LU TP 22-47
August 2022

Phase behavior of mixtures of polyampholytic proteins and RNA: a toy model study

Sonny Nilsson

Department of Astronomy and Theoretical Physics, Lund University

Master thesis supervised by Anders Irbäck

LUND
UNIVERSITY

Abstract

Biomolecular condensates are aggregates formed from liquid-liquid phase separation through the interactions between nucleic acids and multivalent proteins. These condensates are essential for many biochemical processes inside the cell. Therefore, there has been a large effort during the last decade to create models and methods to describe these systems. Mixtures of RNA and proteins are very common in nature, making it conceivable that RNA-protein interactions are important in many biomolecular condensates. In this thesis these biomolecules are modeled as simple chains of charged beads. First, one- and two-component protein systems are investigated, with results that are consistent with previous findings made by other groups. RNA was then added to few different one-component protein systems. It was found that the presence of a few RNA molecules increases the aggregation propensity in the sense that aggregation sets in at a higher temperature. When the amount of RNA in the system was increased past a certain threshold, this trend was reverted.

Populärvetenskaplig beskrivning

Cellen är den fundamentala byggstenen för allt liv vi idag känner till. Inuti celler kompartmentaliseras viktiga funktioner till organeller, vilket är cellernas motsvarighet till organ. Klassiska organeller kan kontrollera sin biokemiska miljö med hjälp av ett membran som avgränsar organellens inre från cytoplasman. Under de senaste åren har forskare dock funnit att membran inte är universell för organeller. Biomolekyler, likt vattenmolekyler, kan attrahera och repellera varandra. Precis som vattenmolekyler har olika aggregationstillstånd, har många biomolekyler också det. Dessa biomolekyler kan under lämpliga förhållanden bildar droppar med mycket högre täthet inuti droppen än utanför. Dropparna har kommit att kallas för biomolekylära kondensat och kan ses som membranlösa organeller.

Huvudkomponenter i biomolekulära kondensat är proteiner och RNA. De proteiner som ingår tillhör ofta den klass av proteiner som, istället för att vika sig till en särskild struktur, är strukturellt oordnande. Precis som ordinära proteiner är de strukturellt oordnade kedjemoleyler med aminosyror som byggstenar. De ingående aminosyrornas fysikaliska egenskaper avgör hur ett protein beter sig. Det är dock inte enbart vilka aminosyror som förekommer i ett protein som spelar roll, utan även deras inbördes ordning längs kedjan är viktig.

Hos majoriteten av de funna biomolekylära kondensaten förekommer också en annan sorts biomolekyl - RNA. Istället för aminosyror är dessa uppbyggda av nukleotider. RNA-molekyler är starkt negativt laddade och attraherar därmed positivt laddade aminosyror som till exempel lysin. En möjlig hypotes är att RNA kan agera som aggregationsfrön, och hjälpa de strukturellt oordnade proteinerna att bilda kondensat.

Att förutsäga en given biomolekyls benägenhet att fasseparera har visat sig vara en utmaning. Analytiska teorier kan ibland användas för att ge en viss förståelse av dessa system. Dock bygger analytiska teorier på grova approximationer. Genom att använda numeriska simuleringar kan en del av dessa approximationer undvikas.

I detta examensarbete undersöker vi förmågan att fasseparera hos korta proteiner med olika laddningsmönster, genom numeriska simuleringar baserade på en enkel modell där varje aminosyra representeras som en kula. Vi studerar system med en eller två proteinsorter. Dessutom undersöker vi hur ett system med en proteinsort påverkas när RNA-molekyler tillförs. Resultaten tyder på att tillförseln av en liten andel RNA ökar systemet benägenhet att aggregera.

Contents

1 Introduction

In recent years, a large body of research has illuminated the importance of membrane-less organelles, also commonly referred to as biomolecular condensates. An extensively studied example is P-granules, which are important for mRNA metabolism [1, 2, 3]. Stress granules have been shown to be biomolecular condensates which are triggered due to stressors applied to the cell, such as temperature changes and oxidation stress. Additionally, clustering of certain proteins in the cell membrane is important for signal transduction, where the clusters are thought to be biomolecular condensates [4]. Biomolecular condensates alter the local biochemical environment within cells by changing the concentration of certain proteins or RNA molecules and thereby the rate of chemical reactions. They also have the ability to change the physical environment such as viscosity [5].

Caenorhabditis elegans is a roundworm often used as a model organism when conducting biomolecular research. In 2009 Brangwynne et al. [6] showed that P-granules in this worm have liquid-like properties such as wettability, surface tension and repeated dissolution/condensation. It was then suggested that the physical mechanism liquid-liquid phase separation (LLPS) is responsible for the formation of these granules. This hypothesis has since been confirmed in multiple experiments [7]. Furthermore, certain proteins with repeating/"blocky" amino acid sequences seem to be very common among the proteins found in biomolecular condensates [7, 8, 9]. These proteins with low-complexity domains are often intrinsically disordered proteins (IDP), which, instead of folding into a specific 3D conformation, populate a broad ensemble of many different conformations. Due to their flexibility, IDPs have many locations where they can interact with other biomolecules. This is thought to be a driving force for LLPS [10].

Understanding the phase behaviour of biomolecules requires the use of statistical mechanics. The complete solution to the problem is encoded into the partition function [11]. However, computing the partition function of non-trivial systems in general impossible. For polymer-solvent mixing, a widely used approximate method is the Flory-Huggins mean field theory from the 1940's [12, 13]. An extension to polyelectrolytes was developed by Voorn and Overbeek in 1957 [14]. However, the mean-field methods are insensitive to the ordering of the monomers written along the polymer chain, which is known to be important in biomolecular LLPS. A more recent theoretical approach is usage of random phase approximation on polyampholytes [15] which does indeed account for the ordering in the polymer sequence, but is approximately valid only for low polymer densities [16].

The interaction energy between non-neighbouring amino acids is thought to be a combination of several different kinds of interaction such as van der Waal, hydrophobic/hydrophillic, electrostatic forces and the Pauli exclusion principle. The 21 different amino acids found in eukaryotic cells have all different structure and potentials.

To understand the properties of polymer systems, numerical simulations are helpful. One category of these simulations are the explicit chain simulations which is what this study will focus on.

A popular strategy is to use coarse-grained (CG) models, where biomolecules are represented by a chain of beads that interacts with other beads through two-body potentials. The loss of resolution is made up by a gain in computer efficiency which makes simulations of larger systems possible, while still preserving qualitative thermodynamical properties [17, 18].

The majority of the well-studied biomolecular condensates contain not only IDPs, but also RNA [19, 20]. The negatively charged RNA molecules interact with proteins through electrostatic as well as other forces [21, 22]. Thus it is conceivable that the addition of RNA molecules can function as "seeds" that IDPs can start to aggregate around, potentially leading to the formation of a biomolecular condensate.

In this thesis, we investigate an IDP-RNA system, where the molecules are modeled as chains of charged beads, interacting through a simplified piecewise constant potential. For comparison, we explore some one- and two-component IDP systems that have been studied before, using other models. These studied found that the aggregation temperature of one-component systems and the demixing propensity of the two-component system depend on the charge distribution along the chains, as measured by the "blockiness"/charge patterning called κ [23, 24, 25]. Using our model, we find qualitatively similar results, which indicates that the properties studied are largely insensitive to the precise form of the interactions.

Having verified this, we explore how RNA effects the aggregation propensity of an IDP. Here RNA is modeled as a polymer consisting of negatively charged beads. We find that adding a small amount of RNA increases the aggregation propensity of the system.

2 Theory and Methods

2.1 Biophysical Model

In this study amino acids contained in the proteins are modeled as hard beads with a step potential. The simulations are run with a constant number of beads N and in a box of volume V, yielding a bead density of $\rho = N/V$. Each sequential bead within all chains are separated by a constant distance of b. The chains of l beads reside inside a cubic box of length L b with a periodic boundary.

Let the ith bead have the position,

$$\vec{r}_i = (x_i^1, x_i^2, x_i^3) \quad i = 1, 2, 3, ..., N$$

where x_i^k are the kth Cartesian component of $\vec{r}_i$. Since a box with a periodic boundary is used, interactions can occur over the boundary. Let r_{ij} be the distance between beads of index i, j.

