The role of assembly signals in the self-assembly of linear viruses

Melle Punter

Institute for Theoretical Physics
Department of Physics and Astronomy
University of Utrecht

Supervisor: Prof. dr. ir. Paul van der Schoot

June 2015

Abstract

Linear artificial viruses (AVs) can be formed through a self-assembly process which is strongly influenced by assembly signals on their genetic material. In this thesis the description of self-assembly with multiple assembly signals and variable nucleation cost is assessed. It shows that a combination of the two determines in which regime the self-assembly is. An assembly signal, position entropy and nucleation entropy dominated regime can be distinguished. Furthermore, in the zipper regime an account is given of the assembly kinetics with finite as opposed to infinite protein concentration. Finite concentration gives rise to overshoots and undershoots. Finally, for a self-competing assembly system a universal curve is given which can be a valuable tool in determining the strength of an assembly signal from two measurements.

Deo gratias

Contents

Chapter 1

Introduction

At the end of the nineteenth century the experiments of Louis Pasteur pointed towards the existence of a disease-causing agent which multiplied within organims. In 1892 it was Dimitri I. Ivanovski who showed in St. Petersburg that the tobacco mosaic disease was caused by an ultrafiltrable agent, that is, one whose size is significantly smaller than that of bacteria [1, Chapter 1]. Afterwards, in 1898 Martinus W. Beijerinck showed that a contagious fluid was the cause of the detriment occurring at the leaves of tobacco plants [2]. This contagious fluid turned out to contain the *tobacco mosaic virus* (TMV).

With the discovery of TMV the field of virology was born and ever since viruses have captured the attention of biologists and doctors. First, they were mainly considered as hazardous infectious particles but with the rise of genetic manipulation they started to receive attention for being potentially useful in, for example, medical applications. With the coming of genetic manipulation and the progress of synthetic science the dream of creating an artificial virus (AV), a particle which resembles a natural virus, became a reality. Such an AV can either be a modified natural virus or a completely synthesized virus like particle. In the next section some general aspects of viruses will be considered.

1.1 General aspects

Up to date at least 1000 different species of natural viruses which infect man and about 1500 plant viruses have been identified. Nevertheless, it is speculated that many more virus species exist [3, 4]. The latin name *virus*, which refers to poison [5], is well chosen since they are known for they infectious and possibly deadly powers. For example, the virus family of *Potyviridae*, which makes up about 20% of the known plant viruses, inflicts about half the crop damage worldwide [3].

The common feature of all natural viruses is that they are composed of, at least one, template - either a RNA or DNA molecule - and of, at least one kind of, proteins which form a capsid that surrounds the template. These proteins are folded polypeptides. The proteins have a limited number of folds, what gives rise to a limited number of possible structures of the capsid. The capsid usually has, for linear and spherical capsids respectively, a helical or icosahedral symmetry [6], see figure 1.1. The experimental knowledge of viruses we have is obtained through the following techniques: transmission electron microscopy, cryo-electron microscopy, cryo-electron tomography, X-ray crystallography, nuclear magnetic resonance spectroscopy, atomic force microscopy and possible combinations of the techniques mentioned.

Viruses do not encode their own protein synthesis machinery, nor their own energy-generating pathways. Instead, viruses use living host cells for their survival. They are able to use the host for the optimal execution of their own reproduction [1, Chapter 2]. This happens in the following way. Once a virus enters a host it disassembles through a number of processes. For example, its capsid is broken down. This exposes its template to the surroundings. Subsequently, it encodes for, usually one, protein which enables in some complicated way the use of the host cell for the reproduction of the template and the

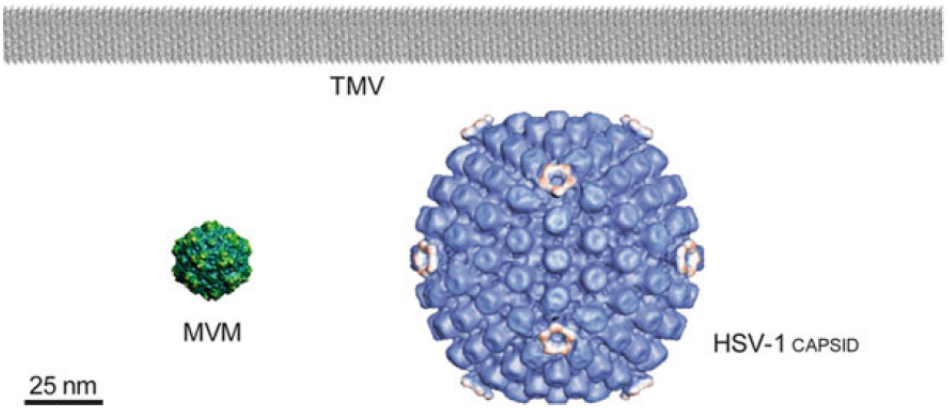

Figure 1.1: Different capsid symmetries of natural viruses. *Top*: the capsid of the *tobacco mosaic virus* (TMV) exhibits helical symmetry. *Bottom left*: that of the *parvovirus* (MVM) shows simple icosahedral symmetry. *Bottom right*: the capsid of the *herpesvirus* (HSV-1) has icosahedral symmetry but is more complex than that of the MVM. Taken from [6]

capsid proteins. New viruses assemble through the coverage of newly created templates with a capsid and subsequent maturation. Finally, the reproduction of viruses in the host can kill the host [6].

From in vitro experiments it is shown that the assembly and disassembly of viruses is a complex process. Among the different processes involved, self-assembly may occur, which refers to the spontaneous building up or breaking down of the capsid. This self-assembly is the main subject of this thesis and will be introduced below. in the paragraph on viral self-assembly.

As noted above, artificial viruses (AVs) can be created nowadays. These enter, in principle, host cells as natural viruses do, but they do not need to replicate. This is because their template does not necessarily contain information for replication, for it is synthesized. Furthermore, they can, in principle, encode for a protein or chemical at will. This enables, for example, their use as targeted drug deliverers, as will be explained in the section on AVs. The next section gives a more detailed account of the assembly step in the generation of a virus. Specifically, it focuses on the subject of this thesis: viral self-assembly.

1.2 Self-assembly

The generation of a virus is a highly complex process which is, for most species, poorly understood. Nevertheless, in general we can distinguish three stages in the generation: capsid assembly, template packaging and virus maturation. In this thesis we focus specifically on capsid assembly.

The assembly of a capsid may comprise allosteric switches and irreversible steps. In general, three assembly strategies are recognized: 1) self-assembly, which merely requires capsid proteins that spontaneously assemble to form the capsid, 2) scaffolding protein-assisted assembly, where besides capsid proteins the assistance of scaffold proteins is required, and 3) template assisted assembly which requires the simultaneous interaction of the template and capsid proteins such that in a condensation process the capsid is formed besides the template being packed. Here, we focus on the first strategy.

It turns out that during self-assembly, in general, capsid building blocks (CBBs) are formed which in turn assemble into the capsid. The pathways through which the CBBs assemble into the capsid is hard to trace. Nevertheless, some important features of self-assembly are known. It is known that 1) capsids can self-assemble through nucleation and growth from CBBs where the last happens through a cascade of second order reactions, 2) it appears that a certain critical concentration exists below which (practically) no assembly occurs and above which (nearly) all capsids fully formed, and 3) the assembly kinetics can be represented by a sigmoidal curve which exhibits a lag phase [6]. These observations need to be accounted for and we will refer to them in the course of this thesis.

One of the best studied self-assembly processes among viruses is that of the *tobacco mosaic virus* (TMV). The model we use in this thesis was originally derived to describe this virus its self-assembly. In the next section we will introduce this virus in more detail.

1.3 *Tobacco mosaic virus*

As noted above, natural viruses come in two kinds: spherical and linear. In this thesis we focus on linear viruses, which make up about 10% of all known natural virus families [6]. In particular, we focus on a linear plant virus, the *tobacco mosaic virus* (TMV), as well as artificial viruses (AVs) which are inspired by TMV. The reason for this is that earlier research has focused on TMV [7], besides that we have access to experimental data of AVs whose design is inspired by TMV. This research background will be further explained in the next section.

TMV is a linear helical virus and, as noted above, due to its early discovery the best studied virus. Its capsid consists of about 2000 capsid proteins and approximately 49 proteins are present within three turns of the helix [3]. The capsid surrounds its ssRNA which consists of 6395 nucleotides. Each capsid protein binds to three nucleotides via hydrophobic and electrostatic interactions, as well as hydrogen-bonding [7]. The capsid exhibits, due to the folding of the capsid proteins, a helical symmetry. A helix is characterized by its pitch, which is the product of the number of capsid proteins per helix turn and the axial rise per capsid protein. For TMV, the pitch is 2.3 nm. TMV is a stiff, right-handed, rod-like structure of about 300 nm in length and 18 nm in diameter, with a central hole of 4 nm in diameter [8]. Moreover, the RNA is located about 4 nm from the axis of the helix, see figure 1.2. Because of its stiffness, an aggregate of them

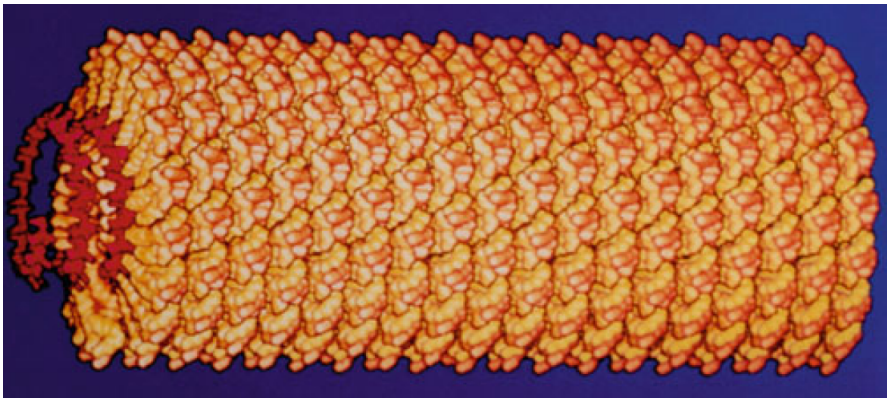

Figure 1.2: A visualization of the *tobacco mosaic virus*. Capsid proteins are yellow and the RNA is red. The capsid shows a helical symmetry and the RNA is extended to see its helical shape. Taken from [3].

can form a liquid crystal.

TMV exhibits a hierarchical self-assembly, that is, the capsid is assembled through a sequence of, in energetic cost decreasing, self-assembly steps. In the first step capsid proteins interact to form larger aggregates and discs. First, a two layered disc (20S) binds to the origin of assembly located at about $\frac{1}{6}$-th of the RNA its length. This origin of assembly is an assembly signal. It is defined by a special sequence of nucleotides which is such that the capsid proteins favourably bind to it. Subsequently, the disc transforms into a short helix and incorporates the RNA between the capsid protein layers. There has been some discussion on the precise nature of the 20S aggregate, but its precise nature is not important to the form of the assembly process. Afterwards, energetically less costly steps are the elongation of the capsid. At one end, the capsid is completed via stepwise addition of further discs. At the other end capsid protein mono- and oligomers complete the capsid, though at a much slower pace [3, 7, 9]. This is visualised in figure 1.3. Because the self-assembly of TMV is relatively well understood is has provided inspiration for the development of artificial capsid proteins. This has been done and presents one of the motivations of this thesis, as will be outlined in the next section. However, in the next section we discuss the possible beneficial applications of artificial viruses.

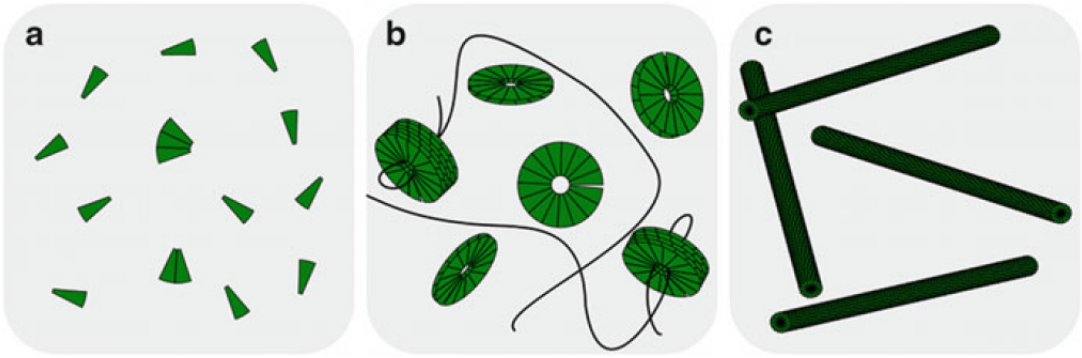

Figure 1.3: Different self-assembly processes of the *tobacco mosaic virus*. First, in (**a**) the capsid proteins form small aggregates. Afterwards, in (**b**) these assemble to form oligomers, discs and the 20S aggregate which binds to the origin of assembly. Subsequently, the capsid is completed through bidirectional assembly, as shown in (**c**). One end is completed through the addition of further discs, the other through the addition of mono- and oligomers. Taken from [9].

