm

[15] Emerson IA, Amala A. Protein contact maps: A binary depiction of protein 3D structures. *Physica A: Statistical Mechanis and its Applications.* 2017; 465: 782-791.
DOI: https://doi.org/10.1016/j.physa.2016.08.033.

[16] Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Folding and Design.* 1997; 2 (5): 295-306.
DOI: https://doi.org/10.1016/S1359-0278(97)00041-2.

[17] O'Neill B. Chapter 3 – Euclidean Geometry. *Elementary Differential Geometry.* 2006; 2: 100-129.
DOI: https://doi.org/10.1016/B978-0-12-088735-4.50007-9.

[18] Bartoli L, Capriotti E, Fariselli P, Martelli PL, Casadio R. The Pros and Cons of Predicting Protein Contact Maps. *Protein Structure Prediction.* 2008; 413: 199-217.
DOI: https://doi.org/10.1007/978-1-59745-574-9_8

[19] NIH center for Macromolecular Modeling & Bioinformatics. VMD Visual Molecular Dynamics [Internet]. Illinois: University of Illinois at Urbana-Champaign; 2020 [updated 2020 Nov 24; cited 2020 Dec 14]. Available from: https://www.ks.uiuc.edu/Research/vmd/

[20] Malmberg CG, Maryott AA. Dielectric Constant of Water from 0° to 100° C. *J. Res. Nat. Bureau of Standards.* 1956 Jan; 56: 1-8.

[21] Luty BA, van Gunsteren WF. Calculating Electrostatic Interactions Using the Particle-Particle Particle-Mesh Method with Nonperiodic Long-Range Interactions. *The journal of Physical Chemistry.* 1996; 100 (7): 2581-2587.
DOI: https://doi.org/10.1021/jp9518623. DOI: 10.1021/jp9518623

[22] Shang X, Kröger M. Time Correlation Functions of Equilibrium and Nonequilibrium Langevin Dynamics: Derivations and Numerics Using Random Numbers. *SIAM Rev.* 2020; 62: 901-935.
DOI: 10.1137/19M1255471

[23] Andersen HC. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* 1980; 72: 2384.
DOI: https://doi.org/10.1063/1.439486

[24] Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophysical Journal.* 2013; 105(4): 962-974.
DOI: 10.1016/j.bpj.2013.07.020

[25] Alberti S, Gladfelter A, Mittag T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell*. 2019 Jan 24; 176 (3): 419-434.
DOI: 10.1016/j.cell.2018.12.035.

[26] Han M, Chen P, Yang X. Molecular dynamics simulation of PAMAM dendrimer in aqueous solution. *Polymer.* 2005; 46(10): 3481-3488.
DOI: https://doi.org/10.1016/j.polymer.2005.02.107

[27] Corporate Finance Institute. Autocorrelation [Internet]. CFI Education Inc; 2015 [updated 2021; cited 2021 Jan 20]. Available from:
https://corporatefinanceinstitute.com/resources/knowledge/other/autocorrelation/

[28] Receveur-Bréchot V, Durand D. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr Protein Pept Sci*. 2012; 13(1): 55-75.
DOI:10.2174/138920312799277901

[29] Theillet FX, Kalmar L, Topma P, Han KH, Selenko P, Dunker AK, et al. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disord Proteins*. 2013; 1(1) :e24360.
DOI: 10.4161/idp.24360

Appendices

Appendix 1 (i-v)

Topology files. One topology file for each simulation.

