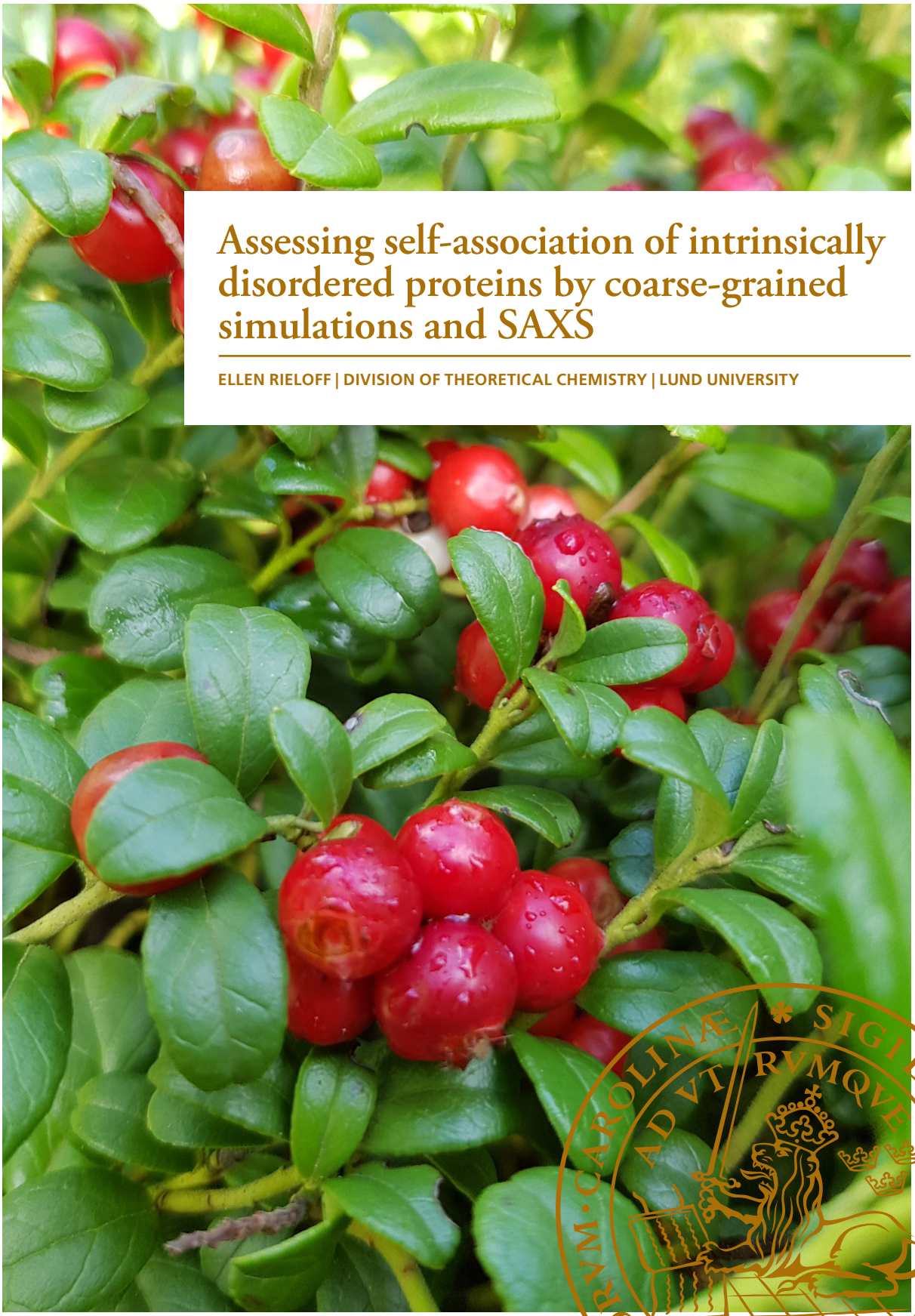
Assessing self-association of intrinsically disordered proteins by coarse-grained simulations and SAXS

Rieloff, Ellen

2019

Document Version:
Publisher's PDF, also known as Version of record



Assessing self-association of intrinsically disordered proteins by coarse-grained simulations and SAXS

ELLEN RIELOFF | DIVISION OF THEORETICAL CHEMISTRY | LUND UNIVERSITY

Assessing self-association of intrinsically disordered proteins

by coarse-grained simulations and SAXS

Ellen Rieloff

Lund University

LICENTIATE THESIS

by due permission of the Faculty of Science, Lund University, Sweden.
To be defended on Friday 29th of March 2019, 13.15 in lecture hall F, Center for Chemistry and Chemical Engineering, Naturvetarvägen 14, Lund.

Faculty opponent
Dr. Ann Terry
MAX IV Laboratory
Lund University

Organization LUND UNIVERSITY Division of Theoretical Chemistry Department of Chemistry P.O. Box 214 SE-221 00 Lund, Sweden	**Document name** LICENTIATE THESIS **Date of issue** 2019-03-29
Author(s) Ellen Rieloff	Sponsoring organization

Title and subtitle

Assessing self-association of intrinsically disordered proteins by coarse-grained simulations and SAXS

Abstract

This research investigates the behavior of intrinsically disordered proteins (IDPs) in solution, especially the self-associating saliva protein Statherin, by a combined computational and experimental approach. For the computational part, a bead necklace model previously parameterised for Histatin 5 was used. This model was shown to be applicable to a range of monomeric intrinsically disordered proteins and regions where the intra-chain interactions are dominated by electrostatic interactions. At high ionic strength the radius of gyration of the proteins fit nicely to the exponential law for polymers, with an exponent of 0.59 indicating self-avoiding random walk behaviour. For the longer proteins in this study (≥ 73 amino acids) a significant response to changes in the ionic strength was shown, depending on the charge distribution in the protein.

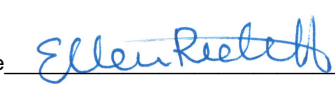
Statherin was characterised experimentally by small angle X-ray scattering and circular dichroism spectroscopy. With an additional short-ranged interaction to mimic the effect of hydrophobic interaction, the model was shown to capture the experimental trends in self-association, in regard to temperature, ionic strength and protein concentration. The combined experimental and computational approach allowed for an assessment of the intermolecular interactions contributing to the self-association. The decrease in self-association with increased temperature is considered to be an effect of mainly entropic origin, while the hydrophobic interaction was shown to be the main driving force for the self-association.

Key words

Intrinsically disordered proteins, self-association, coarse-grained modelling, Monte Carlo simulations, small angle X-ray scattering.

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language** English
ISSN and key title		**ISBN** 978-91-7422-636-2 (print) 978-91-7422-637-9 (digital)
Recipient's notes	**Number of pages** 97	Price
	Security classification	

Signature  Date 2019-02-14

Assessing self-association of intrinsically disordered proteins

by coarse-grained simulations and SAXS

Ellen Rieloff

Lund University

Licentiate Thesis
2019

Assessing self-association of intrinsically disordered proteins by coarse-grained simulations and SAXS

Front cover: Lingonberries. Photo by Ellen Rieloff

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University

ISBN 978-91-7422-636-2 (print), 978-91-7422-637-9 (digital)

Printed in Sweden by Media-Tryck, Lund University
Lund 2019

Media-Tryck is an environmentally certified and ISO 14001:2015 certified provider of printed material.
Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Abstract

This research investigates the behaviour of intrinsically disordered proteins in solution, especially the self-associating saliva protein Statherin, by a combined computational and experimental approach. For the computational part, a bead necklace model previously parameterised for Histatin 5 was used. This model was shown to be applicable to a range of monomeric intrinsically disordered proteins and regions where the intra-chain interactions are dominated by electrostatic interactions. At high ionic strength the radius of gyration of the proteins fit nicely to the exponential law for polymers, with an exponent of 0.59 indicating self-avoiding random walk behaviour. For the longer proteins in this study ($\geq$ 73 amino acids) a significant response to changes in the ionic strength was shown, depending on the charge distribution in the protein.

Statherin was characterised experimentally by small angle X-ray scattering and circular dichroism spectroscopy. With an additional short-ranged interaction to mimic the effect of hydrophobic interaction, the model was shown to capture the experimental trends in self-association, in regard to temperature, ionic strength and protein concentration. The combined experimental and computational approach allowed for an assessment of the intermolecular interactions contributing to the self-association. The decrease in self-association with increased temperature is considered to be an effect of mainly entropic origin, while the hydrophobic interaction was shown to be the main driving force for the self-association.

Contents

Papers

Populärvetenskaplig sammanfattning

För att våra kroppar ska fungera är proteiner en nyckelkomponent. De ansvarar för många livsnödvändiga funktioner, såsom transport av näringsämnen och syre, försvar mot främmande virus och bakterier, samt muskelrörelser. Proteiner är uppbyggda som långa kedjor av aminosyror med olika karaktär. Det finns laddade och polära aminosyror som båda trivs bra i vatten, men även hydrofoba aminosyror som helst undviker vatten. Hur aminosyrasekvensen ser ut är avgörande för vilken struktur proteinet har. Länge trodde man att proteiner behövde en fix tredimensionell struktur för att vara funktionella, och att det var den tredimensionella strukturen som avgjorde funktionen. Detta ifrågasattes dock, när det konstaterades att en betydande del av alla proteiner faktiskt saknar väldefinierad tredimensionell struktur. Dessa så kallade oordnade proteiner har visat sig vara inblandade i många viktiga biologiska processer och även i sjukdomar som till exempel Alzheimers. En biologisk komponent som innehåller många oordnade proteiner är saliv, där proteinerna bidrar till salivens funktionalitet, vilket bland annat är att skydda tandemalj och slemhinnor, verka antibakteriellt och påbörja matsmältningen. Ett av proteinerna är Statherin, vars främsta funktion är att binda kalciumsalter, så att det finns tillgängligt när tandemaljen behöver byggas upp, men inte i för stora mängder så att det bildas utfällningar.

För att förstå sambandet mellan struktur och funktion för oordnade proteiner är en viktig del att förstå hur de beter sig i lösning och vad som kontrollerar beteendet, bland annat vilka interaktioner som är med och styr. Förutom att använda experimentella metoder för att studera detta är datorsimuleringar ett viktigt komplement, då de kan bidra med mer detaljerad information på en molekylär nivå. För att göra simuleringar behöver man en modell som kan beskriva systemet. Vi har använt en grovkornig modell som istället för att ha med alla atomer i proteinet, ser ett protein som ett pärlband, där varje pärla motsvarar en aminosyra. I den första artikeln har vi visat att den här modellen fungerar bra för flera olika oordnade proteiner. Dessutom visade vi att laddningsfördelningen

i proteinet samt kedjelängden spelar roll för proteinets form och storlek, när salthalten ändras.

I den andra artikeln fokuserade vi på Statherin, som självassocierar med ökad koncentration. I experiment såg vid hur storleken på de självassocierade komplexen ändrades med proteinkoncentration, salthalt, temperatur och tillsats av urea, en molekyl som är känd för att bryta hydrofob interaktion. En uppdaterad version av den grovkorniga modellen visade sig kunna beskriva de experimentella trenderna och kunde därför hjälpa till att avkoda vilka interaktioner som är viktiga i självassocieringen. Vi konstaterade att hydrofob interaktion är en viktig drivkraft för att självassocieringen ska ske, men att även entropi spelar roll, bland annat när temperaturen ändras.

List of Publications

This thesis is based on the following papers, which will be referred to by their Roman numerals in the text.

I **Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions**
C. Cragnell, E. Rieloff, M. Skepö.
Journal of Molecular Biology, 430, **2018**, pp. 2478–2492.

II **Assessing the Intricate Balance of Intermolecular Interactions upon Self-Association of Intrinsically Disordered Proteins**
E. Rieloff, M. D. Tully, M. Skepö.
Journal of Molecular Biology, 431, **2019**, pp. 511–523.

Publications not included in this thesis

Structural Characterization of Bubbles Formed in DNA Melting: A Monte Carlo Simulation Study
E. Rieloff, S. C. C. Nunes, A. A. C. C. Pais, M. Skepö.
ACS Omega, 2, **2017**, pp. 1915-1921.

Contribution Report

I **Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions**
I performed the experiments and part of the simulations and analysis, took part in discussions and contributed to the writing of the paper.

II **Assessing the Intricate Balance of Intermolecular Interactions upon Self-Association of Intrinsically Disordered Proteins**
I planned the study together with my supervisor, performed the experiments and simulations and implemented cluster moves and analyses. I analysed the data with input from the co-authors, and wrote the manuscript with support from the co-authors.

Chapter 1

Biological background

The research performed in this thesis investigates the solution behaviour of intrinsically disordered proteins (IDPs), especially the saliva protein Statherin. This chapter will introduce the concept of IDPs and describe their biological relevance to put the work into context. Since the main focus is on Statherin, it and its natural environment is described in more detail.

1.1 Intrinsically disordered proteins

There are four different levels of structure in a protein, illustrated in Figure 1.1. The *primary structure* is the sequence of amino acids, held together by peptide bonds. Local parts of the chain can arrange into regular structures, referred to as *secondary structure*. α-helices and β-sheets are typical examples of such structures, held together by hydrogen bonds [1]. The whole protein can then fold into a three-dimensional shape, referred to as the *tertiary structure*. The major driving force behind the folding is the hydrophobic interaction, trying to hide hydrophobic residues from the surrounding water [2]. In addition, a protein can consist of several different chains, each having a three-dimensional structure and

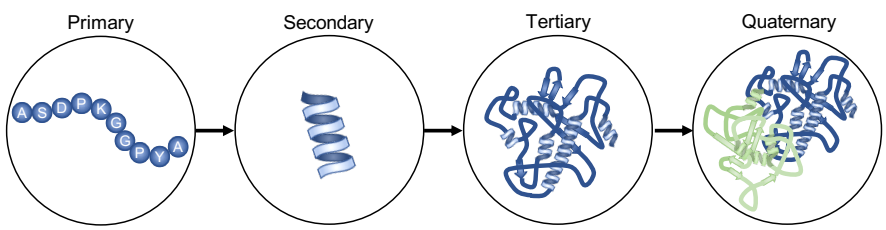


Figure 1.1: Illustration of the different levels of protein structure.

making up a subunit of the complete protein. The arrangement of the subunits is called the *quaternary structure*.

Intrinsically disordered proteins lack well-defined tertiary structure under physiological conditions, which means that they are much more flexible than other proteins and adopt many different conformations. The group is however rather heterogenous, including less or more compact proteins with different degrees of secondary and tertiary structure [3, 4]. Often can protein disorder be recognised already in the primary sequence. IDPs often have a low sequence complexity and are generally enriched in charged and polar amino acids, with a low content of bulky hydrophobic amino acids [5, 6].

For a long time it was believed that the protein function was strongly coupled to the three-dimensional structure, meaning that a protein was required to have a fixed tertiary structure to be functional. Hence, IDPs were regarded as non-functional and of no importance. Later on it was discovered that approximately 10–20% of the eukaryotic proteins are intrinsically disordered, and even more proteins contain long disordered regions [7–10]. In addition, it has been shown that IDPs are indeed functional, being involved in many biological processes [11–14], and that the lack of folded structure might actually be related to their functions [9, 15].

One part of understanding the relationship between structure and function for IDPs is characterisation of the structure in solution. As stated, IDPs are flexible and therefore an average structure is not really representative. Instead, they are better described by an ensemble of different conformations. The ensemble will change with environmental conditions such as temperature, pH and ionic strength, and therefore affect the behaviour of the protein. Most experimental techniques only provide information averaged over both time and a large number of molecules. For this reason, modelling and simulations are of high importance to the field, as they give access to the full ensemble and information at the molecular level. However, experiments still play a vital role as the timescale and system size are limited in simulations. Furthermore, experimental data is required for validation of theoretical models.

1.2 Saliva

Saliva is a complex fluid of great importance to our oral health, even though it consists of 99.5% water. Approximately 0.2% are proteins [16] and despite being such a small constituent, they are responsible for many different functions, as presented in Figure 1.2. Note that many of the proteins are multi-functional and intrinsically disordered.

Other constituents of saliva involves inorganic components such as sodium, potassium, calcium, chloride, and organic components such as lipids and carbohydrates. The composition, ionic strength and pH varies with a lot of different factors, for example time of day and food intake. Diseases and medication can

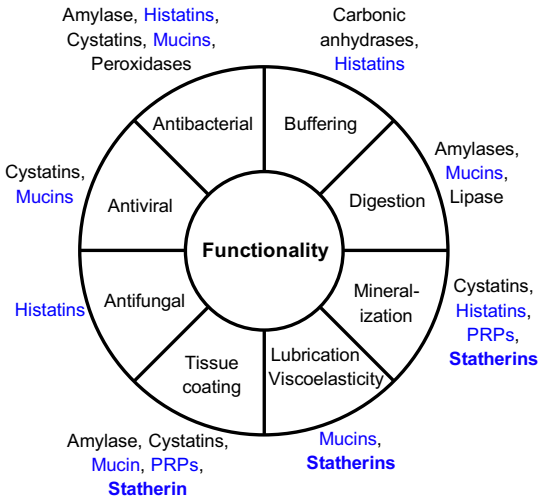


Figure 1.2: Proteins responsible for functionality of saliva. Intrinsically disordered proteins are marked in blue. The figure is adapted from Levine [17].

affect the saliva production and hence change the environment for the proteins [16].

1.3 Statherin

Statherin is an intrinsically disordered saliva protein, and as was shown in Figure 1.2 it is multi-functional. The main function is to prevent spontaneous precipitation of calcium phosphate salts in saliva, in order to maintain a supersaturated environment [18, 19], which helps with remineralisation after dental erosion [20]. Statherin has also been shown to be involved in lubrication [21] and bacterial interactions [22].

Statherin is a rather small protein, only 43 amino acids long with a molecular weight of 5.38 kDa, which makes it suitable for modelling. It has a distinct charge distribution evident from the primary sequence in Figure 1.3, where nine out of ten charged residues are located among the first 13 amino acids in the N-terminal.

```
+DSSEEKFLRRIGRFGYGYGPYQPVPEQPLYPQPYQPQYQQYTF-
```

Figure 1.3: The primary sequence of Statherin [23]. Amino acids that have a negatively charged side chain at pH 8 are marked in red, and those with a positively charged side chain are marked in blue. The phosphorylated serines (marked in dark red) have a charge of $-2e$ each at pH 8.

Overall the hydrophobicity is rather low (based on the hydropathy values in the Kyte-Doolittle scale [24]), which is typical for IDPs. However, residues 15–43 contain seven tyrosines, whose aromatic side-chains have been established to be of importance for liquid-liquid phase separation [25, 26]. Statherin has been shown to self-associate upon increased protein concentration, such that several protein chains merge to a larger complex. Self-association is further described in section 2.2.

Chapter 2

Intermolecular interactions and self-association

2.1 Intermolecular interactions

Intermolecular interactions are generally weak compared to covalent bonds. However, they are highly important to the behaviour of molecular systems, such as how proteins fold or interact. This chapter is mostly based on the book by Israelachvili [27], which can be referred to for a more thorough description.

The interactions described below have different distance dependence, making some short-ranged and others long-ranged. The decay of potentials with different distance dependence is shown in Figure 2.1.

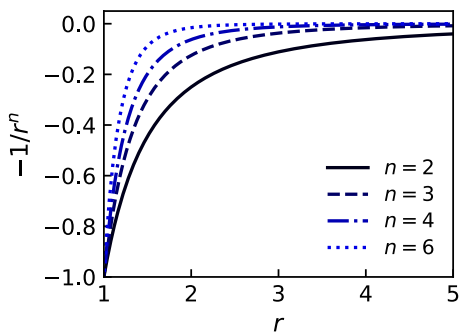


Figure 2.1: Illustration of the decay of potentials with different distance dependence.

2.1.1 Coulomb interaction

The interaction between two charges is described by the Coulomb law

$$F(r) = \frac{Q_1 Q_2}{4\pi\varepsilon_0\varepsilon_\mathrm{r}} \frac{1}{r^2}, \tag{2.1}$$

where F is the electrostatic force between two atoms with charges Q_1 and Q_2 separated by the distance r, ε_0 is the vacuum permittivity and ε_r is the relative permittivity of the surrounding medium. The interaction free energy, $w(r)$, between the two charges is given by

$$w(r) = \frac{Q_1 Q_2}{4\pi\varepsilon_0\varepsilon_\mathrm{r}} \frac{1}{r}. \tag{2.2}$$

The interaction is long-ranged, but if the charges are surrounded by ions as in an aqueous salt solution, the interaction is screened, which reduces the range of the interaction. According to the Debye–Hückel theory, a screened Coulomb potential can be expressed as

$$V(r) = \frac{Q_1 Q_2}{4\pi\varepsilon_0\varepsilon_\mathrm{r}} \frac{1}{r} \exp(-\kappa r), \tag{2.3}$$

where $V(r)$ is the potential energy and κ^{-1} is called the Debye length, defined by

$$\kappa^{-1} = \sqrt{\frac{\varepsilon_0\varepsilon_\mathrm{r} kT}{2N_\mathrm{A} e^2 I}}, \tag{2.4}$$

where k is the Boltzmann constant, T is the temperature, N_A the Avogadro constant, e the elementary charge, and I refers to the ionic strength, defined as

$$I = \frac{1}{2} \sum_{i=1}^{n} c_i Z_i^2. \tag{2.5}$$

Here, n is the number of different ion species, and c_i is the concentration of ion i with charge number Z_i.

2.1.2 Charge–dipole

Most molecules have no net charge; however, they often possess an electric dipole, caused by an asymmetric distribution of electrons in the molecule. The dipole moment is defined as

$$\boldsymbol{\mu} = q\mathbf{l}, \tag{2.6}$$

where $\mathbf{l}$ is the distance vector between the two charges $-q$ and $+q$. When a charge and a dipole interact at a distance $r >> l$, the potential energy is given by

$$V(r,\theta) = -\frac{Q\mu\cos\theta}{4\pi\varepsilon_0\varepsilon_\mathrm{r}} \frac{1}{r^2}, \tag{2.7}$$

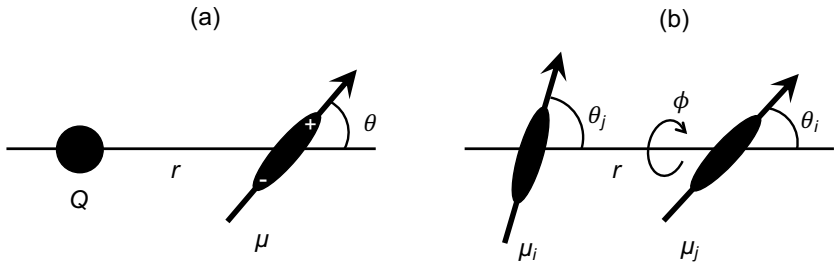


Figure 2.2: Schematic representation of the (a) charge–dipole and (b) dipole–dipole interaction, where r is the distance between the interacting species, θ is the polar angle and ϕ the azimuthal angle.

where the polar angle, θ, is the angle between the distance vector and the dipole (see Figure 2.2a). If the charge is positive, maximum attraction occurs when the dipole points away from the charge ($\theta = 0°$). At large separation or in a medium with high relative permittivity, the angle dependence of the interaction can fall below the thermal energy kT, which allows the dipole to rotate more or less freely. However, conformations allowing for attractive interactions will still be more favourable, so the angle-averaged potential will not be zero. The interaction free energy between a freely rotating dipole and a charge is given by

$$w(r) \approx -\frac{Q^2\mu^2}{6(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2 kT}\frac{1}{r^4} \text{ for } kT > \frac{Q\mu}{4\pi\varepsilon_0\varepsilon_\mathrm{r} r^2}. \tag{2.8}$$

Note that this changes the distance dependence of the potential, making it more short-ranged.

2.1.3 Dipole–dipole

The interaction energy between two stationary dipoles i and j can be described by the following potential

$$V(r,\theta_i,\theta_j,\phi) = -\frac{\mu_i\mu_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r}}\frac{1}{r^3}(2\cos\theta_i\cos\theta_j - \sin\theta_i\sin\theta_j\cos\phi), \tag{2.9}$$

where ϕ is the azimuthal angle between the dipoles (see Figure 2.2b). Also in this case can the dipoles rotate, so the angle-averaged interaction free energy is

$$w(r) = -\frac{\mu_i^2\mu_j^2}{3(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2 kT}\frac{1}{r^6} \text{ for } kT > \frac{\mu_i\mu_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r} r^3}. \tag{2.10}$$

This interaction is usually referred to as the *Keesom interaction* and is a part of the total van der Waals interaction described in section 2.1.6.

2.1.4 Charge–induced dipole

All molecules and atoms, even nonpolar ones, are polarised by an external electric field, which means that the electron cloud in the molecule is displaced. Hence, the electric field exhibited by a charge will induce a dipole moment in the nonpolar molecule. The potential between the charge and the induced dipole is expressed as

$$V(r) = -\frac{-Q^2\alpha}{2(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2}\frac{1}{r^4}, \tag{2.11}$$

where α is the polarisability of the molecule.

2.1.5 Dipole–induced dipole

Similarly to the charge–induced dipole interaction, a nonpolar molecule can gain an induced dipole moment in the field from a permanent dipole. The interaction is described by the following potential,

$$V(r) = -\frac{\mu^2\alpha}{(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2}\frac{1}{r^6}. \tag{2.12}$$

Notice that this potential is already angle-averaged, since the interaction normally is not strong enough to mutually orient the molecules. This interaction is usually referred to as the *Debye interaction* and is a part of the total van der Waals interaction due to the $1/r^6$-dependence.

2.1.6 van der Waals interaction

The total van der Waals interaction includes three different types of interactions, which all have a $1/r^6$-dependence: Keesom, Debye and London (dispersion), of which Keesom and Debye have been described above (section 2.1.3 and 2.1.5). The Keesom interaction is only present between permanent dipoles and the Debye interaction when one of the molecules is a permanent dipole. The last interaction, the *London dispersion interaction* is however present between all types of molecules. It is of quantum mechanical origin, although we can think of it in a simpler manner. For a nonpolar atom (or molecule) the time averaged dipole moment is zero, although at any instant it exists a finite dipole moment caused by an uneven electron distribution around the nucleus. This instantaneous dipole generates an electric field that induces a dipole in another nearby atom (or molecule), leading to an attractive interaction.

2.1.7 Hydrogen bond

In the previous chapter it was mentioned that hydrogen bonds are of specific importance for protein structure. A hydrogen bond can occur between a highly electronegative atom, such as nitrogen, oxygen or fluorine, and a hydrogen covalently bond to another such electronegative atom. It is of predominantly

electrostatic origin and can be seen as an especially strong dipole–dipole interaction. Unlike normal dipole-dipole interactions it is fairly directional and can be described by a $1/r^2$-dependence, similar to the charge–dipole interaction.

2.1.8 Exchange repulsion (excluded volume)

At very small interatomic distances, when electron clouds overlap, a strong repulsive interaction of quantum mechanical origin occurs, which limits how close two atoms can come. The repulsion increases steeply with decreased distance and is therefore often modelled with a hard sphere potential which goes directly from 0 to infinity, or with a soft core potential of $1/r^{12}$-dependence.