1.4 Artificial viruses

Viruses have several outstanding characteristics. Among others, they feature: capsid self-assembly; the targeting of cells through precise molecular recognition; chemical and mechanical actions for delivery of its genome into the host; a precise nanoscale structure which is geometrically well defined; a variety of shapes; an adjustable template; and they can undergo mass production [6, 9]. This make them especially suitable to be used in fields like medicine and nanotechnology. They can be put to use by either genetically modifying a natural virus or by synthesizing all components. In both cases one acquires an artificial virus (AV). Now we will first comment on some major applications of artificial viruses.

First, AVs have great biomedical potential [10–12]. They are often very stable against changes in pH, temperature, ionic strength and solvent. This gives a broad range of conditions for their isolation, storage and use. Moreover, they have regular surface properties which are, in principle, alterable [9]. Therefore, one needs little imagination to see that through the alteration of the surface properties of the virus it could be made cell specific. This could give rise to gene therapy [13–15], which comprises the delivery of therapeutic genes into specific cells. Moreover, they can be used as targeted drug deliverers. That is, they could be used as nanocarriers, where they bring a certain chemical to a specific cell type [6, 16]. Also, their genetic material could be manipulated in such a way that they can encode for a particular chemical needed in a cell.

A concrete example of an AV in biomedical use would be one which enters a (potential) cancer cell. It could make a chemical which can be readily measured for detection of the cell. Otherwise, it could encode for a chemical which kills the cell. The first possibility implies that cancer could possibly be diagnosed in an earlier stage. The second possibility implies that chemotherapy would become cell-specific. Thereby the required dose would be lowered and the well-known side effects of chemotherapy would decrease drastically. Finally, many therapeutic chemicals are not in use at the moment due to their systemic toxicity when used in cell non-specific treatments. Possibly, these could be put to use with a cell-specific treatment [17].

Another large field of application of AVs are nanomaterials [18]. This is an upcoming field of material science at which the material is designed at the nanoscale. AVs have properties which make them well-suited as scaffolds in the design of a nanomaterial. For example: the capsid exhibits constrained internal cavities which are accessible to small molecules but not for larger ones [9]. A possible application of AVs would be to enhance the conduction properties of a material [19]. This could be done by growing metal particles in a regular manner at the surface of the AV. This is already possible for TMV [20]. In case of linear AVs which exhibit a nematic phase, the particles could be aligned and thereby enhance the conductivity of the material. With these beneficial applications in mind we introduce the concrete

research background of this thesis in the following section.

1.5 Research background

As pointed out above the *tobacco mosaic virus* (TMV) was at the cradle of the field of virology. Consequently, much research has been conducted after its characteristics, in particular, after the assembly and disassembly pathways it follows. Quite recently research has started on how self-assembly of TMV occurs in solution as a function of time by Kraft et al [7]. This thesis builds on their work an extends the scope of their analysis, as will be explained below. Although this work was done in the context of TMV it is applicable to self-assembly processes which are similar to that of TMV.

The idea to describe the self-assembly as a function of time is to model the template as a one dimensional structure with a number of binding sites available for capsid building blocks (CBBs). A number of templates are in solution together with a number of CBBs. The precise form of the CBBs depends on the experimental system one considers, but is irrelevant for the description. In time, the system relaxes to equilibrium such that the CBBs are distributed over the templates. This gives rise to a certain distribution. That is, some fraction of the templates will be fully encapsulated, while others do not have a fully grown capsid. This is depicted in figure 1.4.

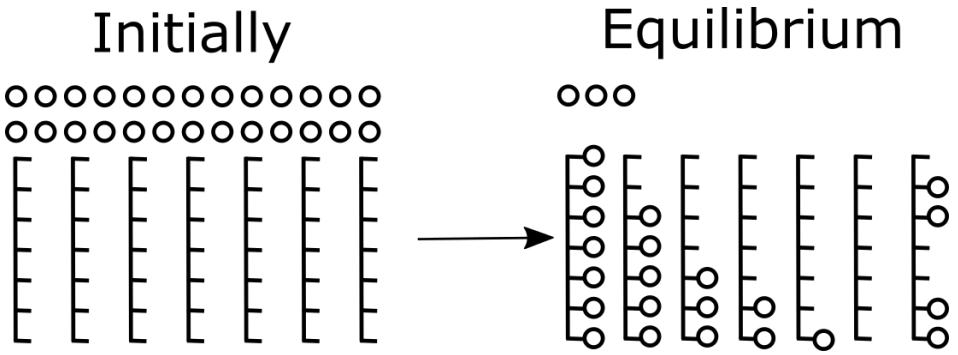

Figure 1.4: Schematic representation of a self-assembly process. Initially a number of templates, the rods, and a number of capsid building blocks (CBBs), the circles, are in solution. The templates have a number of binding sites for the CBBs available. After the system is relaxed into equilibrium the CBBs are distributed in a definite way among the templates. Some templates have one nucleation, others can have two. In principle, every binding site can have a CBB bound independently of its neighbours. This gives rise to the equilibrium distribution of the proteins.

In the paper of Kraft et al. a statistical equilibrium model, the zipper model, is proposed to describe the binding of proteins to the RNA template of TMV. This gives rise to a quantitative prediction of the formation of viral capsids. That is, the distribution and average coverage of the proteins over the templates is probed. Subsequently, they use this model to derive dynamical equations which describe the self-assembly as a function of time. These dynamical equations were derived for a special case - an infinite supply of proteins - wherefore the following question remains open: **how does assembly occur with a finite supply of proteins?** This question will be addressed in this thesis.

The research question of finite protein supply became more prominent due to a large collaborative research project which focused on the creation of an artificial virus made of artificial dsDNA and artificial capsid proteins [21]. The design of this AV was inspired by TMV. Thus the model of Kraft et al. [7], the zipper model, can be applied readily. Within this research project measurements on the assembly of AVs were made. Hereby providing data with which the theoretical predictions of the assembly dynamics could be compared. This puts forward another question we will cover: **how do the dynamics of the zipper model compare to experiment?**

Next, a motivation of this research comes from the fact that assembly signals play a major role in self-assembly. An assembly signal is a certain sequence of bases in the template which is such that a

protein has energetic advantage upon binding to it. These signals occur naturally in TMV and can be built into artificial templates. The artificial dsDNA of Hernandez-Garcia et al. [21] seems to have an assembly signal at the end of the template. This is one reason wherefore we expect the zipper model, as noted above, to describe the assembly of this artificial DNA.
Nevertheless, multiple assembly signals can in principle be built into an artificial template. To describe their effect an extended statistical equilibrium model is needed which allows for multiple assembly signals. Therefore, we propose the Ising-S model which we calculate for at most three assembly signals. Furthermore, the Ising-S model accounts for the occurrence of nucleations at a normal binding site. A normal binding site being defined as one which does not contain an assembly signal. These kind of nucleations also occur and should be taken into account, for they imply an entropic contribution. The question we will answer with regard to the Ising-S model is: **what is the influence of assembly signals on the distribution of capsid proteins over the templates in equilibrium?**

Finally, in experiments is competition often important. For example, energy rich assembly states can compete with entropy rich states on a single template in case of self-competition. Otherwise, a binding energy rich template species can compete for the available CBBs with an entropy rich species. Therefore we answer the question: **what is the influence of competition on the formation of capsids?**

With these main questions in mind we give in the next section a detailed outline of what will be covered.

1.6 Outline

For convenience, we give an outline of what will be covered in this thesis.

In chapter two we first introduce the statistical framework in which we describe the assembly process in equilibrium and the equilibrium quantities which can be derived. Also, we introduce the zipper model, as proposed by Kraft et al. [7], which should account for the equilibrium characteristics of TMV as well as the AVs of Hernandez-Garcia et al. [21]. Besides, we introduce the models describing self-competition and species competition. Next, we introduce the Ising-S model which will be analysed in chapter four. This model accounts for the occurrence of multiple assembly signals on a template together with entropic effects. Finally, the chemical kinetics framework used to describe the dynamics of zipper self-assembly in chapter three is presented.

In chapter three we focus on the zipper model. First, we calculate the partition function and relevant equilibrium quantities. Second, we derive the dynamical equations which govern the assembly dynamics. Then, before we describe the dynamics, we give an overview of the equilibrium properties of the zipper model. With that in mind we show the rich dynamical behaviour which the zipper model comprises by numerically solving the dynamic equations. Finally, we compare the theoretical predictions of the zipper model with experimental data and therefrom give an estimate of the parameter values.

In chapter four we first calculate the general partition function of the Ising-S model. Second, as a check, we show how for a single assembly signal, short templates and a strong protein-protein interaction the Ising-S model reduces to the zipper model. Next, we focus on the distribution of the proteins on the templates. This is a key quantity because it gives what fraction of templates is fully covered and thus what fraction of templates forms a fully covered AV. We do this for two and three assembly signals respectively.

In chapter five we focus on competition. We show how self-competition gives rise to a universal curve. Moreover, we give a very short note on the essence of species competition.

Finally, in chapter six we summarize our results, draw conclusions, discuss the validity of our results and provide recommendations for future research.

Chapter 2

Theory

In this chapter we will outline the theory needed to answer the questions we have put forward in the introduction. The first two questions concern the zipper model which was proposed by Kraft et al [7], so we define it in the third section. Next, we introduce the Ising-S model which we propose to answer the third question posed in the introduction on the effect of entropy and the existence of multiple assembly signals on a template. Finally, we present chemical kinetics to describe the dynamics of zipper self-assembly. However, in the following we first outline the statistical mechanical theory which describes the self-assembly of viruses in a solution. We will show how one can calculate the equilibrium distribution of capsid proteins over the templates and give how one can predict the average occupation of binding sites. Also, we will give the physical motivation for the zipper and Ising-S model by discussing the energetic aspects of the binding of capsid proteins to the templates in the second section.

2.1 Statistical self-assembly description

Based on the existing body of research we use the following theory to describe the solution in which the self-assembly of viruses takes place [7, 22]. We have a solution of volume V in which N_P capsid proteins - or simply proteins - and a number of templates are dissolved. At any template $n \equiv \sum_{i=1}^{q} n_i$ proteins can be bound where n is an integer from 0 up to and including q, $n_i = 0, 1$ and q the total number of binding sites on a single template. If $n_i = 0$ there is no protein bound on the i-th site and if $n_i = 1$ there is a protein unit bound. So, $n = 0$ corresponds to a completely uncovered - an empty - template and $n = q$ refers to a fully capsulated templated. The proteins are bound to the template in a certain configuration, denoted by $\{n_i\}_{i \in P} \equiv \{n_i\}$, where P is the position vector defined as $P \equiv \{1, 2, ..., q\}$. This is visualised in figure 2.1. It shows a template whose binding sites all have a certain occupation number. Together, the occupation numbers define a configuration. We denote the number of templates with a given protein configuration $\{n_i\}$ on them as $N_T(\{n_i\})$. We wish to describe this system in the grand canonical ensemble, so we write down the grand potential, Ω'. One might be bothered by a description in the grand canonical ensemble for the self-assembly systems we want to describe are not in contact with a particle bath. Therefore, although in the thermodynamic limit the grand canonical and canonical descriptions are equivalent, we show in appendix A how to describe the system in the canonical ensemble. In the grand canonical ensemble we have as independent variables V, T and the chemical potential of the proteins and the templates, μ'_P and μ'_T respectively. So, $\Omega' = \Omega'(V, T, \mu'_P, \mu'_T)$ and both $N_T(\{n_i\})$ and N_P are implicit functions of V,T,μ'_P and μ'_T. We may write

$$\Omega' = F' - \mu'_T \sum_{\{n_i\}} N_T(\{n_i\}) - \mu'_P N_P - \mu'_P \sum_{\{n_i\}} n N_T(\{n_i\}),$$

where F' is the Helmholtz free energy of the system and $\sum_{\{n_i\}}$ is the sum over all allowed configurations that the interaction model allows. With interaction model we concretely mean either the zipper or the

Ising-S model, but it may be any model which describes the energies of the different configurations it allows. To write down F' we make the simplifying assumption that the components in the solution have ideal solution statistics. For every template with $\{n_i\}$ proteins bound to it the system gains an interaction energy $E'_{int}(\{n_i\})$. This interaction energy will be left unspecified for now for its details are not necessary to derive the equilibrium properties we are interested in. At least, as long as the interaction does not depend on the densities of the proteins and/or the templates. These considerations give

$$\begin{aligned} F' &= F'_{id} + F'_{int,tot} \\ &= F'_{int,tot} + k_BTN_P\Big[\ln\Big(\frac{N_P}{V}V_P\Big) - 1\Big] + \\ &\quad k_BT\sum_{\{n_i\}} N_T(\{n_i\})\Big[\ln\Big(\frac{N_T(\{n_i\})}{V}V_{T(\{n_i\})}\Big) - 1\Big] \\ F'_{int,tot} &= \sum_{\{n_i\}} N_T(\{n_i\})E'_{int}(\{n_i\}), \end{aligned} \tag{2.1}$$

where $V_{T(\{n_i\})}$ is the typical volume scale of a template in the solution with $\{n_i\}$ proteins bound to it and V_P is the typical volume scale of a protein in the solution. These typical volume scales depend on the solvent and the effective volume which a molecule occupies after integrating out the interactions with the solvent. There is discussion about the precise nature and method to calculate this volume scale. Nevertheless, as far as we are concerned we take the typical volume scales for all molecules in the solution to be equal, so $V_{T(\{n_i\})} = V_P \equiv V_{mol}$.