Reference simulation:

[defaults]

; nbfunc	comb-rule	gen-pairs	fudgeLJ	fudgeQQ
1	1	no	1.0	1.0

[atomtypes]

;Name	bond_type	mass (u)	charge	ptype	V(c6)	W(c12)
ASP		126.5	-1	A	8.8318e-02	4.6907e-04
SER		126.5	0	A	8.8318e-02	4.6907e-04
HIS		126.5	0	A	8.8318e-02	4.6907e-04
ALA		126.5	0	A	8.8318e-02	4.6907e-04
LYS		126.5	1	A	8.8318e-02	4.6907e-04
ARG		126.5	1	A	8.8318e-02	4.6907e-04
GLY		126.5	0	A	8.8318e-02	4.6907e-04
TYR		126.5	0	A	8.8318e-02	4.6907e-04
PHE		126.5	0	A	8.8318e-02	4.6907e-04
GLU		126.5	-1	A	8.8318e-02	4.6907e-04
Cl		126.5	-1	A	8.8318e-02	4.6907e-04

[nonbond_params]

;i	j	func	V(c6)	W(c12)
ASP	SER	1	8.8318e-02	4.6907e-04
ASP	HIS	1	8.8318e-02	4.6907e-04
ASP	ALA	1	8.8318e-02	4.6907e-04
ASP	LYS	1	8.8318e-02	4.6907e-04
ASP	ARG	1	8.8318e-02	4.6907e-04
ASP	GLY	1	8.8318e-02	4.6907e-04
ASP	TYR	1	8.8318e-02	4.6907e-04
ASP	PHE	1	8.8318e-02	4.6907e-04
ASP	GLU	1	8.8318e-02	4.6907e-04
ASP	ASP	1	8.8318e-02	4.6907e-04
SER	SER	1	8.8318e-02	4.6907e-04
SER	HIS	1	8.8318e-02	4.6907e-04
SER	ALA	1	8.8318e-02	4.6907e-04
SER	LYS	1	8.8318e-02	4.6907e-04
SER	ARG	1	8.8318e-02	4.6907e-04
SER	GLY	1	8.8318e-02	4.6907e-04
SER	TYR	1	8.8318e-02	4.6907e-04
SER	PHE	1	8.8318e-02	4.6907e-04
SER	GLU	1	8.8318e-02	4.6907e-04
HIS	HIS	1	8.8318e-02	4.6907e-04
HIS	ALA	1	8.8318e-02	4.6907e-04
HIS	LYS	1	8.8318e-02	4.6907e-04

HIS	ARG	1	8.8318e-02	4.6907e-04
HIS	GLY	1	8.8318e-02	4.6907e-04
HIS	TYR	1	8.8318e-02	4.6907e-04
HIS	PHE	1	8.8318e-02	4.6907e-04
HIS	GLU	1	8.8318e-02	4.6907e-04
ALA	ALA	1	8.8318e-02	4.6907e-04
ALA	LYS	1	8.8318e-02	4.6907e-04
ALA	ARG	1	8.8318e-02	4.6907e-04
ALA	GLY	1	8.8318e-02	4.6907e-04
ALA	TYR	1	8.8318e-02	4.6907e-04
ALA	PHE	1	8.8318e-02	4.6907e-04
ALA	GLU	1	8.8318e-02	4.6907e-04
LYS	LYS	1	8.8318e-02	4.6907e-04
LYS	ARG	1	8.8318e-02	4.6907e-04
LYS	GLY	1	8.8318e-02	4.6907e-04
LYS	TYR	1	8.8318e-02	4.6907e-04
LYS	PHE	1	8.8318e-02	4.6907e-04
LYS	GLU	1	8.8318e-02	4.6907e-04
ARG	ARG	1	8.8318e-02	4.6907e-04
ARG	GLY	1	8.8318e-02	4.6907e-04
ARG	TYR	1	8.8318e-02	4.6907e-04
ARG	PHE	1	8.8318e-02	4.6907e-04
ARG	GLU	1	8.8318e-02	4.6907e-04
GLY	GLY	1	8.8318e-02	4.6907e-04
GLY	TYR	1	8.8318e-02	4.6907e-04
GLY	PHE	1	8.8318e-02	4.6907e-04
GLY	GLU	1	8.8318e-02	4.6907e-04
TYR	TYR	1	8.8318e-02	4.6907e-04
TYR	PHE	1	8.8318e-02	4.6907e-04
TYR	GLU	1	8.8318e-02	4.6907e-04
PHE	PHE	1	8.8318e-02	4.6907e-04
PHE	GLU	1	8.8318e-02	4.6907e-04
GLU	GLU	1	8.8318e-02	4.6907e-04
Cl	Cl	1	8.8318e-02	4.6907e-04
Cl	GLU	1	8.8318e-02	4.6907e-04
Cl	PHE	1	8.8318e-02	4.6907e-04
Cl	TYR	1	8.8318e-02	4.6907e-04
Cl	GLY	1	8.8318e-02	4.6907e-04
Cl	ARG	1	8.8318e-02	4.6907e-04
Cl	LYS	1	8.8318e-02	4.6907e-04
Cl	ALA	1	8.8318e-02	4.6907e-04
Cl	HIS	1	8.8318e-02	4.6907e-04
Cl	SER	1	8.8318e-02	4.6907e-04
Cl	ASP	1	8.8318e-02	4.6907e-04