2.1.9 Hydrophobic interaction

Water is a special solvent due to the possibility to form many hydrogen bonds, which makes the water–water interaction strong. Therefore, the water molecules much rather interacts with other water molecules than non-polar molecules. For small non-polar molecules the water can arrange around the non-polar molecule in such a way that no hydrogen bonds are broken. However, this arrangement is more ordered and therefore comes at an entropic cost, which makes it more favourable to separate the non-polar molecules from the water molecules. For large non-polar molecules it is not possible to retain hydrogen bonds, which instead leads to an energy driven separation. Therefore, the cause of separation between water and non-polar molecules can be both mostly entropic or mostly energetic, however, the net result can always be seen as an effective attraction between non-polar molecules, called a hydrophobic interaction [28].

2.1.10 Effect of conformational entropy

A restriction of available conformations for a flexible polymer leads to a decrease in conformational entropy. This can occur when the polymer approaches a surface or other polymers. If the restriction is large enough, the result will be an effective repulsion of entropic origin.

2.2 Self-associating systems

Self-association is the spontaneous formation of larger structures from smaller constituents. A typical example of self-association is the micelle formation of surfactants. Surfactants usually consist of a hydrophobic tail and a polar head-group, which means that they are *amphiphilic*. Driven by the hydrophobic interaction the surfactants arrange into spherical structures called micelles, hiding the hydrophobic tails in the interior, as shown in Figure 2.3. This only happens above a certain surfactant concentration, named the *critical micelle concentration* (CMC).

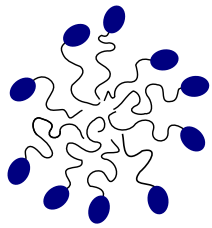


Figure 2.3: A schematic illustration of a micelle formed of surfactants having polar head-groups and hydrophobic tails.

It is the intermolecular interactions, such as van der Waals interactions, hydrogen bonding, hydrophobic interaction and screened electrostatic interactions, that govern self-association. Since these forces are generally weak, at least compared to covalent bonds, the self-association process is highly affected by solution conditions such as pH and ionic strength. Both the interactions between and within self-assembled structures are affected by changes in the solution conditions, therefore the size and shape of the self-assembled complexes can be modified [27].

Large molecules such as amphiphilic block-copolymers can also form micelles, however, due to their much larger size and sometimes more pronounced amphiphilic nature, the behaviour can differ from surfactants. Proteins can also self-associate, which the intrinsically disordered milk protein β-casein is a good example of. The C-terminal part of β-casein contains many hydrophobic residues, while the N-terminal part has several phosphorylated residues that contributes to a net charge, giving the protein chain an amphiphilic structure. Many studies, only a few mentioned here, have been devoted to the β-casein micelle formation and have shown that the micelle size and shape, as well as CMC are sensitive to the solution conditions such as temperature, pH and protein concentration [29–33].

Chapter 3

Theoretical methods

In theoretical chemistry computational methods are used to solve chemical problems. One approach is to set up a model of the system of interest, in my case intrinsically disordered proteins, and perform computer simulations to calculate properties of the system. In this chapter the simulation method and model that I have used will be described, after an introduction to the theory behind.

3.1 Statistical mechanics

Statistical mechanics provides a connection between macroscopic properties, such as temperature and pressure, and microscopic properties related to the molecules and their interactions. The aim is to provide means to both predict macroscopic phenomenas and understand macroscopic phenomenas on a molecular level. Statistical mechanics applied for explaining thermodynamics is usually referred to as statistical thermodynamics. Here I will provide a brief introduction to the key concepts, while a more in-depth description can be found in for example the book by Hill [34].

A central concept in statistical mechanics is *ensembles*. An ensemble is an imaginary collection of a very large number of systems, each being equal at a thermodynamic (macroscopic) level, but differing on the microscopic level. Ensembles can be classified according to the macroscopic system that they represent, and some are outlined below.

Microcanonical ensemble (NVU)
The microcanonical ensemble represents an isolated system, in which the number of particles (N), the volume (V) and the internal energy (U) are constant. Hence, the systems in the ensemble all have the same N, V, and U, and share the same environment, however, they correspond to different microstates.

Canonical ensemble (NVT)
In the canonical ensemble the number of particles (N), the volume (V), and the temperature (T) are constant. Hence, the system is closed and isothermal.

Grand canonical ensemble (μVT)
The grand canonical ensemble represents an open isothermal system, by having the chemical potential (μ), the volume (V) and the temperature (T) kept constant.

When an experimental measurement is performed, a time average is taken over the observable of interest. If we instead want to calculate the observable from molecular properties, we would need to deal with both a large number of molecules and the requirement to observe them for a sufficiently long time to smear out molecular fluctuations. In practice this would be extremely complicated, however, a different approach is possible due to the *first postulate* of statistical mechanics: a (long) time average of a mechanical variable in a thermodynamic system is equal to the ensemble average of the variable in the limit of an infinitely large ensemble, provided that the ensemble replicate the thermodynamic state and environment. Stated differently, this postulate says that instead of using a time average, we can obtain the same result by performing an ensemble average, given that the ensemble is sufficiently large. This is valid for all ensembles and provides the basis for Monte Carlo simulations. There is also a *second postulate* of statistical mechanics which states that for an infinitely large ensemble representing an isolated thermodynamic system, the systems of the ensemble are distributed uniformly over the possible states consistent with the specified values of N, V and U. This postulate is also referred to as the principle of equal *a priori* probabilities, as it says that in the microcanonical ensemble, all microscopic states are equally probable.

Since the simulations performed in this thesis are all in the canonical ensemble, this is the only one that will be considered henceforth. In the canonical ensemble it can be shown that the probability to find the system in a particular energy state U_i is

$$P_i(N,V,T) = \frac{\exp[-U_i(N,V)/kT]}{Q(N,V,T)}, \tag{3.1}$$

where Q is the canonical partition function, given by

$$Q(N,V,T) = \sum_i \exp[-U_i(N,V)/kT], \tag{3.2}$$

where $\exp[-U_i(N,V)/kT]$ is known as the Boltzmann weight. The partition function describes the equilibrium statistical properties of the system and can be used to express the (for the canonical ensemble) characteristic function, Helmholtz free energy, A, according to

$$A = -kTlnQ. \tag{3.3}$$

Other thermodynamic variables, such as the entropy, pressure and total energy can be derived from the Helmholtz free energy.

Here we have introduced statistical mechanics in a quantum mechanical formulation with discrete energy states. However, many simulation methods are based on classical mechanics, in which the microstates are so close in energy that they are approximated as a continuum. In a classical treatment the canonical partition function becomes

$$Q_{\text{class}} = \frac{1}{N!h^{3N}} \int \exp[-H(\mathbf{p}^N, \mathbf{r}^N)/kT] d\mathbf{p}^N \mathrm{d}\mathbf{r}^N, \tag{3.4}$$

where h is Planck's constant and the integration is performed over all momenta $\mathbf{p}^N$ and all coordinates $\mathbf{r}^N$ for all N particles. $H(\mathbf{p}^N, \mathbf{r}^N)$ is the Hamiltonian of the system, having one kinetic energy part (dependent on the temperature) and one potential energy part (dependent on the interactions). The kinetic part can be integrated directly, simplifying the partition function to

$$Q_{\text{class}} = \frac{Z_N}{N!\Lambda^{3N}}, \tag{3.5}$$

where

$$Z_N = \int_V \exp[-U_{\text{pot}}(\mathbf{r}^N)/kT] \mathrm{d}\mathbf{r}^N \tag{3.6}$$

is the configurational integral calculated from the potential energy, U_{pot}, and

$$\Lambda = \frac{h}{(2\pi mkT)^{1/2}} \tag{3.7}$$

is the de Broglie wavelength, where m is the mass. The configurational integral is of importance for calculating the ensemble average of an observable B,

$$\langle B(\mathbf{r}^N) \rangle = \frac{\int_V B(\mathbf{r}^N) \exp[-U_{\text{pot}}(\mathbf{r}^N)/kT] \mathrm{d}\mathbf{r}^N}{Z_N}. \tag{3.8}$$

However, solving the integrals is normally a rather challenging problem that requires numerical solution tools, such as the Monte Carlo method that will be discussed in the next section.

3.2 Metropolis Monte Carlo simulations

As stated above, we can obtain an ensemble average of an observable $B(\mathbf{r}^N)$ by solving the expression in Equation 3.8. In the simplest Monte Carlo technique, often referred to as *random sampling*, this would be done by evaluating $B(\mathbf{r}^N)$ at a large number of random points in phase space and multiplying the result with the Boltzmann factor. Each point in phase space corresponds to a configuration. However, a lot of the generated configurations would give a negligible contribution to the average, by having a really small Boltzmann factor. An example of such configurations are ones with overlapping particles, as this would result

in a very high (or infinite) potential energy. Metropolis et al. [35] presented a more efficient scheme for evaluating a ratio of integrals for obtaining $\langle B(\mathbf{r}^N)\rangle$. In this scheme the sampling is based on the Boltzmann factor, so that the sampling is focused more around configurations with a larger Boltzmann factor. This is a type of *importance sampling* and implies that the number of configurations needed for getting a good result is reduced, which makes the simulations faster. A Metropolis Monte Carlo algorithm is outlined below [36]:

(i) Generate a starting configuration.

(ii) Calculate the interaction energy within the system, U_{old}.

(iii) Choose a particle at random and a type of trial move (see section 3.2.1).

(iv) Generate a new configuration by performing the trial move.

(v) Calculate the energy of the new configuration, U_{new}.

(vi) Compare the energy of the old and the new configuration to determine if the new configuration is accepted. The probability of acceptance is given by:

$$p_{\text{acc}} = \begin{cases} 1 & \text{if } U_{\text{new}} \leq U_{\text{old}} \\ \exp[-\frac{1}{kT}(U_{\text{new}} - U_{\text{old}})] & \text{if } U_{\text{new}} \geq U_{\text{old}} \end{cases}. \tag{3.9}$$

(vii) If the new configuration is rejected, restore the old one.

(viii) Repeat from step (ii).

3.2.1 Trial moves

An advantage with Monte Carlo simulations is that unphysical moves can be used to speed up the exploration of the configurational space. For polymers or proteins modelled as bead-necklaces it is common to use four different moves, described below. For self-associating systems it is also advantageous to add a cluster move. One type of cluster move has been implemented for the Statherin simulations in this thesis and is described below.

Single particle translation
A single bead in the chain, or a counterion, is moved to a new, randomly chosen, position. The length of the translation is limited by an input parameter defined in the simulation.

Slithering move
In the slithering move, also known as reptation, one of the end beads is displaced to a random position within a bond length. The other beads are moved forward in the chain along the old configuration, as illustrated in Figure 3.1a.

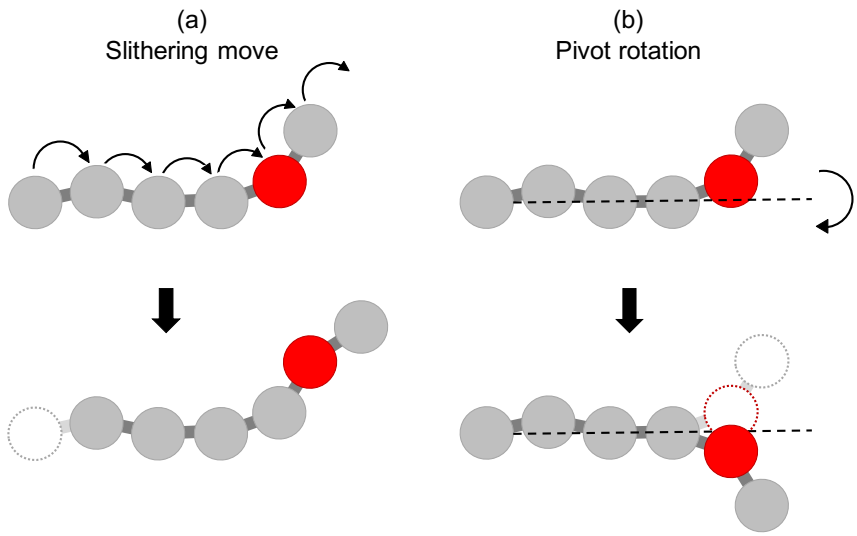


Figure 3.1: Illustration of two types of Monte Carlo moves: (a) slithering and (b) pivot rotation.

Pivot rotation
One end of the chain is rotated around an axis defined by a randomly selected bond, see Figure 3.1b.

Chain translation
A whole chain is translated. This move does not change the conformation of the chain, only the position in relation to other chains.

Cluster move
A translation of a group of chains. The group includes the chain that the selected particle belongs to and all other chains whose center of mass is less than a predefined distance away. If the number of chains in the cluster changes during the displacement, the move is automatically rejected, as this violates detailed balance*.

3.3 The coarse-grained model

For the purpose of simulating intrinsically disordered proteins we use a bead-necklace model based on the primitive model, in which each amino acid is described as a hard sphere (bead), connected by harmonic bonds. The N- and

*Detailed balance implies that the probability of making a move and reversing it should be the same.

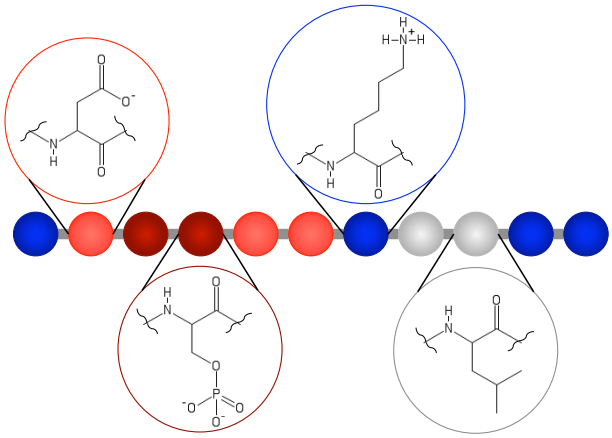


Figure 3.2: A schematic description of the coarse-grained model, showing the N-terminal fragment of Statherin. Apart from the first blue sphere representing the N-terminal, each hard sphere represent an amino acid residue, where blue spheres are positively charged residues, bright red spheres are negatively charged residues and dark red spheres correspond to phosphorylated residues with a charge $Z = -2e$. Gray spheres are neutral amino acids. The four amino acid structures that exemplifies the coarse-graining are aspartic acid, phoshorylated serine, lysine, and leucine (from left to right). This figure was first published in Journal of Molecular Biology (Paper I).

C-termini are modelled explicitly as charged spheres in each end of the protein chain, so the full length corresponds to the number of amino acids+2. Each bead has a fixed point charge of $+1e$, $-1e$, $-2e$ or 0, corresponding to the state of the amino acid side chain at the desired pH. For an illustration of the model, see Figure 3.2. The counterions are included explicitly, while the solvent (water) and salt is treated implicitly. The exact parameterisation of the model used for paper I was performed for the IDP Histatin 5 [37].

The basic model contains contributions from excluded volume, electrostatic interactions, and a short-ranged attraction mimicking van der Waals-interactions. The total potential energy is divided into bonded and non-bonded interactions, according to

$$U_{\text{tot}} = U_{\text{bond}} + U_{\text{nonbond}} = U_{\text{bond}} + U_{\text{hs}} + U_{\text{el}} + U_{\text{short}}, \tag{3.10}$$

where U_{hs} is a hard-sphere potential, U_{el} the electrostatic potential, and U_{short} a short-ranged attraction. The non-bonded energy is assumed pairwise additive, according to

$$U_{\text{nonbond}} = \sum_{i<j} u_{ij}(r_{ij}), \tag{3.11}$$

where u_{ij} is the interaction between two particles, $r_{ij} = |\mathbf{R}_i - \mathbf{R}_j|$ is the center-to-center distance between the two particles, and $\mathbf{R}$ refers to the coordinate vector.

A harmonic bond represents the bonded interaction,

$$U_{\text{bond}} = \sum_{i=1}^{N-1} \frac{k_{\text{bond}}}{2}(r_{i,i+1} - r_0)^2. \tag{3.12}$$

Here, N denotes the number of beads in the protein, k_{bond} is the force constant having a value of 0.4 N/m, and $r_{i,i+1}$ is the center-to-center distance between two connected beads, with the equilibrium separation r_0 = 4.1 Å.

The excluded volume is accounted for by a hard sphere potential,

$$U_{\text{hs}} = \sum_{i<j} u_{ij}^{\text{hs}}(r_{ij}), \tag{3.13}$$

where the summation extends over all beads and ions. Here, u_{ij}^{hs} represents the hard sphere potential between two particles, according to

$$u_{ij}^{\text{hs}}(r_{ij}) = \begin{cases} 0, & r_{ij} \geq R_i + R_j \\ \infty, & r_{ij} < R_i + R_j \end{cases}, \tag{3.14}$$

where R_i and R_j denote the radii of the particles (2 Å). The electrostatic potential energy is given by an extended Debye–Hückel potential,

$$U_{\text{el}} = \sum_{i<j} u_{ij}^{\text{el}}(r_{ij}) = \sum_{i<j} \frac{Z_i Z_j e^2}{4\pi\varepsilon_0\varepsilon_{\text{r}}} \frac{\exp[-\kappa(r_{ij} - (R_i - R_j))]}{(1+\kappa R_i)(1+\kappa R_j)} \frac{1}{r_{ij}}. \tag{3.15}$$

Hence, the salt in the system is treated implicitly as a screening of the electrostatic interactions.

The short-ranged attractive interaction between the beads is included through an approximate arithmetic average over all amino acids, namely,

$$U_{\text{short}} = -\sum_{i<j} \frac{\varepsilon_{\text{short}}}{r_{ij}^6}. \tag{3.16}$$

Here, $\varepsilon_{\text{short}}$ reflects the amino acid polarisability and sets the strength of the attraction. In this model $\varepsilon_{\text{short}}$ is $0.6 \cdot 10^4$ kJ Å/mol, which corresponds to an attraction of 0.6 kT at closest contact.

For the study of Statherin self-association (paper II), an additional short-ranged interaction is required to make the protein chains associate, which would correspond to a hydrophobic interaction. This interaction is applied between all neutral amino acids, according to

$$U_{\text{hphob}} = -\sum_{\text{neutral}} \frac{\varepsilon_{\text{hphob}}}{r_{ij}^6}, \tag{3.17}$$

where $\varepsilon_{\text{hphob}}$ is $1.32 \cdot 10^4$ kJ Å/mol, which corresponds to an attraction of $1.32\,kT$ at closest contact. The value of $\varepsilon_{\text{hphob}}$ was set by comparing the average association number with experimental results.

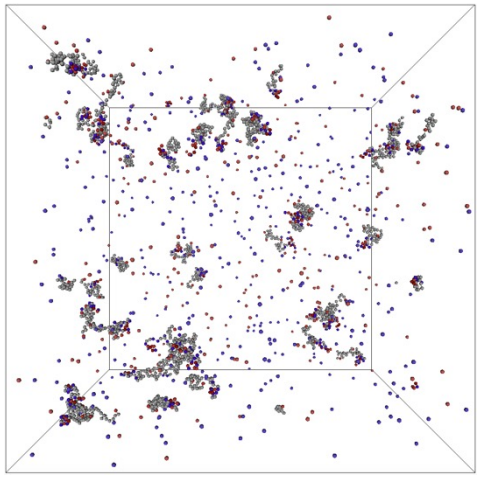


Figure 3.3: A snapshot showing the simulation box containing 45 protein chains and positive and negative counterions.

3.4 Simulation box

For the simulations, the proteins are enclosed in a cubic box with a fixed volume, fixed number of particles, and fixed temperature, corresponding to the canonical (NVT) ensemble. Included are also explicit counterions. For the simulation of a single chain, a very large box is used. For the simulations of several chains, the box length is calculated to give the system a certain concentration depending on the number of proteins inside the box. To ensure that the system size (number of chains) and box size were not too small, several different sizes were tested. A snapshot from a typical simulation in Paper II showing the simulation box is displayed in Figure 3.3.

3.4.1 Periodic Boundary Conditions

In this thesis we are interested in bulk properties of IDPs, that is, their behaviour in solution. Experimentally, for my most dilute samples, even a small sample volume such as 0.1 mL contains about 10^{15} protein molecules, which is way out of reach for computer simulations. The simulated system needs to be much smaller; typically a simulation box has a box length of a few hundred Ångströms. However, this causes a large part of the molecules to be in contact with the walls. Therefore, to simulate bulk properties a different approach is needed, namely, periodic boundary conditions (PBC). The simulation box can be thought of as replicated in all directions to create an infinite lattice, as illustrated in Figure 3.4. In practice, this is achieved by letting a particle that leaves from one side of the box enter again from the opposite side. With this approach there are no walls in the system, hence it resembles the bulk.

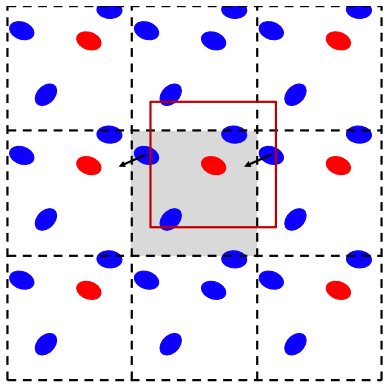


Figure 3.4: A schematic illustration of periodic boundary conditions in two dimensions, where the gray box is replicated in all directions. The arrows represents movement over a border. Applying the minimum image convention implies a cut-off corresponding to the red square for the particle marked in red.

3.4.2 Minimum image convention

When dealing with an infinite system such as when using PBC, adding all the interactions in the system would lead to an infinite sum, since there is an infinite number of particles. Therefore, in practice, the interactions need to be truncated. One approach is to use the minimum image convention, which restricts each molecule to interact only with the closest image of the other molecules. In practice, it corresponds to a cubic cut-off as illustrated in Figure 3.4.

3.5 Analyses

To characterise the simulated protein systems and to be able to compare with experiments, we have performed different analyses that are described below.

3.5.1 Structural analyses

Radius of gyration
The radius of gyration is a measure of the size of the protein and is defined as

$$R_g \equiv \langle N^{-1} \sum_{i=1}^{N} (\mathbf{r}_i - \mathbf{r}_{\text{com}})^2 \rangle^{1/2}, \tag{3.18}$$

where $\langle \cdots \rangle$ refers to an ensemble average, N is the number of beads in the chain, and $\mathbf{r}_{\text{com}}$ corresponds to the center of mass. The radius of gyration can be determined experimentally by small angle X-ray scattering (SAXS), which makes it possible to directly compare simulations and experiments.

End-to-end distance
This corresponds to the distance between the N- and C- terminal, given by

$$R_{ee} \equiv \langle |\mathbf{r}_1 - \mathbf{r}_N|^2 \rangle^{1/2}. \tag{3.19}$$

For an equilibrated simulation, the distribution of both R_g and R_{ee} should be close to gaussian for the protein chains. Hence, besides providing information about the proteins, these distributions are also a good way to assess if the simulations have been run for a sufficiently long time.

Structure factor
For a direct comparison between experiments and simulations, scattering curves are measured by SAXS and corresponding curves are calculated in the simulations. The theory behind SAXS can be found in section 4.1. In the calculations, each particle (bead) is regarded as a point scatterer. For a system containing N identical scattering objects, the total structure factor is expressed as

$$S(\mathbf{q}) = \left\langle \frac{1}{N} \left| \sum_{j=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_j) \right|^2 \right\rangle, \tag{3.20}$$

where $\mathbf{q}$ is the scattering vector. $S(\mathbf{q})$ can be further decomposed into partial structure factors given by

$$S_{jk}(\mathbf{q}) = \left\langle \frac{1}{(N_j N_k)^{1/2}} \left[\sum_{j=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_j) \right] \left[\sum_{k=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_k) \right] \right\rangle, \tag{3.21}$$

where j and k are particle types. The total and partial $S(\mathbf{q})$ are related through

$$S(\mathbf{q}) = \sum_{j=1}^{N_j} \sum_{k=1}^{N_k} \frac{(N_j N_k)^{1/2}}{N} S_{jk}(\mathbf{q}). \tag{3.22}$$

The scattering intensity can be expressed as a product of the form factor and the structure factor, where the form factor corresponds to intra-particle interference and the structure factor to inter-particle interference. For a point scatterer, the form factor is constant, inferring that the scattering intensity is proportional to the structure factor. Consequently, the calculated structure factor for the point scatterers corresponds to the system scattering intensity, only lacking a constant scaling factor. If the system is composed of a single protein chain, the calculated scattering profile comes only from intra-chain interference, hence, it is the protein form factor. For comparison with experiments an approximate effective particle form factor needs to be accounted for. This can be solved by dividing both the experimental and calculated scattering profile by their forward scattering, I_0.

Pair distance distribution function
This is another property that can be obtained from experimental SAXS data. In the simulations it is created as a histogram over all distances within the protein chain.

3.5.2 Complex analyses

In paper II, the study of Statherin self-association, several analyses are performed to characterise result of the self-association, that is, the formed complexes. In these analyses, two chains are regarded as being part of the same complex if the center-to-center distance between a bead in each chain is less than a certain cut-off. This geometric condition is also used for defining if two beads are in contact.

Complex size distribution and average association number
The complex size probability distribution is calculated according to

$$P_n = \frac{n\left\langle N_n^{\text{complex}}\right\rangle}{\sum_n n\left\langle N_n^{\text{complex}}\right\rangle}, \tag{3.23}$$

where $\left\langle N_n^{\text{complex}}\right\rangle$ is the average number of complexes consisting of n chains, and $\sum_n n\left\langle N_n^{\text{complex}}\right\rangle$ is equal to the number of chains in the system, since the number of chains are constant. The average association number is calculated from the complex size probability distribution, as

$$N_{\text{assoc}} = \sum_n nP_n. \tag{3.24}$$

This property can be compared to experimental results.

Contact probability
To understand which residues are responsible for the self-association, the contact probability profile along the chain is calculated. From the geometric condition mentioned above it can be decided for each bead in a chain if it is in contact with another chain. To set the strength of the short-ranged hydrophobic interaction, in addition to comparing the average association number with experimental results, the number of contacts for each chain was monitored along the simulation. The purpose of that was to avoid a too large interaction, which would have prevented chains in complexes from separating.

Radial number density profile
The internal structure of the complexes is investigated through the radial number density profile for different types of residues. This is defined as the number of beads of a certain type that are located within a shell at different distances from the center-of-mass of the complex core, divided by the shell volume. The analysis can hence provide information on if a certain bead type is more common in the core or on the surface of the complex.

Principal moments of the gyration tensor and asphericity
To determine the shape of the complex the principal moments of the gyration tensor are compared. For a perfect sphere, all three principal moments are equally large. The gyration tensor is calculated from the x, y and z-coordinates according to

$$S = \frac{1}{N} \begin{pmatrix} \sum_i^N X_i^2 & \sum_i^N X_i Y_i & \sum_i^N X_i Z_i \\ \sum_i^N X_i Y_i & \sum_i^N Y_i^2 & \sum_i^N Y_i Z_i \\ \sum_i^N X_i Z_i & \sum_i^N Y_i Z_i & \sum_i^N Z_i^2 \end{pmatrix}, \tag{3.25}$$

where $X_i = (x_i - x_{\text{com}})$ and similarly for Y and Z, and N is the number of beads in the complex. Through a transformation to a principal axis system such that

$$S = \text{diag}(R_1^2, R_2^2, R_3^2) \tag{3.26}$$

S is diagonalised and $R_1^2 \geq R_2^2 \geq R_3^2$ are the eigenvalues of S, also called the principal moments of the gyration tensor. In the simulations the ensemble averages of the eigenvalues are calculated for each complex size separately.