By making the following definitions: $\Omega \equiv \frac{\Omega'}{k_BT}\frac{V_{mol}}{V}$, $\mu_T \equiv \frac{\mu'_T}{k_BT}$, $\mu_P \equiv \frac{\mu'_P}{k_BT}$, $E_{int}(\{n_i\}) \equiv \frac{E'_{int}(\{n_i\})}{k_BT}$, $\rho_P \equiv \frac{N_P}{V}V_{mol}$, $\rho_T(\{n_i\}) \equiv \frac{N_T(\{n_i\})}{V}V_{mol,P}$, we obtain the following expression

$$\Omega = \rho_P\Big[\ln\rho_P - 1 - \mu_P\Big] + \sum_{\{n_i\}}\rho_T(\{n_i\})\Big[\ln\Big(\rho_T(\{n_i\})\frac{V_{mol,T(\{n_i\})}}{V_{mol,P}}\Big) - 1 - \mu_T - n\mu_P + E_{int}(\{n_i\})\Big].$$

In equilibrium the grand potential is minimized by the densities, so we consider Ω to be a function of the densities $\Omega = \Omega(\rho_P, \{\rho_T(\{n_i\})\})$, where $\{\rho_T(\{n_i\})\}$ is the set of the densities of templates with allowed configurations. Minimizing gives

$$\begin{aligned} \rho_{T,eq}(\{n_i\}) &= \exp[-E_{int}(\{n_i\}) + \mu_T + n\mu_P], \\ \rho_{P,eq} &= \exp[\mu_P]. \end{aligned} \tag{2.2}$$

The hessian of the grand potential around this point is diagonal and has only positive eigenvalues. Therefore, we conclude that the grand potential is indeed minimal at this values. Furthermore, the total dimensionless protein density at equilibrium, $\phi_P = \frac{N_{P,tot}}{V}V_{mol}$ with $N_{P,tot} = N_P + \sum_{\{n_i\}} nN_T(\{n_i\})$, may be calculated as

$$\phi_P = \rho_{P,eq} + \sum_{\{n_i\}} n\rho_{T,eq}(\{n_i\}) \tag{2.3}$$

Now that we have found the equilibrium densities of the components in the solution we can focus on the quantities which describe the self-assembly.

2.1.1 Equilibrium quantities

In a self-assembly process the proteins are distributed among the templates in a certain way. Some templates will have a fully grown capsid, others have only a partially grown capsid. The distribution of the proteins is of the utmost importance for it determines what fraction of the templates will be fully covered. Therefore, we define the fraction of templates having n proteins bound to it as $P(n) \equiv \frac{\rho_T(n)}{\sum_{n=0}^{q}\rho_T(n)}$. Another important quantity to know is what fraction of the available binding sites is occupied by proteins.

We define this quantity as $\langle\theta\rangle = \sum_{n=0}^{q} \frac{n}{q} P(n)$. These definitions hold in and outside of equilibrium. With the equilibrium results found above we can write

$$\begin{aligned}
\langle\theta\rangle_{eq} &\equiv \frac{\langle n(\{n_i\})\rangle_{eq}}{q} = \frac{1}{q}\frac{\sum_{\{n_i\}} n(\{n_i\})\rho_{T,eq}(\{n_i\})}{\sum_{\{n_i\}} \rho_{T,eq}(\{n_i\})} = \frac{1}{q}\frac{\sum_{\{n_i\}} n \exp[-E_{int}(\{n_i\}) + n\mu_P]}{\sum_{\{n_i\}} \exp[-E_{int}(\{n_i\}) + n\mu_P]}, \\
P_{eq}(n) &\equiv \frac{\sum_{\{n_i\}} \delta_{\sum_{i=1}^{q} n_i, n}\, \rho_{T,eq}(\{n_i\})}{\sum_{\{n_i\}} \rho_{T,eq}(\{n_i\})} = \frac{\sum_{\{n_i\}} \delta_{\sum_{i=1}^{q} n_i, n} \exp[-E_{int}(\{n_i\}) + n\mu_P]}{\sum_{\{n_i\}} \exp[-E_{int}(\{n_i\}) + n\mu_P]},
\end{aligned}$$

where we sum over all possible configurations since different configurations can have the same number of proteins bound. These expressions inspire us to define the, so called, semi-grand partition function

$$\Xi \equiv \sum_{\{n_i\}} \exp[-E_{int}(\{n_i\}) + n\mu_P], \tag{2.4}$$

where $E_{int}(\{n_i\})$ is the dimensionless interaction energy for a template with protein configuration $\{n_i\}$ and $n = \sum_{i=1}^{q} n_i$. Interestingly, this function is exactly the grand canonical partion function of a single template in contact with a heat and particle bath with a chemical potential μ_P, see figure 2.1.

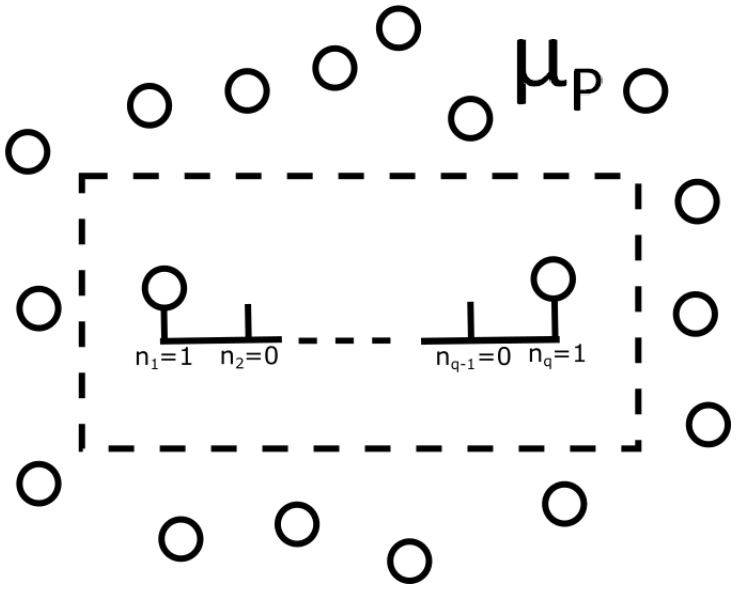

Figure 2.1: The system described by the semi-grand partition function. A single template is in contact with a heat and particle bath with a chemical potential μ_P. On the template a number of proteins can be bound which define a configuration given by the set of the occupation numbers $\{n_i\}$. The partition function is found by summing the Boltzmann factors of all configurations. From it the equilibrium quantities can found.

Using this function we obtain

$$\langle\theta\rangle_{eq} = \frac{1}{q}\frac{\partial \ln \Xi}{\partial \mu_P}, \tag{2.5}$$

$$P_{eq}(n) = \frac{\sum_{\{n_i\}} \delta_{\sum_{i=1}^{q} n_i, n} \exp[-E_{int}(\{n_i\}) + n\mu_P]}{\Xi}. \tag{2.6}$$

These relations shows that the crucial function for our interests is the semi-grand partition function, Ξ and the relative sizes of the terms where it consists of. This function may be calculated exactly for both the zipper and the Ising-S model which we will introduce further below. The following section introduces two important processes which are found during the binding of proteins to the template. This gives the physical background of the models introduced thereafter.

2.2 Allostery and cooperativity

In this section we will outline the energetic model we use to describe the process of capsid protein binding in self-assembly. This binding to the template is complicated and depends on the kind of template, coat protein etc. The self-assembly of the tobacco mosaic virus (TMV) as described in the introduction is, for example, different from the way the AVs of Hernandez-Garcia et al [21] bind to the templates. One difference is that TMV has a RNA template while the AVs have an dsDNA template. Another aspect is that the capsid building blocks (CBBs) for TMV are discs made of 17 capsid proteins, while for the AVs the CBB is simply one capsid protein. For simplicity we refer in the rest of this thesis to the CBBs as proteins. These two aspects imply that the precise assembly pathways of the two systems should differ very much. Nevertheless, to describe self-assembly we only need to consider two processes which involve energetic differences and which are found in many self-assembly processes. The precise nature of these processes is unnecessary to know for we only describe energetic differences. The magnitude of these differences may be determined through experiment.

In order to introduce the processes involved we consider a template, as given in figure 2.1, which is in contact with a heat and particle bath. First, suppose we have an empty template, as pictured in figure 2.2,

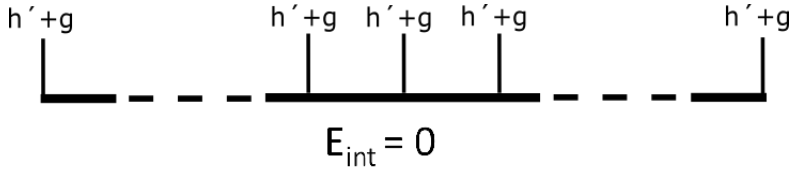

Figure 2.2: An empty template with identical binding sites. At every binding the system has to pay a conformational switching cost $h' > 0$ upon binding of a protein while it also gains a free energy $g < 0$ because of attractive interactions between the template and the protein.

where all binding sites are identical. The reason we assume that all sites are identical is to illustrate the two binding processes. More realistically the sites are typically not identical. For example, an assembly signal implies a site which is different from all others. The role of assembly signals in this decription will be outlined below with regard to the zipper and the Ising-S model. At every site the energetic change of the interaction energy upon binding of a protein is depicted.

The first protein to bind pays a free energy $h' > 0$ which factors in that the template and/or the protein typically have to make some kind of conformational switch. A conformational switch implies an energy barrier. Therefore, h' accounts for this energy barrier or any other process which represents a hurdle for nucleation. We will refer to this as allostery because, for TMV, the conformational switching is an allosteric process.

Furthermore, at every site the first protein gains a free energy $g < 0$ because, disregarding the energy barrier, the protein and the template generally have an attractive interaction. This can be a result of electrostatic, hydrophobic or any other kind of interaction.

After the first protein binds a number of changes occurs for the binding of the next protein, as shown in figure 2.3. The second protein to bind still requires a free energy $h' + g$, except for the sites adjacent to the first protein. On these sites it has no energy barrier. This is called cooperativity: the second protein does not have to pay the conformational switching cost at sites adjacent to an already bound protein.

With the binding of a second protein adjacent to the first, as given in figure 2.4, the interaction energy is changed by $\epsilon + g$. The two proteins which are bound adjacent to each other have an attractive protein-protein interaction wherefore the system gains a free energy $\epsilon < 0$. The next protein to bind at the empty site adjacent to the second protein also changes the energy by $\epsilon + g$, this can in principle proceed indefinitely. This gives rise to a preference of the system for sequential binding. The state with one protein bound, $n = 1$, is a high energy state due to h'. The state with $n = 2$ has a lower interaction energy than the $n = 1$ state. All subsequent states reached through cooperative binding have a lower interaction energy than their precursor. Therefore, with regard to the interaction energy, the $n = q$ state is

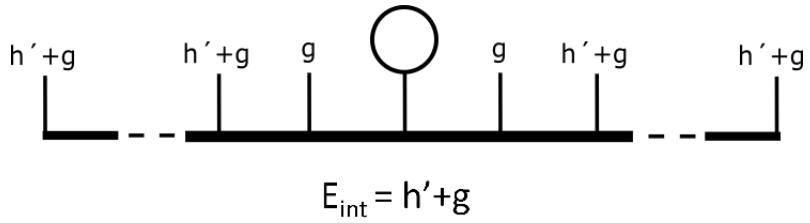

Figure 2.3: A template with one protein bound. Another protein can bind cooperatively or not. Cooperative binding is next to the first protein bound and gives a free energy $g < 0$ plus an interaction free energy with the first protein of $\epsilon < 0$. Non-cooperative binding gives a free energy $h' + g$ to the system, like for the first protein to bind.

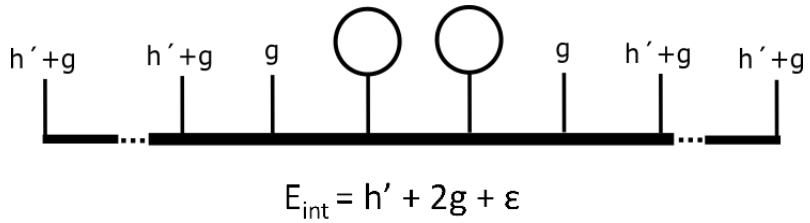

Figure 2.4: A template with two proteins cooperatively bound. Due to cooperatively can a third protein bind cooperatively next to the already bound proteins such that the system gains a free energy $\epsilon + g$. At all other sites the system has to pay a conformational switching cost $h' > 0$ upon binding and does it not gain an protein-protein interaction free energy $\epsilon < 0$.

most favourable. As will be shown in chapter three, cooperative binding gives rise to realization of the law of mass action for the zipper model.

From the above one might guess that the state where all sites are occupied due to cooperative binding is the most favourable state. However, this is not necessarily true because taking a protein out of the particle bath costs an energy μ_P, as depicted in figure 2.1. This chemical potential arises from the fact that the proteins have translational and mixing entropy when they are unbound in the solution. In chapter three we will show how cooperative binding of a protein gives rise to the following Boltzmann factor $s \equiv e^{\mu_P - \epsilon - g}$. Therefore, if $s > 1$ the binding of a protein is energetically favourable for the system, while it is not if $s < 1$. Moreover, for $s = 1$ there is no energetic difference in having protein in the particle bath or cooperatively bound. Therefore, one would expect a phase transition to happen from completely empty templates to completely filled templates upon increasing s in the limit that $q \to \infty$. In this limit, for $s > 1$ the state with $n = q$ would become infinitely advantageous.