Then r_{ij} is calculated with:

$$r_{ij} = \sqrt{\sum_{k=1}^{3} \left[\min(|\Delta x_{ij}^k|, |L - \Delta x_{ij}^k|) \right]^2}.$$

The two-body potential between beads used is similar to the one used in Daniel Nilsson's thesis [26], but instead of a hydrophobic potential, here an electrostatic step potential with hard spheres is used:

$$E_{ij} = \begin{cases} \infty, & \text{if } r_{ij} < 0.75b \\ \epsilon Q_i Q_j, & \text{if } 0.75b < r_{ij} \leq 2b \\ 0, & \text{else} \end{cases}, \quad E = \sum_{i<j} E_{ij}$$

where $Q = \pm 1$ is the charge of residue i, j and ϵ is the interaction strength. The total system energy E is the sum of all the interaction energies.

2.2 Monte Carlo Simulation

Monte Carlo is a class of simulations where random sampling from a probability distribution are performed to obtain numerical results. In this case, we desire to sample configurations of the system from the canonical (NVT) ensemble, meaning the probability to observe a configuration S is:

$$p_\beta(S) = \frac{\exp(-\beta E(S))}{Z(\beta)} \tag{2.1}$$

where Z is the canonical partition function and $\beta = 1/(k_b T)$. A common technique used to sample from this distribution is the Metropolis-Hastings Algorithm (MHA) [27]. This algorithm is a Markov chain Monte Carlo technique, meaning that every generated state, S', depends only on the current state S and the system parameters.

Let $W(S, S')$ be the conditional probability to transition to S', given that the system is in S. To ensure that MHA samples from a given (desired) probability distribution $P(S)$, the two following two conditions are sufficient:

- That $P(S)$ is a stationary distribution, meaning that $\sum_S P(S)W(S, S') = P(S')$
- The system is ergodic, meaning that every state can from reached by any other.

Even if the second criterion is fulfilled in theory, the expected time it takes to go between states can be significantly larger than simulation time. Detailed balance is a sufficient condition to guarantee that the first criteria is fulfilled:

$$P(S)W(S, S') = P(S')W(S', S). \tag{2.2}$$

In MHA the transition probability is decomposed into a proposal probability, $F(S, S')$, and an acceptance probability, $A(S, S')$, such that $W(S, S') = A(S, S')F(S, S')$ for $S \neq S'$. Inserting this form into eq. 2.2, one finds

$$\frac{A(S, S')}{A(S', S)} = \frac{P(S')F(S', S)}{P(S)F(S, S')} = \exp(-\beta(E' - E))\frac{F(S', S)}{F(S, S')}$$

$A(S, S')$ and $A(S', S)$ are not uniquely determined. A common choice is

$$A(S, S') = \min(1, e^{-\beta\Delta E}\frac{F(S', S)}{F(S, S')}).$$

If now the proposition probabilities are symmetric ($\frac{F(S',S)}{F(S,S')} = 1$) one further obtains

$$A(S, S') = \min(1, e^{-\beta\Delta E}). \tag{2.3}$$

Now this scheme will only yield the true distribution if the simulation have run long enough to "forget" the initial states because these early states will not be directly sampled from the Boltzmann distribution. This means a "burn-in" period has to be initially conducted. The proposed transitions, also referred here to simply as "updates", will influence how fast the system will converge to the right distribution [28]. This will be the topic of the next subsection.

2.3 MHA Updates

2.3.1 Single Chain Updates

The first class of updates are the rotation of one/two bead. To rotate a bead of index $i + 1$, a vector is drawn between the two neighbouring beads $i, i + 2$. The bead is then rotated a randomly chosen angle, θ, around this vector, acting as the axis of rotation. Another similar move is the rotation of two neighbouring beads with indices $(i + 1, i + 2)$ simultaneously which is performed similarly, but around the vector between the beads i and $i + 3$.

A classical move that is included here is the pivot which is a "folding" of a chain. This is performed by selecting a random bead i, then creating a unit vector, $\vec{v}$, of random direction from that bead. Then either neighbouring bead $i - 1$ or $i + 1$ is randomly selected. All of the beads on the neighbour's side are then rotated around $\vec{v}$ going through bead i by a randomly determined angle.

The last two updates for single chains are rigid-body rotation and translation. Chain rotation is performed in a similar way as the previous rotations, but now the axis of rotation goes through the center of mass of the chain.

2.3.2 Cluster Updates

One of the issues sampling near a critical point is that the correlation time of many-particle systems grows faster than linear size of the system. To reduce correlation time it was suggested by Robert H. Swendsen and Jian-Sheng Wang in 1987 that certain cluster updates can be performed, called Swendsen-Wang cluster algorithm [28]. They showed that when introducing these types of updates, the correlation time in a system decreased significantly. The main idea is that a cluster of chains is constructed, followed by a translation/rotation of said cluster. The procedure to construct a cluster is:

1. Select a random chain i in the system.
2. Let E_{ij} denote the interaction energy between chains i and j. Chain j is added to the cluster with probability $p = \max(1 - e^{\beta E_{ij}}, 0)$.
3. Repeat step 2 for all new chains added to the cluster. Stop when no new interactions can be found. The cluster obtained this way is then translated or rotated as a rigid body.

In a system containing only negative interactions, it turns out that the new state can always be accepted ($A(S, S') = 1$). However when positive interactions are present, negative contributions from the max function in step 2 above must be compensated for, in order for detailed balance to hold. Let $E_{0,+}/E_{1,+}$ be the sum of all positive interactions between the cluster and surrounding chains before/after the proposed update. Then

$$A(S, S') = \min\left(1, e^{\beta(E_{0,+} - E_{1,+})}\right).$$

The relative frequencies of the different updates is presented in Table 1, and are the same as in Ref. [26]. The program used in present study is a variant of that used in Ref. [26], modified so as to allow for positive interaction energies.

Table 1: Distribution of the proposition probabilities for the updates used in all of the simulation performed here.

Update	Probability (%)
1-bead Rotation	20
2-bead Rotation	20
Pivot	20
Chain Translation	15
Chain Rotation	15
Cluster Translation	5
Cluster Rotation	5

2.4 Phase Transitions

For a system in the canonical ensemble, the Helmholtz free energy

$$F = E - TS \tag{2.4}$$

where S is the entropy, is minimized. When two states have similar free energies, the system can spend significant time in both states. In LLPS the two states corresponds to a dilute phase, g, and a mixed phase, l. Figure 1 illustrates these two states with snapshots from a simulation.

Coexistence of these two types of states is observed along the so called bimodal line in the ρT phase diagram, as illustrated in Figure 2.

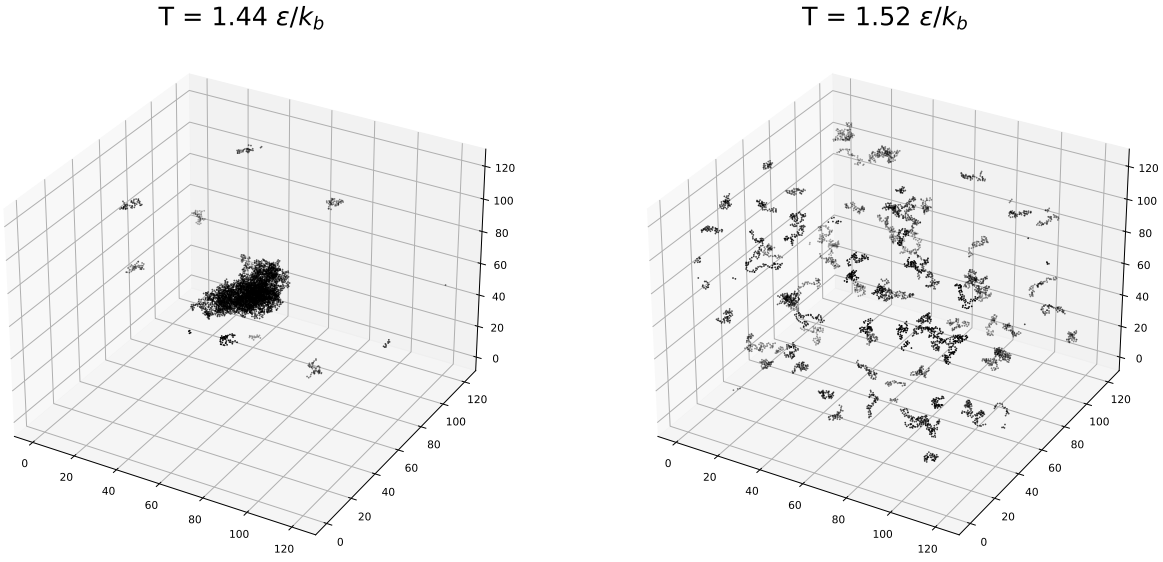


Figure 1: Snapshots of equilibrated particle configurations with the sv1 protein. Simulations ran at low temperatures result in a mixed states and high temperatures in a dilute states.