[bondtypes]

; i	j	func	b0	kb
ASP	SER	1	0.418	1.79e+06

ASP	HIS	1	0.418	1.79e+06
ASP	ALA	1	0.418	1.79e+06
ASP	LYS	1	0.418	1.79e+06
ASP	ARG	1	0.418	1.79e+06
ASP	GLY	1	0.418	1.79e+06
ASP	TYR	1	0.418	1.79e+06
ASP	PHE	1	0.418	1.79e+06
ASP	GLU	1	0.418	1.79e+06
ASP	ASP	1	0.418	1.79e+06
SER	SER	1	0.418	1.79e+06
SER	HIS	1	0.418	1.79e+06
SER	ALA	1	0.418	1.79e+06
SER	LYS	1	0.418	1.79e+06
SER	ARG	1	0.418	1.79e+06
SER	GLY	1	0.418	1.79e+06
SER	TYR	1	0.418	1.79e+06
SER	PHE	1	0.418	1.79e+06
SER	GLU	1	0.418	1.79e+06
HIS	HIS	1	0.418	1.79e+06
HIS	ALA	1	0.418	1.79e+06
HIS	LYS	1	0.418	1.79e+06
HIS	ARG	1	0.418	1.79e+06
HIS	GLY	1	0.418	1.79e+06
HIS	TYR	1	0.418	1.79e+06
HIS	PHE	1	0.418	1.79e+06
HIS	GLU	1	0.418	1.79e+06
ALA	ALA	1	0.418	1.79e+06
ALA	LYS	1	0.418	1.79e+06
ALA	ARG	1	0.418	1.79e+06
ALA	GLY	1	0.418	1.79e+06
ALA	TYR	1	0.418	1.79e+06
ALA	PHE	1	0.418	1.79e+06
ALA	GLU	1	0.418	1.79e+06
LYS	LYS	1	0.418	1.79e+06
LYS	ARG	1	0.418	1.79e+06
LYS	GLY	1	0.418	1.79e+06
LYS	TYR	1	0.418	1.79e+06
LYS	PHE	1	0.418	1.79e+06
LYS	GLU	1	0.418	1.79e+06
ARG	ARG	1	0.418	1.79e+06
ARG	GLY	1	0.418	1.79e+06
ARG	TYR	1	0.418	1.79e+06
ARG	PHE	1	0.418	1.79e+06
ARG	GLU	1	0.418	1.79e+06
GLY	GLY	1	0.418	1.79e+06
GLY	TYR	1	0.418	1.79e+06
GLY	PHE	1	0.418	1.79e+06
GLY	GLU	1	0.418	1.79e+06

TYR	TYR	1	0.418	1.79e+06
TYR	PHE	1	0.418	1.79e+06
TYR	GLU	1	0.418	1.79e+06
PHE	PHE	1	0.418	1.79e+06
PHE	GLU	1	0.418	1.79e+06
GLU	GLU	1	0.418	1.79e+06