Taking into account this critical behaviour and recalling from the previous section that $\rho_{P,eq} = e^{\mu_P}$, it makes sense to define a critical density $\phi_c \equiv e^{\epsilon + g}$ such that $s = \frac{\rho_{P,eq}}{\phi_c}$. So, the concentration of unbound proteins in the solution determines for finite q roughly and for $q \to \infty$ exactly whether we have empty or filled templates.

In chapter three will be shown that the nucleation of the capsid on the template gives rise to a Boltzmann factor of $\sigma = e^{-h+\epsilon}$. So, a state which has one or more proteins bound - a state with a nucleated capsid - has factor σ in the semi-grand partition function. This includes in the statistics that both the conformational cost and the protein-protein interaction which the first protein does not have, impede the nucleation.

2.3 Zipper model

As explained in section 2.1 the crucial ingredient in determining the semi-grand partition function is the interaction energy E_{int}. In this section we give the motivations and explicit form of E_{int} for the zipper

model. By doing so we will refer to the previous section for the concepts of allostery and cooperativity.

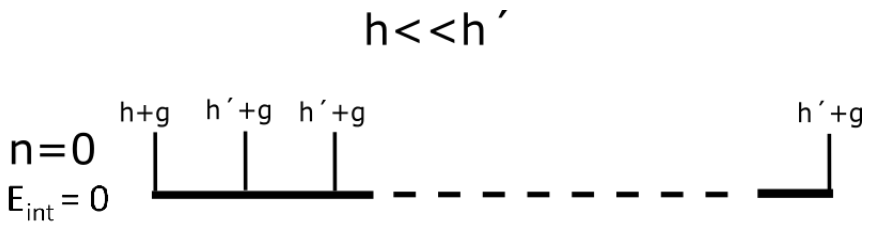

Figure 2.5: A zipper template with an assembly signal at the end of the template. This causes the energy barrier to be equal to $h \ll h'$. This causes the first protein to bind at the assembly signal. Thereafter, the other states are accessed through cooperative binding.

The zipper model describes the self-assembly of the *tobacco mosaic virus* (TMV) and that of artificial viruses (AV) [21] which is sequential for both. The sequential binding starts at the assembly signal which is approximately located at one end of the template. Afterwards, the self-assembly cooperatively proceeds up to $n = q$, which corresponds to a whole capsid. The start of the assembly at the end of the template can be modeled by assuming a greater attraction between the protein and the template at the assembly signal. This can be factored in by assuming that the energy barrier for the first protein to bind at the assembly signal is lowered by an amount $h - h'$, where h' is the energy barrier at any normal site - that is a site where no assembly signal is - and h is the lowered energy barrier. This is represented in figure 2.5. The difference between h and h' must be such that the entropic advantage of binding at a normal site is negligible compared to binding at the assembly signal. The entropic free energy of nucleation at a normal site goes as $\ln q - 1 \approx \ln q$ because the protein has $q - 1$ possibilities for binding at a normal site and q is assumed to be large. So, we require $\ln q \ll h' - h$. This condition also ensures that no second nucleation is favourable. The reason is that for the n-th protein to bind cooperatively, $\ln(q - n) \ll h' - g$ is required where $g < 0$ is the attractive interaction between a protein and the template. Under these conditions we expect only cooperative binding starting at the end of the template to occur. We will refer to this as zipper-like assembly. Zipper-like because it resembles the zipping of a jacket. From these considerations we can define the zipper states, which were first proposed by Kittel [23], in the following way

$$E_{int}(\{n_i\}) = E_{Zip}(n) = h + \epsilon(n-1) + gn, \text{ for } 1 \leq n \leq q \tag{2.7}$$

$$E_{int}(\{n_i\}) = E_{Zip}(0) = 0, \text{ for } n = 0, \tag{2.8}$$

$$\Xi_{zip} = \sum_{n=0}^{q} \exp[-E_{Zip}(n) + n\mu_P] = 1 + \sigma \sum_{n=1}^{q} s_{eq}^n, \tag{2.9}$$

with $g < 0$ the protein-template interaction which is generally attractive, $\epsilon < 0$ the protein-protein interaction which is also generally attractive, $s_{eq} \equiv e^{\mu_P - \epsilon - g} = \frac{\rho_{P,eq}}{\phi_c}$, $\sigma \equiv e^{\epsilon - h}$, $\phi_c \equiv e^{\epsilon + g}$ the critical concentration and $\rho_{P,eq}$ as defined in section 2.1. What is striking is that ϵ and h make up one parameter in the partition sum. It implies that they have the same effect: they both impede nucleation, that is, binding of the first protein. This is to be understood because the first protein has, compared to the other proteins, an energy barrier h besides not gaining a protein-protein interaction energy ϵ upon binding. In the previous section we noted that the law of mass action dictates that the density of a template with n proteins bound scales as s_{eq}^n. This is indeed the case for the zipper model. With E_{int} defined and the semi-grand partition function written down, all equilibrium properties of the system can be calculated, as will be done in chapter three.

As noted in the introduction AVs can be made which have more than one assembly signal. Moreover, the condition $\ln q \ll h' - h$ may be violated by large q and/or small energetic advantages of binding at an assembly signal. To model these possibilities we propose the Ising-S model as introduced in the next section.

2.3.1 Competition

Above we introduced the zipper model. It accounts for sequential binding of proteins to a template. The condition for this to occur is $\ln q \ll h' - h$. In relation to this model there is an interesting phenomenon to consider: competition. This comes in two kinds: self-competition and species competition. The first kind implies that for one kind of template not only zipper states are available but also competitor states. This would be due to a partial relaxation of the above mentioned zipper condition. The second comprises two kinds of templates competing for proteins from the same pool. That is, the two kinds of templates are in the same solution. We consider a species with assembly signal to be in competition with a species without assembly signal. In chapter five these two kinds of competition will be covered. Below we will introduce the energy states and semi-grand partition function for both kinds of competition.

For species competition one species would have energy states as defined by the zipper model and the other would have competitor states. The competitor states are assumed to have only one nucleation, like the zipper model, but the nucleation can be at any of the sites, for they are all equivalent. This implies that these states have a multiplicity factor since a cluster with n proteins can have $q - n + 1$ positions. With these considerations we obtain

$$E_{int}(\{n_i\}) \equiv E_{comp}(n) = h' + \epsilon(n-1) + gn, \text{ for } 1 \leq n \leq q \tag{2.10}$$

$$E_{int}(\{n_i\}) = E_{comp}(0) = 0, \text{ for } n = 0, \tag{2.11}$$

$$\Xi_{comp}(q) = 1 + \sigma' \sum_{n=1}^{q} (q - n + 1) s_{eq}^n, \tag{2.12}$$

where $\sigma' = e^{-h'+\epsilon}$ and $s_{eq} = e^{\mu_P - g - \epsilon}$. The energy barrier is equal to h' since no assembly signal is present.

For self-competition the competitor states contribute to the same partition function as the zipper states because they are states of the same kind of template. The self-competition partition function can thus be written as

$$\Xi_{sc} = 1 + (\Xi_{zip}(q) - 1) + (\Xi_{comp}(q-1) - 1). \tag{2.13}$$

So, the states where nucleation is not at the assembly signal, thus having $q - 1$ binding possible binding sites, are competitor states while all others are zipper states. The self-competition partition function will be encountered in section 4.2. There will be shown that the first order approximation of the Ising-$\{1\}$ model is equal to Ξ_{sc}. Here, $\{1\} = S$ because S is the set of special positions of the Ising-S model. This will be introduced further in the next section.

For both kinds of competition mass conservation determines the value of s_{eq}. In this way the connection between, respectively, the different species and states is made. In chapter five a short analysis of the effects of competition will be made.

2.4 Ising-S model

As noted in the introduction we propose the Ising-S model to account for multiple assembly signals and the effect of entropy. Below we will introduce this model.

Like for the zipper model we assume a free energy barrier for nucleation and the possibility of cooperative binding. Unlike the zipper model we do not impose a strong assembly signal at the end of the template. Instead, we let the energetic differences and the number of assembly signals be unconstrained. In this way the energetic advantage of the assembly signal(s) can be tuned as well as the entropic effects.

As for the zipper model we take protein-template interaction between one protein and the template to be given by $g < 0$ and the protein-protein interaction of two proteins to be given by $\epsilon < 0$. The energy barrier upon binding at a normal binding site - where there is no assembly signal - is $h' > 0$ and the barrier at the i-th assembly signal is h_i such that at the assembly signal the energy is lowered by $h_i - h'$. This might seem a needless complicated definition but it will turn out to be useful when considering the reduction of the Ising-S model to the zipper model.

For reference we group the special sites as $S = \{p_1, p_2, ..., p_m\}$ where $p_i \in P \equiv \{1, 2, ..., q\}$ and m is the number of assembly signals. So, in principle the assembly signals can be anywhere located at the template. In general we refer to this model as the Ising-S model but for a particular case, say $p_1 = 1$ and $p_2 = q$, we have $S = \{1, q\}$ and thus one can refer to the Ising-$\{1, q\}$ model.

From these considerations we can write down the interaction energy

$$E_{int}(\{n_i\}) = \epsilon \sum_{i=1}^{q-1} n_i n_{i+1} + gn + h' \sum_{i=0}^{q-1} (1 - n_i) n_{i+1} + \sum_{i \in S} (h_i - h') n_i, \tag{2.14}$$

where $n_0 \equiv 0$ because otherwise the first cluster is not counted. For the i-th binding site is n_i unity if occupied by a protein and zero if unoccupied. The first term accounts for the protein-protein interactions, the second for the protein-template interactions, the third for the nucleation costs by counting the number of protein cluster and the fourth gives the energy lowering due to the occupation of assembly signals. We can write down the semi-grand partition function as

$$\Xi^S = \sum_{\{n_i\}} \exp\Big[- h' \sum_{i=0}^{q-1} (1 - n_i) n_{i+1} - \epsilon \sum_{i=1}^{q-1} n_i n_{i+1} + (\mu_P - g) n - \sum_{i \in S} (h_i - h') n_i \Big].$$

For any configuration with a given n and a given number of protein clusters, $k \equiv \sum_{i=0}^{q-1} (1 - n_i) n_{i+1}$, is the Boltzmann factor fully determined. This can be seen as follows. Every cluster of n proteins gives $n - 1$ protein-protein interactions. Therefore, with the n proteins divided into k clusters there are in total $n - k$ protein-protein interactions. The Boltzmann weight can thus be written as

$$\exp\Big[- h'k - \epsilon(n - k) + n(\mu_P - g) - \sum_{i \in S} (h_i - h') n_i \Big] = \sigma^k s_{eq}^n \chi_{p_1}^{n_{p_1}} \chi_{p_2}^{n_{p_2}} ... \chi_{p_m}^{n_{pm}},$$

with $s_{eq} \equiv e^{\mu_P - \epsilon - g} = \frac{\rho_{P,eq}}{\phi_c}$, $\sigma \equiv e^{\epsilon - h'}$, $\chi_i \equiv e^{h' - h_i}$, $\phi_c \equiv e^{\epsilon + g}$ the critical concentration and $\rho_{P,eq}$ as defined in section 2.1. This expression of the Boltzmann factor is well suited for the cluster expansion of the partition function because it gives the multiplicity of states with n proteins and k clusters. This expansion will be used in the calculation of the equilibrium distribution in section 4.3. This form of the Boltzmann weights shows that at least three independent parameters which determine it, instead of two for the zipper model. The reason is that nucleation can take place at any site, whether it is special or not. Therefore are the energy gain of binding at an assembly signal and the energy barrier of nucleation not necessarily connected. To clarify this we consider the $S = \{1\}$ model, where we have $\sigma^k s_{eq}^n \chi_1^{n_1}$. In the configurations of n proteins bound adjacent to each other starting at the special site we have for the Boltzmann factor $\sigma s_{eq}^n \chi_1 = e^{-(h' + h_1) - n(\epsilon - 1) - gn + n\mu_P}$. Compare this to equation (2.7) taking into account that the Boltzmann factor is equal to $e^{-E_{int} + n\mu_P}$ and one sees that the Ising-$\{1\}$ has, among others, the same configurations as the zipper model. These zipper configurations can be dominating, but this depends on the value of σ. If σ is very small, the configurations with more than one cluster will be suppressed for they scale as σ^k with k the number of clusters. Suppose, for example, that the n proteins are divided into two clusters. One of the clusters starts at the special site, the other at any other site. These configurations will have a Boltzmann factor of $\sigma^2 s_{eq}^n \chi_1$. Therefore, if σ is small these configurations will be suppressed. This hints that if $\sigma \to 0$ and if χ_1 is large, such that the entropic gain of not binding at the assembly signal is negligible, the Ising-$\{1\}$ reduces to the zipper model. We will show in section 4.2 that this is indeed the case. The great advantage of this model is that for small σ one can see exactly whether entropy is of any effect. The effect of entropy depends on q, because if it is large enough one would expect that due to entropy there will be nucleation at non-special sites. We will show in chapter four that the question whether entropic effects are important can be roughly answered considering the correlation length ξ. This is the typical length of a protein cluster for some given conditions. If $\xi \gg q$ one would not expect entropic effects to be relevant while for $\xi \ll q$ they would be relevant.