The asphericity is another measurement of shape calculated from the principal moments according to

$$\alpha_s = \frac{\left(\langle R_1^2 \rangle - \langle R_2^2 \rangle\right)\left(\langle R_2^2 \rangle - \langle R_3^2 \rangle\right)\left(\langle R_3^2 \rangle - \langle R_1^2 \rangle\right)}{2\left(\langle R_1^2 \rangle + \langle R_2^2 \rangle + \langle R_3^2 \rangle\right)^2}. \tag{3.27}$$

It ranges between 0 and 1, the values for a perfect sphere and a rod, respectively.

Chapter 4

Experimental methods

To validate the simulation model it is necessary to compare with experimental results. This chapter will introduce the theory behind the experimental methods I have used to study Statherin and describe how the data is analysed.

4.1 Small-angle X-ray scattering

SAXS is a commonly used technique for obtaining information on size, shape and structure for macromolecules in solution. The development of the technique started in the 1930's and the first monograph on the technique was published in 1955 [38]. Scattering at small angles normally contains information in the nanometer length scale [39].

4.1.1 Basic principle

In a SAXS experiment, a narrow beam of X-rays is sent through a sample. The X-rays interact with the electrons in the atoms, which causes the atoms to emit spherical scattered waves. The scattered waves interfere with the incoming waves, which gives rise to an interference pattern at the detector, from which structural information can be extracted. A typical set-up showing the main parts of a SAXS instrument is found in Figure 4.1.

Scattering can occur with or without the loss of energy, however, it is elastic scattering, that occurs without energy loss, that is of importance for SAXS. Both the incident beam and the scattered beam can be considered as planar waves defined by a wave vector, $\mathbf{k}_i$ and $\mathbf{k}_s$, respectively. The momentum transfer, usually referred to as the scattering vector, $\mathbf{q}$, is defined as the difference between the incident and scattered wave vector, as illustrated in Figure 4.2. The magnitude of the wave vector is $|\mathbf{k}_i| = 2\pi/\lambda$, where λ is the wavelength of the incident beam. Since there is no loss of energy in elastic scattering, $|\mathbf{k}_s| = |\mathbf{k}_i|$, hence, the

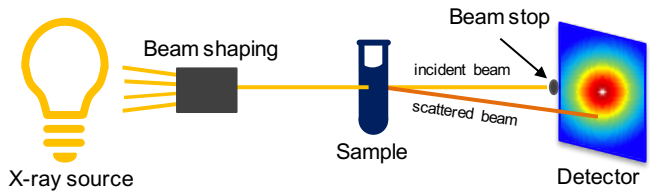


Figure 4.1: A schematic representation of the main components in a SAXS instrument. The beam stop hinders the incident beam from reaching the detector and overshadowing the sample scattering.

magnitude of **q** can be expressed as

$$q = \frac{4\pi}{\lambda} \sin(\theta), \tag{4.1}$$

where 2θ is the angle between the incident and scattered wave vector [39].

Since the X-rays are scattered due to interactions with electrons, the more electrons a sample contains, the stronger the scattering signal will be [39]. The difference in electron density throughout the sample is therefore responsible for creating the contrast. Biological macromolecules contain mostly light elements such as hydrogen and carbon, and therefore the difference in electron density compared to the aqueous solution is small. Hence, the resulting signal will be especially weak. Therefore, for biological samples, it can be advantageous to use X-rays produced from a synchrotron, a type of large circular accelerator, instead of a lab source. The synchrotron produces X-rays with much higher brilliance, which means that the exposure time needed for detecting a useful signal is much shorter, often a few seconds compared to hours. I have performed my SAXS experiments at the European synchrotron facility, ESRF, in Grenoble, France.

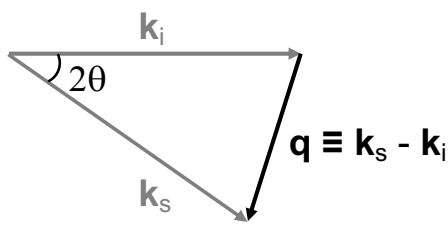


Figure 4.2: A schematic representation of the scattering vector **q**, defined by the incident wave vector $\mathbf{k}_i$ and the scattered wave vector, $\mathbf{k}_s$.

4.1.2 The scattering intensity

The detector records the scattering intensity at positions in two dimensions, however, since thermal motion causes the particles orientation to be random in respect to the incident beam, the scattering signal is a spherical average and can therefore be reduced to one dimension. The scattering intensity is usually presented as a function of q, to be independent of the wavelength. Since the scattering intensity contains information on both the single particle (intraparticle interference) and relation between different particles (interparticle interference), it is often expressed as

$$I(q) = P(q) \cdot S(q), \tag{4.2}$$

where $P(q)$ is the form factor and $S(q)$ is the structure factor. From the form factor the size and shape of the individual particle can be determined, while the structure factor contains information on the distance between particles, which can show if the particles are repelling or attracting each other. Attraction will increase the scattering curve at low q and repulsion will decrease it. In dilute systems and weakly interacting systems no structure is formed in the solution, meaning that the structure factor is a constant. Hence, at such conditions the form factor can be determined.

IDPs exhibit many different conformations, hence the scattering pattern will correspond to an average over the different conformations. Likewise, if the sample is polydisperse, so that particles of different sizes are present, the scattering curve will also provide the average.

4.1.3 Data analysis

For proteins some standard analyses which do not require any modelling are usually performed. Three relevant for IDPs are outlined below.

The Guinier approximation
The Guinier approximation provides a relation between the scattering curve at low q and the object size given by the radius of gyration, R_g, according to [40]

$$\ln I(q) = \ln I_0 - (R_g q)^2/3, \tag{4.3}$$

where I_0 is the forward scattering (the scattering signal extrapolated to $q = 0$). Usually $\ln I(q)$ is linear with respect to q^2 at small q, normally in the region $qR_g < 1.3$ for well-folded proteins, while for IDPs this region is normally reduced to $qR_g < 0.8$ [41]. If the Guinier plot shows an upswing at low q this indicates considerable aggregation in the sample, while a downswing corresponds to intermolecular repulsion. In both cases detailed analysis of the data should be avoided.

The forward scattering is related to the molecular weight by

$$M_w = \frac{I_0 \cdot N_A}{c([\rho_p - \rho_s]\nu_p)} \tag{4.4}$$

where the forward scattering I_0 is given in absolute units (cm^{-1}) and c is the protein concentration. The electron density of the protein, ρ_p, the electron density of the solvent, ρ_s, and the partial specific volume of the protein, ν_p, can all be calculated theoretically. The forward scattering is measured in arbitrary units that differs between detectors, but can be transformed to absolute units, for example by measuring the scattering of water. Normally a difference less than 10% between the measured and the theoretical weight is regarded as good [33, 42]. For self-associating proteins such as Statherin, the average association number can be calculated from the measured molecular weight. If the sample is polydisperse, it is however important to remember that this average is not a number average. The scattering from a sphere can be expressed analytically, from which it can be shown that in the $q \to 0$ limit, $I \propto R^6$, where R is the sphere radius [39]. Hence, large particles contribute more to the average than small particles.

Pair distance distribution function

The pair distance distribution function, $P(r)$ provides information on shape, since it shows the distribution of pair distances within the protein. It is expressed in real space, compared to the scattering pattern that contains information in inverse space. $I(q)$ and $P(r)$ are related by a Fourier transform, according to [39]

$$P(r) = \frac{1}{2\pi^2} \int_0^\infty I(q) qr \sin(qr) \mathrm{d}q. \tag{4.5}$$

Since $I(q)$ is not known over the full interval $0 \leq q \leq \infty$, $P(r)$ can not be obtained directly, hence an indirect Fourier transformation method [43, 44] is often used instead.

The $P(r)$ provides easy differentiation between globular and unfolded proteins, such as IDPs. For a globular protein, the $P(r)$ is a symmetric bell-shaped curve, while for unfolded proteins the $P(r)$ shows an extended tail. If a protein has multiple domains it can be detected in the $P(r)$ as two different peaks.

R_g and I_0 can also be calculated from $P(r)$, by using the equations below

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 P(r) \mathrm{d}r}{2 \int_0^{D_{max}} P(r) \mathrm{d}r} \tag{4.6}$$

$$I_0 = 4\pi \int_0^{D_{max}} P(r) \mathrm{d}r, \tag{4.7}$$

where D_{max} is the maximum distance within the protein [41]. Since the Guinier method only uses a small region of the scattering curve, especially for IDPs, while $P(r)$ is based on more or less the whole curve, R_g obtained from $P(r)$ is usually more reliable [45].

Kratky plot

To assess the flexibility of a protein and differentiate between globular and

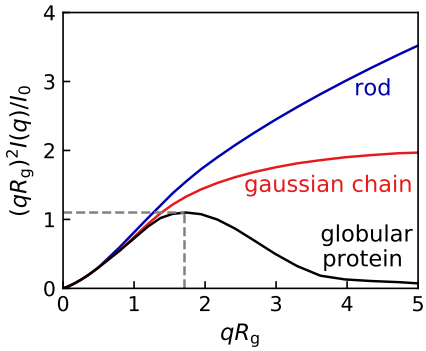


Figure 4.3: A dimensionless Kratky plot showing the different behaviour of a rigid rod, a gaussian chain and a globular protein. A globular protein shows a maximum at approximately $(\sqrt{3}, 1.1)$.

disordered proteins the Kratky plot is useful. A dimensionless Kratky plot allows for comparison between proteins of different sizes, and is constructed as $(qR_g)^2 I(q)/I_0$ vs qR_g [46]. Figure 4.3 illustrates the different behaviour of a rigid rod, a gaussian chain and a globular protein. An intrinsically disordered protein usually exhibits a plateau as the gaussian chain, while the actual slope varies depending on for example the amount of partial structure.

4.1.4 Size exclusion chromatography-coupled SAXS

A size exclusion chromatography (SEC) column is used for separating a sample according to size. A SEC column usually contains porous beads that allow small molecules to travel into the bead pores, while large objects only moves in between the beads. Hence, smaller objects travel a longer route and will be eluted later than large objects. A SEC column can therefore be used in-line with SAXS to separate the sample according to size and measure SAXS on it directly as it is eluted. For poly-disperse samples one should therefore be able to obtain SAXS spectra for the different sized objects individually and hence obtain a size distribution. SEC-SAXS is also useful if the sample is very prone to aggregate, since large aggregates will be eluted first and the monomeric protein afterwards.

4.2 Circular dichroism spectroscopy

Circular dichroism (CD) spectroscopy is a technique based on the adsorption of polarised light and reveals information on the secondary structure content in proteins.

4.2.1 Basic principle

Light is a form of electromagnetic radiation that comprises an electric field and a magnetic field. These fields oscillate in perpendicular planes, and are also perpendicular to the direction of propagation. Normally light is unpolarised, which means that it oscillate in all possible directions. In linearly polarised light, the oscillations are restricted to only one direction, as illustrated in Figure 4.4. In circularly polarised light, the electric vector rotates around the direction of propagation, undergoing a full revolution per wavelength. Clockwise rotation corresponds to right circularly polarised light, and counterclockwise to left circularly polarised light [47].

Linearly polarised light can be viewed as being made up by two components of circularly polarised light of equal magnitude and phase, rotating in opposite directions (left and right), as illustrated in Figure 4.5a. If the two components are of different amplitudes, the light will be elliptically polarised, as the electric vector instead will trace an ellips, see Figure 4.5b. This is what happens during a CD spectroscopy experiment, as an optically active sample will absorb the left and right circularly polarised light to different extents. In an experiment the difference in absorption of left and right circularly polarised light is monitored for different wavelengths [48].

To obtain a CD signal from a sample, there must occur absorption. A group that absorbs light is called a chromophore and for the sample to be optically active the chromophores need to be either chiral, covalently linked to a chiral centre or situated in a chiral environment due to the three-dimensional structure of the molecule, which is the case for the peptide backbone. In a protein, the chromophores of largest interest are the peptide bond, aromatic amino acid side chains and the disulphide bond. Different secondary structure give rise to characteristic patterns in the far-UV region (approximately 170-250 nm), which is where mostly the peptide bond absorbs. For the peptide bond two electronic transitions are possible, one around 190 nm and one around 220 nm [47].

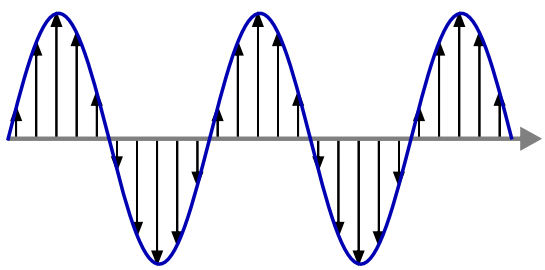


Figure 4.4: An illustration of linearly polarised light. The grey arrow corresponds to the direction of propagation and the black arrows represent the electric vector at different points along the propagation. The magnetic field (not shown) is perpendicular to the electric field.

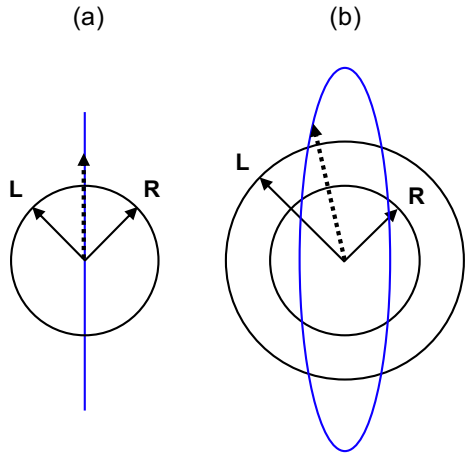


Figure 4.5: (a) Linearly polarised light made up by two components of circularly polarised light L and R rotating in opposite direction. The dashed arrow represents the electric vector corresponding to the sum of the two components and is always oriented along the blue line. (b) Different amplitude of the two components causes the electric vector (dashed arrow) to trace an ellipse (outlined in blue).

4.2.2 Data analysis

The magnitude of the CD signal depends on the sample concentration and the path length. For historic reasons the spectrum is usually presented in terms of ellipticity, which has the unit degrees, and not as a difference in absorbance. The ellipticity is calculated from the major and minor axes of the resulting ellipse and has a numerical relationship to the absorbance. To be able to compare different measurements, the signal needs to be normalised with respect to concentration and path length. One common way to do that is by expressing the signal as the mean residue ellipticity (unit: deg·cm^2·dmol^{-1}), calculated as

$$[\theta]_{\mathrm{MRW}} = \theta \cdot \mathrm{MRW}/(10 \cdot d \cdot c), \tag{4.8}$$

where θ is the observed ellipticity (mdeg), d the path length of the cell (cm) and c the protein concentration (mg/mL). The mean residue weight, MRW, is the molecular weight (Da) divided by the number of peptide bonds [48].

Different secondary structure shows different CD spectra, which is illustrated in Figure 4.6 [48, 49]. There are several available algorithms for assessing the secondary structure composition, that are usually based on the approximation that a given protein CD spectrum can be expressed as a linear combination of spectra of different secondary structure components [49]. A certain secondary structure still have structural variability, so the result varies with the used reference proteins, especially for the assessment of random coil. Since random coil is not a well-defined structure, CD spectra of mostly random coil structure can

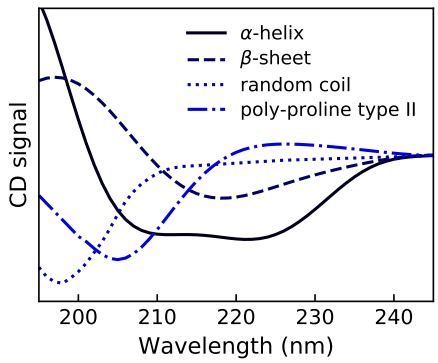


Figure 4.6: A schematic sketch of CD spectra for different secondary structure.

vary a lot. Hence, the determined composition might be highly uncertain and therefore it can be enough to only make a crude assessment from the shape of the CD spectra, especially for IDPs.

Chapter 5

The research

Below the main results of the two papers included in this thesis are summarised. Paper I describes a coarse-grained model and its applications for monomeric IDPs, while Paper II provides an extension of the model for the investigation of Statherin, a self-associating IDP. In both papers small angle x-ray scattering is the main experimental technique used for validation of the model.

5.1 Paper I

Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions

In this paper we tested the performance of the coarse-grained model, parameterised for Histatin 5, on a range of intrinsically disordered proteins and regions with an overall low average hydrophobicity. We showed that the agreement between the radius of gyration determined in simulations and by SAXS experiments reported in the literature was good for this group of proteins, although a high number of phosphorylations caused the simulated proteins to be overly contracted. In the simulations a preference for short distances between positively charged and phosphorylated residues were detected.

In addition, it was shown that the radius of gyration for the proteins fit nicely to the exponential law for polymers, with an exponent of 0.59 indicating behaviour similar to a self-avoiding random walk. Moreover, the proteins with a length of 73 amino acids or more were shown to have a significant response to changes in ionic strength, and the behaviour was controlled by the charge distribution, such that some proteins behaved as polyelectrolytes and others as polyampholytes. Both the chain length and ionic strength affected the conformational ensemble of the protein.

Another aspect worth highlighting is that this paper presents the form factor for monomeric Statherin and shows that the simulated and experimental form factor is in excellent agreement, which is of importance for paper II.

5.2 Paper II

Assessing the Intricate Balance of Intermolecular Interactions upon Self-Association of Intrinsically Disordered Proteins

This study is focused around the self-association process of Statherin. The overall aim was to investigate the balance of interactions in an attractive IDP system and develop the coarse-grained model to account for more complex systems. Statherin was regarded as a good model system for its amphiphilic character and relatively short chain length, which makes modelling feasible. We used the model presented in paper I since it was shown to work well for monomeric Statherin, although an additional short-ranged attraction was added to induce self-association. In addition, we performed a more detailed investigation of the monomeric behaviour by SAXS and CD spectroscopy. It showed that urea increases stiffness in the protein by increasing the amount of poly-proline II structure and that increased temperature causes a loss of poly-proline II structure.

For the self-association, the original model was adjusted by comparing to SAXS measurements. Both SAXS and simulations showed that electrostatic repulsion and increased temperature repress complex growth. Hydrophobic interaction was regarded as the main driving force, since high concentrations of urea terminated the self-association. By performing simulations without electrostatic interactions, the temperature effect was suggested to be of entropic origin. The simulations also provided a description of the size distribution in the system, showing that the monomeric specie dominates. In addition, the shape of the complexes and the effect of mutations were investigated through the use of simulations, showing the benefits of a combined experimental and theoretical approach.

5.3 Outlook

The overall objective of the research performed in this thesis has been to contribute to the development of models suitable for IDPs. Modelling and simulations are an important complement to experimental techniques in understanding the behaviour and functions of IDPs. The focus has been on bulk behaviour, more specifically for Statherin. Statherin acts a good model system, but are also of individual interest, since it is a saliva protein with functions important for oral health. Since Statherin has been shown to bind to hydroxyapatite, the main constituent of tooth enamel, and moreover, is one of the main contributors to the acquired enamel pellicle [50–52], my continued research will also expand into

surface adsorption of Statherin. For the self-association, I would like to continue investigating the role of specific amino acids by studying different mutations. Obtaining an experimental measurement of the size distribution is also of interest.

References

1. Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Biochemistry*, international 7th ed.; W. H. Freeman and Company: New York, USA, 2011.

2. Dill, K. A. Dominant forces in protein folding. *Biochemistry* **1990**, *29*, 7133–7155.

3. Dunker, A. K. et al. Intrinsically disordered protein. *J. Mol. Graphics Modell.* **2001**, *19*, 26–59.

4. Uversky, V. N. Unusual biophysics of intrinsically disordered proteins. *Biochim. Biophys. Acta, Proteins Proteomics* **2013**, *1834*, 932–951.

5. Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Sequence complexity of disordered protein. *Proteins: Struct., Funct., Bioinf.* **2001**, *42*, 38–48.

6. Vucetic, S.; Brown, C. J.; Dunker, A. K.; Obradovic, Z. Flavors of protein disorder. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 573–584.

7. Dunker, A. K.; Romero, P.; Obradovic, Z.; Garner, E. C.; Brown, C. J. Intrinsic Protein Disorder in Complete Genomes. *Genome Inform.* **2000**, *11*, 161–171.

8. Romero, P.; Obradovic, Z.; Kissinger, C.; Villafranca, J.; Garner, E.; Guilliot, S.; Dunker, A. Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* **1998**, *3*, 437–448.

9. Ward, J.; Sodhi, J.; McGuffin, L.; Buxton, B.; Jones, D. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337*, 635–645.

10. Xue, B.; Dunker, A. K.; Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **2012**, *30*, 137–149.

11. Wright, P. E.; Dyson, H. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321 – 331.

12. Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.

13. Uversky, V. N.; Dunker, A. K. Understanding protein non-folding. *Biochim. Biophys. Acta, Proteins Proteomics* **2010**, *1804*, 1231 – 1264.

14. Tompa, P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.* **2012**, *37*, 509 – 516.

15. Liu, J.; Faeder, J. R.; Camacho, C. J. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19819–19823.

16. Edgar, M., Dawes, C., O'Mullane, D., Eds. *Saliva and Oral Health*, 3rd ed.; British Dental Association: London, UK, 2004.

17. Levine, M. J. Development of artificial salivas. *Critical Reviews in Oral Biology & Medicine* **1993**, *4*, 279–286.

18. Moreno, E.; Zahradnik, R. Demineralization and Remineralization of Dental Enamel. *JJ. Dent. Res.* **1979**, *58*, 896–903.

19. Hay, D.; Smith, D.; Schluckebier, S.; Moreno, E. Basic Biological Sciences Relationship between Concentration of Human Salivary Statherin and Inhibition of Calcium Phosphate Precipitation in Stimulated Human Parotid Saliva. *J. Dent. Res.* **1984**, *63*, 857–863.

20. Buzalaf, M. A.; Hannas, A. R.; Kato, M. T. Saliva and dental erosion. *J. Appl. Oral Sci.* **2012**, *20*, 493–502.

21. Douglas, W. H.; Reeh, E. S.; Ramasubbu, N.; Raj, P. A.; Bhandary, K. K.; Levine, M. J. Statherin: A major boundary lubricant of human saliva. *Biochem. Biophys. Res. Commun.* **1991**, *180*, 91 – 97.

22. Gibbons, R. J.; Hay, D. I. Human salivary acidic proline-rich proteins and statherin promote the attachment of Actinomyces viscosus LY7 to apatitic surfaces. *Infect. Immun.* **1988**, *56*, 439–445.

23. Schlesinger, D. H.; Hay, D. I. Complete Covalent Structure of Statherin, a Tyrosine-rich Acidic Peptide Which Inhibits Calcium Phosphate Precipitation from Human Parotid Saliva. *J. Biol. Chem.* **1977**, *252*, 1689–1695.

24. Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.

25. Lin, Y.; Currie, S. L.; Rosen, M. K. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J. Biol. Chem.* **2017**, *292*, 19110–19120.

26. Pak, C. W.; Kosno, M.; Holehouse, A. S.; Padrick, S. B.; Mittal, A.; Ali, R.; Yunus, A. A.; Liu, D.; Pappu, R. V.; Rosen, M. K. Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. *Mol. Cell* **2016**, *63*, 72–85.

27. Israelachvili, J. N. *Intermolecular and Surface Forces*, 3rd ed.; Academic Press, Elsevier: Oxford, UK, 2011.

28. Chandler, D. Hydrophobicity: Two faces of water. *Nature* **2002**, *417*, 493–502.

29. Evans, M. T. A.; Phillips, M. C.; Jones, M. N. The conformation and aggregation of bovine β-casein A. II. Thermodynamics of thermal association and the effects of changes in polar and apolar interactions on micellization. *Biopolymers* **1979**, *18*, 1123–1140.

30. Takase, K.; Niki, R.; Arima, S. A sedimentation equilibrium study of the temperature-dependent association of bovine β-casein. *Biochim. Biophys. Acta, Proteins Proteomics* **1980**, *622*, 1 – 8.

31. O'Connell, J.; Grinberg, V.; de Kruif, C. Association behavior of β-casein. *J. Colloid Interface Sci.* **2003**, *258*, 33 – 39.

32. Portnaya, I.; Cogan, U.; Livney, Y. D.; Ramon, O.; Shimoni, K.; Rosenberg, M.; Danino, D. Micellization of Bovine β-Casein Studied by Isothermal Titration Microcalorimetry and Cryogenic Transmission Electron Microscopy. *J. Agric. Food Chem.* **2006**, *54*, 5555–5561, PMID: 16848545.

33. Moitzi, C.; Portnaya, I.; Glatter, O.; Ramon, O.; Danino, D. Effect of Temperature on Self-Assembly of Bovine β-Casein above and below Isoelectric pH. Structural Analysis by Cryogenic-Transmission Electron Microscopy and Small-Angle X-ray Scattering. *Langmuir* **2008**, *24*, 3020–3029.

34. Hill, T. L. *An Introduction to Statistical Thermodynamics*, 2nd ed.; Addison-Wesley Publishing Company: Reading, MA, USA, 1962.

35. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

36. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, CA, USA, 2002.

37. 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: Struct., Funct., Bioinf.* **2016**, *84*, 777–791.

38. Guinier, A.; Gérard, F. *Small-Angle Scattering of X-rays*; John Wiley & Sons, Inc.: New York, USA, 1955.

39. Svergun, D. I.; Koch, M. H. J.; Timmins, P. A.; May, R. P. *Small Angle X-ray and Neutron Scattering from Solutions of Biological Macromolecules*, 1st ed.; Oxford University Press: Oxford, UK, 2013.

40. Guinier, André, La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. *Ann. Phys.* **1939**, *11*, 161–237.

41. Receveur-Bréchot, V.; Durand, D. How Random are Intrinsically Disordered Proteins? A Small Angle Scattering Perspective. *Curr. Protein Pept. Sci.* **2012**, *13*, 55–75.

42. Orthaber, D.; Bergmann, A.; Glatter, O. SAXS experiments on absolute scale with Kratky systems using water as a secondary standard. *J. Appl. Crystallogr.* **2000**, *33*, 218–225.

43. Glatter, O. Data evaluation in small angle scattering: calculation of the radial electron density distribution by means of indirect Fourier transformation. *Acta Phys. Austriaca* **1977**, *47*, 83–102.

44. Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **1992**, *25*, 495–503.

45. Jacques, D. A.; Trewhella, J. Small-angle scattering for structural biology—Expanding the frontier while avoiding the pitfalls. *Protein Sci.* **2010**, *19*, 642–657.