Since ρ_g and ρ_l corresponds to different system energies, there will be a signature behaviour in the run-time history of the energy and energy histogram, see Figure 3. In the run-time history there occurs transitions between two states corresponding to $E_g(T)$ and $E_l(T)$, separated by a distance ΔE. As seen in the histogram, the energy associated with the two states forms a bimodal distribution, which can be approximated by two separate Gaussian distributions. The transition temperature $T_b(\rho)$ may be defined as the temperature at which the areas under these distributions are the same.

The heat capacity, C_v, is an useful property of the system when looking for phase transitions. It is can be written in two ways:

$$C_v(T) = \frac{d\langle E\rangle_t(T)}{dT} = \frac{\delta^2 E}{k_b T^2} \tag{2.5}$$

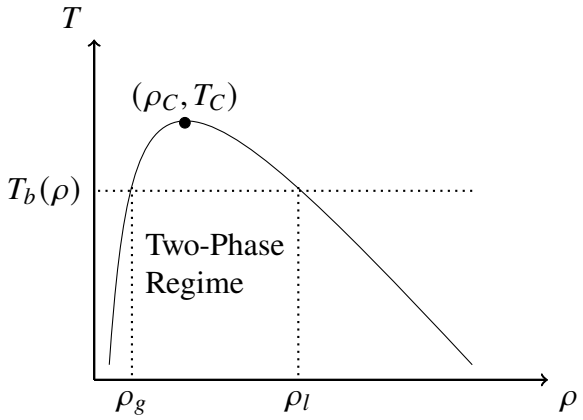


Figure 2: Schematic phase diagram for a single-component protein system. The bimodal curve (full line) defines the shape of the mixed two-phase regime. The critical temperature T_C is the highest temperature at which phase seperation is observed.

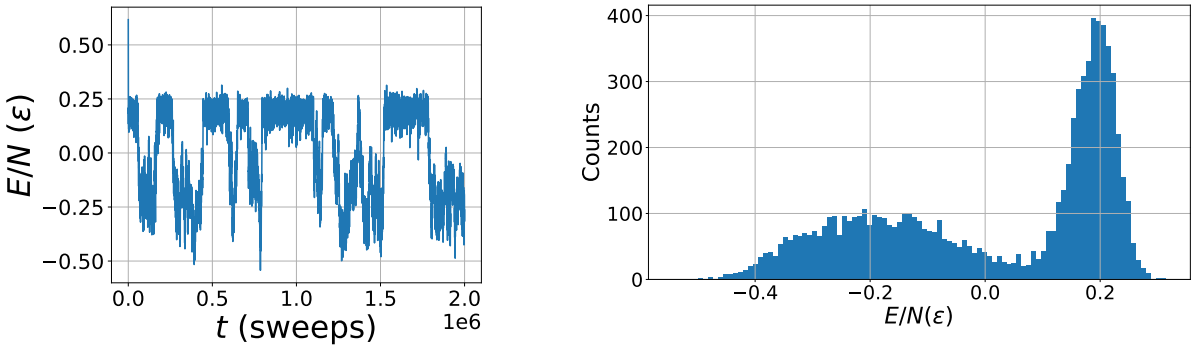


Figure 3: Time series (left) and distribution (right) of the energy from a simulation near $T = T_b$ with sv1 (30 chains). At this temperature the system spontaneously transition between a dilute and a mixed state.

where $\langle..\rangle$ means ensemble average and $\delta^2 E$ the variance of E. Given that the energies associated with each state are sufficiently large apart, C_v is at its largest when $T = T_b$, written as $C_v(T_b) = C_v^{\max}$. At $T = T_b(\rho)$ we can then write:

$$\begin{aligned} C_v^{\max} &= \frac{\delta^2 E_g(T_b) + \delta^2 E_l(T_b)}{2k_b T_b^2} + \frac{(\Delta E(T_b))^2}{4k_b T_b^2} \\ &= \frac{C_v^g(T_b) + C_v^l(T_b)}{2} + \frac{(\Delta E(T_b))^2}{4k_b T_b^2} \end{aligned} \tag{2.6}$$

where C_v^g and C_v^l are computed from the variances of the energy distributions associated with the two states. When the two distributions are similar in energy, the distribution seizes to be bimodal and becomes essentially one peak. Judging whether the distribution is unimodal or bimodal in

the large-system limit is not necessarily easy based on simulations of finite systems. Therefore, it is important to investigate how the shape evolves with increasing system size.

2.5 Simulation Details

The simulation code is entirely written in C. Initially every chain is placed in uniformly random positions. The system is then updated following the MHA procedure described in subsection 2.3. One "sweep" consists of N (the number of beads in the system) updates of the system. The number of sweeps for each system was varied depending on the length a "burn-in" period. For a system close to $T = T_b$, the increase in time correlations has to also be considered. In practice, this was determined such that there was at least a couple of transitions between the mixed and dilute states. As N increases the free energy barrier between these states grows, leading to an increasing correlation time. The systems used here were run for $2 \cdot 10^6$ to 10^7 sweeps, where the larger simulations took about a week in real time. A way to circumvent this limitation is to run multiple smaller simulations of the same system (but with different seeds) which can be performed in parallel and then combined.

Simulations obtained at different temperatures near $T = T_b(\rho)$, can be combined with reweighting of the energy distribution, which gives a more complete and accurate description of the specific heat. In appendix A.1 it is fully described how this procedure works. Briefly one could say that since all samples follows a canonical distribution it is possible to use the information about the energy landscape that is sampled, in order to estimate the shape of the energy landscape at a "close-by" temperature. The further away a temperature is in relation to the simulation data, the less of the energy landscape is known, leading to reweighting becoming less accurate.

The simulations are performed with $N/l = 30, 60, 90$ chains as will be indicated. To prohibit chains from interacting with themselves across the periodic simulation box, the density was chosen such that $L > l\ b = 50\ b$ for every simulation. Thus, $\rho = 0.0025 b^{-3}$ was chosen for all systems.

Six different protein sequences and one RNA sequence will be studied in this thesis. The six protein sequences shown in Figure 4 are taken from the thirty net-neutrally charged polyampholytes studied by Das and Pappu in 2008 [24], and have been studied by several other groups. These polyampholytic sequences have a different degree of "blockiness", quantified by a charge decoration parameter κ (defined in A.2). The patterning parameter goes from $\kappa \approx 0$ for a alternating sequence (sv1) to $\kappa = 1$ for a sequence with all the beads of one type on one end of the chain, with the other type on the other end. The number in the name of each sequence indicates the value of κ. Keeping the density and volume constant, we take the highest temperature at which aggregation sets in as a measure of the propensity of the system to aggregate.

It was shown by Das and Pappu that sequences with larger κ can aggregate more easily than sequences with low κ [23]. However, they used two-body potential composed of a LJ and a

Coulomb potential in their model. Here we will first test whether similar results are obtained with our step-wise potential introduced in Section 2.1. In a study by Pal et. al. [25] they studied demixing of protein-mixtures with two types. Similar potentials were used as in Ref. [23], and sv28 was mixed together with a second protein with smaller κ. It was shown that an increase in the difference in charge decoration between the two proteins tends to decrease the proteins propensity to mix.

To calculate the uncertainty of measured properties, Jackknife resampling will be performed. How this resampling is performed is described in Appendix A.3. In short, it works by constructing new data sets by removing parts of the original data set. These new partitions of the original data set together makes up a statistic which can be used to estimate errors.

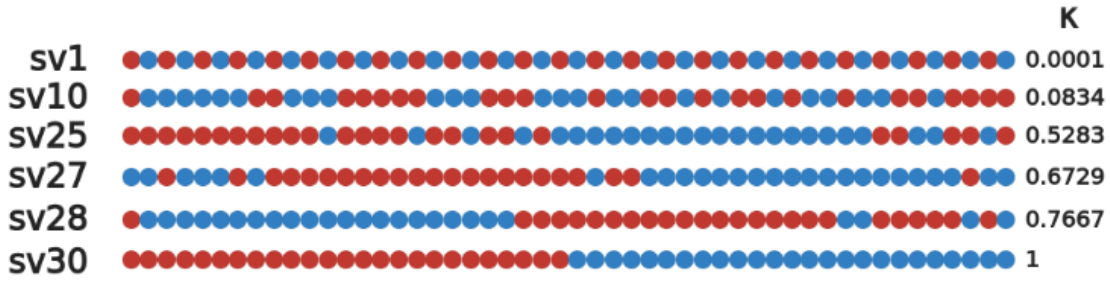


Figure 4: Sequences of amino acids in the proteins used. Higher number after sv indicates the size of the blockiness. The color blue/red corresponds to a charge of $Q = +1, -1$.