In chapter four we will calculate the semi-grand partition function of this model and we will give expressions for $\langle \theta \rangle$ and $P_{eq}(n)$. In the following section the kinetic theory used for calculating the dynamics of the zipper model will be explained.

2.5 Dynamical equations

In the above sections we outlined the zipper and Ising-S model which describe self-assembly. These models can predict the equilibrium properties of the system. However, in what way the system reaches equilibrium is a wholly different question. To answer this question we use the theory of chemical kinetics. The binding of a protein to a template can be viewed as a chemical reaction. The way to describe the evolution of a system where a number of reactions takes place between its components is by means of rate equations. We will introduce this chemical theory below.

Suppose we have N species of molecules taking part in a number of reactions with X_i denoting the i-th species. Furthermore, we assume we have a total number of R reversible reactions, since we assume to have only reversible reactions in self-assembly. Also, we define the coefficient of species i in the j-th forward reaction as a_{ij} and in the j-th backward reaction as b_{ij}. This gives for the system of reactions

$$\mathrm{a_{1j}X_1 + a_{2j}X_2 + ... + a_{Nj}X_N} \quad \underset{\mathrm{k_-(j)}}{\overset{\mathrm{k_+(j)}}{\rightleftharpoons}} \quad \mathrm{b_{1j}X_1 + b_{2j}X_2 + ... + b_{Nj}X_N}, \tag{2.15}$$

where $j = 1, 2, ..., R$. To write the rate equations we infer the law of mass action for both the forward and backward reactions. To denote this, we define the vector containing the concentrations of all components as $[\vec{X}] \equiv \{[X_1]\,[X_2]\,...[X_N]\}$. We can write the flux of respectively the forward (+) reaction, the backward (-) reaction and the net flux as

$$f_j^+([\vec{X}]) = k_j^+ \prod_{i=1}^{N} [X_i]^{a_{ij}}, \tag{2.16}$$

$$f_j^-([\vec{X}]) = k_j^- \prod_{i=1}^{N} [X_i]^{b_{ij}}, \tag{2.17}$$

$$f_j([\vec{X}]) \equiv f_j^+([\vec{X}]) - f_j^-([\vec{X}]), \tag{2.18}$$

where k_j^+ and k_j^- are, respectively, the forward and backward rate constants of the j-th reaction. Also, we may define the a matrix M as $M_{ij} = b_{ij} - a_{ij}$, which gives the net consumption and production of each component in all reactions. Finally, we can write, with the Einstein summation convention, the rate equations

$$\begin{aligned} \partial_t[X_i] &\equiv M_{ij} f_j, \\ &= \sum_{j=1}^{R} M_{ij} (k_j^+ \prod_{l=1}^{N} [X_l]^{a_{lj}} - k_j^- \prod_{l=1}^{N} [X_l]^{b_{lj}}). \end{aligned} \tag{2.19}$$

In these equations we may take all concentrations to be dimensionless by multiplying them with time-independent quantities and redefining the rate constants. This poses no problem since the rate constants are experimentally determined and may always be multiplied with some constant. In chapter three we will introduce what chemical reactions are occurring at self-assembly and we will derive the dynamical equations.

Chapter 3

Zipper model

In this chapter we will first calculate the semi-grand partition function of the zipper model and subsequently derive the $\langle\theta\rangle_{eq}$ and $P_{eq}(n)$: the average fraction of occupied binding sites and the equilibrium distribution respectively. Afterwards, the previously found equilibrium quantities will be used to derive the dynamical equations which describe the zipper self-assembly as a function of time. Next, we give after a short overview of the equilibrium properties of the system a description of the dynamics which the zipper model comprises. Moreover, we note some possible approximations to the dynamical equations which can serve as inspiration for future research. Finally, acquired data will be fit to the dynamical model to see whether the model fits reality.

3.1 Equilibrium quantities

As outlined in section 2.1 we can calculate $\Xi(\mu_p)$ and subsequently $\langle\theta\rangle_{eq}$ and $P_{eq}(n)$. From equation (2.9), $\Xi(\mu_p)$ is calculated with the geometric sum formula as

$$\Xi_{zip} = 1 + \sigma s_{eq}\frac{1-s_{eq}^q}{1-s_{eq}}, \tag{3.1}$$

where $\sigma \equiv e^{\epsilon-h}$ and $s_{eq} \equiv e^{-\epsilon-g+\mu_P} = \rho_{P,eq}e^{-\epsilon-g} = \frac{\rho_{P,eq}}{\phi_c}$ with ϕ_c the critical concentration, $\rho_{P,eq}$ the equilibrium concentration of unbound proteins and σ the Boltzmann factor for nucleation. The value of s_{eq} roughly determines whether the system is assembled, $s_{eq} > 1$, or disassembled, $s_{eq} < 1$. The value of σ strongly influences what fraction of templates is nucleated. So, which fraction has $n > 0$ in equilibrium. The reason for defining s_{eq} instead of simply s is, as will be shown in section 3.2, that s is time dependent. From equation (2.5) we calculate the average occupation of the templates

$$\langle\theta\rangle_{eq} = \frac{\sigma}{q}\frac{s_{eq}}{(1-s_{eq})}\frac{1-(q+1)s_{eq}^q+qs_{eq}^{q+1}}{1-s_{eq}+\sigma s_{eq}(1-s_{eq}^q)}, \tag{3.2}$$

in the same way as Kraft et al. [7]. Furthermore, with equation (2.6) we may find the equilibrium distribution of the templates

$$P_{eq}(n) = \begin{cases} \frac{1}{\Xi}, & \text{for } n = 0 \\ \frac{\sigma s_{eq}^n}{\Xi}, & \text{for } 1 \leq n \leq q \end{cases} \tag{3.3}$$

This distribution gives what fraction of templates has n proteins bound to it. This fraction - and thus concentration - follows the law of mass action since it is proportional to $s_{eq}^n = \left(\frac{\rho_{P,eq}}{\phi_c}\right)^n$. In the next section we will use these equilibrium quantities to derive the dynamical equations.

3.2 Dynamical equations

To derive the equations wich describe zipper self-assembly we use the chemical theory of section 2.5 and the equilibrium quantities of the previous section.

We first note that the molecule species which react in the system are proteins and templates with a given number of proteins bound to them. This can be summarized into the species vector $\vec{X} = \{\mathrm{TEM}_0, \mathrm{TEM}_1, ..., \mathrm{TEM}_\mathrm{q}, P\}$. Here, P denotes an unbound protein unit in the solution which may bind at one binding site on the template, TEM_0 denotes an uncovered template in the solution, TEM_1 denotes a template with one protein unit bound, etc. We assume only unbound proteins to react with templates, so we have no template-template reactions. The reactions taking place may be written down as follows

$$\mathrm{P} + \mathrm{TEM}_{\mathrm{n-1}} \quad \underset{\mathrm{k\_(n)}}{\overset{\mathrm{k_+(n)}}{\rightleftharpoons}} \quad \mathrm{TEM}_\mathrm{n}, \tag{3.4}$$

for $1 \leq n \leq q$. From these reactions we can write the a, b and M matrices, as defined in section 2.5, which are $(q+2) \times q$ matrices since we have $q+2$ components and q reactions. Furthermore, we write $[\vec{X}](t) = \{\rho_T(0,t), \rho_T(1,t), ..., \rho_T(N,t), \rho_P(t)\}$, with $\rho_P(t)$ the concentration of unbound protein units in the solution at time t, $\rho_T(0,t)$ the concentration of TEM_0 molecules, etc. We can write the fluxes as

$$f_n^+([\vec{X}])(t) = k_n^+ \prod_{m=1}^{N} [X_m]^{a_{mn}} = k_n^+ \rho_T(n-1,t)\rho_P(t), \tag{3.5}$$

$$f_n^-([\vec{X}])(t) = k_n^- \prod_{m=1}^{N} [X_m]^{b_{mn}} = k_n^- \rho_T(n,t), \tag{3.6}$$

$$f_n([\vec{X}])(t) \equiv f_n^+([\vec{X}])(t) - f_n^-([\vec{X}])(t), \tag{3.7}$$

with $n = 1, 2, ...q$. Now we can write the rate equations using the M matrix and we drop for simplicity the $[\vec{X}]$ arguments

$$\frac{\partial \rho_T(0,t)}{\partial t} = -f_1(t), \tag{3.8}$$

$$\frac{\partial \rho_T(n,t)}{\partial t} = f_n(t) - f_{n+1}(t), \tag{3.9}$$

$$\frac{\partial \rho_T(q,t)}{\partial t} = f_q(t), \tag{3.10}$$

$$\frac{d\rho_P(t)}{dt} = -\sum_{j=1}^{q} f_j(t), \tag{3.11}$$

with $1 \leq n \leq q-1$. In order to write the equations in a more convenient way we define the total density of templates $\rho_T \equiv \sum_{n=0}^{q} \rho_T(n,t)$ and the distribution of the templates as a function of time $P(n,t) \equiv \frac{\rho_T(n,t)}{\rho_T}$. Furthermore, as noted in section 2.2 in equilibrium the concentration of the unbound proteins relative to the critical concentration, defined as $s_{eq} = \frac{\rho_{P,eq}}{\phi_c}$ with $\phi_c \equiv e^{\epsilon+g}$, determines the probability of having a protein cooperatively bound to the template. Since the concentration of unbound proteins is now a function of time we define $s(t) \equiv \frac{\rho_P(t)}{\phi_c}$. Also, we define the total concentration of proteins relative to the critical concentration as $S \equiv \frac{\phi_P}{\phi_c}$. Finally, the ratio of the number of available binding sites to the total concentration of proteins present in the system should be important in the

dynamics, so we define $\lambda \equiv \frac{q\rho_T}{\phi_P}$. With these definitions we may write

$$\begin{aligned}
\frac{\partial P(0,t)}{\partial t} &= -v_1(t), && (3.12)\\
\frac{\partial P(n,t)}{\partial t} &= v_n(t) - v_{n+1}(t), && (3.13)\\
\frac{\partial P(q,t)}{\partial t} &= v_q(t), && (3.14)\\
\frac{1}{S}\frac{ds(t)}{dt} &= -\frac{\lambda}{q}\sum_{n=1}^{q} v_n(t), && (3.15)
\end{aligned}$$

where $v_n(t) \equiv \frac{f_n(t)}{\rho_T} = v_n^+(t) - v_n^-(t)$, $v_n^+(t) \equiv \frac{f_n^+(t)}{\rho_T} = k_n^+ P(n-1,t)\rho_P(t) = k_n^{+\prime} P(n-1,t)\frac{s(t)}{S}$, $v_n^-(t) \equiv \frac{f_n^-(t)}{\rho_T} = k_n^- P(n,t)$ and $k_n'^+ \equiv k_n^+ \phi_P$. We assume the total mass of the proteins to be conserved, so for all times

$$\begin{aligned}
\phi_P &= \rho_P(t) + \sum_{n=0}^{q} n\rho_T(n,t), && (3.16)\\
1 &= \frac{s(t)}{S} + \frac{\lambda}{q}\sum_{n=0}^{q} nP(n,t). && (3.17)
\end{aligned}$$

To check that this relation is respected by the rate equations we take the time derivative and by using equation (3.12) up to and including (3.15) we see that

$$\sum_{n=0}^{q} n\,\frac{\partial P(n,t)}{\partial t} = \sum_{j=1}^{q} v_j(t),$$

so the rate equations indeed conserve the protein mass in time. One may also check that ρ_T, and thus the probability of the distribution, is conserved by the rate equations.
A constraint on the reaction rate constants can be found by considering the following. If t goes to infinity $\lim_{t\to\infty} P(n,t) = P_{eq}(n)$ and $\lim_{t\to\infty} \frac{\partial P(n,t)}{\partial t} = 0$, for $0 \le n \le q$. This gives the relation $k_+(n+1)\frac{s_{eq}}{S}P_{eq}(n) = k_-(n+1)P_{eq}(n+1)$ for $0 \le n \le q-1$, where for notational convenience we let $k_+' \to k_+$. This allows to write the equilibrium constants as

$$\begin{aligned}
K_1 &= \frac{k_+(1)}{k_-(1)} = \sigma S, && (3.18)\\
K_n &= \frac{k_+(n)}{k_-(n)} = S, && (3.19)
\end{aligned}$$

for $2 \le n \le q$. Furthermore, like Kraft et al. [7], we make the simplifying assumption that $k_+(1) = \kappa k_+$ and $k_+(n) = k_+$, with κ a measure for the kinetic probability that nucleation occurs. We define $\tau \equiv k_+ t$, $y(\tau) \equiv \frac{s(\tau)}{s_{eq}}$, $f(n,\tau) \equiv \frac{P(n,\tau)}{P_{eq}(n)}$ for $0 \le n \le q$ and use the equilibrium relations from section 3.1 to obtain

the following dynamical equations

$$\frac{\partial f(0,\tau)}{\partial \tau} = -\frac{\kappa s_{eq}}{S}\Big(y(\tau)f(0,\tau) - f(1,\tau)\Big), \tag{3.20}$$

$$\frac{\partial f(1,\tau)}{\partial \tau} = -\frac{s_{eq}}{S}\Big(y(\tau)f(1,\tau) - f(2,\tau)\Big) + \frac{\kappa}{\sigma S}\Big(y(\tau)f(0,\tau) - f(1,\tau)\Big), \tag{3.21}$$

$$\frac{\partial f(n,\tau)}{\partial \tau} = -\frac{s_{eq}}{S}\Big(y(\tau)f(n,\tau) - f(n+1,\tau)\Big) + \frac{1}{S}\Big(y(\tau)f(n-1,\tau) - f(n,\tau)\Big), \tag{3.22}$$

$$\frac{\partial f(q,\tau)}{\partial \tau} = \frac{1}{S}\Big(y(\tau)f(q-1,\tau) - f(q,\tau)\Big), \tag{3.23}$$

$$\frac{dy(\tau)}{d\tau} = -\frac{\lambda}{q\Xi_{eq}}\Big[\kappa\Big(y(\tau)f(0,\tau) - f(1,\tau)\Big) + \sigma\sum_{n=1}^{q-1} s_{eq}^{n}\Big(y(\tau)f(n,\tau) - f(n+1,\tau)\Big)\Big]. \tag{3.24}$$

These equations imply that if $\lambda \to 0$ we obtain equation (16) up to and including (19) from Kraft et al. [7]. In this limit there is an infinite supply of proteins and from mass conservation we have $y(\tau) = 1$, so $s_{eq} = S$. For $\lambda > 0$ we can not solve this set of equations exactly for they are non-linear and coupled. However, they can be solved numerically as will be done in the following section.