[moleculetype]
; name nrexcl
Hist5 1

[atoms]

;nr	type	resnr	residu	atom	cgnr	charge(q)	m (u)
1	ASP	1	Hist5	C01	1	-1	126.5
2	SER	2	Hist5	C02	2	0	126.5
3	HIS	3	Hist5	C03	3	0	126.5
4	ALA	4	Hist5	C04	4	0	126.5
5	LYS	5	Hist5	C05	5	1	126.5
6	ARG	6	Hist5	C06	6	1	126.5
7	HIS	7	Hist5	C07	7	0	126.5
8	HIS	8	Hist5	C08	8	0	126.5
9	GLY	9	Hist5	C09	9	0	126.5
10	TYR	10	Hist5	C10	10	0	126.5
11	LYS	11	Hist5	C11	11	1	126.5
12	ARG	12	Hist5	C12	12	1	126.5
13	LYS	13	Hist5	C13	13	1	126.5
14	PHE	14	Hist5	C14	14	0	126.5
15	HIS	15	Hist5	C15	15	0	126.5
16	GLU	16	Hist5	C16	16	-1	126.5
17	LYS	17	Hist5	C17	17	1	126.5
18	HIS	18	Hist5	C18	18	0	126.5
19	HIS	19	Hist5	C19	19	0	126.5
20	SER	20	Hist5	C20	20	0	126.5
21	HIS	21	Hist5	C21	21	0	126.5
22	ARG	22	Hist5	C22	22	1	126.5
23	GLY	23	Hist5	C23	23	0	126.5
24	TYR	24	Hist5	C24	24	0	126.5

[bonds]

;	ai	aj
	1	2
	2	3
	3	4
	4	5
	5	6
	6	7
	7	8
	8	9

9	10
10	11
11	12
12	13
13	14
14	15
15	16
16	17
17	18
18	19
19	20
20	21
21	22
22	23
23	24

[system]
Histatin5 in Water

[molecules]
Hist5 1
Cl 5

End of the topology file for the reference simulation.

For the other simulations with changes of the parameters the topology file is identical to the topology file for the reference simulation except for some small differences.

For the simulation with the LJ potential scaled down to a third, the values for V and W changed to:

V(c6)	W(c12)
2.9439e-02	1.5636e-04

For the simulation with a decrease in the radius of the residues, the values for V and W changed to:

V(c6)	W(c12)
4.9911e-02	1.4980e-04

For the simulation with an increase in the bond length is the value of b0 changed to:
b0 = 0.500

For the simulation with an increase the protein concentration is the value for the number of chains and counter ions changed to:

[molecules]
Hist5 200
Cl 1000

Appendix 2 (v-vii)

Structure files.

Structure file for all simulations except the simulation with decreased radius of the residues and the simulation with an increased bond length:

```
Histatin 5
  24
  1          HIST5          C01    1  0.000   0.000   0.000
  2          HIST5          C02    2  0.418   0.000   0.000
  3          HIST5          C03    3  0.835   0.000   0.000
  4          HIST5          C04    4  1.253   0.000   0.000
  5          HIST5          C05    5  1.671   0.000   0.000
  6          HIST5          C06    6  2.089   0.000   0.000
  7          HIST5          C07    7  2.506   0.000   0.000
  8          HIST5          C08    8  2.924   0.000   0.000
  9          HIST5          C09    9  3.342   0.000   0.000
 10          HIST5          C10   10  3.759   0.000   0.000
 11          HIST5          C11   11  4.177   0.000   0.000
 12          HIST5          C12   12  4.595   0.000   0.000
 13          HIST5          C13   13  5.012   0.000   0.000
 14          HIST5          C14   14  5.430   0.000   0.000
 15          HIST5          C15   15  5.848   0.000   0.000
 16          HIST5          C16   16  6.266   0.000   0.000
 17          HIST5          C17   17  6.683   0.000   0.000
 18          HIST5          C18   18  7.101   0.000   0.000
 19          HIST5          C19   19  7.519   0.000   0.000
 20          HIST5          C20   20  7.936   0.000   0.000
 21          HIST5          C21   21  8.354   0.000   0.000
 22          HIST5          C22   22  8.772   0.000   0.000
 23          HIST5          C23   23  9.189   0.000   0.000
 24          HIST5          C24   24  9.607   0.000   0.000
 29.23900   29.23900   29.23900
```