In numerical simulations, the systems are limited in size. Therefore, it is important to consider how the simulation results depend on system size. In this study, we compare results attained for different proteins or pair of proteins using one and the same system size. Relative differences in such comparisons are likely to show a relatively weak system size dependence. To illustrate how the estimated shape of the phase diagram may depend on the system size, Figure 5 shows the specific heat, C_v/N, for two sv1 systems with 30 and 90 chains, respectively, at the same bead density $\rho = 0.0025\ b^{-3}$. In this example, a three-fold increase in system size leads to a 4% positive shift of the transition temperature T_b.

3 Results and Discussion

3.1 Phase Behaviour of Single-Component Systems

To investigate the propensity to aggregate for each of the sequences presented in Figure 4, we performed MC simulations at a series of temperatures. When the temperature is low enough, all systems investigated exhibit a single dominant condensate. Aggregation sets in at $T = T_b$ (Figure 2). The transition is signaled by a peak in the specific heat, C_v, which is caused by

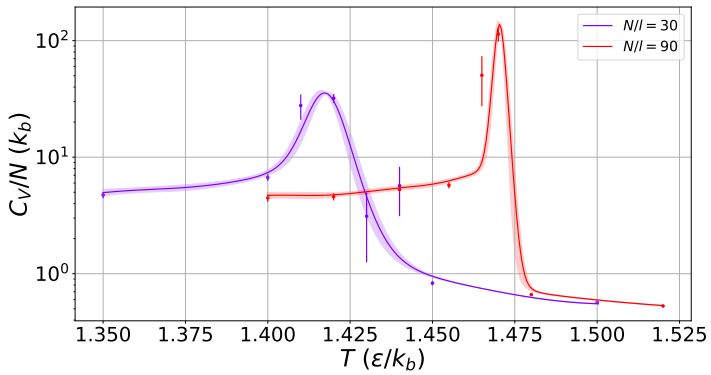


Figure 5: Two systems with equal density shifts T_b and has a sharper aggregation transition as N is increases due to finite size scaling.

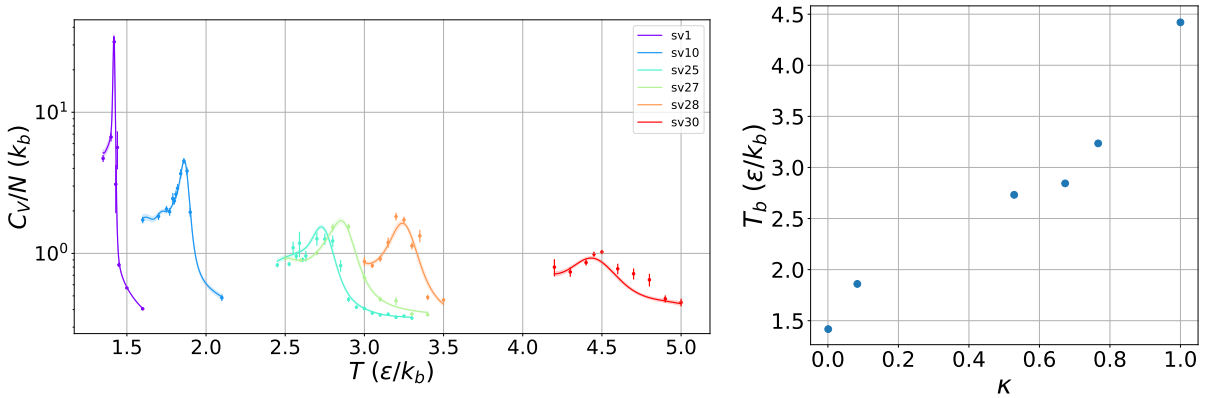
the coexistence of states with and without a condensate. Every simulated system contains 30 chains and have a bead density of $\rho = 0.0025\ b^{-3}$. In Figure 6 (a) the specific heat can be seen for these systems. As the charge decoration κ (Section A.2) increases, the size and location of the peak in the specific heat change. In particular, T_b increases with κ, meaning that aggregation sets in at a higher temperature for blocky sequences. This increase in T_b is illustrated in Figure 6b. This conclusion is in agreement with the results obtained by Das et al. [23], who studied the same sequences but used a different model that was based on LJ and Coulomb interactions.

The energy distribution is expected to be bimodal at $T = T_b$ if phase separation occurs, due to the coexistence of states with and without a condensate. A bimodal energy distribution leads to a high energy variance, which in turn leads to a high C_v. Figure 6(a) shows that the height of the peak, $C_v(T_b)$, decreases as κ increases, indicating that the two existing states becomes less separated in energy. This observation suggests that, especially for large κ, a more extensive analysis is required in order to firmly conclude whether or not LLPS occurs. A systematic approach to this problem would be to carry out a finite-size scaling analysis [26].

3.2 Mixing of Two-Component Systems

Having studied single-component systems of the sequences in Figure 4, we now turn to two-component systems containing pairs of these sequences. We find that these mixed systems, like the one-component systems, form a single dominant aggregate if the temperature is sufficiently low. Our goal is to explore the structure of these aggregates. In particular, we wish to find out whether or not the two components are well mixed within the aggregates.

Consider a system with two sequences, labeled p and q. To study the degree of mixing of the two sequences, we construct radial (bead) distribution functions (RDFs), denoted by $g_{pq}(r)$, $g_{pp}(r)$ and $g_{qq}(r)$. To obtain g_{pp}, given a chain p bead in the system, b_0, we compute the distribution of distances to beads belonging to other p chains. The final g_{pp} is obtained by averaging



(a) (b)

Figure 6: (a) Temperature dependence of the specific heat, C_v/N, in one-component systems. Dots represents the specific heat from individual simulations, while the line is calculated by reweighting the energy histograms obtained from the same simulations. The reweighting procedure is described in Appendix A.1 (b) T_b versus the charge decoration of the sequences in (a). Errors are calculated from the reweighting in (a) but are smaller than the points.

over all possible choices of b_0. g_{qq} is analogously computed. Finally, in constructing g_{pq}, we consider the distribution of chain p beads around a given chain q bead. Specifically, g_{pq} may be written as

$$g_{pq} = \frac{\langle \rho_p(r) \rangle_q}{\rho_p} \tag{3.7}$$

Here, the normalization is such that $g_{pq}(r)$ is unity for all r if the chain p beads are uniformly distributed in the system.

For all systems we used 30+30 chains of the two types of proteins. The total bead density remained the same as in previous section, $\rho = 0.0025\ b^{-3}$. The temperature $T = 1.3\ \epsilon/k_b$ was used as it is well under $T = T_b$ for all single-component systems with 30 chains. All simulations ran for 10^7 sweeps. In the single-component systems, there was a large difference in transition temperature T_b between the sequences. Consequently, the low temperature used in the mixed systems yields large burn in periods. Therefore, only configurations from the final 30% of the simulation time were used. All combinations of proteins mentioned in the previous section were used here, except for combinations between exclusively sv25,sv27,sv28 and sv30 due to the burn in period exceeding 10^7 sweeps.