46. Durand, D.; Vivès, C.; Cannella, D.; Pérez, J.; Pebay-Peyroula, E.; Vachette, P.; Fieschi, F. NADPH oxidase activator $p67^{phox}$ behaves in solution as a multidomain protein with semi-flexible linkers. *J. Struct. Biol.* **2010**, *169*, 45 – 53.

47. Miles, A.; Wallace, B. In *Biophysical Characterization of Proteins in Developing Biopharmaceuticals*; Houde, D. J., Berkowitz, S. A., Eds.; Elsevier: Amsterdam, 2015; pp 109 – 137.

48. Kelly, S. M.; Jess, T. J.; Price, N. C. How to study proteins by circular dichroism. *Biochim. Biophys. Acta, Proteins Proteomics* **2005**, *1751*, 119 – 139.

49. Sreerama, N.; Woody, R. W. *Numerical Computer Methods, Part D*; Methods in Enzymology; Academic Press, 2004; Vol. 383; pp 318 – 351.

50. Hay, D. The isolation from human parotid saliva of a tyrosine-rich acidic peptide which exhibits high affinity for hydroxyapatite surfaces. *Arch. Oral Biol.* **1973**, *18*, 1531 – 1541.

51. Hay, D. The interaction of human parotid salivary proteins with hydroxyapatite. *Arch. Oral Biol.* **1973**, *18*, 1517 – 1529.

52. Smith, A. V.; Bowen, W. In situ studies of pellicle formation on hydroxyapatite discs. *Arch. Oral Biol.* **2000**, *45*, 277 – 291.

Acknowledgements

First I want to thank my supervisor Marie, who first introduced me to the world of simulations, by taking me in as a student for a summer project. Thank you for all the guidance and support during the years. I also want to thank all the members in our group, for good discussions and nice chatting. A special thanks to Carolina for teaching me how to work experimentally with the proteins and introducing me to SAXS.

Thank you to all colleagues at the division of Theoretical Chemistry and the division of Physical Chemistry for making this such a nice environment to work in. Especially I am thankful to the people involved in the Teokem Thursday evenings, for the nice times and introducing me to so many board games. I also want to thank Jan for reading and giving feedback on this thesis.

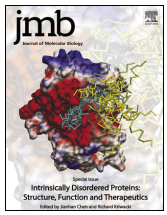
A special thanks to Emil for being a good friend and my personal experimental expert, always willing to answer questions and discuss things. I also want to thank my family for supporting me. Lastly, I am truly grateful to Max for always being by my side, encouraging me and being my programming expert.

Paper I

Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions

C. Cragnell, E. Rieloff, M. Skepö.

Journal of Molecular Biology, 430, **2018**, pp. 2478–2492.



Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions

Carolina Cragnell, Ellen Rieloff and Marie Skepö

Division of Theoretical Chemistry, *Department of Chemistry, Lund University, P.O. Box 124, SE-221 00 Lund, Sweden*

Correspondence to Marie Skepö: *marie.skepo@teokem.lu.se*
https://doi.org/10.1016/j.jmb.2018.03.006
Edited by Jianhan Chen

Abstract

In this study, we have used the coarse-grained model developed for the intrinsically disordered saliva protein (IDP) Histatin 5, on an experimental selection of monomeric IDPs, and we show that the model is generally applicable when electrostatic interactions dominate the intra-molecular interactions. Experimental and theoretically calculated small-angle X-ray scattering data are presented in the form of Kratky plots, and discussions are made with respect to polymer theory and the self-avoiding walk model. Furthermore, the impact of electrostatic interactions is shown and related to estimations of the conformational ensembles obtained from computer simulations and "Flexible-meccano." Special attention is given to the form factor and how it is affected by the salt concentration, as well as the approximation of using the form factor obtained under physiological conditions to obtain the structure factor.

Introduction

Intrinsically disordered proteins and regions (IDPs and IDRs), from now on referred to as IDPs, are characterized by a lack of stable tertiary structure when the proteins exist as isolated polypeptide chains under physiological conditions *in vitro* [1,2]. More recently, it has been shown that ~30% of all proteins in eukaryotic organisms belong to this group of proteins, and that IDPs are involved in a large number of central biological processes and diseases. This discovery challenged the traditional protein structure paradigm, which stated that a specific well-defined structure was required for the correct function of a protein. Biochemical evidence has since shown that IDPs are functional, and that the lack of folded structures is related to their functions [3,4].

There is a great interest in the research community in the structure–function relationship for IDPs, and one hypothesis is that upon adsorption to surfaces, IDPs might adopt a structure, which gives rise to a function. Hence, for that purpose it is of interest to relate the properties of IDPs in solution with their properties in the adsorbed state, as well as their interaction with biological membranes. To be able to obtain a molecular understanding of macromolecules, it is useful to combine experimental techniques with atomistic and coarse-grained modeling. There have been great advances regarding atomistic simulations of IDPs, with the development and justification of force fields and water models, where the results have been validated against experimental results such as Förster resonance energy transfer, small-angle X-ray scattering (SAXS), and NMR. The reader is referred to the literature for more information [5–10]. The advantages of atomistic simulations are that one uses a full-atom approach and takes the water into account explicitly, whereas the limitation is that one is restricted to relatively short proteins due to the system size and computational power.

To be able to model longer proteins and more complex systems, coarse-grained modeling and Monte Carlo/molecular dynamics simulations are a good alternative. Of course, there will be approximations and simplifications; nevertheless, the approach has been shown to work very well. For more than 30 years, a coarse-grained model based on the primitive model [11], in combination with Monte Carlo simulations, has been used to model

Table 1. Details of the proteins within this study in terms of the length of the amino acid sequence, the number of phosphorylated residues (N_{phos}), the FCR, the NCPR, the percentage of prolines, and the number of hydrophobic residues (N_{hphob}). Furthermore, both the radii of gyration (R_g) obtained from experiments and simulations are included.

	Length	N_{phos}	FCR	NCPR	% Prolines	N_{hphob}	$R_{g,\ SAXS}$ (Å)	$R_{g,\ Sim}$ (Å)
Hst 5_{4-15} [16]	12	0	0.42	+0.42	0	2	9.2 ± 0.1	9.64 ± 0.02
Hst 5 [12]	24	0	0.38	+0.21	0	2	13.8 ± 0.1	13.77 ± 0.44
IB5 [15]	73	0	0.11	+0.08	40	5	27.9 ± 1.0	26.01 ± 0.05
Ash1 [13]	83	0	0.20	+0.18	15	12	28.4 ± 3.4	29.56 ± 0.02
Sic1 [14]	92	0	0.12	+0.12	16	20	28.8 ± 1.2	30.71 ± 0.05
Il-1ng [15]	141	0	0. 19	+0.11	36	2	41.1 ± 1.0	38.24 ± 0.07
RNase E [17]	248	0	0.39	+0.05	6	55	52.6 ± 0.3	48.52 ± 0.11
Phosphorylated IDPs								
Statherin,	43	2	0.28	−0.09	16.3	7	19.3 ± 0.2	18.05 ± 0.05
pAsh1 [13]	83	10	0.45	−0.06	14.5	12	27.5 ± 1.2	21.76 ± 0.02
pSic1 [14]	92	6	0.25	−0.01	16.3	20	32.2 ± 2.2	27.55 ± 0.05

The experimental R_g values for Sic1 and pSic1 were determined using SAXS data obtained from the Protein Ensemble Database [14], and the Guinier approach.

polyelectrolytes and polyampholytes under various conditions. Sometimes this model is also referred to as the bead-necklace model. In this model, each monomer corresponds to a bead of a certain radius that can also have a charge associated with it. The water is always treated as a dielectric continuum.

In this study, we have used the coarse-grained model developed for the intrinsically disordered saliva protein Histatin 5 [12], on an experimental pool of IDPs obtained from different sources [13–18], as well as new experimental SAXS data for Statherin, also a saliva protein. We show that the model is generally applicable when electrostatic interactions dominate the intra-molecular interactions. For consistency, the reader should notice that we restrict our comparisons to experimental data obtained from SAXS. Focus will be on experimental and theoretically calculated SAXS data presented as Kratky plots, as well as comparison with polymer theory and the self-avoiding random walk (SARW) model. Furthermore, the impact of electrostatic interactions is shown and related to estimations of the conformational ensembles obtained from computer simulations and Flexible-meccano [19].

Results and Discussion

Polymer Model

The aim is to investigate if there exists a general coarse-grained model that accurately captures the structural properties of IDPs at both *high* and *low* salt concentrations. To assure the generality, the model developed for Histatin 5 [12] will be utilized on an experimental pool of IDPs covering a sequence length from 12 to 248 amino acids, and we will only compare the finding with experimental SAXS data. The IDPs have been characterized according to Das *et al.* [20], using the concepts: net charge per residue (NCPR), and fraction of charged residues (FCR), where NCPR = ($f_+ - f_-$) and FCR = ($f_+ + f_-$), with f being the fraction of positive/negative charges. According to this approach, polyampholytes and polyelectrolytes can be characterized to be either strong or weak, where FCR $\geq$ 0.3 corresponds to the former and FCR $<$ 0.3 to the latter. Moreover, they can be neutral, that is, NCPR $\approx$ 0, or have a net charge. Polyampholytes have approximately an equivalent fraction of opposite charges; thus, NCPR is low, whereas polyelectrolytes have more of one type of charge. The proteins used in this study are summarized in Table 1. As shown, although the selection of proteins might seem small, a fairly representative pool of IDPs is given with respect to the charges, the number of phosphorylated residues (N_{phos}), the number of hydrophobic amino acids (N_{hphob}), and the proline content. The number of hydrophobic residues is based on the notion that all amino acids with a higher hydropathy value than glycine in the Kyte–Doolittle scale [21], are considered hydrophobic.

The level of compaction/extension has been analyzed by comparing the radius of gyration (R_g) from SAXS with the corresponding analysis obtained from Monte Carlo simulations, that is, comparison of ensemble-averaged estimates as well as the full conformational ensemble through the probability distribution. Fig. 1a displays the radii of gyration from the simulations *versus* the experimental counterparts. As is clearly shown, there is a good correspondence between the ensemble estimates. However, there are proteins that display simulated radii of gyration that are statistically different from the experimental data; moreover, the experimental data are more extended than the model predicts, that is: RNase E, two of the phosphorylated proteins, namely, pAsh1, and pSic1, as well as the proline-rich protein Il-1ng. For RNase E, we hypothesize that it is due to a slight degree of self-association; for pAsh1 and pSic1, we expect it to be

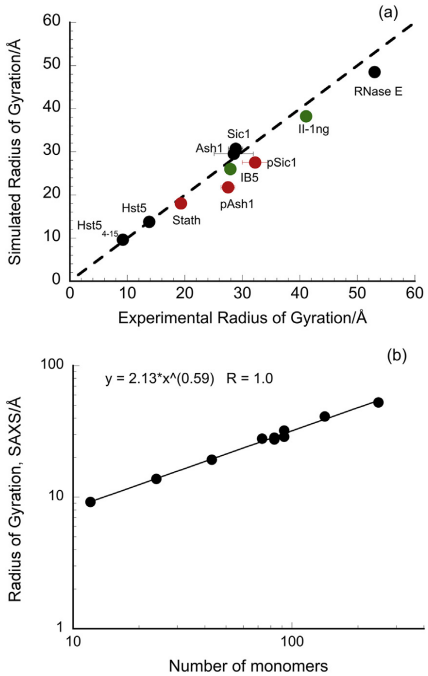


Fig. 1. (a) Radii of gyration obtained from simulations *versus* the radii of gyration obtained from experiments where black filled circles correspond to non-phosphorylated IDPs, red filled circles to phosphorylated proteins where the phoshate group is assumed to have a net charge of $-2e$, and green filled circles to proline-rich proteins. (b) The experimental radii of gyration as a function of the protein sequence length on log–log scale. The ionic strength corresponds to 150 mM, except for IB5 and Il-1ng, where it was 50 mM. For most of the reported values, the precision is smaller than the marker in the plot; hence, the reader is referred to Table 1 for more information.

due to the high number of phosphorylated residues, whereas for Il-1ng it is due to the proline content which, due to the cyclic structure of the amino acid, gives the proline an exceptional conformational rigidity. Nevertheless, the reader should notice that the radii of gyration for the proline-rich proteins do agree remarkably well.

For some polymers, such as the well-known polymer polyethylene glycol, it is possible to define an empirical expression for a simplistic estimation of the R_g [22], according to the power-law $R_g = \rho_0 N^{\upsilon}$. In this context, υ refers to the Flory exponent, which depends on the structural behavior of the polymer chain in the solvent, N refers to the number of monomers in the chain, and ρ_0 is a prefactor. The latter is a function of, among other things, the details of the monomer as the radius, the persistence length, and the bond geometry. This leads to the question: Is it possible to define a similar expression for IDPs as for polyethylene glycol? For a random walk (also denoted ideal chain), the parameter υ is equal to 0.5, whereas it is approximately 0.6 for a SARW [23]. In the latter, the interactions between the chain monomers (or for IDPs, the amino acids), are modeled as excluded volumes, which cause a reduction in the conformational possibilities of the chain, in comparison with a random walk where all bonds and torsion angles are equally probable. In Fig. 1b, the experimentally obtained radii of gyration (from SAXS) of our selection of model proteins are shown as a function of sequence length. From the fit to the curve, υ is estimated to be approximately 0.59, which matches closely the exponent obtained from the computer simulations ($\upsilon = 0.58$), where only excluded volumes are taken into account (data not shown). Hence, it seems that the selection of IDPs used in this study behave as SARWs under the given solution conditions, that is, high ionic strength. This is a reasonable conclusion when electrostatic interactions dominate the intra-chain interactions, which can be highly screened by the large amount of salt present in the solution. This rationale is further verified since the fractions of hydrophobic residues of the used IDPs are rather low, $\leq$20% (see Table 1).

By fitting the experimentally obtained radii of gyration as a function of the number of amino acids for the proteins used in this study, we obtain a prefactor ρ_0 of approximately 2.13, which is in good agreement with the model in the computer simulations where the radius of the amino acids is set to 2 Å. In the literature, the Flory exponent varies between $\upsilon = 0.5$ and 0.6 depending on the technique (Förster resonance energy transfer or SAXS), protein, and solvent used, that is, in the latter with or without denaturing agents [24–30]. This is plausible since the Flory exponent is sensitive to the intramolecular interactions in the protein, thus the amino acid composition. A more hydrophilic protein with a low fraction of hydrophobic amino acids will obtain the higher value of the Flory exponent, whereas the opposite occurs if the fraction of charges is low and the number of hydrophobic amino acids is high, where the latter has been reported by Hofmann *et al.* [27]. It is very interesting to notice though that hydrophobic disordered proteins are expanded in water, as reported for example by Riback *et al.* [31]. In the latter, the authors of this paper hypothesize that the decrease in the Flory exponent might be due to the hydrophobic effect; that is, the final conformational state is driven by the total minimization of the hydrophobic surface, which manifests itself as an effective attractive force. Notice also that the statistical basis in all experimental studies presented is rather low; hence, the shape of

the curve is rather sensitive to the addition of a further IDP.

As is well known, an IDP can exist in an infinite number of spatial states due to its high flexibility and fast dynamics. To obtain more information about the conformational averages, the Monte Carlo simulation technique is invaluable since it gives the Boltzmann-weighted probability of finding a system in a specific state. The properties of IDPs are of course dependent on different parameters such as the amino acid sequence and the temperature, as well as the solution properties. It has been shown in several papers [27,28,30], and above, that the IDPs can be considered to behave as SARWs when only steric interactions are taken into account due to high salt concentration or the presence of a denaturing agent. The next question is: How does the chain length affect the conformational ensemble average under such conditions? For this purpose, we have analyzed the full width half maximum (FWHM) and peak position of the probability distribution function of the radius of gyration and the shape of the adopted conformations using our model protein without charges, that is, considering only steric interactions. As expected and shown in Fig. 2, the ensemble of possible conformations increases as a function of the number of amino acids; cf. R_g spans from 10 to 35 Å, and from 40 to 130 Å, for 50- and 500-amino-acid monomers, respectively. By analyzing the FWHM as a function of the number of amino acids in the protein sequence, an estimate of the conformational entropy of the model protein can be obtained such that the broader the peak, the larger the chain entropy. The FWHM and the peak position as a function of protein length show the same $\upsilon \approx 0.6$ scaling behavior as the radius of gyration (data not shown).

The shape of the IDP can be defined as the ratio of the mean-square end-to-end distance, $\langle R_{ee}^2 \rangle^{1/2}$, and the mean-square radius of gyration $\langle R_g^2 \rangle^{1/2}$ (also denoted R_{ee} and R_g) according to: $r_{shape} = \langle R_{ee}^2 \rangle / \langle R_g^2 \rangle$. In the rod-like limit, $r_{shape} = 12$; for a flexible chain in good solvent, $r_{shape} \approx 6.3$; and for an ideal chain, $r_{shape} = 6$. For all chain lengths, the shape probability distribution is a symmetric bell-shaped function with a broad maximum of only 0.15 at $r_{shape} = 6$. The latter number indicates that a specific average conformation occurs during 15% of the simulation length (data not shown). Hence, there is a relatively high probability to accommodate all the different possible shapes, for example, from a rather contracted chain to a rigid prolate. Notice that $r_{shape} = 1$ does not necessarily indicate that an IDP is a compact globule, rather that the chain is contracted and that the mean-square end-to-end distance and the mean-square radius of gyration are of the same order.

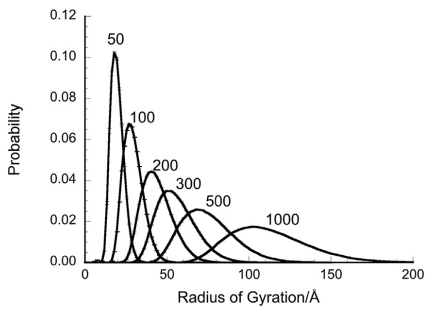


Fig. 2. The conformational ensemble of radius of gyration for different lengths of the model protein, where only steric interactions through excluded volumes are taken into account.

The effect of electrostatic interactions on the single molecular level

The impact of electrostatic interactions at the single molecular level on the conformational ensemble of IDPs, and how it affects the scattering spectra, visualized as Kratky plots, has also been investigated. Of particular interest is when the ionic strength is 150 mM, since that is commonly applied in SAXS experiments to determine the form factor. Here, the study has been divided into two parts: (i) non-phosphorylated and (ii) phosphorylated proteins.

Non-phosphorylated IDPs

Fig. 3 shows the obtained radii of gyration calculated from simulations at 10 mM and 150 mM salt, which corresponds to Debye screening lengths

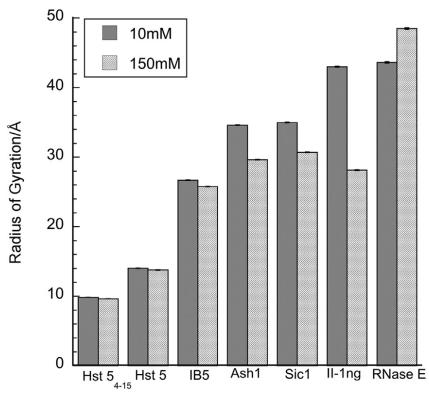


Fig. 3. The simulated radii of gyration of the chosen IDPs at high and low ionic strength (150 and 10 mM).

(κ^{-1}) of approximately 30 and 8 Å, respectively. As shown, it is clearly visible that upon the addition of salt, some proteins attain polyelectrolytic behavior, whereas other proteins exhibit polyampholytic behavior. In the former, the protein contracts, whereas in the latter, it becomes more extended when the salt concentration is increasing. Moreover, a clear trend is also obtained with respect to the chain length; that is, the screening effect is more accentuated for longer proteins, which induces larger discrepancies in the estimated extensions. Hence, in this respect, the charge distribution obtained from the specific amino acid sequence and the protein length due to the higher probability to attain a larger population of conformations are of importance.

The effect of salt on R_g and the conformational ensemble has been further analyzed focusing on the protein $Ash1_{420–500}$ (hereafter referred to as Ash1). This protein has been extensively studied in the paper by Martin *et al.* [13]. Among other things, they showed that Ash1 adopts coil-like conformations that are expanded and well solvated. The R_g for Ash1 from experiments and modeling with and without charges at different ionic strengths are given in Table 2. There is a clear trend in the simulated R_g, which decreases as a function of salt concentration. The SAXS measurements (150 mM salt) gave an R_g of 28.5 ± 3.4 Å, which means that all simulated radii of gyration except the one obtained at 10 mM salt are within the uncertainty. The simulations show that the conformational properties of SARW are reached first upon the addition of 1000 mM salt, that is, when the Debye screening length is shorter than the average bead-to-bead distance in the model, cf. 3.04 Å for the former with 4.1 Å for the latter. The reader should notice that the more dramatic effects occur, of course, in the lower salt regime, for example, between 10 and 150 mM salt. These results are clearly shown in the probability distribution of the conformational ensemble as given in Fig. 4a. Notice that a small change in the ensemble average will affect the conformational ensemble more remarkably, and that the electrostatic interactions within the chain are quite pronounced

Table 2. Conformational properties and the FWHM of the IDR in Ash1 as a function of salt obtained from simulation.

I (mM)	κ^{-1} (Å)	R_g (Å)	R_{ee} (Å)	FWHM (Å)
10	30.4	34.54 ± 0.01	88.43 ± 0.05	13.70 ± 0.10
150	7.9	29.56 ± 0.02	74.33 ± 0.05	12.91 ± 0.18
300	5.6	28.68 ± 0.02	71.99 ± 0.05	12.71 ± 0.19
500	4.3	28.19 ± 0.02	70.69 ± 0.06	12.63 ± 0.20
1000	3.04	27.77 ± 0.01	69.62 ± 0.04	12.58 ± 0.20
SARW	N.A.	27.28 ± 0.04	68.12 ± 0.13	12.47 ± 0.21
SAXS	7.9	28.5 ± 3.4	N.A.	N.A.

Included also is the radius of gyration obtained from SAXS by Martin *et al.* [13] at an ionic strength of 150 mM and the simulated SARW for Ash1.

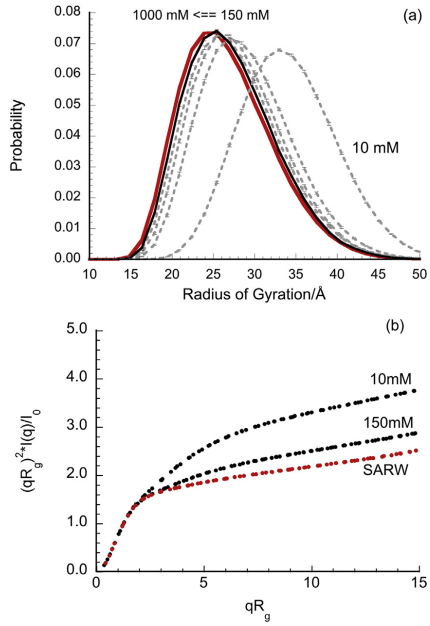


Fig. 4. The probability distribution of the radius of gyration (a), that is, conformational ensemble, and the dimensionless Kratky plot as a function of salt concentration for Ash1 (b). The red function corresponds to the SARW, whereas 10 and 150 mM are shown as black-dotted curves. In panel a, the full black line corresponds to 1000 mM.

even at higher salt concentrations. As shown in Fig. 4a, Ash1 behaves as a polyelectrolyte in the sense that it contracts upon the addition of salt. The FWHMs of the probability distribution of R_g for Ash1 at an ionic strength of 10 and 150 mM are estimated to be 13.70 ± 0.10 and 12.91 ± 0.18 Å, respectively. These numbers confirm that the conformational entropy of Ash1 is decreasing upon salt addition, which is in line with the fact that the preferred shape is more contracted at higher salt concentrations.

The asphericity ranges from 0 for a sphere to 1 for a rod, and have been determined according to the protocol by Angelescu and Linse [32]. The ensemble averages of the asphericity as well as the shape factor indicate that at low ionic strength, that is, 10 mM, Ash1 becomes more extended than a SARW, the values being 0.6 and 6.6, respectively. At increased salt concentrations, the values level off to approximately 0.5 for the asphericity and 6.3 for the shape, clearly indicating conformations resembling a SARW. Hence, at 150 mM and higher ionic strengths, it is possible to model the form factor as a SARW, especially when

taking into account the resolution of SAXS experiments. However, it is important to remember that it is indeed an approximation, as true SARW behavior is reached first at 1000 mM. At low salt concentration, it is not possible to model the form factor as a SARW, and additionally, the differences between 10 and 150 mM are quite pronounced. On the other hand, it is also very difficult to measure the form factor of IDPs at low salt concentrations by SAXS due to the contribution from the structure factor on the scattering curve. An advantage with computer simulations is that it enables discrimination of how intra- and intermolecular interactions affect the form factor. Fig. 4b shows the unitless Kratky plot that qualitatively assesses the overall conformational state and reveals the flexibility/rigidity of the protein. Both the results obtained from simulations at 10 and 150 mM salt, as well as for a SARW, are shown for comparison. In this representation, the salt effect is clearly visible and these results confirm, indeed, that the form factor depends on the salt concentration; that is, it is not accurate to use the same form factor at *high* and *low* ionic strength. This will of course have implications when deriving the structure factor at low ionic strengths using: $I(q) = S(q) \cdot P(q)$, where $P(q)$ often is determined at a higher salt concentration by SAXS. Here $S(q)$ and $P(q)$ correspond to the structure and the form factor, respectively.

Phosphorylated IDPs

Many of the IDPs belong to the family of phosphoproteins; that is, for example, they often contain phosphorylated serines or threonines. In this study, three model proteins have been investigated: Statherin, pSic1, and pAsh1. The first protein, Statherin, contains two phosphorylated serines residing in the N-terminus, possesses an amphiphilic structure, and has a tendency to self-associate. In the second protein, pSic1, there are six phosphorylated groups, whereas in pAsh1, there are ten. The reader is referred to Fig. 5 to achieve an overview of the distribution of the phosphorylated as well as the positively and the negatively charged amino acids. Furthermore, according to Das *et al.* [20], FCR and NCPR (denoted FCR:NCPR) for Statherin, pSic1, and pAsh1 are 0.23:−0.05; 0.25:−0.01; and 0.46:−0.06. Hence, the two former can be considered as weakly charged polyelectrolytes/polyampholytes where pSic1 is almost net neutral, whereas in this context, pAsh1 is strongly charged. As a reminder, the threshold for strongly charged polyelectrolytes is FCR > 0.3.