For later convenience and as a check of the validity of the above dynamical equations we show that equation (3.24) is really mass conservation in disguise. From equation (3.20) up to and including (3.23) we find that

$$y(\tau)f(n,\tau) - f(n+1,\tau) = -\frac{S}{s_{eq}}\Big(\frac{1}{\sigma s_{eq}^{n}}\partial_\tau f(0,\tau) + \sum_{m=1}^{n} s_{eq}^{-(n-m)}\partial_\tau f(m,\tau)\Big), \tag{3.25}$$

for $1 \le n \le q-1$, with which we can write

$$\partial_\tau y(\tau) = \frac{\lambda S}{q}\frac{1}{s_{eq}}\Big(q\partial_\tau P(0,\tau) + \sum_{n=1}^{q}(q-n)\partial_\tau P(n,\tau)\Big),$$

where we used the definition of $f(n,\tau)$. From equation (3.12) up to and including (3.14) we have that $\sum_{n=0}^{q}\partial_\tau P(n,\tau) = 0$ and thus we find

$$\frac{\partial_\tau s(\tau)}{S} = -\lambda\sum_{n=1}^{q}\frac{n}{q}\partial_\tau P(n,\tau) = \lambda\partial_\tau\langle\theta\rangle(\tau).$$

This expression may be integrated from $\tau = 0$ to arbitrary τ to give

$$\frac{s(\tau)}{S} + \lambda\langle\theta\rangle(\tau) = \frac{s(0)}{S} + \lambda\langle\theta\rangle(0) \equiv 1. \tag{3.26}$$

Since the mass conservation equation derives from the dynamical equations we can use it in the approximations to the dynamical equations in appendix C.

3.3 Numerical analysis

In the previous section the dynamical equations which govern the assembly kinetics of the zipper model were derived. Below, an introduction into the richness of the dynamics will be given. In particular, the influence of a finite protein concentration, $\lambda > 0$, on the dynamics will be considered. As a preparation, the equilibrium properties - which determine the end point of the assembly - will be outlined and all parameters will be defined. Afterwards, these will be used in analysing the observed peculiarities of the assembly kinetics.

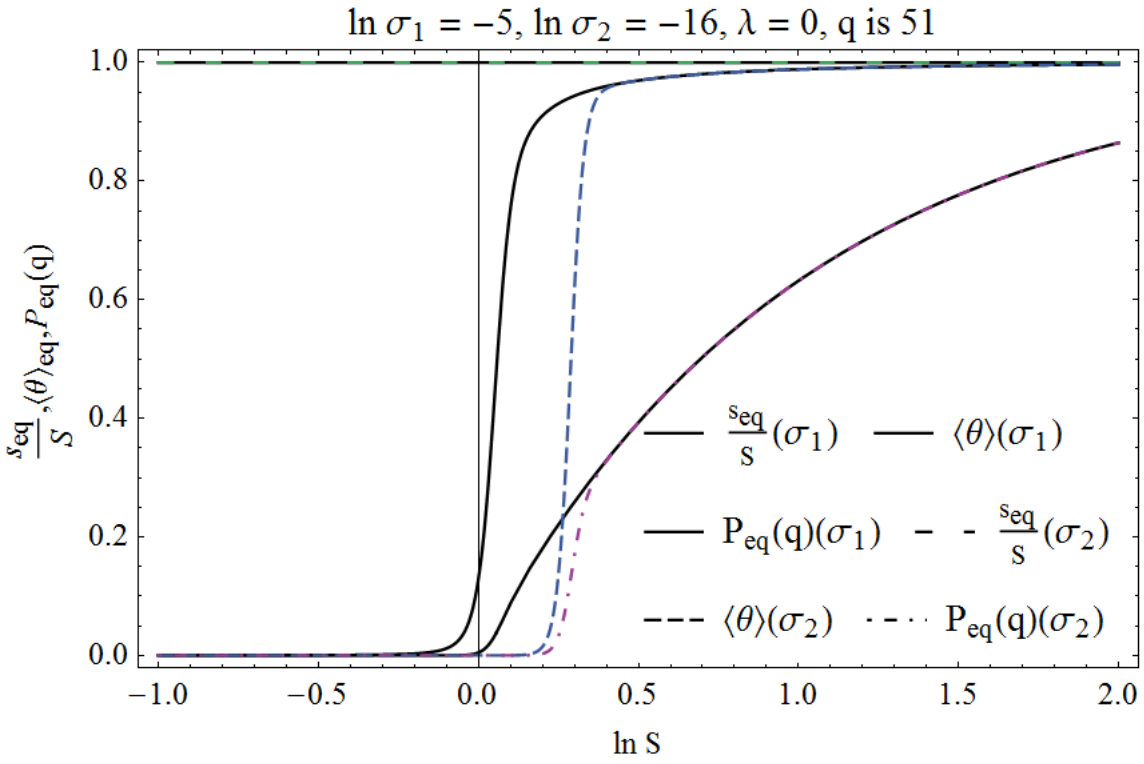

Figure 3.1: For $\lambda = 0$ characteristic equilibrium quantities are given as a function of S for two limiting values of σ. Here, $\frac{s_{eq}}{S}$ is the fraction of unbound proteins relative to the total concentration, $P_{eq}(q)$ is the fraction of templates which are fully covered and $\langle\theta\rangle$ is the average occupation number of the binding sites. At $S < 1$ (practically) no assembly occurs for the (unbound) protein concentration is smaller than the critical concentration. After some value of S does σ not have any influence. The reason is that nearly all templates are nucleated. Finally, from mass conservation it is given that $s_{eq} = S$ with $\lambda = 0$.

3.3.1 Equilibrium properties

The parameters which govern the equilibrium properties of the system are $S = \frac{\phi_P}{\phi_c}, \lambda = \frac{q\rho_T}{\phi_P}, \sigma = e^{-h+\epsilon}$ and q, with S the relative total concentration of proteins, ϕ_P the total concentration of proteins, $\phi_c = e^{\epsilon+g}$ the critical concentration, λ the stoichiometric ration, q the number of bindings sites per template and ρ_T the concentration of templates. Together, they determine $s_{eq} = s_{eq}(S, \lambda, \sigma, q) = \frac{\rho_{P,eq}}{\phi_c}$, the relative concentration of unbound proteins, through mass conservation - $s_{eq} = S(1 - \lambda\langle\theta\rangle(s_{eq}, \sigma, q))$ with $\langle\theta\rangle$ from section 3.1 - which in turn determines the equilibrium properties $\frac{s_{eq}}{S}, \langle\theta\rangle$ and $P_{eq}(q)$. In appendix B is an analysis of the mass conservation equation given for more insight in the function $s_{eq}(S, \lambda, \sigma, q)$. Respectively, the first property gives the fraction of proteins which is unbound, the second what fraction at the binding sites of the templates is covered with proteins, and the third what fraction of templates is completely covered.

In figure 3.1 are the properties given as a function of S, for two limiting values of σ - $\sigma_1 = e^{-5}$ and $\sigma_2 = e^{-16}$ - for $\lambda = 0$ and $q = 51$. As explained below, in this entire section we take $q = 51$, for q mainly influences the lag time but not the essence of the dynamics. The figure shows that $\frac{s_{eq}}{S} = 1$ for all S and σ. This is to be expected since $\lambda = 0$ implies an infinite supply of proteins. Furthermore, the value of $\langle\theta\rangle$ increases from zero to unity for $S > 1$. This is to be expected since $s_{eq} = S$ and only for $s_{eq} = e^{\mu_P - \epsilon - g} > 1$ is cooperative binding energetically favourable. For σ_2 does the increase start at a higher value of S. One can show the cause of this by expanding $s_{eq}(S, \lambda, \sigma, q)$ up to first order in σ. Afterwards, by calculating the value S_* at which $\langle\theta\rangle$ starts to grow, one can see that $S_* \propto \sigma^{\frac{1}{q}}$. This will not be done explicitly though.

Finally, the value of $P_{eq}(q)$ shows, similarly to $\langle\theta\rangle$, a lag for σ_2 and an independence of σ after some value of S. This implies that almost all templates are nucleated for the following reason. From section 3.1 we have for the equilibrium distribution

$$P_{eq}(n) = \frac{\sigma s_{eq}^n}{1 + \sigma \frac{1+s_{eq}^q}{1-s_{eq}}},$$

where in the denominator is the partition sum. The partition sum is made up of two contributions. The first term is for empty templates, the second is for nucleated templates. Obviously, if the second contribution is much greater than the first, the distribution is independent of σ. In other words, if almost all templates are nucleated does σ not have an influence on the distribution.

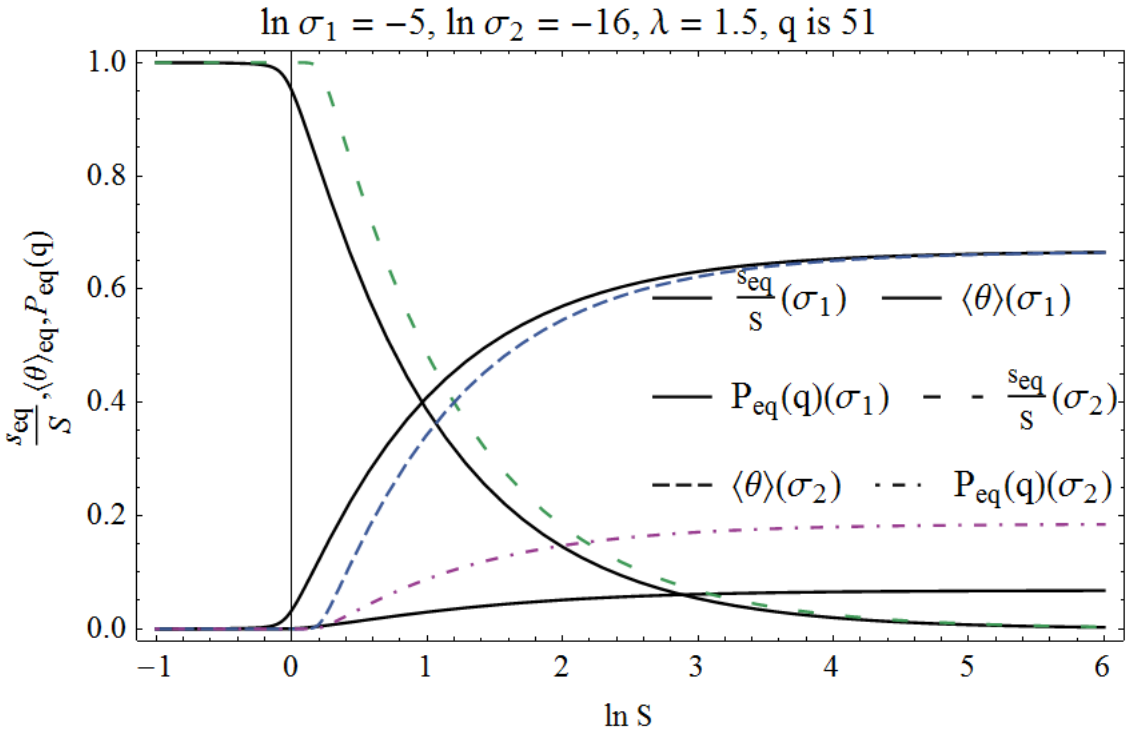

Figure 3.2: For $\lambda = 1.5$ characteristic equilibrium quantities are given as a function of S for two limiting values of σ. At $S < 1$ (practically) no assembly occurs for the (unbound) protein concentration is smaller than the critical concentration. After some value of S does σ not have any influence on $\langle\theta\rangle$. The value of $P_{eq}(q)$ is σ dependent however. The reason is that a significant portion of the templates is uncovered. Finally, $\frac{s_{eq}}{S}$ decays to zero because an ever larger portion of the proteins is bound to the templates.