The simulation with a decreased radius of the residues have the same structure file as the reference simulation except for the box size, the last three numbers:

```
 26.58600   26.58600   26.58600
```

The structure file for the simulation with an increased bond length:

```
Histatin 5
  24
  1          HIST5          C01    1  0.000   0.000   0.000
  2          HIST5          C02    2  0.500   0.000   0.000
  3          HIST5          C03    3  1.000   0.000   0.000
  4          HIST5          C04    4  1.500   0.000   0.000
  5          HIST5          C05    5  2.000   0.000   0.000
  6          HIST5          C06    6  2.500   0.000   0.000
  7          HIST5          C07    7  3.000   0.000   0.000
  8          HIST5          C08    8  3.500   0.000   0.000
```

9	HIST5	C09	9	4.000	0.000	0.000
10	HIST5	C10	10	4.500	0.000	0.000
11	HIST5	C11	11	5.000	0.000	0.000
12	HIST5	C12	12	5.500	0.000	0.000
13	HIST5	C13	13	6.000	0.000	0.000
14	HIST5	C14	14	6.500	0.000	0.000
15	HIST5	C15	15	7.000	0.000	0.000
16	HIST5	C16	16	7.500	0.000	0.000
17	HIST5	C17	17	8.000	0.000	0.000
18	HIST5	C18	18	8.500	0.000	0.000
19	HIST5	C19	19	9.000	0.000	0.000
20	HIST5	C20	20	9.500	0.000	0.000
21	HIST5	C21	21	10.00	0.000	0.000
22	HIST5	C22	22	10.50	0.000	0.000
23	HIST5	C23	23	11.00	0.000	0.000
24	HIST5	C24	24	11.50	0.000	0.000
29.23900	29.23900	29.23900				

Appendix 3 (viii-xii)

Mdp-files. The mdp-files are the same for each simulation except for the simulation when the radius of the residues was decreased. The cut-off for that simulation was set to 2.28.

Mdp-file for the energy-minimization step:

```
integrator      =   steep       ;   Algorithm   (steep=steepest   descent minimization). For EM.
emtol           =   -1          ; Stop minimization when the maximum force < -1 [kJ/mol/nm]
emstep          =   0.001       ; Energy step size [nm]
nsteps          =   200000      ; Maximum number of (minimization) steps to perform

; Neighbor searching
cutoff-scheme   =   Verlet      ; Buffered neighbor searching
nstlist         =   10          ; Frequency to update the neighbor list
ns-type         =   grid        ; Make a grid in the box and only check atoms in neighboring grid cells
pbc             =   xyz         ; Periodic boundary conditions in all directions

;Electrostatic
coulombtype     =   P3M-AD      ; Particle-Particle Particle-Mesh algorithm.
rcoulomb        =   2.5         ; =6a. Distance for the Coulomb cut-off. Short range cut-off.
rvdw            =   2.5         ; =6a. Distance for the Lennard-Jones or cut-off
fourierspacing  =   0.15        ; grid spacing for FFT. Spacing for a PPPM FFT grid
epsilon-r       =   80          ; Dielectric constant.
```