Starting off with Statherin, our SAXS measurements show that despite its tendency to self-associate, it is possible to obtain a form factor for Statherin at low protein concentrations. As shown in Fig. 1a as well as given in Table 1, the experimentally and simulated radii of gyration agree relatively

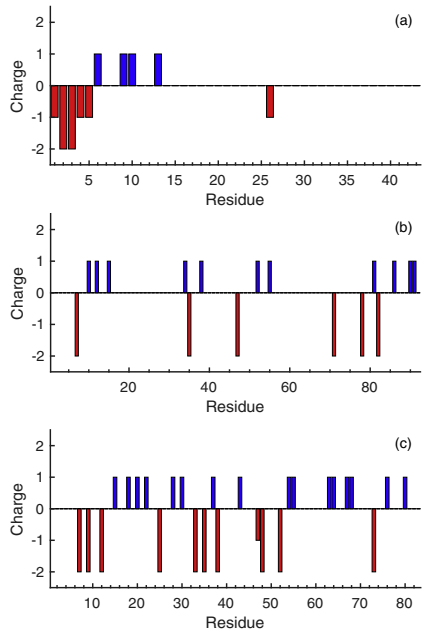


Fig. 5. Charge distribution at pH 7 for Statherin (a), pSic1 (b), and pAsh1 (c), where positive charges are marked in blue and negative charges in red. The N- and C-terminal charges are not included.

well; hence, the two phosphorylated serines at position 2 and 3 do not seem to influence the ensemble average to greater extent in that respect. Fig. 6a shows the dimensionless Kratky plot, and as clearly visible, the profiles from the experiment and the simulation agree very well and display a random coil behavior, that is, a linear rise to a plateau at higher scattering angles. Interestingly, the simulation snapshots indicate that the N-terminus where the two phosphorylated serines reside seems to form a cluster, while the rest of the chain is flexible, as illustrated by Fig. 6b. From the simulations, it is also shown that the R_g is not sensitive to salt (data not shown).

pSic1 on the other hand is twice as long as Statherin and contains six phosphorylated residues at positions 7, 35, 47, 71, 78, and 82, that is, relatively well separated from each other. As shown in Fig. 1a, there is a significant difference in the radii of gyration obtained from the experiment *versus* the simulation, where the former indicates a conformation more expanded than a SARW, and the latter displays a more compact conformation, less expanded than SARW (28.94 ± 0.05 Å). From the simulations, it is

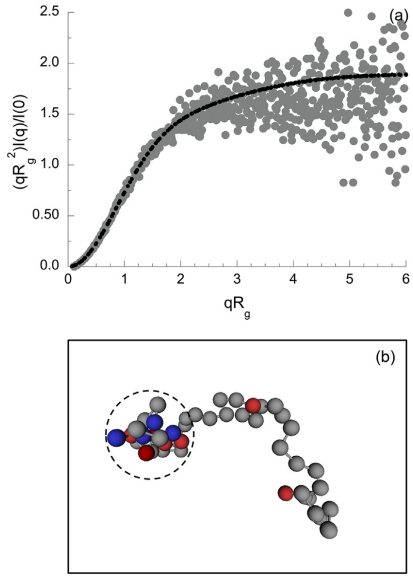


Fig. 6. (a) Dimensionless Kratky plot for experimental data at pH 8.1 (gray filled circles) and for the simulated data (black filled circles) at an ionic strength of 150 mM for Statherin. (b) Representative snapshot of a chain conformation obtained in a simulation at 150 mM salt. Blue spheres are positively charged amino acids, red spheres are negatively charged amino acids, and the dark red spheres represent phosphorylated serines with the charge $Z_{phos} = -2e$, whereas the gray spheres correspond to neutral amino acids. The salt was treated implicitly, and the counterions are omitted for clarity. The dashed line circles the N-terminal part of the chain.

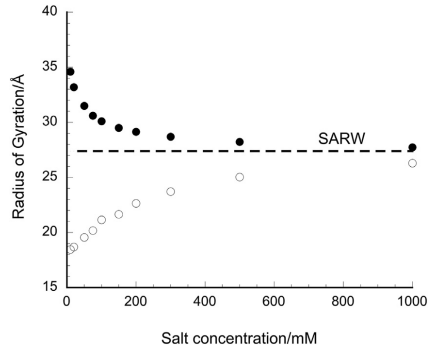


Fig. 7. The ensemble average of the radius of gyration in Å as a function of the salt concentration in mM, for Ash1 in black circles and the 10-sites phosphorylated counterpart pAsh1 in open circles. The salt is assumed to be of 1:1 nature with respect to the charge. The dashed line corresponds to the estimated radius of gyration utilizing the SARW. The reader should keep in mind that the experimentally obtained values of R_g for the two proteins correspond to 28.5 ± 3.4 Å and 27.5 ± 1.2 Å, [13], respectively, which is approximately the same number as obtained from the SARW model. The precision of the data is too small in comparison with the marker to be visible.

also shown that R_g is sensitive to salt and decreases when the salt concentration is increased, from 31.11 ± 0.05 Å to 27.55 ± 0.05 Å at 10 and 150 mM salt, respectively, which advocates the existence of electrostatic attractive interactions within the chain.

The last phosphorylated protein in our study, pAsh1, contains 10 phosphorylated residues, where nine out of ten are within the 52 amino acids in the N-terminal (positions 7, 9, 12, 25, 33, 35, 38, 48, 52, and 74). As shown in Fig. 1a, there is a discrepancy between the experimental and simulated data, where the simulation again advocates a more contracted ensemble average than the experiment as well as SARW (27.28 ± 0.04 Å). Experimentally, it has been shown that upon phosphorylation of Ash1 at ten distinct sites, the global conformational properties of pAsh1 are indistinguishable from those of unphosphorylated Ash1. The obtained ensemble averages of the radii of gyration from SAXS measurements were determined to be 28.4 ± 3.4 Å and 27.5 ± 1.2 for Ash1 and pAsh1, respectively, at 150 mM NaCl [13]. Simulations of the ensemble average of the radius of gyration as a function of salt clearly indicate that Ash1 displays a polyelectrolytic and pAsh1 a polyampholytic behavior (see Fig. 7) and that realistic trends are captured.

Our conclusion is that depending on the number of phosphorylated sites and their distribution, short-ranged attractive electrostatic interactions could influence the conformational properties quite dramatically. For Ash1/pAsh1, the radius of gyration decreases with ≈ 26%, whereas the corresponding numbers for Sic1/pSic1 and Statherin system are 10% and 1%, respectively. Moreover, the shape of the proteins deviates more dramatically when phosphorylated groups are introduced, cf. protein with and without phosphorylation. The effect is enhanced with an increasing number of phosphorylated residues, as visualized in the Kratky plots obtained from simulations in Fig. 8. The dependence of the amino acid distribution is further strengthened by the partial radial distribution function between the positively charged amino acids and the phosphorylated residues in Fig. 9, which emphasizes the effect of short-ranged attractive electrostatic interactions. Moreover, as shown in Fig. 10, a substantial amount of salt is needed to screen this short-ranged attractive electrostatic interaction; that is, κ^{-1} needs to be shorter than the distance between the amino acids within the chain.

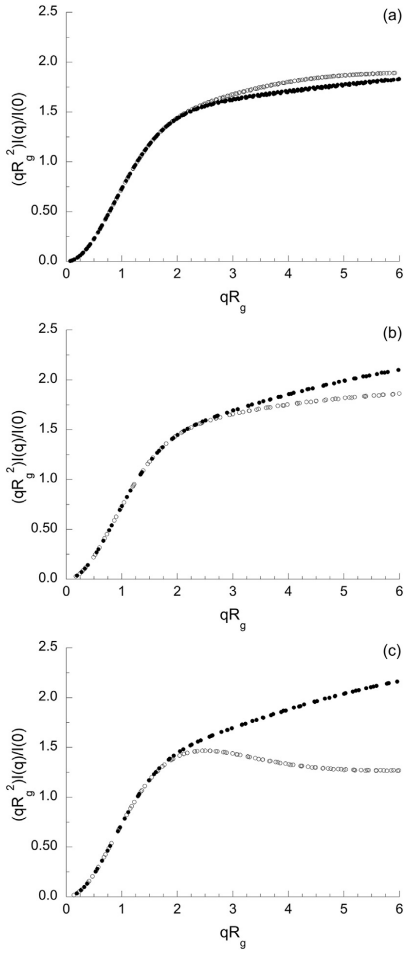


Fig. 8. The simulated dimensionless Kratky plot for Statherin with and without phosphorylated residues (a), Sic1/pSic1 (b), and Ash1/pAsh1 (c), where open circles represent the phosphorylated protein and filled circles the non-phosphorylated counterpart. The reader should notice that the number of phosphorylated groups is increasing from two to six to ten, for the phosphorylated proteins in panels a, b, and c, respectively.

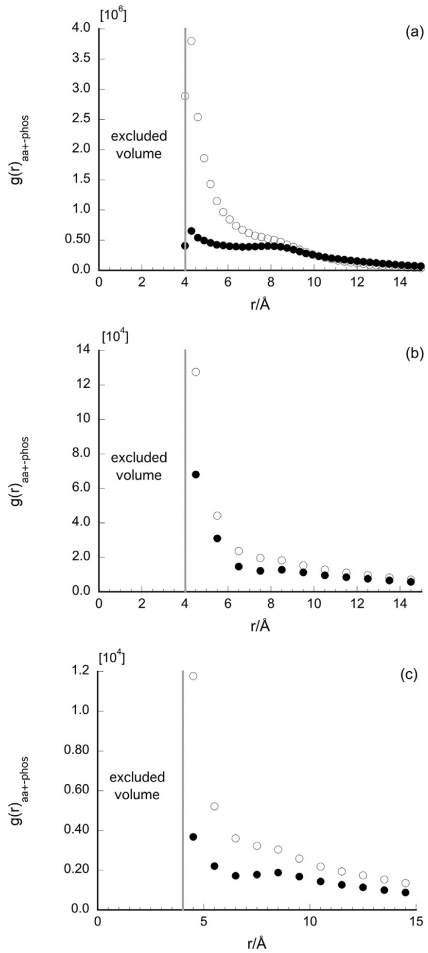


Fig. 9. Partial radial distribution function between positively charged amino acids and phosphorylated residues at 150 mM salt for Statherin (a), pSic1 (b) and pAsh1 (c), where the phosphate groups have the charge –2*e* (open circles) or 0 (filled circles, corresponding to non-phosphorylated protein).

A plausible explanation to the difference between the experimental and simulated radius of gyration for pAsh1 could be due to the physicochemical properties of the phosphorylated residue. Phosphorylation changes the characteristics of the amino acids, especially due to introducing charge. The first pK_a of the phosphate group is below 3, while the second pK_a value is slightly below 6 [33,34], meaning that at physiological pH, the phosphate group should carry a –2*e* charge. However, pK_a values between 6.9 and 7.2 have also been found in Web-based tools for calculating the point of zero charge (see http://scansite.mit.edu/calc_mw_pi.html and ProMoST) [35]. Hence, the radius of gyration has also been determined by simulating the corresponding proteins

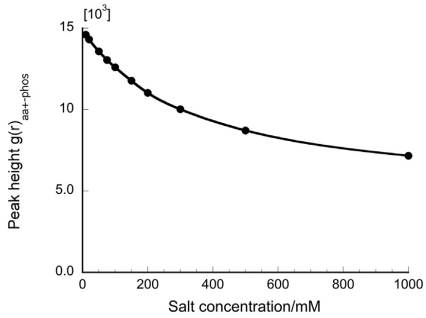


Fig. 10. Peak value of the partial radial distribution function at 4.5 Å between positively charged amino acids and phosphorylated residues as a function of salt concentration, for pAsh1. The precision is within the data marker.

for the phosphorylated proteins where the phosphate group carries a charge $Z_{phos} = -1e$. As shown in Table 3, it gives a much better agreement with the experiments. However, no such interpretation should be made as the phosphorylated residues carry the charge $Z_{phos} = -1e$ at physiological pH. Other possibilities could be that there is a distribution of phosphorylated residues in the experimental sample which does not exist in the model, or that some phosphorylated residues are neutralized due to their binding affinity to, for example, calcium. Monte Carlo simulations provide an exact solution to the model used; hence, traces of other proteins, multivalent ions, and so on, do not exist, which should be kept in mind when comparison are performed with the experimental counterpart.

Model adjustability

The total potential energy of the coarse-grained model presented in this study includes a short-ranged attractive interaction between all amino acids, as well as explicit charges depending on the nature of the amino acid. Moreover, the protein is modeled as totally flexible in the sense that steric interactions are included only through the excluded volume of the amino acid; that is, the chain entropy might be overestimated and the protein too fluidic. This can, of course, be opposed by introducing, for example, an angular potential or increasing the amino acid excluded volume to decrease the flexibility, which is of relevance for the group of proline-rich proteins. Here we compare our modeling results with the non-glycosylated proline-rich saliva proteins, IB5 and II-1ng [15], whose amino acid sequences contain approximately 40% prolines. The experimental and simulated radii of gyration are approximately equivalent, taking the uncertainties into consideration. Although the radius of gyration agrees very well, that might not be the case for the shape. This will be further analyzed by focusing on IB5. As shown in the Kratky plot in Fig. 11, there is a discrepancy between the experimental and simulated curves. From the experimental Kratky profile, one can conclude that the ensemble is biased toward more stiff conformations, in comparison to the unperturbed model (black curve), which, most probably, is an effect of the high proline content.

One possibility to improve the agreement between SAXS and simulations is by introducing an angular potential. The effect of the prolines has been taken into account in the simulations by adding an angular potential of 0.0023 kJ mol^{-1} deg^{-2}; that is, the average angle between three consecutive beads increased from approximately 103° to 141°, that is, a quite dramatic change (see red curve). The resulting radius will then be overestimated but the flexibility/rigidity is more realistic. Another possibility would be to induce a local stiffness within the chain representing

Table 3. Number of phosphorylated residues, N_{phos}, and simulated radii of gyration (R_g) for phosphorylated IDPs, expressed in Å, at 150 mM monovalent salt for phosphorylated residues with the net charge of $Z_{phos} = -1e$ or $Z_{phos} = -2e$

	N_{phos}	$R_{g,\ exp}$ [Å]	$R_{g,\ sim}$ [Å] $Z_{phos} = -1$	$R_{g,\ sim}$ [Å] $Z_{phos} = -2$
Statherin	2	19.3 ± 0.2	18.24 ± 0.04	18.05 ± 0.05
pSic1	6	28.6 ± 0.5	29.00 ± 0.06	27.55 ± 0.05
pAsh1	10	27.5 ± 1.2	25.61 ± 0.08	21.66 ± 0.12

The experimental SAXS data for pAsh1 and pSic1, respectively, are obtained from Martin *et al.* [13] and Mittag *et al.* [39].

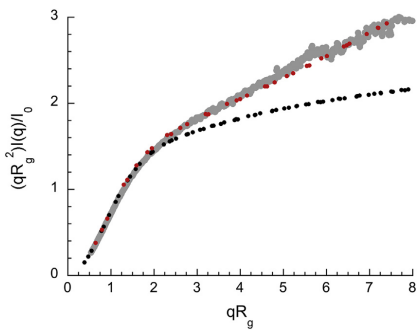


Fig. 11. Dimensionless Kratky representation of IB5 from SAXS measured by Boze *et al.* [15] (gray), the flexible protein model (black), and the model with an additional angular potential, k_{angle} = 0.0023 kJ mol^{-1} deg^{-2} (red).

the segments consisting of several prolines. This is, however, out of the scope for the current study, since we are aiming for a general model, which can be easily adjusted to all IDPs with a few parameters.

Conclusions

To summarize our findings, the coarse-grained model, based on the primitive model, is well applicable for IDPs where the intra-chain interactions are dominated by electrostatic interactions. By extending the model to include, for example, angular potentials, and/or a short-ranged attractive interaction preferably between the hydrophobic amino acids within the chain, in principle it is possible to tune the fitting parameters to obtain an agreement between the simulations and the experimental data for a specific protein.

A popular method for analyzing SAXS spectra of IDPs and to achieve information about the ensemble average of the radius of gyration is by utilizing Flexible-meccano. Comparisons between the results obtained from Monte Carlo simulations and Flexible-meccano agree well. As shown in Fig. 12, this method works well for the unphosphorylated IDPs used in this study and it is definitely a valuable tool to obtain information about the most probable conformations and R_g distributions. The take-home message is that coarse-grained modeling and Monte Carlo simulations can contribute when the aim is to understand the underlying physics and the intricate balance between the different contributions regarding the intra-chain interactions. The model seems to be generally valid when electrostatic interactions dominate, and it can be adjusted to correspond to any IDP/IDR by tuning the intra-chain potentials.

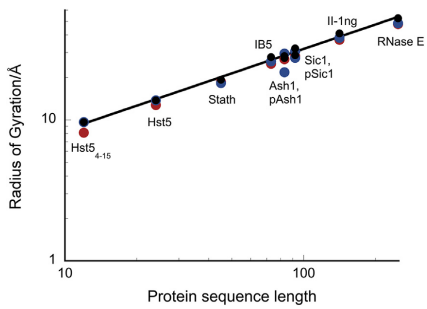


Fig. 12. The ensemble average of radius of gyration as a function of the length of the amino acid sequence in the protein on a log–log scale for the experimental pool of proteins where the full line including black data markers corresponds to a power law fit of the experimental values, the red filled circles to the results obtained from Flexible-meccano, and the blue filled circles from Monte Carlo simulations.

Furthermore, it is possible to use an empirical expression to achieve an estimate of the radius of gyration of the monomeric protein when the dominant intra-chain interactions are electrostatic in nature. This could be of practical importance when performing experiments to achieve a rapid understanding of, for example, the association state of the protein or if there exist residual elements of local structure.

Coarse-grained modeling and Monte Carlo/molecular dynamics simulations are valuable approaches when the aim is to achieve an understanding of how the structure and the inter- and intramolecular interactions are affected by variations in pH, salt concentration, and protein point mutations. It is also useful for studying more complex systems, such as the effect of protein concentration, interaction with other macromolecules (e.g., proteins and surfactants), as well as the interaction with surfaces and biological membranes. In the latter, the distribution and valency of the surface charges, the surface charge density, and the bilayer composition can be evaluated. The information from these simulations can then be correlated with the function.

Model and Method

Coarse-grained model

The monomers of the proteins, that is, the amino acids, are represented by hard spheres (beads) that mimic their excluded volume including the hydration layer and are connected via harmonic bonds. The N- and C-termini are included explicitly to account for the extra charge. The bead radius was set to 2 Å providing a realistic contact separation between the charges and an accurate Coulomb interaction. The non-bonded spheres interact through a short-ranged attractive interaction and electrostatic interactions, where the interparticle electrostatic interactions are described on the Debye–Hückel level. The simulations are performed at constant pH with point charges. Each monomer is negative, positive, or neutral, depending on the amino acid sequence, as illustrated in Fig. 13.

The total potential energy of the simulated system contains bonded and non-bonded contributions, and is given by:

$$U_{\text{tot}} = U_{\text{nonbond}} + U_{\text{bond}} = U_{\text{hs}} + U_{\text{el}} + U_{\text{short}} + U_{\text{bond}} \tag{1}$$

where the non-bonded energy is assumed to be pairwise additive according to:

$$U_{\text{nonbond}} = \sum_{i<j} u_{ij}(r_{ij}), \tag{2}$$

where $r_{ij} = |\mathbf{R}_i - \mathbf{R}_j|$ is the center-to-center distance between two monomers, and $\mathbf{R}$ refers to the

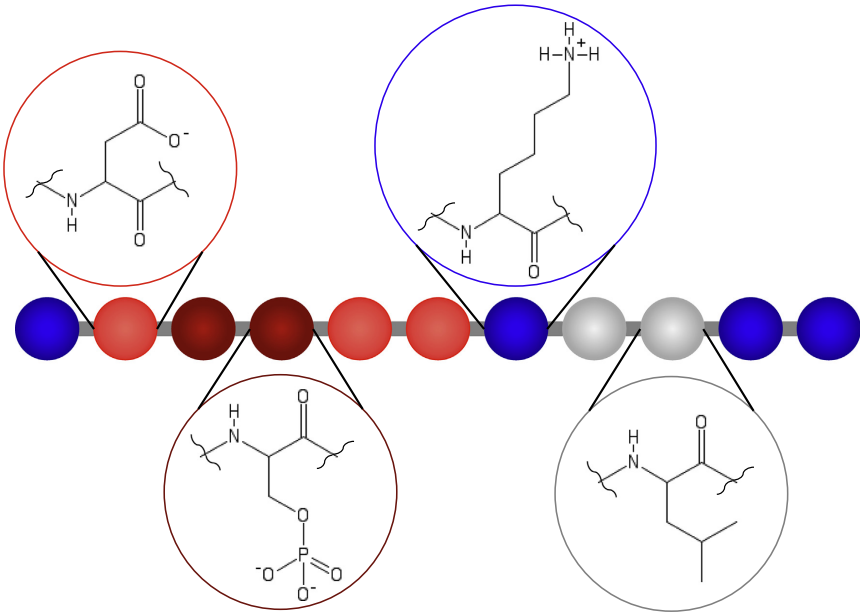


Fig. 13. Schematic description of the coarse-grained model showing the N-terminal fragment of the saliva protein Statherin. Blue spheres have the charge $Z = +1e$; bright red spheres, $Z = -1e$; and dark red spheres, $Z = -2e$. Gray spheres correspond to neutral amino acids. The four structures depicted are aspartic acid, phosphorylated serine, lysine, and leucine. The N-terminal is modeled explicitly as a positively charged sphere.

coordinate vector. The excluded volume is taken into account through the hard-sphere potential, U_{hs}, given by:

$$U_{\text{hs}} = \sum_{i<j} u_{ij}^{\text{hs}}(r_{ij}), \tag{3}$$

which sums up over all amino acids. The hard-sphere potential, $u_{ij}^{\text{hs}}(r_{ij})$, between two monomers in the model is given by:

$$u_{ij}^{\text{hs}}(r_{ij}) = \begin{cases} 0, & r_{ij} \geq R_i + R_j \\ \infty, & r_{ij} < R_i + R_j \end{cases}, \tag{4}$$

where R_i and R_j denote the radii of the beads. The electrostatic potential U_{el}, is given by an extended Debye–Hückel potential according to:

$$\begin{aligned} U_{\text{el}} &= \sum_{ij} u_{ij}^{\text{el}}(r_{ij}) \\ &= \sum_{i<j} \frac{Z_i Z_j e^2}{4\varepsilon_0 \varepsilon_r} \frac{\exp\left[-\kappa\left(r_{ij}-\left(R_i-R_j\right)\right)\right]}{\left(1+\kappa R_i\right)\left(1+\kappa R_j\right)} \frac{1}{r_{ij}}, \end{aligned} \tag{5}$$

where e is the elementary charge, κ denotes the inverse Debye screening length, ε_0 is the vacuum permittivity, and ε_r the dielectric constant for water. The short-ranged attractive interaction between the monomers is included through an approximate arithmetic average over all amino acids, given by:

$$U_{\text{short}} = -\sum_{i<j} \frac{\varepsilon}{r_{ij}^6}, \tag{6}$$

where ε reflects the polarizability of the proteins and thus sets the strength of the interaction. In this model, ε was set to 0.6×10^4 kJ Å^6/mol giving an attractive potential of 0.6 kT at closest contact. The bonded interaction, a harmonic bond, is given by:

$$U_{\text{bond}} = \sum_{i=1}^{N-1} \frac{k_{\text{bond}}}{2} \left(r_{i,i+1}-r_0\right)^2 \tag{7}$$

where $\boldsymbol{r}_{i,i+1}$ denotes the distance between two connected monomers with the equilibrium separation $r_0 = 4.1$ Å, and the force constant $k_{\text{bond}} = 0.4$ N/m, whereas N denotes the number of monomers of the

protein. The proteins are assumed to be totally flexible, except for when the effect of intrinsic stiffness is evaluated. An angular dependent component, expressed below, is then added to the potential:

$$U_{\text{angle}} = \sum_{i=2}^{N-1} \frac{k_{\text{angle}}}{2}(\alpha_i - \alpha_0)^2. \qquad (8)$$

Here, α_i is the angle formed by the vectors $\boldsymbol{r}_{i+1}-\boldsymbol{r}_i$ and $\boldsymbol{r}_{i-1}-\boldsymbol{r}_i$, made by three consecutive beads with the equilibrium angle $\alpha_0 = 180°$ and the force constant k_{angle}. In addition to the angular potential, the electrostatic interactions among the segments as well as the volume of the hard spheres also contribute to the rigidity of the protein.

Simulation aspects

The equilibrium properties of the model systems were obtained applying Monte Carlo simulations in the canonical (NVT) ensemble, that is, constant volume, number of beads, and temperature ($T = 298$ K), utilizing the Metropolis algorithm. The protein chain was enclosed in a cubic box of variable volume, which was dependent on the protein length. Periodic boundary conditions were applied in all directions. The long-ranged Coulomb interactions were truncated using the minimum image convention. Four different types of displacements were allowed: (i) translational displacement of a single bead, (ii) pivot rotation, (iii) translation of the entire chain, and (iv) slithering move, in order to accelerate the examination of the configurational space [36]. The probability of the different trial moves was weighted to enable single-particle moves 20 times more often than the other three. Initially, the protein was randomly placed in the box and an equilibrium simulation of typical 2×10^5 trial moves/bead was performed, whereas the proceeding production run comprised 10^6 passes divided into 10 subdivisions. The radius of gyration and end-to-end distance probability distribution functions of the proteins, that is, the conformational ensembles, were analyzed to confirm that the simulations were sampled accurately. The reported uncertainty of simulated quantities is one standard deviation of the mean. It is estimated from the deviation among the means of the subdivisions of the total number of MC passes according to:

$$\sigma^2(\langle x\rangle) = \frac{1}{n_s(n_s-1)} \sum_{s=1}^{n_s} \left(\langle x\rangle_s - \langle x\rangle\right)^2, \qquad (9)$$

where $\langle x\rangle_s$ is the average of quantity x from one subdivision, $\langle x\rangle$ the average of x from the total simulation, and n_s the number of subdivisions. The simulations were performed by using the integrated Monte Carlo/molecular dynamics/Brownian dynamics simulation package Molsim [37].

Structural analysis

The model was validated by comparing the simulated scattering intensities with the experimental scattering intensities obtained by SAXS. For a system containing N identical scattering objects, the structure factor is given by:

$$S(q) = \left\langle \frac{1}{N} \left| \sum_{j=1}^{N} \exp(i\mathbf{q}\cdot\mathbf{r}_j) \right|^2 \right\rangle. \qquad (10)$$

The total structure factor can further be decomposed into partial structure factors given by:

$$S_{\text{ij}}(q) = \left\langle \frac{1}{(N_i N_j)^{1/2}} \left[\sum_{i=1}^{N_i} \exp(i\mathbf{q}\cdot\mathbf{r}_i)\right] \left[\sum_{j=1}^{N_j} \exp(-i\mathbf{q}\cdot\mathbf{r}_j)\right] \right\rangle. \qquad (11)$$

The total and partial S(q) are related through:

$$S(q) = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \frac{(N_i N_j)^{1/2}}{N} S_{ij}(q). \qquad (12)$$

For a point scatterer, the form factor is constant, inferring that the scattering intensity is proportional to the structure factor. In order to account for an approximate effective particle/residue form factor, the scattering profile further needs an appropriate normalization, such that I_0 coincides with the experimental scattering profile.

FWHM analysis

To obtain the FWHM of the radius of gyration probability distribution, the curve was fitted with a Gaussian function on the form:

$$f(x) = a \cdot \exp\left[-\frac{(x-b)^2}{c^2}\right], \qquad (13)$$

where a, b, and c are fitting parameters. The FWHM was calculated from the parameter c, according to:

$$\text{FWHM} = 2\sqrt{\ln(2)} \cdot c \qquad (14)$$

and is reported with a 95% confidence interval.