In each simulated system, the protein type with smallest and largest charge decoration is denoted p and q, respectively. Consider Figure 7 where the RDFs for the sv1+sv30 (upper panel) and sv10+sv25 (lower panel) systems. For small distances, g_{qq} is typically larger than g_{pp}, indicating that the core of the condensate is at least somewhat dominated by proteins with larger charge decoration, as one might have anticipated from the fact that aggregation sets in at a higher temperature for sequences with large κ (Fig. 6). If this tendency is strong, then the proteins

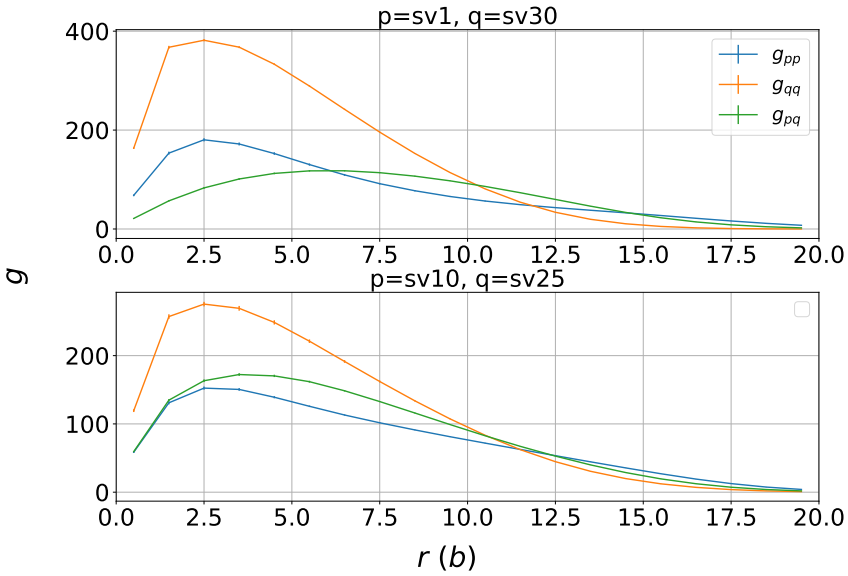


Figure 7: Radial distribution function of systems with two proteins. For sequences with small difference in charge decoration are mixed together, the mixed RDF g_{pq} becomes more similar in shape and size as g_{pp} and g_{qq}. In the upper/lower plot are the RDFs for sv1+sv30/sv10+sv25 respectively.

are demixing. Demixing appears to occur in sv1+sv30, where the core is strongly dominated by protein q. This is evident by the difference in size and shape of the mixed RDF, g_{pq}, and g_{pp} (or g_{qq}). In the sv10+sv25 system, the size and shape of g_{pq} is similar to that of g_{pp}. This means that the density profile of proteins of type p is similar around proteins of both types. In contrast, g_{pq} in the sv1+sv30 system is flattened, meaning that proteins of different types are demixed to a greater extent.

To quantify demixing a mixing parameter $\xi(r)$ will be used, defined as [25]

$$\xi(r) = \frac{2g_{pq}(r)}{g_{pp}(r) + g_{qq}(r)}. \tag{3.8}$$

In Figure 8, $\xi(r)$ is computed for three different mixtures. Large deviations from $\xi = 1$ indicates that the different types of proteins do not mix well. Therefore, $\xi(r)$ is computed for a small single value of $r = R$ for all the systems. $R = 0.5\ b$ was chosen here as it captures the distribution at the smallest distances possible in the system.

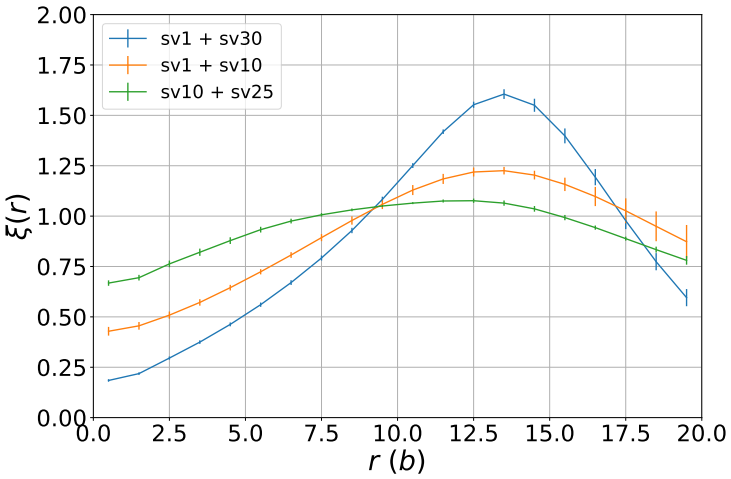


Figure 8: ξ, as defined by equation 8, provides a measure of demixing at distance r. The degree of mixing at small r, as measured by $\xi(r)$, is small when the difference in κ is large (as in the sv1+sv30 system), and larger when this difference is smaller (as in the sv10+sv25 system)

.

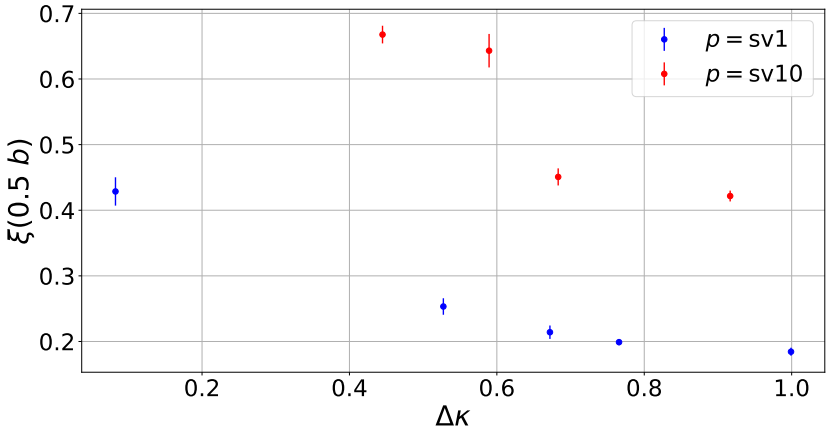


Figure 9: Mixing parameter $\xi(r)$ at $r = 0.5\ b$ plotted against the difference $\Delta\kappa$ in patterning parameter between the two systems. The temperature was set to $T = 1.3\ \epsilon/k_b$, which is well below $T = T_b$ for all systems. Blue/red points are for systems with sv1/sv10 as p component.

A natural question is if charge decoration can be used to predict demixing. We define $\Delta\kappa = \kappa_q - \kappa_p$ where κ_q and κ_p are the charge decorations for the two types of proteins. Plotting $\xi(0.5b)$ versus $\Delta\kappa$ for all systems studied we obtain Figure 9. Here protein type p is either sv1/sv10 as indicated by blue/red. Clearly, $\xi(0.5b)$ is not a simple function of $\Delta\kappa$, as the points with p = sv1 fall below those with p = sv10. It is evident that by only changing protein type q to proteins with higher charge decoration leads to the system being less prone to mix. This is in agreement with what was found in [25] where a Coulomb+LJ potential was used.

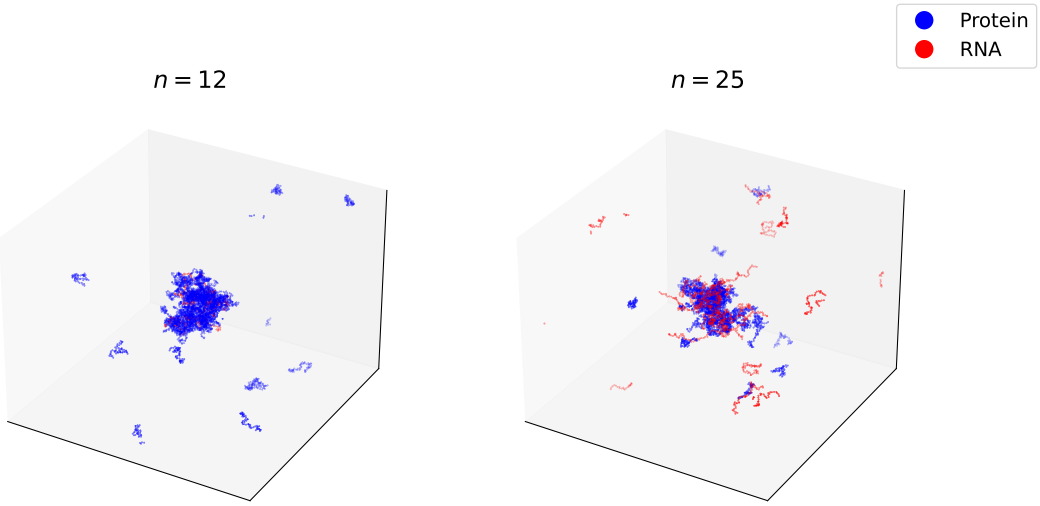
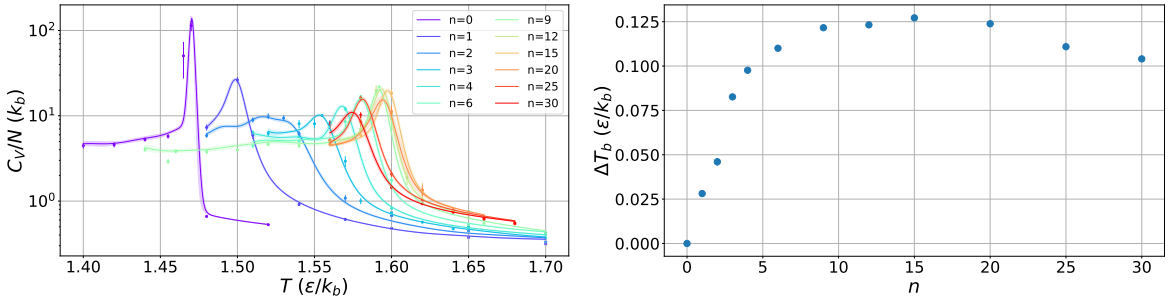


Figure 10: Snapshots of RNA-protein simulations at $T = 1.56\ \epsilon/k_b$. When few RNA $n = 12$ is present (see left) RNA molecules tend to stay inside the high-density phase. When too many RNA molecules are present in the system, additional molecules ends up in a dilute phase outside the aggregate, as in $n = 25$ (see right).