With these properties for $\lambda = 0$ in mind does figure 3.2 give the properties for $\lambda = 1.5$. This figure shows, like figure 3.1, that for $S < 1$ there is barely any assembly. But for $S > 1$ does the fraction of unbound proteins go to zero as S increases. The reason is that 1.5 times more binding sites are available than proteins. Therefore, all proteins which can bind will be bound such that the concentration of unbound proteins will be close to the critical concentration - so $s_{eq} \approx 1$ - for all $S > 1$. Furthermore, $\langle\theta\rangle$ does not go to unity but to $\frac{2}{3}$ since $\lambda = 1.5$. Finally, $P_{eq}(q)$ does depend on σ because, as explained above, a considerable fraction of the templates is uncovered. This is caused by λ being greater than unity. In the next paragraph will these equilibrium properties turn out to be helpful in understanding the dynamical properties.

3.3.2 Dynamical properties

To see the dynamical properties of zipper assembly we will first discuss the role of σ and the role of q. Therefrom, we argue that they can be chosen fixed. Afterwards, we will give a reference assembly graph

for $\lambda = 0$ in order to understand a phase diagram which shows how the dynamics are altered for $\lambda > 0$.

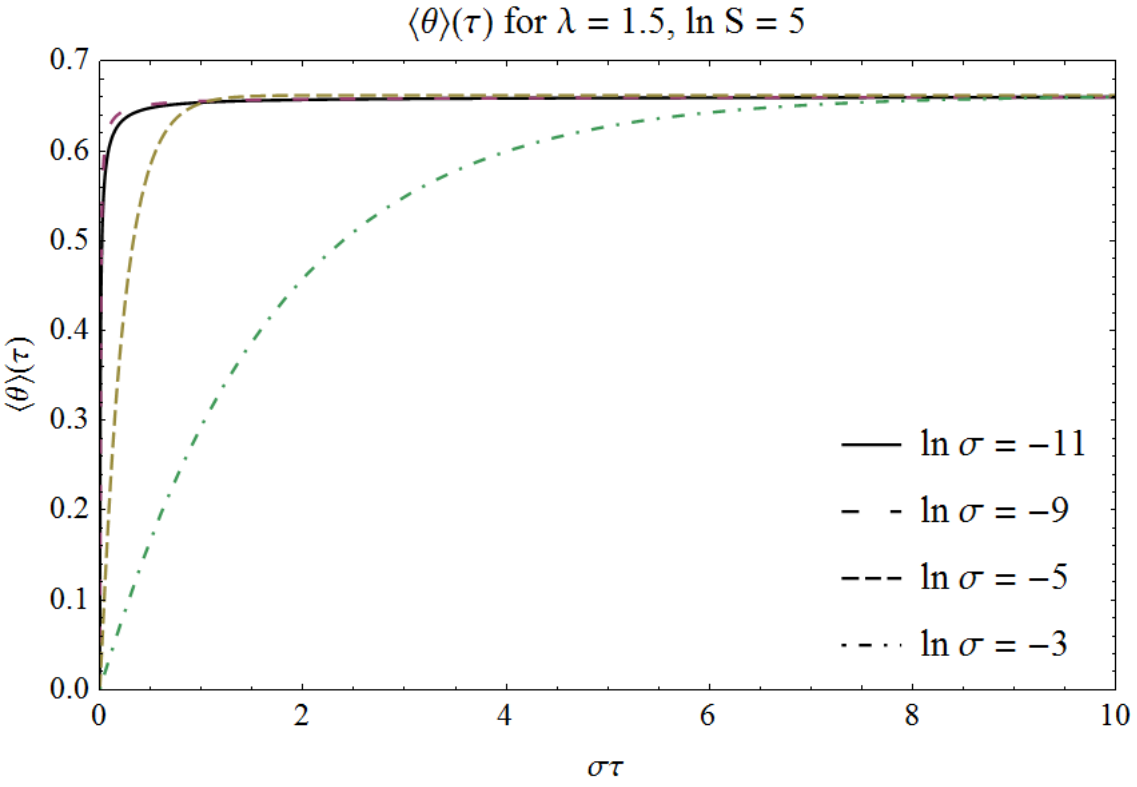

Figure 3.3: The typical timescale of the dynamics is very much influenced by σ. For different values of σ is the assembly time of the same order when expressed in $\sigma\tau$. For $\sigma = e^{-11}$ and $\sigma = e^{-9}$ the dynamics are almost equal. However, this only holds for high S such that $\langle\theta\rangle_{eq}$ is independent of σ.

In the previous paragraph was shown what the influence of σ on the equilibrium properties of the system is. To see its influence on the dynamics we recall from section 3.2 that $K_1 = \sigma S$ and $K_n = S$, with $2 \leq n \leq q$, and K_n being the equilibrium constant of the n-th chemical reaction. Per definition, we have $K_n = \frac{k_+(n)}{k_-(n)}$ with $k_+(n)$ the forward rate constant for obtaining the species on the right hand side of the n-th reaction and $k_-(n)$ vice versa (see equation (3.4)). Since all forward rate constants were taken to be equal and σ much smaller than one, does K_1 imply that $k_-(1)$ is very large as compared to all other $k_-(n)$. Informally, this means that the first reaction, the nucleation reaction, is slow: it is the rate determining step. Next, in all simulations do we take $P(0,0) = 1$, that is, at $\tau = 0$ only empty templates are present. Therefore, all nucleated templates in the assembly arise through the first reaction. This makes the first reaction indeed rate determining. Moreover, this is exactly the expected effect of σ since it is exponentially inversely proportional to the nucleation cost which acts as an energy barrier. Therefore, as shown in figure 3.3, does σ determine the time scale of the dynamics. For the two smallest values does $\langle\theta\rangle$ show universal behaviour, but also for the two greatest values is σ a major time scaling factor. Though, the universal behaviour only arises when the equilibrium value of $\langle\theta\rangle$ is the same for different values of σ. Therefore, in the rest of this analysis we consider only $\sigma = e^{-4} \approx 0.02$ whose dynamics should be representative for all smaller value of σ.

Next, the value of q determines the number of equations and thus how much time is minimally needed to fill a template completely. In fact, the lag time - the time required before the fraction of fully covered templates starts to deviate from zero appreciably - scales with q. However, this lag time is typically much smaller than $\frac{1}{\sigma}$, the typical time scale dictated by σ, wherefore it does not have a great influence on the dynamics. Therefore, we consider only $q = 51$.

In the paper of Kraft et al [7] were the dynamics of the zipper model probed for $\lambda = 0$. Therefore, we focus on how the dynamics of $\lambda > 0$ differs. For reference, is in figure 3.4 as a function of $\sigma\tau$ the value

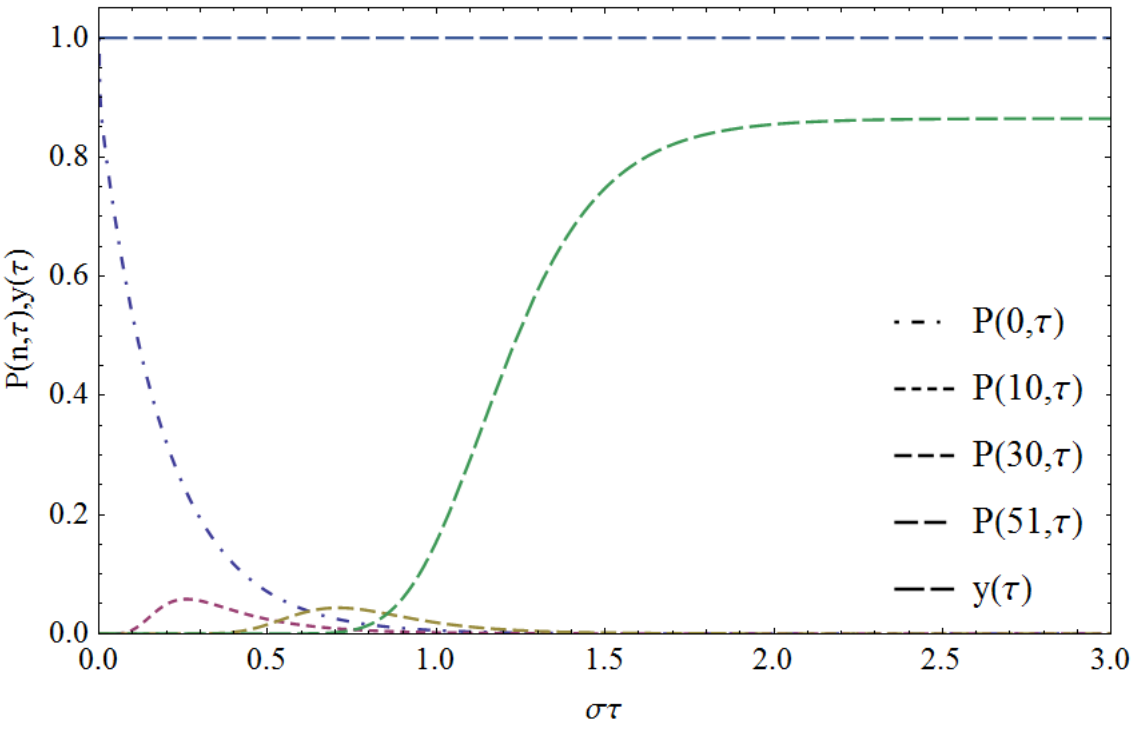

Figure 3.4: Typical assembly kinetics with $\lambda = 0, S = e^2, \sigma = e^{-4}$ and $q = 51$. It shows $P(0,\tau)$ to be an (approximately) exponentially decaying function. The intermediate template occupancies, $P(10,\tau)$ and $P(30,\tau)$, show a peak when the assembly wave passes them. The fully filled template fraction, $P(q,\tau)$, exhibits a sigmoidal increase with the lag time determined by the time the assembly wave needs to reach $n = q$. The concentration of proteins relative to the equilibrium concentration, $y(\tau) = \frac{s(\tau)}{s_{eq}}$, is constant because $\lambda = 0$. However, for $\lambda > 0$ it typically (approximately) decays exponentially to unity.

of a few distribution values and $y(\tau)$ given for $\lambda = 0$. It shows that $P(0,\tau)$ is a monotonically decaying function, $P(q,\tau)$ has a sigmoidal shape with a lag time and $y(\tau) = \frac{s(\tau)}{s_{eq}}$ is constant. If $\lambda > 0$ will $y(\tau)$ not be a constant but typically decrease exponentially to unity. Furthermore, the values of $n = 10$ and $n = 30$ respectively show a maximum. This could be called an assembly wave.

With this typical behaviour of $P(0,\tau)$, $P(q,\tau)$ and $y(\tau)$ specified we consider how it changes for $\lambda > 0$. First thing to note is that the behaviour generally does not change but some peculiarities do arise. In figure 3.5 a phase diagram of peculiarities for $\lambda > 0$ is shown. The peculiarities we consider are: an overshoot in $P(q,\tau)$, an undershoot in $P(0,\tau)$ and an undershoot in $y(\tau)$. The first implies that the system forms too many fully formed templates and needs to dismantle some. The second says that too many templates are nucleated, whereas for the third too many proteins are used and need to be taken from templates again. By determining whether a peculiarity is present or not we use as criterion for the first $\frac{P(q,\tau)}{P_{eq}(q)} > 1.01$, for the second $|P_{min}(0) - P_{eq}(0)| > 0.05$ and for the third $y_{min} < 0.99$, where $P_{min}(0)$ is the minimum value of $P(0,\tau)$ and y_{min} that of $y(\tau)$ during the simulation. These criteria are necessary to distinguish between numerical errors, negligible peculiarities and true peculiarities. In the figure is in the (λ, S) plane depicted which peculiarities are present for a number of points. First, for $0 \leq \lambda < 0.7$ and $S \leq 1$ no peculiarities are observed. For $S \leq 1$ does negligible assembly take place while for $\lambda < 0.7$ apparently the excess of proteins is such that no peculiarities happen. With increasing λ an overshoot first occurs for $S = e^1$ until for $\lambda = 1$ it happens for all values of S. Furthermore, for $\lambda > 1.25$ no overshoot is observed. This clearly shows that the overshoot occurs around $\lambda = 1$. So proteins should neither be too scarce nor too abundant. The reason for this is not easily seen since $\lambda = 1$ is not a particularly special value as judged from the dynamical equations. The cause of the overshoot is probably caused by some kind of 'overshoot momentum' the system has. This momentum is arguably