Flexible-meccano

We have used the program Flexible-meccano [19] with default settings to generate a pool of 10,000 possible polypeptide backbones by randomly selecting specific amino-acid conformations from a library of non-secondary structural elements of high-resolution X-ray crystallographic structures.

Experiments

Sample preparation

Statherin was purchased from Genemed Synthesis, Inc.. A 20 mM Tris [>99.9%, CAS (77-86-1); Saveen Werner AB] buffer with 150 mM NaCl [reagent grade, CAS (7647-14-5); Sharlau] was prepared with Milli-Q water, and the pH was set to 8.1 by dropwise addition of 1 M HCl, and thereafter, it was filtered through a hydrophilic polypropylene 0.2 µm membrane (Pall Corporation). The protein powder was dissolved in buffer by a small addition of NaOH to increase the pH, since the protein powder contained trifluoroacetate. A concentrating cell (Vivaspin 2, 2000 MWCO, Prod. No. VS02H92; Sartorius, Cambridge, United Kingdom) was used to remove low-molecular-weight impurities. The sample was rinsed with buffer corresponding to 30 times the sample volume, by centrifugation at 1600 rpm at 8°C. To ensure an exact background in the SAXS measurements, the sample was dialyzed (Slide-A-Lyzer Dialysis Cassette, 2000 MWCO, Prod. No. 66203; Thermo Scientific, Waltham, MA, USA) overnight at 6°C. Before the SAXS measurements, the sample was centrifuged at 14,000 rpm at 6°C for at least 2 h to remove aggregates. Thereafter, it was diluted to a concentration series, and the protein concentration was determined with a nanodrop spectrometer at the beamline using $\lambda = 280$ nm and $\varepsilon = 8740\ M^{-1}\ cm^{-1}$. The samples were centrifuged in small PCR tubes imminent to the SAXS measurements to remove any bubbles.

SAXS measurements

SAXS experiments were performed at BM29, ESRF-Grenoble, France. The incident beam wavelength was 0.99 Å, and the distance between sample and detector (PILATUS 1M) was set to 2867 mm, giving the scattering vector 0.0039–0.49 $Å^{-1}$. The scattering vector, q, is defined as $q = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the incident beam. Several successive frames of the scattering from the samples were recorded with a 0.5-s exposure time. The scattering from the pure solvent, which was measured before and after each sample for the same exposure times, was subtracted from the sample scattering. All measurements were performed at 20°C, and I_0 was converted to absolute scale by measuring the scattering of water. SAXS data were measured either after passing through a size exclusion chromatography (SEC) column or within a flowing capillary. For the inline SEC-SAXS, 5 mg/mL protein was injected through a 100-µL loop into a Superdex 75 10/300 GL column (GE Healthcare), equilibrated in 20 mM Tris, with 150 mM NaCl and a pH of 8.1. During SEC-SAXS, data were collected with a 1 s exposure time.

SAXS analysis

The SAXS and SEC-SAXS data were extracted and processed using PRIMUS [38] and ScÅtter (available at www.bioisis.net), respectively. Special attention was paid to radiation damage by comparing the

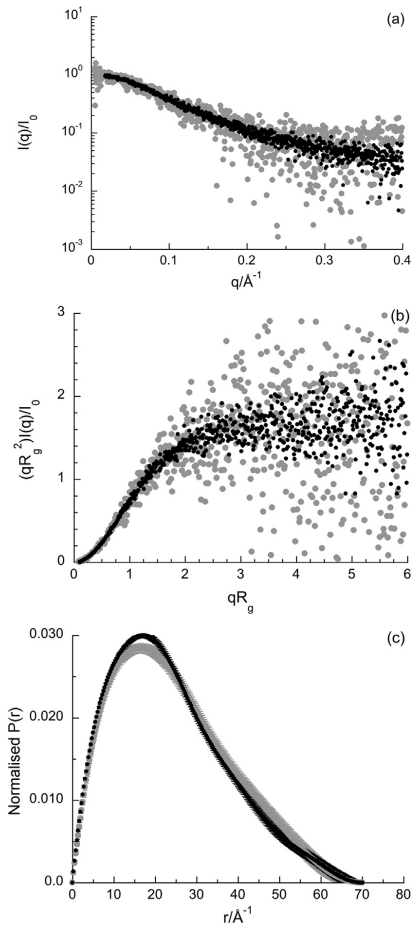


Fig. 14. SAXS data obtained for Statherin at 20 mM Tris and 150 mM NaCl (pH 8.1) at BM29, ESRF-Grenoble, France. Form factor (a), dimensionless Kratky plot (b), and pair distance distribution function, $P(r)$ (c). The black circles correspond to data obtained from SEC in combination with SAXS, and the gray circles refer to continuous flow SAXS. If the precision is not visible, it is within the size of the data marker.

successive frames prior to background subtraction, and any affected data were rejected from further analysis. The form factor was obtained at the protein concentration 0.24 mg/mL, as shown in Fig. 14. From the pair distance distribution, $P(r)$, the radius of gyration, R_g, was determined to be 19.8 ± 0.6 Å. The molecular weight was determined to be 5.29 kDa based on I_0 obtained from $P(r)$. This is in good agreement with the theoretical molecular weight of 5.38 kDa, confirming that monomeric Statherin was obtained. The scattering curve from the peak in SEC-SAXS, also presented in Fig. 14, is in excellent agreement with the curve measured at 0.24 mg/mL, and R_g obtained from $P(r)$ was determined to be 19.3 ± 0.2 Å. Hence, it is consistent with the measurement at 0.24 mg/mL. Since the protein concentration in the eluent from the SEC column was unknown, no molecular weight was obtained. However, due to the perfect agreement between the data obtained from SEC-SAXS and measured at 0.24 mg/mL, the less noisy SEC-SAXS data were used for comparison with simulations.

Acknowledgment

We are grateful to Dr. Mark Tully at the European Synchrotron Radiation Facility (ESRF), Grenoble, for providing assistance in using beamline BM29. This study was supported by the Science Faculty project grant program for research with neutrons and synchrotron light (Lund University Strategic funds for MAX-IV and European Spallation Source). The simulations were performed on resources provided by the Swedish National Infrastructure for Computing at the Center for scientific and technical computing at Lund University (LUNARC). We are also grateful to Boze *et al.* [15] for sharing their IB5 SAXS data.

Received 14 December 2017;
Received in revised form 3 March 2018;
Accepted 12 March 2018
Available online 21 March 2018

Keywords:
intrinsically disordered proteins;
coarse-grained modeling;
Monte Carlo simulations;
electrostatic interactions;
small-angle X-ray scattering

Abbreviations used:
IDPs, intrinsically disordered proteins; IDRs, Intrinsically disordered regions; SAXS, small-angle X-ray scattering; SARW, self-avoiding random walk; FCR, fraction of charged residues; FWHM, full width half maximum; NCPR, net charge per residue.

References

[1] A.K. Dunker, et al., Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.
[2] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (2002) 527–533.
[3] J.T. Liu, J.R. Faeder, C.J. Camacho, Toward a quantitative theory of intrinsically disordered proteins and their function, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19819–19823.
[4] J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337 (2004) 635–645.
[5] M. Kjaergaard, A.-B. Norholm, R. Hendus-Altenburger, S.F. Pedersen, F.M. Poulsen, B.B. Kragelund, Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II? Protein Sci. 19 (2010) 1555–1564.
[6] R.B. Best, W. Zheng, J. Mittal, Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association, J. Chem. Theory Comput. 10 (2014) 5113–5124.
[7] J. Henriques, C. Cragnell, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment, J. Chem. Theory Comput. 11 (2015) 3420–3431.
[8] J. Henriques, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the TIP4P-D water model and the representativeness of protein disorder models, J. Chem. Theory Comput. 12 (2016) 3407.
[9] S. Piana, A.G. Donchev, P. Robustelli, D.E. Shaw, Water dispersion interactions strongly influence simulated structural properties of disordered protein states, J. Phys. Chem. B 119 (2015) 5113–5123.
[10] S. Rauscher, V. Gapsys, M.J. Gajda, M. Zweckstetter, B.L. de Groot, H. Grubmueller, Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment, J. Chem. Theory Comput. 11 (2015) 5513–5524.
[11] D.A. McQuarrie, Statistical Mechanics, 1st ed. University Science Books, Sausalito, Califonia, 2000.
[12] C. Cragnell, D. Durand, B. Cabane, M. Skepo, Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS, Proteins Struct. Funct. Bioinf. 84 (2016) 777–791.
[13] E.W. Martin, A.S. Holehouse, C.R. Grace, A. Hughes, R.V. Pappu, T. Mittag, Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation, J. Am. Chem. Soc. 138 (2016) 15323–15335.
[14] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A.K. Dunker, et al., pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins, Nucleic Acids Res. 42 (2014) D326–D335.
[15] H. Boze, T. Marlin, D. Durand, J. Perez, A. Vernhet, F. Canon, et al., Proline-rich salivary proteins have extended conformations, Biophys. J. 99 (2010) 656–665.
[16] S. Jephthah, J. Henriques, C. Cragnell, S. Puri, M. Edgerton, M. Skepo, Structural characterization of histatin 5–spermidine conjugates: a combined experimental and theoretical study, J. Chem. Inf. Model. 57 (2017) 1330–1341.
[17] H.A. Bruce, D. Du, D. Matak-Vinkovic, K.J. Bandyra, R.W. Broadhurst, E. Martin, et al., Analysis of the natively

unstructured RNA/protein-recognition core in the *Escherichia coli* RNA degradosome and its interactions with regulatory RNA/Hfq complexes, Nucleic Acids Res. 46 (2018) 387–402.
[18] D.P. O'Brien, B. Hernandez, D. Durand, V. Hourdel, A.-C. Sotomayor-Perez, P. Vachette, et al., Structural models of intrinsically disordered and calcium-bound folded states of a protein adapted for secretion, Sci. Rep. 5 (2015).
[19] V. Ozenne, F. Bauer, L. Salmon, J.-R. Huang, M.R. Jensen, S. Segard, et al., Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables, Bioinformatics 28 (2012) 1463–1470.
[20] R.K. Das, K.M. Ruff, R.V. Pappu, Relating sequence encoded information to form and function of intrinsically disordered proteins, Curr. Opin. Struct. Biol. 32 (2015) 102–112.
[21] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol 157 (1982) 105–132.
[22] H. Lee, A.H. de Vries, S.-J. Marrink, R.W. Pastor, A coarse-grained model for polyethylene oxide and polyethylene glycol: conformation and hydrodynamics, J. Phys. Chem. B 113 (2009) 13186–13194.
[23] Flory, Principles of Polymer Chemistry, Cornell Univ. Press, Ithaca, NY, 1953.
[24] P. Bernado, M. Blackledge, A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering, Biophys. J. 97 (2009) 2839–2845.
[25] A. Borgia, W. Zheng, K. Buholzer, M.B. Borgia, A. Schueler, H. Hofmann, et al., Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods, J. Am. Chem. Soc. 138 (2016) 11714–11726.
[26] G. Fuertes, N. Banterlea, K.M. Ruff, A. Chowdhury, D. Mercadante, C. Koehler, et al., Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) E6342–E6351.
[27] H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, B. Schuler, Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 16155–16160.
[28] J.E. Kohn, I.S. Millett, J. Jacob, B. Zagrovic, T.M. Dillon, N. Cingel, et al., Random-coil behavior and the dimensions of chemically unfolded proteins, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 12491–12496.
[29] I.S. Millet, S. Doniach, K.W. Plaxco, Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins, Unfolded Proteins 62 (2002) 241–262.
[30] D.K. Wilkins, S.B. Grimshaw, V. Receveur, C.M. Dobson, J.A. Jones, L.J. Smith, Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques, Biochemistry 38 (1999) 16424–16431.
[31] J.A. Riback, M.A. Bowman, A.M. Zmyslowski, C.R. Knoverek, J.M. Jumper, J.R. Hinshaw, et al., Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water, Science 358 (2017) 238–241.
[32] D.G. Angelescu, P. Linse, Branched-linear polyion complexes investigated by Monte Carlo simulations, Soft Matter 10 (2014) 6047–6058.
[33] C.D. Andrew, J. Warwicker, G.R. Jones, A.J. Doig, Effect of phosphorylation on alpha-helix stability as a function of position, Biochemistry 41 (2002) 1897–1905.
[34] M. Zachariou, I. Traverso, L. Spiccia, M.T.W. Hearn, Potentiometric investigations into the acid-base and metal ion binding properties of immobilized metal ion affinity chromatographic (IMAC) adsorbents, J. Phys. Chem. 100 (1996) 12680–12690.
[35] B.D. Halligan, V. Ruotti, W. Jin, S. Laffoon, S.N. Twigger, E. A. Dratz, ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels, Nucleic Acids Res. 32 (2004) W638-W644.
[36] K. Binder, Monte Carlo and Molecular Dynamics Simulations in Polymer Science, Oxford University Press, New York, 1995.
[37] J. Rescic, P. Linse, MOLSIM: a modular molecular simulation software, J. Comput. Chem. 36 (2015) 1259–1274.
[38] P.V. Konarev, V.V. Volkov, A.V. Sokolova, M.H.J. Koch, D.I. Svergun, PRIMUS: a Windows PC-based system for small-angle scattering data analysis, J. Appl. Crystallogr. 36 (2003) 1277–1282.
[39] T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, et al., Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase, Structure 18 (2010) 494–506.

Paper II

Assessing the Intricate Balance of Intermolecular Interactions upon Self-Association of Intrinsically Disordered Proteins

E. Rieloff, M. D. Tully, M. Skepö.

Journal of Molecular Biology, 431, **2019**, pp. 511–523.

Assessing the Intricate Balance of Intermolecular Interactions upon Self-Association of Intrinsically Disordered Proteins

Ellen Rieloff [1], Mark D. Tully [2] and Marie Skepö [1]

*1 - **Theoretical Chemistry,** Lund University, POB 124, SE-221 00 Lund, Sweden*
*2 - **European Synchrotron Radiation Facility (ESRF),** Grenoble, France*

***Correspondence to Ellen Rieloff and Marie Skepö:** ellen.rieloff@teokem.lu.se, marie.skepo@teokem.lu.se*
https://doi.org/10.1016/j.jmb.2018.11.027
Edited by Monika Fuxreiter

Abstract

Attractive interactions between intrinsically disordered proteins can be crucial for the functionality or, on the contrary, lead to the formation of harmful aggregates. For obtaining a molecular understanding of intrinsically disordered proteins and their interactions, computer simulations have proven to be a valuable complement to experiments. In this study, we present a coarse-grained model and its applications to a system dominated by attractive interactions, namely, the self-association of the saliva protein Statherin. SAXS experiments show that Statherin self-associates with increased protein concentration, and that both an increased temperature and a lower ionic strength decrease the size of the formed complexes. The model captures the observed trends and provides insight into the size distribution. Hydrophobic interaction is considered to be the major driving force of the self-association, while electrostatic repulsion represses the growth. In addition, the model suggests that the decrease of association number with increased temperature is of entropic origin.

Introduction

Intrinsically disordered proteins (IDPs) are characterized by a lack of stable tertiary structure under physiological conditions *in vitro* [1,2] and hence are best described by conformational ensembles [3,4]. Bioinformatic studies have led to the conclusion that 10%–20% of the eukaryotic proteins are intrinsically disordered, and even more proteins contain intrinsically disordered regions (IDRs) [5–8]. It has also been established that IDPs and IDRs are involved in many biological processes and diseases, and that the lack of folded structure is related to their functions [7,9].

Attractive interactions between IDPs can lead to the formation of aggregates, which in the case of diseases such as Parkinson's disease and Alzheimer's disease is harmful [10]. IDP attractions can also be fundamental for a desired outcome, such as in the formation of proteinaceous membrane-less organelles [11–14], which are condensed liquid droplets often enriched in IDPs and IDRs and commonly found in the cell cytoplasm and nucleus [15]. Various pieces of evidence suggest that liquid–liquid phase separation is a driving force for the formation of some proteinaceous membrane-less organelles [11–14], and that the phase separation itself is driven by weak multivalent interactions between disordered proteins [16,17].

For understanding IDPs and their interactions, computer simulations are a useful complement to experiments [18,19]. There have been considerable advances regarding atomistic simulations of IDPs, where development and justification of force fields and water models have been validated against experimental results [20–24]. The full-atom approach and explicit water treatment in atomistic simulations are great advantages for gaining a molecular understanding, however, atomistic simulations are computationally demanding, both regarding execution time and data storage. Hence, this poses limitations on the accessible timescale and system size, and therefore, a coarse-grained approach is a more viable option for studying complex systems, such as the examples above. Recently, a coarse-grained model based on the primitive model,

in combination with Monte Carlo simulations, has proven capable of capturing bulk properties at dilute conditions for a range of IDPs [25]. We aim to develop this model to also account for more complex systems, and first is the investigation of a model system dominated by intermolecular attractions, namely, the self-association of the saliva protein Statherin. Statherin has a distinct amphiphilic character in its primary sequence, shown in Fig. 1. Almost all charges are located in the N-terminal part, starting with a block of negative charges, followed by a block of positive charges. From the hydropathy values in the Kyte–Doolittle scale [26], it is shown that overall the hydropathy is rather low, which is typical for IDPs. However, residues 15–43 contain seven tyrosines, whose aromatic side chains have been established to be of importance for liquid–liquid phase separation [27,28]. Statherin also consists of 16% proline residues, which are denoted as "disorder-promoting" [29].

In this work, Statherin is characterized experimentally at monomeric conditions through the use of small-angle X-ray scattering (SAXS) and circular dichroism (CD), and at self-associating conditions through SAXS experiments and simulations. The simulation model is validated against the experiments and is demonstrated to be useful for describing polydispersity and the interplay between electrostatics, hydrophobic interactions, and entropy in the self-association process.

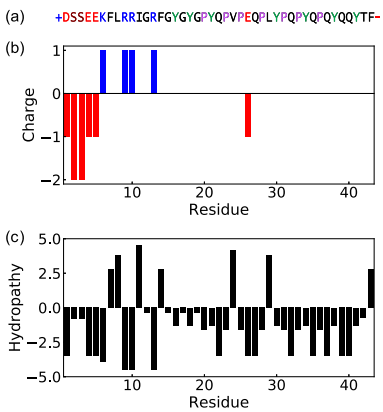


Fig. 1. (a) Amino acid sequence of Statherin with the charge distribution at pH 8 and certain amino acids highlighted. Positive residues are marked in blue, negative in red, phosphorylated serines with the charge $-2e$ in dark red, and prolines in lilac and tyrosines in green. (b) Charge distribution and (c) hydropathy values using the Kyte–Doolittle scale, where −4.5 is the most hydrophilic and +4.5 is the most hydrophobic [26].

Results and Discussion

The experimental results for Statherin at monomeric conditions are presented first, followed by the self-association studied both experimentally and by Monte Carlo simulations.

Monomeric behavior

In Fig. 2a–c, data for monomeric Statherin obtained by SAXS coupled with size-exclusion chromatography (SEC' taken from Ref. [25]) is presented. From regular SAXS measurements at low protein concentration (0.24 mg/mL), the molecular weight was determined to be 5.29 kDa, based on the forward scattering, I_0, obtained from the pair distance distribution function, $P(r)$ [25]. This is in good agreement with the theoretical molecular weight of 5.38 kDa, confirming monomeric conditions. As seen in Fig. 2a, Statherin shows the typical featureless scattering profile of an IDP. The IDP character is also verified by the dimensionless Kratky plot in Fig. 2b, where the profile has an uprise slope and reaches a plateau at higher q values, typical for flexible chains. In addition, the CD data presented in Fig. 2d confirm a random coil behavior with some presence of secondary structure. The global minimum is located at 205 nm, which is slightly higher than the usual 198 nm for random coils; however, it is typical for poly-proline II (PPII) structure. The shallow minimum close to 222 nm might suggest a small presence of α-helix. Several studies of Statherin with CD or NMR have suggested that the charged N-terminal has a propensity for forming α-helix and that a part of the middle adopt PPII structure. Nevertheless, the overall structure is still disordered in aqueous solution [30–34]. Fig. 2d also shows that there are no large differences in structure due to salt concentration.

The radius of gyration for monomeric Statherin in 150 mM NaCl has been reported as 19.3 ± 0.2 Å, based on the $P(r)$ presented in Fig. 2c [25]. With urea, the radius of gyration is increased to 22.1 ± 0.2 Å for 4 M urea and to 23.7 ± 0.3 Å for 8 M urea. The dimensionless Kratky plot, shown in Fig. 3a, also indicates an increase in stiffness when urea is added. From CD measurements it is seen that the mean residue ellipticity ($[\theta]_{MRW}$) at 228 nm, presented in Fig. 3b and c, increases linearly with increased urea concentration and also becomes positive at high urea concentrations. This corresponds to an increase of PPII content, in agreement with the study by Whittington *et al.* [35], reporting that urea promotes PPII formation. PPII conformation is more extended than both random coil and α-helix; hence, this explains the changes observed in the SAXS measurements.

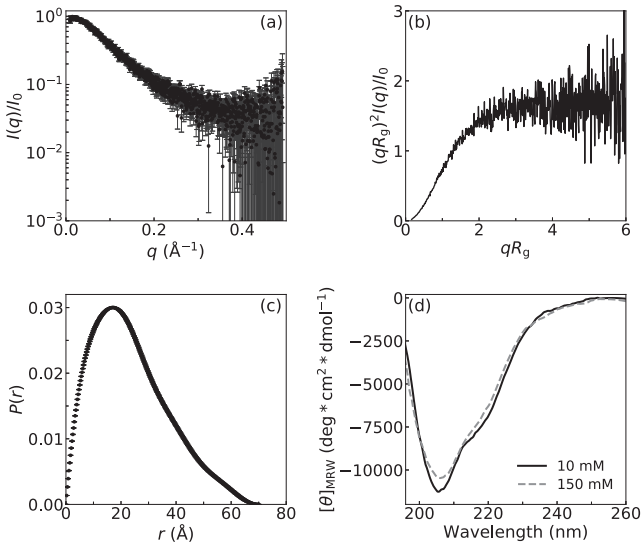


Fig. 2. SAXS data for Statherin obtained by SEC-SAXS, at 150 mM NaCl and 20 mM Tris buffer with pH 8, from Ref. [25]. (a) Form factor, (b) dimensionless Kratky plot, and (c) pair distance distribution function. (d) CD spectra for Statherin in 10 and 150 mM NaF and 20 mM phosphate buffer (pH 8) with a protein concentration of 0.11 and 0.13 mg/mL, respectively, measured at 20 °C.

Temperature also induces changes in secondary structure. With increased temperature, the $[\theta]_{MRW}$ increases at 205 nm and decreases at 228 nm, as shown in Fig. 4, suggesting a loss of PPII as described by Kjaergaard *et al.* [36] for other IDPs. The loss of PPII appears rather proportional to temperature.

Self-association

Experimental results

With increased protein concentration, Statherin self-associates into complexes, which is evident from an increase in forward scattering. The average number of proteins per complex was determined from the forward scattering and is presented against the protein concentration in Fig. 5a for the reference system with 150 mM NaCl. Panels b and d in the same figure present corresponding data from simulations and will be discussed in the next section. The growth is linear with respect to concentration up to 10 mg/mL, and afterward, the slope decreases, which might suggest a maximum size of the Statherin complex. Likewise, the radius of gyration follows the same trend, although a plateau is reached earlier. However, a depression of the forward scattering at higher concentrations due to a structure factor cannot be ruled out, and therefore, the high concentration data should be interpreted with care. Especially since, at 24 mg/mL and higher concentrations, inter-particle interference is visible in the $P(r)$ as a decrease below zero at long distances. The scattering curves, Guinier plots, and I_0 and radius of gyration determined by both Guinier and $P(r)$ are provided in Supplemental information.

The Kratky plot in Fig. 5c shows a transition from flexible chain behavior to more globular when the complexes are formed. The complexes are also more spherical in shape than the free proteins, which is evident from the pair distance distribution function presented in Fig. 6, plotted to enhance the differences compared to a sphere.

Since urea weakens hydrophobic interactions [37], the effect of urea on the Statherin complexes was studied. With 8 M urea, no increase in forward scattering was observed even when reaching 32 mg/mL in protein concentration. The only effect observed was a lowering of the forward scattering due to a structure factor emerging. This indeed suggests hydrophobic interactions as a driving force for the self-association in Statherin. With 4 M urea, it was a downshift at intermediate q when going from 2 to 4 mg/mL and that continued for even higher protein concentrations (data not shown). This in

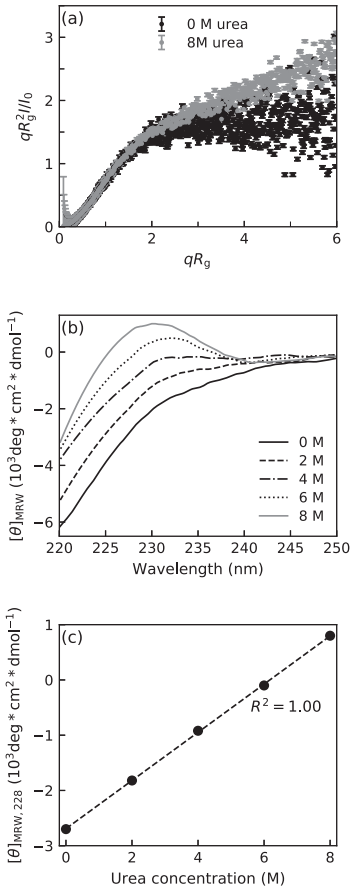


Fig. 3. Effect of urea. (a) Dimensionless Kratky plot for Statherin at 150 mM NaCl measured by SEC-SAXS and with 8 M urea measured by SAXS at a protein concentration of 4 mg/mL, (b) CD spectra and (c) mean residue ellipticity at 228 nm for Statherin (0.12–0.14 mg/mL) *versus* urea concentration, obtained from CD measurements at 20 °C and pH 8.

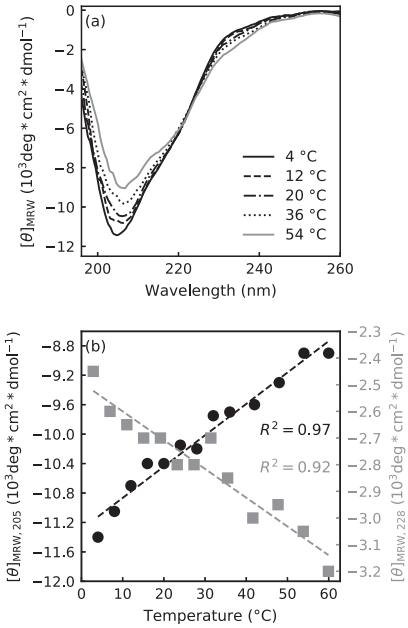


Fig. 4. Temperature dependence of monomeric Statherin (0.13 mg/mL) with 150 mM NaF in 20 mM phosphate buffer at pH 8. (a) CD spectra and (b) mean residue ellipticity at 205 nm (black circles) and 228 nm (gray squares).

combination with a decrease in slope in the Kratky plot with increasing concentration suggests that there are still complexes forming in 4 M urea. For surfactants, both the critical micelle concentration and the micelle size have been reported to change with the concentration of urea [38–40].