3.3 Aggregation of Protein-RNA Systems

In this last part we investigate if the addition of RNA molecules to a protein system alters the system's propensity to aggregate. This is primarily done by comparing the specific heat of systems containing different amount of RNA chains. The RNA molecule is modeled as a chain of 50 negatively charged beads with the same kind of potentials as for the beads in the proteins. Additionally, the distance between neighbouring beads is the same as in the proteins, b. The protein used in the mixtures was chosen to be sv1 as it was the protein with the sharpest evaporation/condensation transition in single-component systems. Each system investigated contains $90 - n$ sv1 and n RNA chains.

Since the beads inside each RNA-molecule are repulsive to each other, the RNA-RNA interaction energies are always positive. Therefore, one would expect that replacing some protein molecules with RNA molecules leads to a system that is less prone to aggregate (if RNA-proteins interactions are weak). However, the question remains if RNA-protein interactions can increase the system's propensity to aggregate. In Figure 10 two snapshots of systems can be seen, for $n = 12$ (left panel) and $n = 25$ (right panel). The temperature in these snapshots is $T = 1.56$ ϵ/k_b, a temperature low enough for both systems to contain an aggregate. When only a few RNA molecules are present in the system, RNA is prone to stay inside the aggregate. However, when too much RNA is added, the aggregate becomes saturated and the dilute region of the system



(a) (b)

Figure 11: Specific heat of sv1-RNA mixtures. Total system size is 90 molecules and n corresponds to the number of RNA molecules. Addition of RNA-molecules shifts T_b in a non-trivial way.(a) Heat capacity as depending on T. (b) The shift in transition temperature $\Delta T_b(n) = T_b(n) - T_b(0)$ from the simulated systems in (a). For a system with n RNA molecules and $90 - n$ proteins, $T_b(n)$ is defined as the temperature at which the specific heat is the largest. the error bars are smaller than markers.

contains both sv1 and RNA molecules.

As mentioned in Section 3.1, the maximum of the specific heat, T_b, is the maximum temperature at which a larger aggregate is observed. How the specific heat changes with n is presented in Figure 11a. From this figure we can confirm that the aggregation depends on the amount of RNA chains in the system. For clarity, the shift $\Delta T_b(n) = T_b(n) - T_b(0)$ is presented in Figure 11b as a function of n. When a system with sv1 proteins is perturbed by a few RNA molecules, there is a large increase in T_b. As further RNA molecules are added, the increase of ΔT diminishes until a threshold is reached, after which adding more RNA decreases T_b, and thus the aggregation propensity. The largest value of T_b occurs when 17% of the chains are RNA molecules.

At an evaporation/condensation transition, a bimodal energy distribution is expected. Figure 12 shows the energy distribution at $T = T_b(n)$ for different n. The observed energy distribution is bimodal for all n except $n = 1$ and $n = 2$. For these two n, the shape of the distribution at nearby temperatures was inspected. For $n = 1$, no sign of bimodality was observed at any temperature. On the other hand, bimodality was observed for $n = 2$ at a temperature slightly above $T_b(2)$, but with a small separation of the two peaks. Those observations indicate that to decide whether or not phase separation occurs, one would have to scale up the system size (while keeping the sv1:RNA ratio fixed).

At $n = 3$, bimodality reappears (Figure 12). The high-energy peak corresponds to dilute states. As n is increased, this peak is shifted to higher energies. This is expected because intermolecular interactions are weak and intramolecular interactions are higher for RNA than for sv1. The low-energy peak corresponds to a mixed state, with a dense droplet in a dilute background. For small n, all the RNA molecules tend to be part of the droplet. At $n \approx 12$, the droplet appears to become saturated in RNA molecules. For $n > 12$, some of the RNA molecules tend to end up in

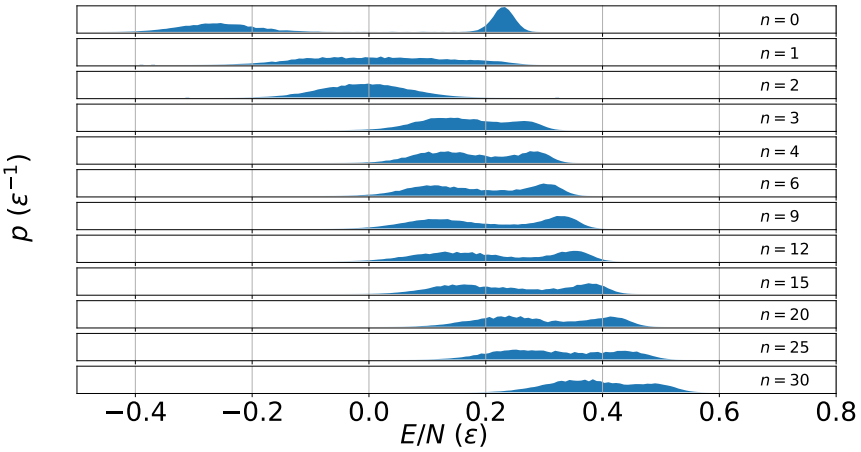


Figure 12: Energy histograms of sv1-RNA mixtures at the transition temperature. The systems contain n RNA and $90 - n$ sv1 molecules. Adding RNA molecules affects the energies associated with the dilute and the mixed states.

the dilute background (see Figure 10).

The size of the droplet formed at T_b depends on the amount of RNA present in the system. To analyze aggregate sizes, we divide a given configuration into clusters of chains, requiring that any pair of chains with negative interaction energy must end up in the same cluster. By analyzing many configurations, we can obtain the probability that a randomly selected chain belongs to a cluster of a given size. Figure 13 shows the cluster size distribution obtained in this way for $n = 0, 1, 4, 9, 15$ and 25. The distribution tend to be bimodal, with one peak corresponding to the droplet and the other peak corresponding to the dilute background. For $n = 0$, the droplet is relatively large and the suppression of intermediate cluster sizes is very strong. However, when RNA is present, intermediate cluster sizes become more common. Since proteins will not aggregate in any significant way by themselves at $T > T_b(0)$, RNA has to be important for these intermediate-size clusters to form. Therefore, RNA appears to act as seeds that proteins can aggregate around.

With only a few RNA molecules present the droplet is significantly smaller than it is in the pure protein system. This is consistent with the fact that $T_b(n) > T_b(0)$ for $n > 0$, which implies that the proteins are unable to form a large cluster on their own at $T_b(n)$. The presence of RNA is therefore crucial for clustering, and with only a few RNA molecules present only a limited number of protein molecules can be recruited to the droplet. For $n = 4$, the largest cluster is around 40% smaller than it is for $n = 0$.