The mdp-file for the first temperature equilibrium step:

```
;Run parameters
integrator      =   sd          ; Leap-frog stochastic dynamics integrator
dt              =   0.001       ; Time step
nsteps          =   5000000     ; 5 ns.

;Output control
nstxout         =   5000        ; Save coordinates every 1.0 ps
nstvout         =   5000        ; Save velocities every 1.0 ps
nstenergy       =   5000        ; Save energies every 1.0 ps
nstlog          =   5000        ; Update log file every 1.0 ps

; Neighbor searching
cutoff-scheme   =   Verlet      ; Buffered neighbor searching
nstlist         =   10          ; Frequency to update the neighbor list
ns-type         =   grid        ; Make a grid in the box and only check atoms in neighboring grid cells
pbc             =   xyz         ; Periodic boundary conditions in all directions

;Electrostatic
coulombtype     =   P3M-AD      ; Particle-Particle Particle-Mesh algorithm.
rcoulomb        =   2.5         ; =6a. Distance for the Coulomb cut-off. Short range cut-off.
```

```
rvdw                =       2.5             ; =6a. Distance for the Lennard-Jones or cut-off
fourierspacing      =       0.33            ; grid spacing for FFT. Spacing for a PPPM FFT grid.
epsilon-r           =       122.2           ; Dielectric constant.
pme-order           =       4
DispCorr            =       Ener

; Pressure coupling is off
pcoupl              =       no               ; no pressure coupling in NVT

;Initial velocities
gen-vel             =       yes             ; Generate velocities
gen-temp            =       200

;Langevin dynamics
bd-fric             =       5.5             ; friction coefficient
ld-seed             =       -1

;Temperature coupling is Langevin dynamics.
tc-grps             =       System          ; groups to couple to separate temperature baths
tau-t               =       2.3             ; Time constant [ps]. 2.3
ref-t               =       200             ; Temperature [K].

;Bond parameters
continuation        =       no              ; Apply constraints to the start configuration
constraints         =       none            ; No constraints except those defined in the top.file
morse               =       no              ; Harmonic potential
```

The mdp-file for the second temperature equilibrium step:

```
;Run parameters
integrator          =       sd              ; Leap-frog stochastic dynamics integrator.
dt                  =       0.001           ; Time step
nsteps              =       5000000         ; 5 ns.

;Output control
nstxout             =       5000            ; Save coordinates every 1.0 ps
nstvout             =       5000            ; Save velocities every 1.0 ps
nstenergy           =       5000            ; Save energies every 1.0 ps
nstlog              =       5000            ; Update log file every 1.0 ps

; Neighbor searching
cutoff-scheme       =       Verlet          ; Buffered neighbor searching
nstlist             =       10              ; Frequency to update the neighbor list
ns-type             =       grid            ; Make a grid in the box and only check atoms in
neighboring grid cells
pbc                 =       xyz             ; Periodic boundary conditions in all directions

;Electrostatic
coulombtype         =       P3M-AD          ; Particle-Particle Particle-Mesh algorithm.
rcoulomb            =       2.5             ; =6a. Distance for the Coulomb cut-off. Short
range cut-off.
rvdw                =       2.5             ; =6a. Distance for the Lennard-Jones or cut-off
fourierspacing      =       0.33            ; grid spacing for FFT.
```

```
epsilon-r       =       97.4            ; Dielectric constant.
pme-order       =       4
DispCorr        =       Ener

; Pressure coupling is off
pcoupl          =       no              ; no pressure coupling in NVT

;Langevin dynamics
bd-fric         =       5.5             ; friction coefficient
ld-seed         =       -1

;Temperature coupling is Langevin dynamics.
tc-grps         =       System          ; groups to couple to separate temperature baths
tau-t           =       2.3             ; Time constant [ps].
ref-t           =       250             ; Temperature [K].

;Bond parameters
continuation    =       yes             ; Apply constraints to the start configuration
constraints     =       none            ;No constraints except those defined in the top.file
morse           =       no              ;       Harmonic        potential
```