Self-association has been observed no matter the salt concentration, which supports hydrophobic interactions being the major driving force. However, the average association number appears to increase with increased ionic strength, as presented in Fig. 7a. Due to the possibility of structure factor influence on the scattering data at lower ionic strength, the effect of electrostatic interactions is further discussed within the framework of the simulations (data presented in Fig. 7b).

Changing the temperature also affects the self-association, as shown by a decrease in association number with increased temperature in Fig. 8. The average radius of gyration follows the same trend (data not shown). The decrease of the association number with temperature has also been observed for surfactants with ionic or zwitterionic headgroups [41], while non-ionic surfactants have shown the opposite temperature dependence [41,42]. For the intrinsically disordered milk-protein β-casein, the association number increases with increased temperature at neutral pH [43], as for non-ionic surfactants. Although β-casein and Statherin have similar block structures, the overall hydrophobicity is higher in β-casein. Hence, it is not unreasonable that the temperature dependence is different.

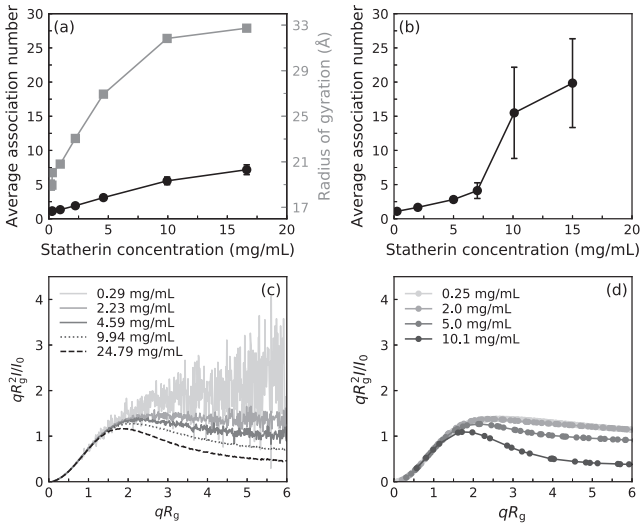


Fig. 5. (a) Average number of proteins per complex (black circles) and radius of gyration (gray squares) *versus* protein concentration determined from SAXS. (b) Average number of proteins per complex *versus* protein concentration from simulations. (c) Dimensionless Kratky plot from experiments. (d) Dimensionless Kratky plot from simulations. The data is reported for the reference system (experimental conditions: 20 mM Tris, 150 mM NaCl, pH 8, 20 °C; simulation conditions: 150 mM implicit salt, 20 °C). In panel a, the error bars on the association number represent a 10% uncertainty.

Simulation results

We have simulated the Statherin system using a modified version of the coarse-grained model presented in Ref. [25]. Therein it was shown that the coarse-grained model works well for Statherin at monomeric conditions. However, to capture the self-association, an additional attractive interaction is needed. We have implemented a short-ranged potential corresponding to 1.32 kT at closest contact between neutral amino acids, mimicking a smeared hydrophobic interaction, which causes the proteins to associate upon increased concentration. For the reference system, 150 mM salt, the simulation data follow the linear trend described in experimental data up to approximately 7 mg/mL, according to Fig. 5b. Then it deviates, by forming large complexes, which shall be interpreted as that the model is reliable only at lower protein concentrations. The model is able to capture the experimentally established transition to a more globular state with increased protein concentration in the Kratky plot, c.f. Fig. 5d and c, although the single chain is too compact due to the extra attraction. To capture the behavior at both monomeric conditions and higher protein concentrations, an angular potential can be included as well. However, since the goal with this model is to capture general trends, an exact matching with the experimental Statherin data is not important, and hence, the results of the model without further modifications are presented.

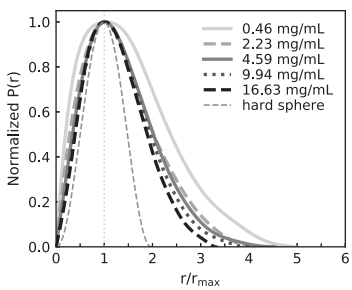


Fig. 6. Pair distance distribution function normalized to enhance deviations in shape from a homogeneous hard sphere, where r_{max} corresponds to the value of r where $P(r)$ has its maximum, for the reference system (20 mM Tris, 150 mM NaCl, pH 8, 20 °C).

The simulations show that the complexes are polydisperse; see the complex size probability distribution in Fig. 9a. At 7 mg/mL and lower concentrations,

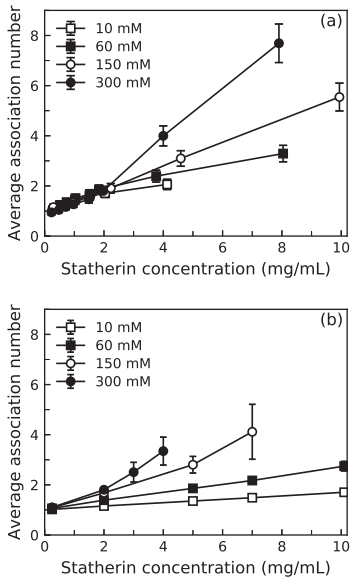


Fig. 7. Average association number determined (a) by SAXS and (b) from simulations, as a function of Statherin concentration for different concentrations of NaCl, at 20 °C. The error bars in panel a represent a 10% uncertainty.

the monomer is the dominating specie and the amount of the different species decreases with increasing size. The polydispersity and monomeric dominance is also evident from the snapshot in Fig. 9b, which furthermore suggests that it is the middle and C-terminal part that forms the core of the complex and that the charged N-terminal part is located on the surface of the complex.

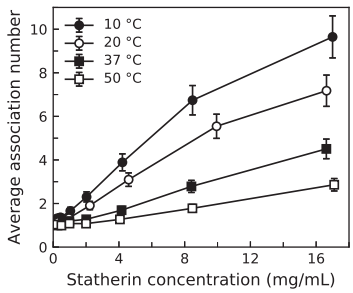


Fig. 8. Average number of proteins per complex determined by SAXS *versus* protein concentration at 150 mM NaCl for 10–50 °C. The error bars represent a 10% uncertainty. The data at 20 °C correspond to the data at 150 mM NaCl in Fig. 7a.

The contact probability between residues of different chains is presented in Fig. 9c and confirms indeed that it is the neutral amino acids that are mostly in contact with other chains. In Fig. 9d, the radial number density distribution from the complex center of mass is presented. It again confirms that the core consists of neutral residues. The negatively charged residue 26 is also part of the core of the complex. The other charged residues are located closer to the surface of the complex.

The experimental $P(r)$ in Fig. 5d shows that the complexes are more spherical than the monomers, due to the change with increasing concentration. However, the experiments only provide the average over all different complex sizes. In the simulations, we have calculated the principal moments of the gyration tensor and from that the asphericity for the complexes of different sizes. It indeed confirms that the monomers are not spherical, having an asphericity value of 0.41. The asphericity decreases with increasing association number until six, where it stabilizes around 0.13 also for larger complexes. If the asphericity is less than 0.1, the object is normally considered spherical [44]. The decrease in asphericity agrees with the experimental results and furthermore shows that the complexes are close to the spherical limit. However, for complexes consisting of seven protein chains, $\langle R_1^2 \rangle$, $\langle R_2^2 \rangle$ and $\langle R_3^2 \rangle$ were 323.5 ± 7.1 Å^2, 158.2 ± 1.2 Å^2, and 91.1 ± 0.5 Å^2, respectively, showing that the instantaneous shapes of the complexes are still not spherical.

The increase of size of the complexes with increased ionic strength observed in SAXS experiments is also captured by the simulations, as seen in Fig. 7b, even if the effect is slightly overestimated compared to experiments (Fig. 7a). This confirms that although the hydrophobic interaction is the major driving force for self-association, electrostatic repulsion stabilizes the system and depresses the growth. To further investigate the electrostatic effect, we performed simulations without phosphorylated serines, which increases the net charge from −4 to 0. This shifts the complex size probability distribution toward larger sizes, depicted in Fig. 10. The overall contact probability also increases from 0.36 ± 0.03 with phosphorylated serines to 0.41 ± 0.01 without phosphorylations at a protein concentration of 2 mg/mL, while the contact profile remains similar in shape. This demonstrates that phosphorylations indeed affect the electrostatic interactions and that it is of importance for the self-association.

Another mutation that illustrates the importance of electrostatics is the point mutation of residue 26, glutamic acid, changing the negatively charged residue located in the middle of the neutral block to a neutral residue. Already in a simulation at 2 mg/mL, the majority of the chains join in one large complex, while for comparison, the reference system rarely exhibits complexes larger than tetramers at the same

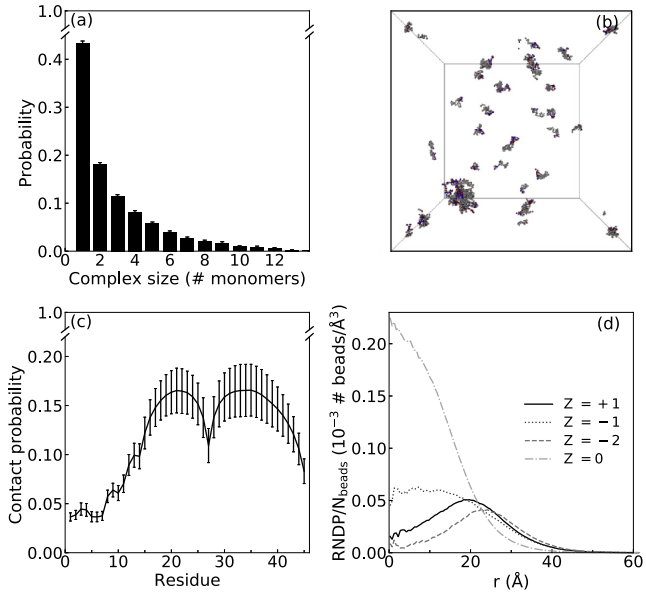


Fig. 9. Simulation data at 5 mg/mL with 150 mM implicit salt. (a) Complex size probability distribution. (b) Snapshot with excluded counterions, where gray beads represent neutral residues, red beads represent negatively charged residues, and blue beads represent positively charged residues. (c) Chain contact probability profile. (d) Radial number density for different bead types, normalized by the number of beads of each type in the protein, as a function of distance from the core center of mass, for complexes consisting of seven proteins. Z represents the charge of each bead type.

concentration. This shows that specific residues can make a great difference for the self-association (results not shown).

With increased temperature, the average association number, displayed in Fig. 11, decreases, again in accordance with experimental results. Since Statherin has a net charge of $-4e$, the overall electrostatic interaction is repulsive. Increased temperature enhances electrostatic interactions, and hence, it would counteract self-association by enhancing the net electrostatic repulsion between Statherin monomers. In addition, the effect of entropy, also opposing self-association, increases with temperature as well. Note that the hydrophobic interaction is regarded temperature-independent in this model. Simulations of the Statherin system without charges at a concentration of 4 mg/mL show a decrease in average association number between 20 and 50 °C, from 3.06 ± 0.63 to 1.39 ± 0.01, compared to 2.24 ± 0.15 to 1.40 ± 0.01 for the same system with charges. This suggests entropy as the main contribution to the temperature effect.

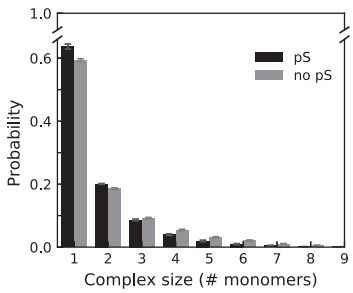


Fig. 10. Complex size probability distribution for 2 mg/mL Statherin with and without phosphorylated serines at 150 mM ionic strength.

Temperature also affects the structure of the complexes. Overall, the asphericity increases as a function of temperature for complexes of the same size, as seen in Fig. 11b. In addition, the radius of gyration also shows the same trend, for example, for complexes of seven proteins, the R_g goes from 22.8 ± 0.1 to 29.8 ± 0.2 Å when temperature changes from 15 to 50 °C. These changes reflect an

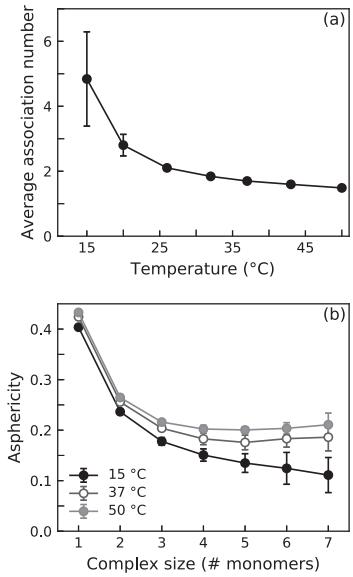


Fig. 11. (a) Average association number as a function of temperature at 5 mg/mL. (b) Asphericity *versus* association number at 15, 37 and 50 °C.

increased flexibility in the complexes, which is expected due to the entropy increase. Although it was shown in the monomeric section that the structure of the individual protein chain changes upon temperature increase, it is expected to be of minor importance for the self-association process, due to the model capturing the trends without including such detail.

Model limitations and improvements

From the simulations, it is apparent that the model breaks down at higher concentrations. The exact concentration depends on the conditions, especially temperature and ionic strength. At the lower-salt concentrations (10 and 60 mM), no breakdown is observed even at 20 mg/mL. The breakdown can be connected with the implicit treatment of salt, since simulations with 150 mM explicit salt and 20 mg/mL protein or more still give an average size less than 10 chains/complex. Hence, an explicit treatment of electrostatics is suggested to provide better results, although at a high computational cost. In the model, the hydrophobic interaction, mimicking the effect of both the enthalpic contribution and the entropic effect on the water molecules, is regarded temperature independent. Including temperature dependence would change the exact values to a certain extent, although the trend would remain. Hence, it would not affect the conclusion that entropy in the system is the largest contributor to the temperature effect for this protein.

Conclusions

A modified version of the coarse-grained model in Ref. [25] have been shown capable to describe the Statherin complexes at lower concentration and provide extra insight regarding the structure of the complexes, as well as aiding in explaining the effect of external conditions on the self-association, in terms of a balance between different interactions and entropy. The findings are summarized in Fig. 12. Hydrophobic interaction is shown to be the major driving force for the self-association, due to urea inhibiting complex formation. The size decrease as a result of increased temperature is regarded as an entropic effect, while electrostatic interactions were

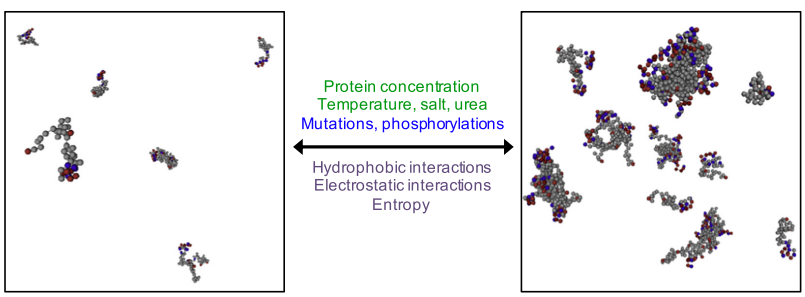


Fig. 12. Summary of what was shown to affect the Statherin association state. External factors are printed in green, chain characteristics in blue, and energetic and entropic factors in purple. In the snapshots, gray beads represent neutral residues; blue, positively charged residues; and red, negatively charged residues. The phosphorylated serines are marked in dark red. Counterions are omitted for clarity.

still shown to be of importance by balancing the hydrophobic attraction. In addition, it was demonstrated that mutations affecting the charge distribution can have a major effect on the self-association.

The self-association of Statherin is only one example of an IDP system dominated by intermolecular attractions; however, the similarities to micelle formation suggest that the established interactions are common for many systems, although with varying balance. It is therefore of interest to apply this model to other interacting IDPs in the future, as well as to continue the development for studies of systems with a higher complexity. Computational studies of IDP systems are advantageous in that it allows for separation of different contributions and a faster screening of mutations. In combination with experiments, it opens up for a deeper understanding of the function and behavior of IDPs.

Methods and Model

SAXS

Sample preparation

The buffers, all containing 20 mM Tris [>99.9%, CAS (77-86-1); Saveen Werner AB], and varying concentrations of NaCl [reagent grade, CAS (7647-14-5); Sharlau] and urea [ReagentPlus ≥99.5%, CAS (57-13-6); Sigma-Aldrich] were prepared with Milli-Q water, and by dropwise addition of 1 M HCl, the pH was set at room temperature to correspond to 8.1 at the measuring temperature. Thereafter, the buffers were filtered through a hydrophilic polypropylene 0.2 μm membrane (Pall Corporation).The Statherin powder (purchased from Genemed Synthesis, Inc.) was dissolved in buffer with a small addition of NaOH to increase the pH, since the protein powder contained trifluoroacetate. Concentrating cells (Vivaspin 2, 2000 MWCO, Prod. No. VS02H92; Sartorius, Cambridge, United Kingdom) were used to remove low-molecular-weight impurities. The samples were rinsed with buffer corresponding to 30 times the sample volume, by centrifugation at 358*g* at 8 °C. To ensure an exact background in the SAXS measurements, the samples were dialyzed (Slide-A-Lyzer Dialysis Cassette, 2000 MWCO, Prod. No. 66203 or Slide-A-Lyzer MINI Dialysis Unit, 2000 MWCO, Prod. No. 69580; Thermo Scientific, USA) overnight at 6 °C. Before the SAXS measurements, the samples were centrifuged at 18,400*g* at 6 °C for at least 2 h to remove impurities. Thereafter, they were diluted to a concentration series, and the protein concentration was determined with a nanodrop spectrometer using λ = 280 nm and ε = 8740 M^{-1} cm^{-1}. The samples were centrifuged in small PCR tubes imminent to the SAXS measurements to remove any bubbles.

Measurements and analysis

SAXS experiments were performed at BM29, ESRF-Grenoble, France. The incident beam wavelength was 0.99 Å, and the distance between sample and detector (PILATUS 1M) was set to 2867 mm, giving the scattering vector 0.0039 – 0.49 $Å^{-1}$. The scattering vector, q, is defined as $q = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the incident beam. Several successive frames of the scattering from the samples were recorded with an exposure time of 0.5 or 1 s, depending on concentration and system. The scattering from the pure solvent, which was measured before and after each sample for the same exposure times, was subtracted from the sample scattering. Measurements were performed at 10, 20, 37 and 50 °C at 150 mM NaCl, and the forward scattering, I_0, was converted to absolute scale by water calibration. At 20 °C measurements were also performed for 10, 60 and 300 mM NaCl and 4 and 8 M urea. The data were processed and analyzed using the ATSAS package [45]. Special attention was paid to radiation damage by comparing the successive frames prior to background subtraction, and any affected data were rejected from further analysis. Both I_0 and R_g were determined from $P(r)$, although the Guinier approach was also used for comparison. The molecular weight used for calculating the association number was determined from I_0 (see Supplemental information). Considering standard uncertainties of the used values, the uncertainty of the association number can be estimated as approximately 10% [43,46].

For a description of the SEC inline with SAXS, used for obtaining the form factor of monomeric Statherin, we refer to Ref. [25].

CD

Protein was dissolved in and purified with 20 mM phosphate buffer (sodium phosphate dibasic dihydrate [Reag. Ph. Eur., CAS (10028-24-7); Sigma-Aldrich] and sodium phosphate monobasic monohydrate [ACS reagent, CAS (10049-21-5); Sigma-Aldrich]) at pH 8, using a concentrating cell, as described for the SAXS samples. The protein was diluted to approximately 0.13 mg/mL using 20 mM phosphate buffer with 10 or 150 mM NaF [≥99%, CAS (7681-49-4); Sigma-Aldrich] and for the 150 mM NaF with 0–8 M urea [ReagentPlus ≥99.5%, CAS (57–13-6); Sigma-Aldrich]. The samples were filtered using a 0.22-μm Millex–GV filter (Merk Millipore Ltd). CD spectra between 190 and 260 nm at temperatures 4 – 60 °C were recorded on a JASCO J-715 instrument with a PTC-348WI Peltier type cell holder for temperature control, averaging over three spectra for each sample, using a quartz cuvette with a 1-mm path length (HellmaAnalytics) and 20-nm/min scanning speed, 2-s response time, 1-nm band width, and 100-mdeg

sensitivity. At 20 °C, further measurements were performed for samples with 150 mM NaF and 2–8 M urea. The ellipticity reported is the mean residue ellipticity, defined as

$$[\theta]_{\mathrm{MRW}} = \theta \cdot \mathrm{MRW}/(10 \cdot d \cdot c), \qquad (1)$$

where θ is the observed ellipticity (mdeg), d the path length of the cell (cm), and c the protein concentration (mg/mL). The mean residue weight, MRW, is the molecular weight (Da) divided by the number of peptide bonds. The spectra were smoothed using a Savitzky–Golay filter. The effect of the Savitzky–Golay filter is presented in Fig. S4 in Supplemental information.

Coarse-grained model

We have employed a coarse-grained model in which each amino acid is modeled as a hard sphere, further described in Ref. [25]. For the inclusion of hydrophobic interaction, a short-ranged potential is added to the model:

$$U_{\mathrm{hphob}} = -\sum_{\mathrm{neutral}} \frac{\varepsilon_{\mathrm{hphob}}}{r_{ij}^{6}} \qquad (2)$$

where the summation extends over all neutral amino acids, $r_{ij} = \mid \mathbf{R}_i - \mathbf{R}_j \mid$ is the center-to-center distance between two beads and $\mathbf{R}$ refers to the coordinate vector. $\varepsilon_{\mathrm{hphob}}$ is $1.32 \cdot 10^4$ kJ Å/mol, which corresponds to an attraction of 1.32 kT at closest contact, determined by comparing the average complex size with experimental results on the reference system.

Simulation aspects

The equilibrium properties of the model systems were obtained by Metropolis Monte Carlo simulations in the canonical (NVT) ensemble, utilizing the simulation package Molsim [47], version 4.8.8. Forty-five protein chains were enclosed in a cubic box of varying volume, dependent on the protein concentration. Periodic boundary conditions were applied in all directions. The long-ranged Coulomb interactions were truncated using the minimum image convention.

To accelerate the examination of the configurational space, five different types of displacements were allowed: (i) translational displacement of a single bead, (ii) pivot rotation [48,49], (iii) translation of the entire chain, (iv) slithering move [50], and (v) cluster displacements. Counterions were only moved individually by translation. The cluster displacement was performed as a translational displacement of the chain of a selected particle as well as all chains whose center of mass were less than 40 Å away from the selected particle. The cluster displacement was automatically rejected if the number of particles within the cluster changed, that is, if the displacement caused two clusters to merge. The probability of the different trial moves was weighted so that 80% of the trial moves were single bead displacements, 5% were pivot rotations, 5% were chain displacements, 3% were slithering moves, and 7% were cluster moves. Initially, the proteins were randomly placed in the box and an equilibrium simulation of typically $3 \cdot 10^5$ trial moves/bead was performed. The proceeding production run comprised at least 10^6 passes divided into subdivisions of 10^5 passes. To ensure accurately sampled simulations, the contact probability of each chain individually and the variations of contact number along the propagation of the simulation were analyzed (data not shown).

For all simulated quantities except the average association number, the reported uncertainty is one standard deviation of the mean. It is estimated from the deviation among the means of the subdivisions of the total number of MC passes, according to

$$\sigma^2(\langle x \rangle) = \frac{1}{n_s(n_s-1)} \sum_{s=1}^{n_s} \left(\langle x \rangle_s - \langle x \rangle\right)^2, \qquad (3)$$

where $\langle x \rangle_s$ is the average of quantity x from one subdivision, $\langle x \rangle$ the average of x from the total simulation, and n_s the number of subdivisions. For the average association number, the reported uncertainty is the standard deviation of the means of all subdivisions.

Analyses

The calculation of the scattering profile from simulation is described in Ref. [25]. In the analyses of complexes, two chains were assigned to the same complex if the center-to-center distance between two beads in the two different chains was less than 5 Å. The same geometric condition was used for defining if a bead was in contact with another chain, which was the basis for monitoring the variations of contact number along the propagation, and calculating the contact probability for beads along the chain. Contact probability for the beads is defined as the number of passes in which the bead is in contact with at least one bead from another chain, divided by the total number of passes in the simulation. Similarly, contact probability for a chain is calculated as the number of passes in which the chain is in a complex divided by the total number of passes in the simulation and the overall contact probability is the average over all chains. The complex size probability distribution was calculated according to

$$P_n = \frac{n\langle N_n^{\mathrm{complex}} \rangle}{\sum_n n\langle N_n^{\mathrm{complex}} \rangle}, \qquad (4)$$

where $\langle N_n^{\text{complex}} \rangle$ is the average number of complexes consisting of n chains, and $\sum_n n\langle N_n^{\text{complex}} \rangle$ is equal to the number of chains in the system, due to chain conservation. Note that P_n is weighted by the number of chains in a complex. The average association number was calculated from the complex size probability distribution, as

$$N_{\text{assoc}} = \sum_n nP_n. \quad (5)$$

The radial number density profile was calculated for each complex size and bead type individually. The radial number density at each distance is defined as the number of beads within a shell at that distance from the center-of-mass of the complex core, divided by the shell volume. The complex core was defined to consist of the beads 15–44 in each chain.

The shape of the complexes was quantified by the principal moments of the gyration tensor and the asphericity. The gyration tensor was defined as

$$S = \frac{1}{N}\begin{pmatrix} \sum_i^N X_i^2 & \sum_i^N X_i Y_i & \sum_i^N X_i Z_i \\ \sum_i^N X_i Y_i & \sum_i^N Y_i^2 & \sum_i^N Y_i Z_i \\ \sum_i^N X_i Z_i & \sum_i^N Y_i Z_i & \sum_i^N Z_i^2 \end{pmatrix}, \quad (6)$$

where $A_i = (a_i - a_{\text{com}})$ for $a = x, y, z$, and N is the number of beads in the complex. Transformation to a principal axis system such that

$$S = \text{diag}(R_1^2, R_2^2, R_3^2) \quad (7)$$

diagonalizes S and $R_1^2 \geq R_2^2 \geq R_3^2$ are the eigenvalues of S, also called the principal moments of the gyration tensor. In the simulations, the ensemble averages of the eigenvalues were calculated for each complex size separately. The asphericity, defined as

$$\alpha_s = \frac{(\langle R_1^2 \rangle - \langle R_2^2 \rangle)(\langle R_2^2 \rangle - \langle R_3^2 \rangle)(\langle R_3^2 \rangle - \langle R_1^2 \rangle)}{2(\langle R_1^2 \rangle + \langle R_2^2 \rangle + \langle R_3^2 \rangle)^2}, \quad (8)$$

ranges between 0 for a perfect sphere and 1 for a rod.