In a simple two-state picture, the transition temperature is given by $T_b = \Delta E/\Delta S$, where ΔE and ΔS are the energy and entropy difference between the two states. With RNA present in the

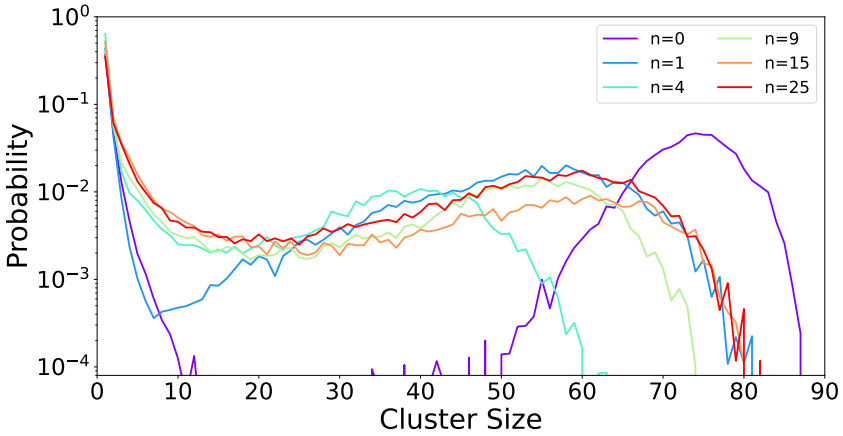


Figure 13: Probability of molecules being in clusters of different sizes at $T_b(n)$, for $n = 0, 1, 4, 9, 15$ and 25. A cluster is defined as the connected network of chains with negative interaction energies. Cluster sizes are reduced in systems containing RNA. The cluster size are smallest when $n = 4$, but grows again as more RNA molecules are added.

systems, we find that T_b increases (Figure 11b) whereas ΔE decreases (Figure 12). This indicates that ΔS is larger in the pure protein system, in which a large fraction of the molecules aggregate to a relatively large structure upon droplet formation. Therefore, the increase in T_b when adding a few RNA molecules is consistent with the reduced droplet size observed in these systems

4 Conclusion and Summary

In this project, we have used computer simulations to explore how the distribution of charge along polymer chains influence the aggregation behavior in coarse-grained systems where aggregation is driven by electrostatic forces. Here biomolecules were modeled as chains of hard spheres with electrostatic interactions given by a simple step potential. We studied a set of zero net charge IDPs with different degree of blockiness, as measured by the charge patterning parameter κ. The project was divided into three parts.

In the first part, we studied one-component systems with 30 chains. In particular, we determined the temperature at which aggregation sets in, T_b. A positive correlation was found between T_b and the charge patterning κ. This finding is in agreement with a previous study [23] of the same sequences based on a model with Coulomb interactions rather than a step potential, which indicates that the conclusion is insensitive to model details.

In the second part, we studied two-component systems with 30 chains of each type. Using radial distribution functions, the degree of mixing of the two components was investigated. In systems with a large difference in κ between the two components, demixing was observed. This observation is also in agreement with previous work based on a different model [25].

Finally, we also explored how the presence of RNA might affect the aggregation behaviour of an IDP, using one of the sequences studied previously, the alternating sequence sv1. RNA was modeled as a negatively charged homopolymer, using the same model as for the IDP sequences. The results suggest that RNA helps sv1 to aggregate, in the sense that aggregation sets in at a higher temperature when RNA is present. Also, the aggregation transition becomes less sharp when adding RNA, and the droplet size is reduced, at least when the amount of RNA is small. However, in order to draw any firm conclusions about the aggregation transition, further studies are required. In particular, the system size dependence of the results should be adressed.

Acknowledgments

I would like to thank Anders Irbäck for shedding light of the behaviour of some of the simulated systems, as well as being a great mentor and supervisor. The simulation program provided by ex-PhD student Daniel Nilsson was of great help and have saved me a great deal of programming time.

A Appendix

A.1 Reweighting Technique

To use multiple simulation to calculate some function of β, a reweighting technique [29] can be used. Suppose we are to calculate an function $g(\beta)$ in the canonical ensemble given set of simulation has performed at $\beta = \beta_1, \beta_2, ..., \beta_R$, each yielding a series of state configurations S. The probability, $p_\beta(S)$, to observe a particular configuration with energy E is

$$p_\beta(S) = \frac{\exp(-\beta E(S))}{Z(\beta)}.$$

For a given system there exists an energy density, $n(E)$, which is the sum of all configurations corresponding to energy E. Thus the probability, $p_\beta(E)$, to observe a particular value of internal energy must be

$$p_\beta(E) = n(E)\frac{\exp(-\beta E)}{Z(\beta)} = n(E)\exp(-\beta E)\exp(-\log(Z)) = n(E)\exp(-\beta E + f(\beta))$$

where $f(\beta)$ is is free energy of the system at the inverse temperature β. Let i be the index corresponding to a particular simulation with the inverse temperature $\beta = \beta_i$. The probability distribution can be approximated with a histogram of N_i observations:

$$p(E, \beta_i) = n(E)\exp(-\beta E + f_i) \approx \frac{h_i(E)}{N_i}$$

Thus the density of states can be approximated from a single simulation with:

$$n(E) \approx \frac{h_i(E)\exp(\beta_i E - f_i)}{N_i}$$

The R simulations performed can be combined to get a better estimate of $n(E)$. Every simulation yields an approximation of $n(E)$, and can be linearly combined to:

$$n(E) = \sum_{i=1}^{R} r_i(E)p_i(E)\exp(\beta_i E - f_i), \quad \sum_{i=1}^{R} r_i(E) = 1 \tag{A.9}$$

To minimize the residual sum of squares $\delta^2 n$ with respect to $r_i(E)$.

$$\delta^2 n = \delta^2\Big(\sum_{i=1}^{R} r_i(E)p_i(E)\exp(\beta_i E - f_i)\Big) = \sum_{i=1}^{R} r_i^2(E)\frac{\delta^2(h_i(E))}{N_i^2}\exp(2(\beta_i E - f_i))$$

The number of counts in each bin of the histogram can be seen as a sample from a Poisson distribution ($\lambda = h_i(E)$). If each measurement are sampled with a fixed interval and there exists a correlation step length of τ_i we further get

$$\delta^2 h_i(E) = (1 + 2\tau_i)h_i(E) = g_i h_i(E) = g_i N_i n(E)\exp(f_i - \beta_i E)$$

Thus,

$$\delta^2 n(E) = \sum_{i=1}^{R} r_i^2(E) \frac{n(E) g_i}{N_i} \exp(\beta_i E - f_i)$$

To minimize $\delta^2 n(E)$ we use Lagrange multipliers and get:

$$\frac{\partial}{\partial r_i}\Big(\delta^2 n(E) - \lambda\big(\Big[\sum_{j=1}^{R} r_j\Big] - 1\big)\Big) = 2 r_i \frac{n(E) g_i}{N_i} \exp(\beta_i E - f_i) - \lambda = 0$$

which means that

$$r_i = \frac{\lambda}{2n(E)} N_i \exp(-\beta_i E + f_i) g_i^{-1}.$$

Using the constraint from A.9 we further get:

$$\frac{\lambda}{2n(E)} = \frac{1}{\sum_{i=1}^{R} N_i \exp(-\beta_i E + f_i) g_i^{-1}}.$$

Now the exponential of the free energy (the partition function) can be written (from $\sum_E p_i(E) = 1$):

$$\exp(f_i) = \frac{1}{\sum_E n(E) \exp(-\beta_i E)}$$

so now we can finally write:

$$n(E) = \frac{\sum_{i=1}^{R} g_i^{-1} h_i(E)}{\sum_{i=1}^{R} \frac{N_i \exp(-\beta_i E) g_i^{-1}}{\sum_{E'} n(E') \exp(-\beta_i E')}}$$

where this equation can simply be solved with recursion. The density of states usually scales exponentially with temperature. Therefore for a computer, the logarithm of the density of states might be more practical to compute:

$$\log\big(n(E)\big) = \log\Big(\sum_{i=1}^{R} h_i(E) g_i^{-1}\Big) - \log\Big(\sum_{i=1}^{R} \frac{\exp\big(\log(N_i g_i^{-1}) - \beta_i E\big)}{\sum_{E'} \exp\big(\log(n(E')) - \beta_i E'\big)}\Big).$$

Now an energy probability distribution for a given temperature can be calculated with:

$$p_\beta(E) = \frac{\exp\big(\log(n(E)) - \beta E\big)}{\sum_{E'} \exp\big(\log(n(E')) - \beta E'\big)}$$

from which a specific heat and average energy can be calculated for any β. However, reweighting too far away from an energy landscape that have not been thoroughly explored will yield large errors. The temperatures used for NVT simulations should therefore be chosen such that the energy histograms are not too far apart from each other.