The mdp-file for the third temperature equilibrium step:

```
;Run parameters
integrator      =       sd              ; Leap-frog stochastic dynamics integrator.
dt              =       0.001           ;Time step
nsteps          =       20000000        ; 20 ns.

;Output control
nstxout         =       5000            ; Save coordinates every 1.0 ps
nstvout         =       5000            ; Save velocities every 1.0 ps
nstenergy       =       5000            ; Save energies every 1.0 ps
nstlog          =       5000            ; Update log file every 1.0 ps

; Neighbor searching
cutoff-scheme   =       Verlet          ; Buffered neighbor searching
nstlist         =       10              ; Frequency to update the neighbor list
ns-type         =       grid            ; Make a grid in the box and only check atoms in
neighboring grid cells
pbc             =       xyz             ; Periodic boundary conditions in all directions

;Electrostatic
coulombtype     =       P3M-AD          ; Particle-Particle Particle-Mesh algorithm.
rcoulomb        =       2.5             ; =6a. Distance for the Coulomb cut-off. Short
range cut-off.
rvdw            =       2.5             ; =6a. Distance for the Lennard-Jones or cut-off
fourierspacing  =       0.33            ; grid spacing for FFT
epsilon-r       =       77.6            ; Dielectric constant.
pme-order       =       4
DispCorr        =       Ener

; Pressure coupling is off
pcoupl          =       no              ; no pressure coupling in NVT
```

```
;Langevin dynamics
bd-fric             =   5.5             ; friction coefficient
ld-seed             =   -1

;Temperature coupling is Langevin dynamics.
tc-grps             =   System          ; groups to couple to separate temperature baths
tau-t               =   2.3             ; Time constant [ps].
ref-t               =   300             ; Temperature [K].

;Bond parameters
continuation        =   yes             ; Apply constraints to the start configuration
constraints         =   none            ;No constraints except those defined in the top.file
morse               =   no              ; Harmonic potential
```

The mdp-file for the production run:

```
;Run parameters
integrator          =   sd              ; Leap-frog stochastic dynamics integrator.
dt                  =   0.001           ; Time step
nsteps              =   40000000        ; 40 ns.

;Output control
nstxout             =   0               ; suppress bulky .trr file by specifying
nstvout             =   0               ; 0 for output frequency of nstxout,
nstfout             =   0               ; nstvout, and nstfout
nstenergy           =   5000            ; Save energies every 1.0 ps
nstlog              =   5000            ; Update log file every 1.0 ps
nstxout-compressed=     5000            ; save compressed coordinates every 10.0 ps
compressed-x-grps   =   System          ; save the whole system

; Neighbor searching
cutoff-scheme       =   Verlet          ; Buffered neighbor searching
nstlist             =   10              ; Frequency to update the neighbor list
ns-type             =   grid            ; Make a grid in the box and only check atoms in
neighboring grid cells
pbc                 =   xyz             ; Periodic boundary conditions in all directions

;Electrostatic
coulombtype         =   P3M-AD          ; Particle-Particle Particle-Mesh algorithm.
rcoulomb            =   2.5             ; =6a. Distance for the Coulomb cut-off. Short
range cut-off.
rvdw                =   2.5             ; =6a. Distance for the Lennard-Jones or cut-off
fourierspacing      =   0.33            ; grid spacing for FFT.
epsilon-r           =   77.6            ; Dielectric constant.
pme-order           =   4
DispCorr            =   Ener

; Pressure coupling is off
pcoupl              =   no              ; no pressure coupling in NVT

;Langevin dynamics
bd-fric             =   5.5             ; friction coefficient
```

```
ld-seed                 =       -1

; Temperature coupling
tc-grps                 =       System          ; groups to couple to separate temperature baths
tau-t                   =       2               ; Time constant [ps]. One per group. 2.3
ref-t                   =       300             ; Temperature [K].

;Bond parameters
continuation            =       yes             ; Apply constraints to the start configuration
constraints             =       none
morse                   =       no              ; Harmonic potential
```