Acknowledgments

The authors thank the European Synchrotron Radiation Facility (ESRF) for providing beamtime, Dr. Bart Van Laer at ESRF for providing assistance in using beamline BM29, and Magnus Bergvall's foundation for financial support. The study was supported by the Science Faculty project grant program for research with neutrons and synchrotron light (Lund University Strategic funds for MAX-IV and European Spallation Source). The simulations were performed on resources provided by the Swedish National Infrastructure for Computing at the Center for scientific and technical computing at Lund University (Lunarc).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2018.11.027.

Received 10 September 2018;
Received in revised form 12 November 2018;
Accepted 28 November 2018
Available online 7 December 2018

Keywords:
intrinsically disordered proteins;
SAXS;
self-association;
Monte Carlo simulations;
coarse-graining

Abbreviations used:
IDPs, intrinsically disordered proteins; IDRs, intrinsically disordered regions; SAXS, small-angle X-ray scattering; SEC, size-exclusion chromatography; PPII, poly-proline II.

References

[1] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (10) (2002) 527–533, https://doi.org/10.1016/S0968-0004(02)02169-2.
[2] A. Dunker, J. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (1) (2001) 26–59, https://doi.org/10.1016/S1093-3263(00)00138-8.
[3] A.K. Dunker, J. Gough, Sequences and topology: intrinsic disorder in the evolving universe of protein structure, Curr. Opin. Struct. Biol. 21 (3) (2011) 379–381, https://doi.org/10.1016/j.sbi.2011.04.002.
[4] J. Habchi, P. Tompa, S. Longhi, V.N. Uversky, Introducing protein intrinsic disorder, Chem. Rev. 114 (13) (2014) 6561–6588, https://doi.org/10.1021/cr400514h.
[5] A.K. Dunker, P. Romero, Z. Obradovic, E.C. Garner, C.J. Brown, Intrinsic protein disorder in complete genomes, Genome Inform. 11 (2000) 161–171, https://doi.org/10.11234/gi1990.11.161.
[6] P. Romero, Z. Obradovic, C. Kissinger, J. Villafranca, E. Garner, S. Guilliot, A. Dunker, Thousands of proteins likely

to have long disordered regions, Pac. Symp. Biocomput. 3 (1998) 437–448.
[7] J. Ward, J. Sodhi, L. McGuffin, B. Buxton, D. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337 (3) (2004) 635–645, https://doi.org/10.1016/j.jmb.2004.02.002.
[8] B. Xue, A.K. Dunker, V.N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, J. Biomol. Struct. Dyn. 30 (2) (2012) 137–149, https://doi.org/10.1080/07391102.2012.675145.
[9] J. Liu, J.R. Faeder, C.J. Camacho, Toward a quantitative theory of intrinsically disordered proteins and their function, Proc. Natl. Acad. Sci. U. S. A. 106 (47) (2009) 19819–19823, https://doi.org/10.1073/pnas.0907710106.
[10] V.N. Uversky, Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders, Front. Aging Neurosci. 7 (2015) 18, https://doi.org/10.3389/fnagi.2015.00018.
[11] S.F. Banani, H.O. Lee, A.A. Hyman, M.K. Rosen, Biomolecular condensates: organizers of cellular biochemistry, Nat. Rev. Mol. Cell Biol. 18 (2017) 285–298.
[12] S. Weber, C. Brangwynne, Inverse size scaling of the nucleolus by a concentration-dependent phase transition, Curr. Biol. 25 (5) (2015) 641–646, https://doi.org/10.1016/j.cub.2015.01.012.
[13] C.P. Brangwynne, C.R. Eckmann, D.S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, A.A. Hyman, Germline p granules are liquid droplets that localize by controlled dissolution/condensation, Science 324 (5935) (2009) 1729–1732, https://doi.org/10.1126/science.1172046.
[14] L.-P. Bergeron-Sandoval, N. Safaee, S. Michnick, Mechanisms and consequences of macromolecular phase separation, Cell 165 (5) (2016) 1067–1079, https://doi.org/10.1016/j.cell.2016.05.026.
[15] A.L. Darling, Y. Liu, C.J. Oldfield, V.N. Uversky, Intrinsically disordered proteome of human membrane-less organelles, Proteomics 18 (5–6) (2018) 1700193, https://doi.org/10.1002/pmic.201700193.
[16] V.N. Uversky, Protein intrinsic disorder-based liquid–liquid phase transitions in biological systems: complex coacervates and membrane-less organelles, Adv. Colloid Interf. Sci. 239 (2017) 97–114, https://doi.org/10.1016/j.cis.2016.05.012.
[17] V.N. Uversky, Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder, Curr. Opin. Struct. Biol. 44 (2017) 18–30, https://doi.org/10.1016/j.sbi.2016.10.015.
[18] S. Rauscher, R. Pomès, Molecular simulations of protein disorder, Biochem. Cell Biol. 88 (2) (2010) 269–290, https://doi.org/10.1139/O09-169.
[19] V.M. Burger, T. Gurry, C.M. Stultz, Intrinsically disordered proteins: where computation meets experiment, Polymers 6 (10) (2014) 2684–2719, https://doi.org/10.3390/polym6102684.
[20] R.B. Best, W. Zheng, J. Mittal, Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association, J. Chem. Theory Comput. 10 (11) (2014) 5113–5124, https://doi.org/10.1021/ct500569b.
[21] S. Piana, A.G. Donchev, P. Robustelli, D.E. Shaw, Water dispersion interactions strongly influence simulated structural properties of disordered protein states, J. Phys. Chem. B 119 (16) (2015) 5113–5123, https://doi.org/10.1021/jp508971m.
[22] J. Henriques, C. Cragnell, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment, J. Chem. Theory Comput. 11 (7) (2015) 3420–3431, https://doi.org/10.1021/ct501178z.
[23] S. Rauscher, V. Gapsys, M.J. Gajda, M. Zweckstetter, B.L. de Groot, H. Grubmüller, Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment, J. Chem. Theory Comput. 11 (11) (2015) 5513–5524, https://doi.org/10.1021/acs.jctc.5b00736.
[24] J. Henriques, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the tip4p-d water model and the representativeness of protein disorder models, J. Chem. Theory Comput. 12 (7) (2016) 3407–3415, https://doi.org/10.1021/acs.jctc.6b00429.
[25] C. Cragnell, E. Rieloff, M. Skepö, Utilizing coarse-grained modeling and Monte Carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions, J. Mol. Biol. 430 (16) (2018) 2478–2492, https://doi.org/10.1016/j.jmb.2018.03.006.
[26] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1) (1982) 105–132, https://doi.org/10.1016/0022-2836(82)90515-0.
[27] Y. Lin, S.L. Currie, M.K. Rosen, Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs, J. Biol. Chem. 292 (46) (2017) 19110–19120, https://doi.org/10.1074/jbc.M117.800466.
[28] C. Pak, M. Kosno, A. Holehouse, S. Padrick, A. Mittal, R. Ali, A. Yunus, D. Liu, R. Pappu, M. Rosen, Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein, Mol. Cell 63 (1) (2016) 72–85, https://doi.org/10.1016/j.molcel.2016.05.042.
[29] R. Williams, Z. Obradovic, V. Mathura, W. Braun, E. Garner, J. Young, S. Takayama, C. Brown, A. Dunker, The protein non-folding problem: amino acid determinants of intrinsic order and disorder, Pac. Symp. Biocomput. 2001 (6) (2001) 89–100.
[30] G.A. Naganagowda, T.L. Gururaja, M.J. Levine, Delineation of conformational preferences in human salivary statherin by 1H, 31P NMR and CD studies: sequential assignment and structure-function correlations, J. Biomol. Struct. Dyn. 16 (1) (1998) 91–107, https://doi.org/10.1080/07391102.1998.10508230.
[31] G.A. Elgavish, D.I. Hay, D.H. Schlesinger, 1H and 31P nuclear magnetic resonance studies of human salivary statherin, Int. J. Pept. Protein Res. 23 (3) (1984) 230–234, https://doi.org/10.1111/j.1399-3011.1984.tb02714.x.
[32] G. Goobes, R. Goobes, W.J. Shaw, J.M. Gibson, J.R. Long, V. Raghunathan, O. Schueler-Furman, J.M. Popham, D. Baker, C.T. Campbell, P.S. Stayton, G.P. Drobny, The structure, dynamics, and energetics of protein adsorption—lessons learned from adsorption of statherin to hydroxyapatite, Magn. Reson. Chem. 45 (S1) (2007) S32–S47, https://doi.org/10.1002/mrc.2123.
[33] N. Ramasubbu, L.M. Thomas, K.K. Bhandary, M.J. Levine, Structural characteristics of human salivary statherin: a model for boundary lubrication at the enamel surface, Crit. Rev. Oral Biol. Med. 4 (3) (1993) 363–370, https://doi.org/10.1177/10454411930040031501.
[34] P.A. Raj, M. Johnsson, M.J. Levine, G.H. Nancollas, Salivary statherin. Dependence on sequence, charge, hydrogen bonding potency, and helical conformation for adsorption to hydroxyapatite and inhibition of mineralization, J. Biol. Chem. 267 (9) (1992) 5968–5976.

[35] S.J. Whittington, B.W. Chellgren, V.M. Hermann, T.P. Creamer, Urea promotes polyproline II helix formation: implications for protein denatured states, Biochemistry 44 (16) (2005) 6269–6275, https://doi.org/10.1021/bi050124u.
[36] M. Kjaergaard, A.-B. Nørholm, R. Hendus-Altenburger, S.F. Pedersen, F.M. Poulsen, B.B. Kragelund, Temperature-dependent structural changes in intrinsically disordered proteins: formation of α-helices or loss of polyproline II? Protein Sci. 19 (8) (2010) 1555–1564, https://doi.org/10.1002/pro.435.
[37] M. Abu-Hamdiyyah, The effect of urea on the structure of water and hydrophobic bonding, J. Phys. Chem. 69 (8) (1965) 2720–2725, https://doi.org/10.1021/j100892a039.
[38] W. Bruning, A. Holtzer, The effect of urea on hydrophobic bonds: the critical micelle concentration of *n*-dodecyltrimethylammonium bromide in aqueous solutions of urea1, J. Am. Chem. Soc. 83 (23) (1961) 4865–4866, https://doi.org/10.1021/ja01484a044.
[39] U. Thapa, K. Ismail, Urea effect on aggregation and adsorption of sodium dioctylsulfosuccinate in water, J. Colloid Interface Sci. 406 (2013) 172–177, https://doi.org/10.1016/j.jcis.2013.06.009.
[40] J. Broecker, S. Keller, Impact of urea on detergent micelle properties, Langmuir 29 (27) (2013) 8502–8510, https://doi.org/10.1021/la4013747.
[41] A. Malliaris, J. Le Moigne, J. Sturm, R. Zana, Temperature dependence of the micelle aggregation number and rate of intramicellar excimer formation in aqueous surfactant solutions, J. Phys. Chem. 89 (12) (1985) 2709–2713, https://doi.org/10.1021/j100258a054.
[42] R. Zana, C. Weill, Effect of temperature on the aggregation behaviour of nonionic surfactants in aqueous solutions, J. Phys. Lett. 46 (20) (1985) 953–960.
[43] C. Moitzi, I. Portnaya, O. Glatter, O. Ramon, D. Danino, Effect of temperature on self-assembly of bovine β-casein above and below isoelectric pH. Structural analysis by cryogenic-transmission electron microscopy and small-angle x-ray scattering, Langmuir 24 (7) (2008) 3020–3029, https://doi.org/10.1021/la702802a.
[44] M. Kenward, M.D. Whitmore, A systematic Monte Carlo study of self-assembling amphiphiles in solution, J. Chem. Phys. 116 (8) (2002) 3455–3470, https://doi.org/10.1063/1.1445114.
[45] D. Franke, M.V. Petoukhov, P.V. Konarev, A. Panjkovich, A. Tuukkanen, H.D.T. Mertens, A.G. Kikhney, N.R. Hajizadeh, J.M. Franklin, C.M. Jeffries, D.I. Svergun, *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions, J. Appl. Crystallogr. 50 (4) (2017) 1212–1225, https://doi.org/10.1107/S1600576717007786.
[46] D. Orthaber, A. Bergmann, O. Glatter, SAXS experiments on absolute scale with Kratky systems using water as a secondary standard, J. Appl. Crystallogr. 33 (2) (2000) 218–225, https://doi.org/10.1107/S0021889899015216.
[47] J. Reščič, P. Linse, MOLSIM: a modular molecular simulation software, J. Comput. Chem. 36 (16) (2015) 1259–1274, https://doi.org/10.1002/jcc.23919.
[48] M. Lal, Monte Carlo computer simulation of chain molecules. I, Mol. Phys. 17 (1) (1969) 57–64.
[49] N. Madras, A.D. Sokal, The pivot algorithm: a highly efficient Monte Carlo method for the self-avoiding walk, J. Stat. Phys. 50 (1) (1988) 109–186, https://doi.org/10.1007/BF01022990.
[50] F.T. Wall, F. Mandel, Macromolecular dimensions obtained by an efficient Monte-Carlo method without sample attrition, J. Chem. Phys. 63 (11) (1975) 4592–4595.

Supplemental information for Assessing the intricate balance of intermolecular interactions upon self-association of intrinsically disordered proteins

Ellen Rieloff[a,*], Mark Tully[b], Marie Skepö[a,*]

[a] *Theoretical Chemistry, Lund University, POB 124, SE-221 00 Lund, Sweden*
[b] *European Synchrotron Radiation Facility (ESRF), Grenoble, France*

Analysis of Small-angle X-ray scattering data

Here we present collected SAXS curves and additional information regarding the determination of forward scattering and radius of gyration for the data collected at 20 °C with 10 and 150 mM NaCl. The data at other salt concentrations and temperatures were treated in the same way. Figure S1 shows the scattering curves for Statherin with increasing protein concentration measured at 20 °C, for 150 and 10 mM NaCl. At higher concentrations than presented in the figure, a clear depression at low q was shown, and therefore such data was excluded from analysis. The forward scattering and radius of gyration were determined by both the Guinier method and from the pair distance distribution function, P(r). Guinier plots with fits to the used range are presented in Figure S2 for the data at 150 mM NaCl and in Figure S3 for the data at 10 mM NaCl. The used range in the Guinier method was limited to $qR_g < 0.8$, or extended to $qR_g < 1.0$ when appropriate, since that is usually the linear region for an IDP [1]. The figures also include the fits in the P(r) analysis. The resulting values are presented in Table S1 and Table S2. Overall the agreement between the two methods are good, although the radius of gyration from the pair distance distribution is slightly larger. Since it is

*Corresponding author
Email addresses: ellen.rieloff@teokem.lu.se (Ellen Rieloff), marie.skepo@teokem.lu.se (Marie Skepö)

known that the Guinier law is less appropriate for describing an unfolded chain and therefore can underestimate the size of intrinsically disordered proteins, we have presented the values from the pair distribution function in the article.

The molecular weight, M_{w}, was calculated using the following equation:

$$M_{\mathrm{w}} = \frac{I_0 \cdot I_{0\mathrm{w,ref}} \cdot N_{\mathrm{A}}}{I_{0\mathrm{w,meas}} \cdot c([\rho_{\mathrm{p}} - \rho_{\mathrm{s}}]\nu_{\mathrm{p}})} \tag{1}$$

where the forward scattering I_0 is given in arbitrary units, $I_{0\mathrm{w,ref}}$ is the absolute scattering of water, N_{A} is the Avogadro constant, $I_{0\mathrm{w,meas}}$ the measured scattering of water in arbitrary units, and c the protein concentration. The electron density of the protein, ρ_{p}, was determined from the number of electrons in the protein and the molecular weight, while the electron density of the solvent, ρ_{s}, was calculated with MulCh [2] based on the Tris and NaCl concentrations. The partial specific volume of the protein, ν_{p}, was calculated from the amino acid sequence using Sednterp [3], assuming no effect from phosphorylations.

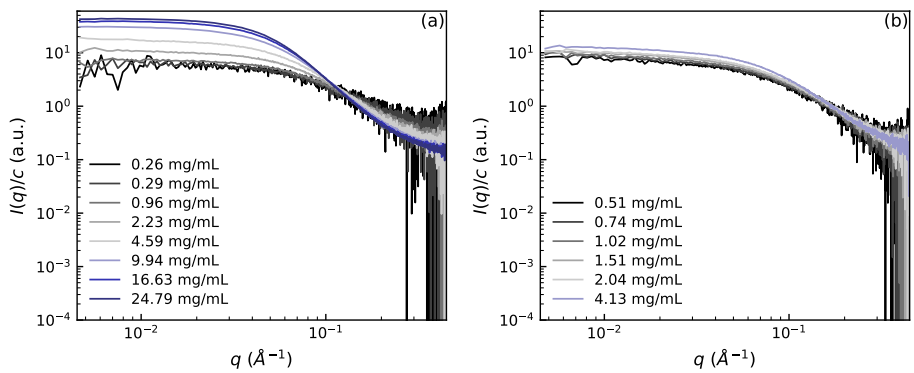


Figure S1: Overlayed scattering curves for Statherin with (a) 150 mM NaCl and (b) 10 mM NaCl, and 20 mM Tris, pH 8.1, at 20 °C.

Table S1: Forward scattering, I_0, and radius of gyration, R_g, determined both by the Guinier approximation and from the pair distribution function, for the data at 150 mM NaCl and 20 °C.

c (mg/mL)	$I_{0,\mathrm{Guinier}}/c$ (a.u.)	$I_{0,\mathrm{P(r)}}/c$ (a.u.)	$R_{\mathrm{g,Guinier}}$ (Å)	$R_{\mathrm{g,P(r)}}$ (Å)
0.26	5.9 ± 0.1	6.0 ± 0.1	17.1 ± 0.6	19.0 ± 0.4
0.29	6.5 ± 0.1	6.4 ± 0.1	20.7 ± 0.9	20.1 ± 0.3
0.96	7.3 ± 0.1	7.4 ± 0.1	20.0 ± 0.2	20.8 ± 0.2
2.23	10.5 ± 0.1	10.5 ± 0.1	22.5 ± 0.2	23.1 ± 0.2
4.59	17.0 ± 0.1	17.1 ± 0.1	25.8 ± 0.2	26.9 ± 0.3
9.94	30.6 ± 0.1	30.7 ± 0.1	31.4 ± 0.8	31.8 ± 0.1
16.63	39.5 ± 0.1	39.7 ± 0.1	32.2 ± 0.3	32.7 ± 0.1
24.79	44.4 ± 0.1	45.4 ± 0.1	31.9 ± 0.6	33.2 ± 0.1

Table S2: Forward scattering, I_0, and radius of gyration, R_g, determined both by the Guinier approximation and from the pair distribution function, for Statherin at 10 mM NaCl and 20 °C.

c (mg/mL)	$I_{0,\mathrm{Guinier}}/c$ (a.u.)	$I_{0,\mathrm{P(r)}}/c$ (a.u.)	$R_{\mathrm{g,Guinier}}$ (Å)	$R_{\mathrm{g,P(r)}}$ (Å)
0.51	6.7 ± 0.1	6.8 ± 0.1	19.8 ± 0.9	21.9 ± 0.8
0.74	7.5 ± 0.1	7.5 ± 0.1	22.7 ± 0.5	24.2 ± 0.7
1.02	8.0 ± 0.1	8.0 ± 0.1	22.0 ± 0.3	23.1 ± 0.4
1.51	8.8 ± 0.1	9.0 ± 0.1	22.2 ± 0.3	23.9 ± 0.3
2.04	9.4 ± 0.1	9.5 ± 0.1	21.9 ± 0.3	23.4 ± 0.3
4.13	11.3 ± 0.1	11.5 ± 0.1	21.8 ± 0.2	23.1 ± 0.1

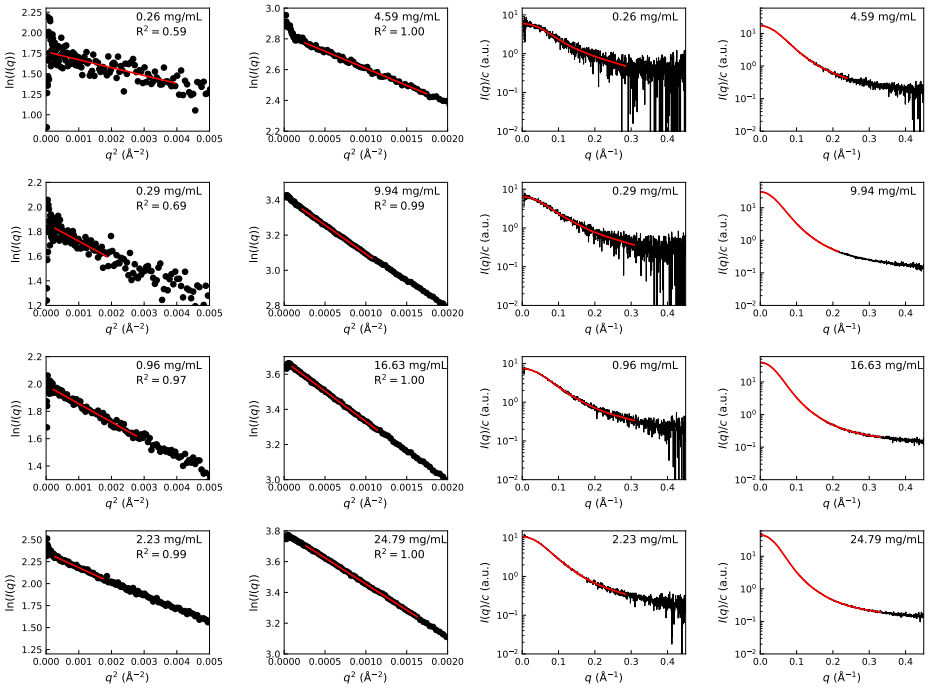


Figure S2: Guinier plots (the two left columns) and SAXS curves with the fits obtained in the P(r) analysis (the two right columns) for the reference system, obtained with 150 mM NaCl, 20 mM Tris, pH 8.1, at 20 °C. The red straight lines in the Guinier plots are the Guinier fits in the used range. The red curves are obtained in the indirect transform for obtaining P(r), using the ATSAS package [4].

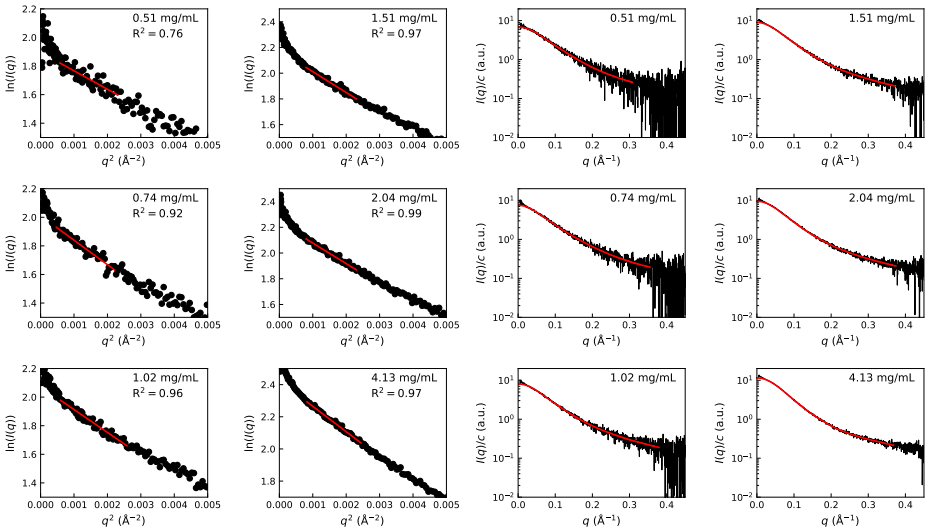


Figure S3: Guinier plots with red lines corresponding to the Guinier approximation in the used range (the two left columns) and SAXS curves with the fits obtained in the P(r) analysis given in red (the two right columns) for Statherin with 10 mM NaCl, 20 mM Tris, pH 8.1, at 20 °C. The red straight lines in the Guinier plots are the Guinier fits in the used range. The red curves are obtained in the indirect transform for obtaining P(r), using the ATSAS package [4].

Circular Dichroism data

To provide an estimate of the variation in the circular dichroism data, Figure S4 shows how the smoothened data achieved by applying a Savitzky–Golay filter relates to the raw data for two replicates at 4 and 28 °C. For each replicate a new sample was prepared and the measurements of the different replicates were made on different days. At 4 °C the agreement between the two replicates is excellent, while there is a small difference between the replicates at 28 °C. Factors contributing to the variation involves noise as well as uncertainties in the measured concentration.

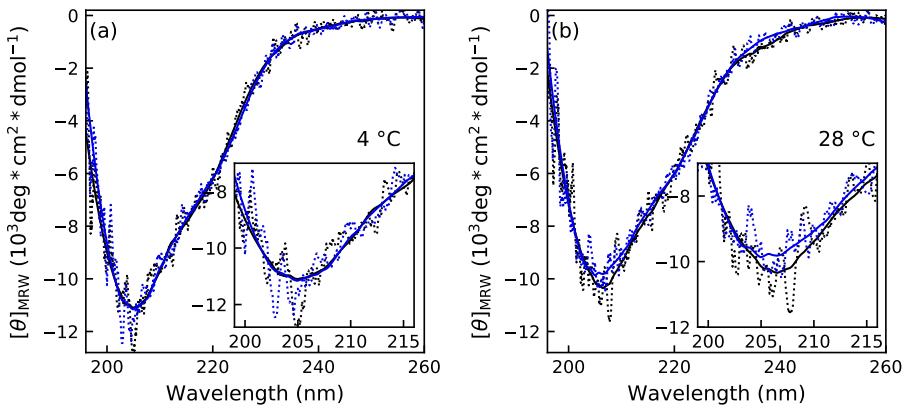


Figure S4: Raw data (dotted lines) and smoothened data (solid lines) from two different circular dichroism measurements (blue and black) for Statherin at (a) 4 °C and (b) 28 °C, in 20 mM phosphate buffer, 150 mM NaF, pH 8. The insets are enlargements of the data around the minimum.

References

[1] V. Receveur-Brechot, D. Durand, How random are intrinsically disordered proteins? a small angle scattering perspective, Curr. Protein Pept. Sci. 13 (1) (2012) 55–75. doi:doi:10.2174/138920312799277901.

[2] A. E. Whitten, S. Cai, J. Trewhella, *MULCh*: modules for the analysis of small-angle neutron contrast variation data from biomolecular assemblies, J. Appl. Crystallogr. 41 (1) (2008) 222–226. doi:10.1107/S0021889807055136.

[3] T. Hurton, A. Wright, G. Deubler, B. Bashir, Sedimentation interpretation program, 20120828 BETA, based on the original program by D. B. Hayes and T. Laue and J. Philo. Available at http://rasmb.org/sednterp/.

[4] D. Franke, M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, D. I. Svergun, *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions, J. Appl. Crystallogr. 50 (4) (2017) 1212–1225. doi:10.1107/S1600576717007786.

Printed by Media-Tryck, Lund 2019 NORDIC SWAN ECOLABEL 3041 0903

ISBN: 978-91-7422-636-2

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University