A.2 Charge Patterning Parameter

To quantify the "charge-blockiness" of a polyampholyte sequence, a patterning parameter κ can be utilized [24]. Let

$$f_{+,-} = \frac{Q_{+,-}}{N}$$

where N is the number of amino acids in the sequence and $Q_{+,-}$ be the number of positive/negative charges in the sequence respectively. Then the charge asymmetry will be:

$$\sigma = \frac{(f_+ - f_-)^2}{f_+ + f_-}.$$

The sequence is then partitioned into n segments such that each following segment is shifted one amino acid to the side. The size of each partition will then $N_{blob} = N - n$. For each partition new charge fractions $f_{+,i}$, $f_{-,i}$ are calculated for $i = 1, 2, ..., n$. The charge asymmetry for each segment will be:

$$\sigma_i = \frac{(f_{+,i} - f_{-,i})^2}{f_{+,i} + f_{-,i}}$$

Now the unnormalized patterning parameter is defined as:

$$\delta(N_{blob}) = \frac{\sum_{i=1}^{n}(\sigma_i - \sigma)^2}{n}$$

which will depend on how large the partitions are. By convention, δ for a given sequence is defined as the average of $\delta(5)$ and $\delta(6)$. The normalized patterning parameter is defined as:

$$\kappa = \frac{\delta}{\delta_{max}}$$

where δ_{max} is the calculated from the sequence of length N that has the highest patterning parameter, which turns out to be a sequence with $N/2$ negative charges followed by $N/2$ positive charges for an overall neutral sequence.

A.3 Jackknife Resampling

Assume that from a measured set of N samples, $\hat{X} = X_1, X_2, ..., X_n$, it is of interest to estimate some parameter $g(X)$. Then it might not be trivial to find an analytic expression of the variance and bias of the estimator. In these situations the jackknife resampling method might be utilized to produce a new set of jackknife samples as a "one-size fits all" solution [30].

To construct a jackknife sample, $\vec{X}_i$, remove one element of the original sample set:

$$\vec{X}_i = \{X_1, ..., X_{i-1}, X_{i+1}, ..., X_n\}$$

This sample can then be used to construct g_{-i} defined as

$$g_{-i} = g(\hat{X}_i)$$

where $\hat{X}_i$ is an estimation of $\vec{X}_i$. Furthermore we define the so called "Jackknife replicates":

$$g_i = n\hat{g} - (n-1)g_{-i}$$

where $\hat{g} = E\{g(\hat{X})\}$. This is done so that it is further possible to obtained a less biased estimator by eliminating the $O(1/n)$ term from the expansion of g with:

$$\tilde{g} = \frac{1}{n}\sum_{j=1}^{n} g_i = n\hat{g} - \frac{n-1}{n}\sum_{j=1}^{n} g_{-i}$$

References

[1] Altmeyer M., Neelsen K.J., Teloni F., Pozdnyakova I., Pellegrino S., Grøfte M., Rask M.-B.D., Streicher W., Jungmichel S., Nielsen M.L., and Lukas J. Liquid demixing of intrinsically disordered proteins is seeded by poly(adp-ribose). *Nature Communications*, 6:1, 2015.

[2] Parker M.W., Bell M., Mir M., Kao J.A., Darzacq X., Botchan M.R., and Berger J.M. A new class of disordered elements controls dna replication through initiator self-assembly. *eLife*, 8:48562, 2019.

[3] Sabari B.R., Dall'Agnese A., Boija A., Klein I.A., Coffey E.L., Shrinivas K., Abraham B.J., Hannett N.M., Zamudio A.V., Manteiga J.C., Li C.H., Guo Y.E., Day D.S., Schuijers J., Vasile E., Malik S., Hnisz D., Lee T.I., Cisse I.I., Roeder R.G., Sharp P.A., Chakraborty A.K., and Young R.A. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361:3958, 2018.

[4] Khuloud Jaqaman and Jonathon A. Ditlev. Biomolecular condensates in membrane receptor signaling. *Current Opinion in Cell Biology*, 69:48, 2021.

[5] Shana Elbaum-Garfinkle, Younghoon Kim, Krzysztof Szczepaniak, Carlos Chih-Hsiung Chen, Christian R. Eckmann, Sua Myong, and Clifford P. Brangwynne. The disordered p granule protein laf-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences*, 112:7189, 2015.

[6] Clifford P. Brangwynne, Christian R. Eckmann, David S. Courson, Agata Rybarska, Carsten Hoege, Jöbin Gharakhani, Frank Jülicher, and Anthony A. Hyman. Germline p granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, 324:1729, 2009.

[7] Masato Kato, Tina W. Han, Shanhai Xie, Kevin Shi, Xinlin Du, Leeju C. Wu, Hamid Mirzaei, Elizabeth J. Goldsmith, Jamie Longgood, Jimin Pei, Nick V. Grishin, Douglas E. Frantz, Jay W. Schneider, She Chen, Lin Li, Michael R. Sawaya, David Eisenberg, Robert Tycko, and Steven L. McKnight. Cell-free formation of rna granules: Low complexity sequence domains form dynamic fibers within hydrogels. *Cell*, 149:753, 2012.

[8] Bálint Mészáros, István Simon, and Zsuzsanna Dosztányi. The expanding view of protein–protein interactions: complexes involving intrinsically disordered proteins. *Physical Biology*, 8:035003, 2011.

[9] Carolyn J Decker, Daniela Teixeira, and Roy Parker. Edc3p and a glutamine/asparagine-rich domain of lsm4p function in processing body assembly in saccharomyces cerevisiae. *The Journal of cell biology*, 179:437, 2007.

[10] Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nature reviews Molecular cell biology*, 18:285, 2017.

[11] David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, 1987.

[12] Paul J. Flory. Thermodynamics of high polymer solutions. *The Journal of Chemical Physics*, 10:51, 1942.

[13] Maurice L. Huggins. Solutions of long chain compounds. *The Journal of Chemical Physics*, 9:440, 1941.

[14] J. T. G. Overbeek and M. J. Voorn. Phase separation in polyelectrolyte solutions. theory of complex coacervation. *Journal of Cellular and Comparative Physiology*, 49:7, 1957.

[15] A Johner J Wittmer and J. F Joanny. Random and alternating polyampholytes. *Europhysics Letters*, 24:263, 1993.

[16] Julie D. Forman-Kay Yi-Hsuan Lin and Hue Sun Chan. Theories for sequence-dependent phase behaviors of biomolecular condensates. *Biochemistry*, 57:2499–2508, 2018.

[17] Gregory L. Dignon, Wenwei Zheng, Young C. Kim, Robert B. Best, and Jeetain Mittal. Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Computational Biology*, 14:1, 2018.

[18] Russell DeVane, Wataru Shinoda, Preston B Moore, and Michael L Klein. Transferable coarse grain nonbonded interaction model for amino acids. *Journal of chemical Theory and Computation*, 5:2115, 2009.

[19] Benjamin R. Sabari, Alessandra Dall'Agnese, and Richard A. Young. Biomolecular condensates in the nucleus. *Trends in Biochemical Sciences*, 45:961, 2020.

[20] Phase separation of c9orf72 dipeptide repeats perturbs stress granule dynamics. *Molecular Cell*, 65:1044, 2017.

[21] William M. Aumiller, Fatma Pir Cakmak, Bradley W. Davis, and Christine D. Keating. Rna-based coacervates as a model for membraneless organelles: Formation, properties, and interfacial liposome assembly. *Langmuir*, 32:10042, 2016.

[22] William M Aumiller and Christine D Keating. Phosphorylation-mediated rna/peptide complex coacervation as a model for intracellular liquid organelles. *Nature chemistry*, 8:129, 2016.

[23] Suman Das, Alan N Amin, Yi-Hsuan Lin, and Hue Sun Chan. Coarse-grained residue-based models of disordered protein condensates: Utility and limitations of simple charge pattern parameters. *Physical Chemistry Chemical Physics*, 20:28558, 2018.

[24] Rahul K. Das and Rohit V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110, 2013.

[25] Tanmoy Pal, Jonas Wessén, Suman Das, and Hue Sun Chan. Subcompartmentalization of polyampholyte species in organelle-like condensates is promoted by charge-pattern mismatch and strong excluded-volume interaction. *Physical Review E*, 103:042406, 2021.

[26] Daniel Nilsson and Anders Irbäck. Finite-size scaling analysis of protein droplet formation. *Physical Review E*, 101:022413, 2020.

[27] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97, 1970.

[28] H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in monte carlo simulations. *Physical Review Letters*, 58:86, 1987.

[29] Alan M Ferrenberg and Robert H Swendsen. Optimized monte carlo data analysis. *Computers in Physics*, 3:101, 1989.

[30] M. Quenouille. Notes on bias in estimation. *Biometrika*, (43), 1956.