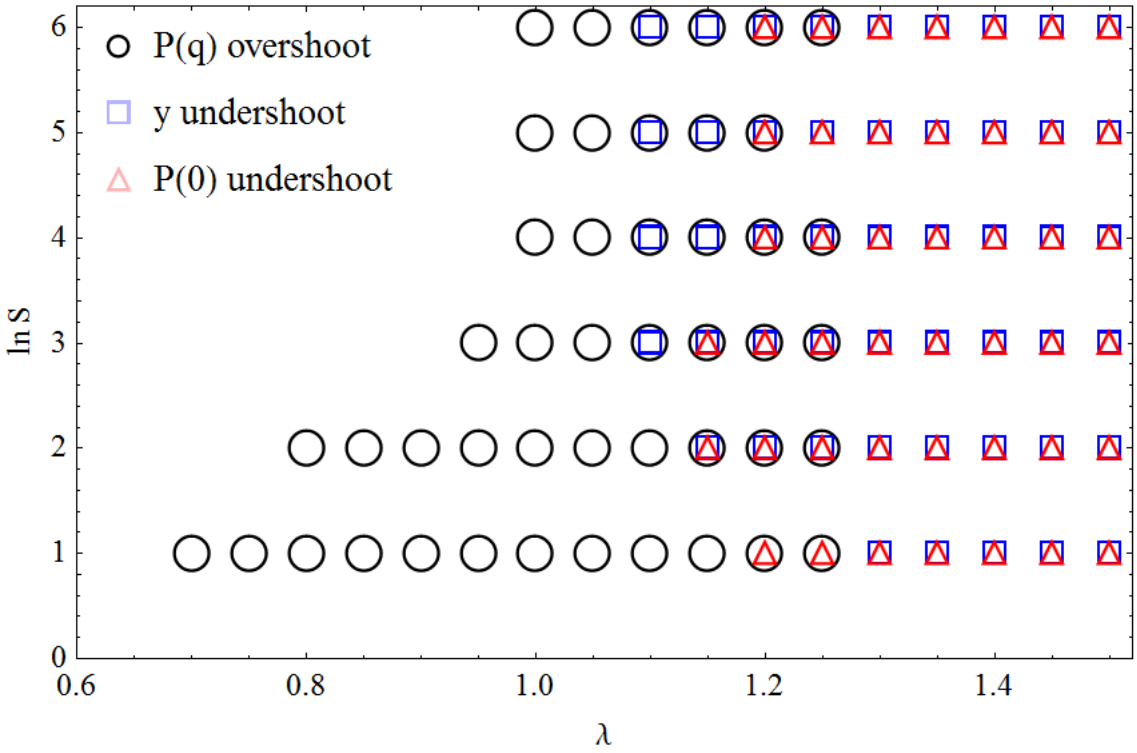

Figure 3.5: A phase diagram of peculiarities occuring in the dynamics for $\lambda > 0$ with $\sigma = e^{-4}$ and $q = 51$. An overshoot in $P(q, \tau)$ arises for λ values around unity and $S > 1$. For $\lambda > 1$, when a significant portion of templates remains uncovered, undershoots in $y(\tau)$ and $P(0, \tau)$ set in. Intuitively, the cause of these over- and undershoots is the system having a kind of inertia. For $\lambda > 0$ it acts for small times as if $\lambda = 0$ and has to correct for an over- or undershoot later on.

proportional to $y(0)$ for all forward rates in the dynamical equations are proportional to it at small times. The value $y(\tau)$ at $\tau = 0$ is important because the cause of the overshoot should lie at small times. Thus a greater $y(0)$ should give a greater momentum for overshooting. Furthermore, as $P_{eq}(q)$ becomes small for $\lambda > 1.25$ overshoots disappear. Therefore, this momentum should scale with $P_{eq}(q)$. Furthermore, the value of $P_{eq}(q)$ also should not be too large for if it goes to unity there is no possibility for an overshoot by probability conservation. Therefore, the momentum should also scale as $1 - P_{eq}(q)$. This gives rise to the overshoot momentum function $g(S, \lambda, \sigma, q) = y(0)P_{eq}(q)(1 - P_{eq}(q))$ which is shown in figure 3.6.

Next, for $\lambda > 1.05$ undershoots of both $P(0, \tau)$ and $y(\tau)$ occur. This clearly shows that it only occurs in case of a shortage of proteins. In this case too many templates get nucleated. The possibility of this to happen can be seen as follows. As noted above, is for $\lambda > 1$ the amount of uncovered templates not negligible as compared to the fraction which is covered. Therefore, too many templates can get nucleated in the assembly which later have to be disassembled again.

The cause of the over- and undershoots are, qualitatively speaking, because for small times the system behaves as if $\lambda = 0$. Afterwards, at later times it has to correct for the excessive binding of proteins. To understand more quantitatively where the overshoot and undershoots arise the conditions for them to occur should be probed in a more detailed manner. However, this is left for future research.

3.4 Comparison with experiments

As noted in the introduction we have experimental data available with which we can test the validity of our model. The data comprises the self-assembly of artificial dsDNA with artificial capsid proteins. In total we have three sets of data: SQ10, SQ14 and NP3. The names SQ10 and SQ14 refer to the kind of capsid proteins used. The proteins of SQ14 have a stronger protein-protein interaction. NP3 refers to

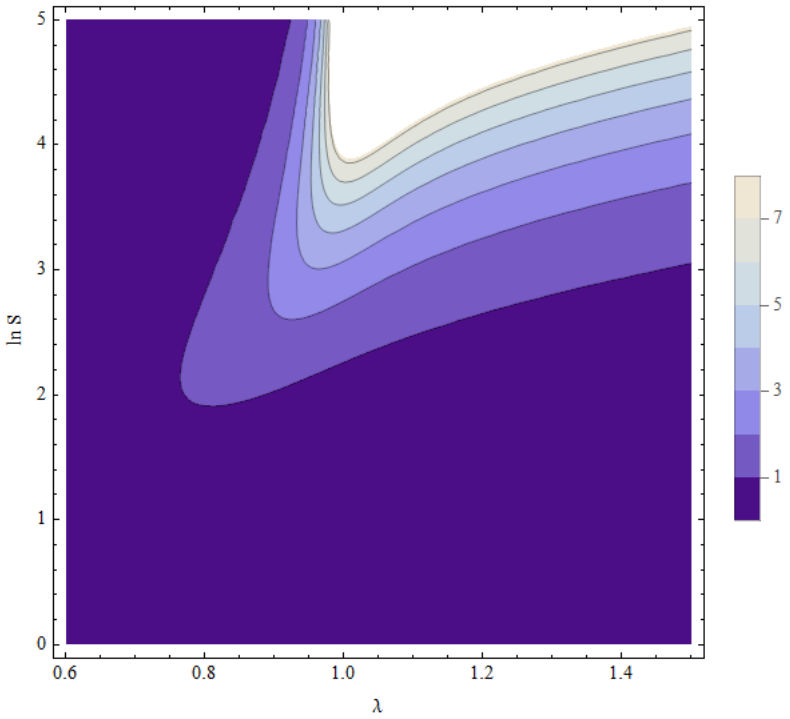

Figure 3.6: A contourplot of the overshoot momentum function, $g(S, \lambda, \sigma, q)$, for $\sigma = e^{-4}$ and $q = 51$. The function gives a rough indication of the appearance of overshoots in $P(q, \tau)$. It approximately agrees with figure 3.5 for $\lambda < 1$. However, for $\lambda > 1$ it fails to explain why overshoots disappear when increasing λ.

the ratio of available positive chargers on all the capsid proteins together and the number of available negative charges on all the dsDNA molecules. Each data set comprises of measurements of the lengths of the capsids at different times. From this a binned distribution of the templates can be derived. Below we will first describe how the data was generated. Afterwards, we will show how we fitted our model to the data and what results it generated.

3.4.1 Data acquisition

The acquisition of the three data sets introduced above has the following in common. The concentration of dsDNA molecules, or simply templates, was $c_T = 1\,\mu\mathrm{g(mL)}^{-1}$, or $0.65\,\mathrm{nM}$. Each template was built of 2500 base pairs. The buffer used was sodium phosphate of 10 mM at pH 7.4. To avoid the formation of disulfide bridges - the capsid proteins carry a cysteine - was dichloordiphenyltrichloroethane (DDT) added in 0.1 mM concentration. The experiments were conducted at room temperature $\approx 20\,^{\circ}\mathrm{C}$. Finally, the number of binding sites per template was identical for the following reason. Every template has $2 \times 2500 = 5000$ bases which all have charge -1. Furthermore, a capsid protein has charge +12. Because the charge of the template needs to be neutralised by that of the proteins for a full capsid we have $q = 5000/12 \approx 417$.

Next, the measurement procedure was identical for all data sets. To make a measurement of the distribution of the templates at a given time, a sample of a few micro liters was taken from the solution. This was put on a silica surface for 2-3 minutes. Afterwards, the sample was rinsed with $1\,\mathrm{mL}$ demiwater such that the unbound capsids proteins were removed. Subsequently, it was dried with nitrogen steam such that the assembly was frozen down. Finally, the lengths of the (partly) assembled capsids in the sample was measured. This measurement was done through atomic force microscopy (AFM). That is, the length of the partly assembled capsids was measured under the microscope.

The variable conditions among the different data sets was the concentration of capsid proteins, c_P, and

the kind of capsid protein used. For SQ10 and SQ14 the concentration of capsid proteins is on purpose almost the same such that the influence of the different capsid proteins on the assembly can be seen.

To make a fit one first needs to solve the dynamical equations. For this a number of parameters has to be determined. Some of these can be known by measurement while others have to be fitted. Below we will give an overview of these parameters and how there value can be found.

3.4.2 Parameter determination

In the previous section we outlined how the measurements were made. Here we will give an overview of how the parameters present in the dynamic equations can be found such that we can fit our model to the data.

The following parameters have to be determined: 1) λ, the ratio of the total number of available binding sites on the templates and the number of capsid proteins, 2) κ, the kinetic barrier for nucleation which we set to 1 since the effect of nucleation is already captured in σ, 3) S, the relative concentration of the total number of capsid proteins in the solution which is determined by s_{eq}, σ and q, 4) s_{eq}, the relative concentration of unbound proteins in equilibrium, 5) σ, the Boltzmann factor of nucleation, 6) k_+, the forward rate constant, and 7) t_0: an offset time which we add to the time of every measurement in order to account for the 2-3 minutes waiting time, as well as to account for the processes which disturb the sample during rinsing and drying. Now we will review all these parameters on how they can be determined.

The stoichiometric ratio λ was different for every experiment for it can be calculated as

$$\lambda = \frac{5000 \times c_T}{12 \times c_P} = \frac{q c_T}{c_P} = \frac{q \rho_T}{\phi_P},$$

with, respectively, c_T and c_P the concentration of templates and total concentration of proteins. These can have arbitrary units. We used the definition of q as explained in the paragraph above and we could write λ in terms of the dimensionless concentrations because both concentrations are made dimensionless by multiplying with the characteristic volume scale of the system V_{mol}. Concluding, we can determine λ directly from the measurement of c_P and c_T.

Furthermore, to calculate the dimensionless density $\phi_P = \frac{N_{P,tot}}{V} V_{mol}$ we need to know what the typical volume scale is of the proteins in the solution. This is quite hard to calculate, but luckily we do not need to know it. If we assume that the capsid proteins have the same volume scale as the solvent, then we can say

$$\phi_P = \frac{N_{P,tot} V_{mol}}{(N_{P,tot} + N_{solvent}) V_{mol}} \approx \frac{N_{P,tot}}{N_{solvent}} = \frac{c_P}{c_{solvent}}.$$

For water molecules $c_{solvent} = 55.6\text{M}$ and the protein concentrations we work with are of the order of a few μM, so the approximation is justified.

The value of S can be found directly from mass conservation

$$S = \frac{s_{eq}}{1 - \lambda \langle \theta \rangle (s_{eq}, \sigma, q)},$$

since it is totally dependent on the other parameters. By combining S and ϕ_P the critical density and a combination of the relevant energies can be found. Since $\phi_c = \frac{S}{\phi_P} = e^{\epsilon + g}$ we can find what $\epsilon + g$ is.

The value of s_{eq} can be determined in the following way. The zipper equilibrium distribution is solely determined by s_{eq}, σ and q. Therefore, by considering a measurement at very late time and by fitting it to a zipper equilibrium distribution one can find information on s_{eq}. However, the distribution of equation (3.3) can not be used because the uncovered templates have not been counted in the measurements. To correct for this we consider

$$\frac{\sum_{n=1}^{q} P_{eq}(n)}{1 - P_{eq}(0)} = 1,$$

where we used $\sum_{n=0}^{q} P_{eq}(n) = 1$ which is implied by equation (3.3). Thus, if we define $P_{eq,cor}(n) = \frac{P_{eq}(n)}{1 - P_{eq}(0)}$ for $1 \leq n \leq q$, we have a normalised corrected distribution which can be written explicitly as

$P_{eq,cor}(n) = \frac{s_{eq}^n}{s_{eq}\frac{1-s_{eq}^q}{1-s_{eq}}}$. So, σ has no influence on the corrected distribution and we are free to determine s_{eq}. The independence of $P_{eq,cor}(n)$ on σ is to be expected because σ only influences which part of the templates is nucleated in equilibrium. Since $P_{eq,cor}(n)$ gives the distribution of nucleated templates only it should not depend on σ.

Finally, σ, k_+ and t_0 can only be found through trial and error, educated guesses and choosing the relevant features of the distribution to fix a fit at. Therefore, we will refer to s_{eq}, σ, k_+ and t_0 as the fitting parameters. Since there are three fitting parameters, three features of the distribution can always be accounted for. The value of the fit is in how well it describes other features.

3.4.3 Fits

In the previous paragraph we showed what parameter values can be readily determined from a data set and what are the fitting parameters. In the following subparagraphs we will give the fits made of the three data sets which are at our disposal. For SQ10 we will show in full detail which procedure we followed. This will not be done for SQ14, for the procedure is almost exactly the same. Therefore, we only mention what differs from SQ10. Also for NP3 do we only note what differs from SQ10.

SQ10

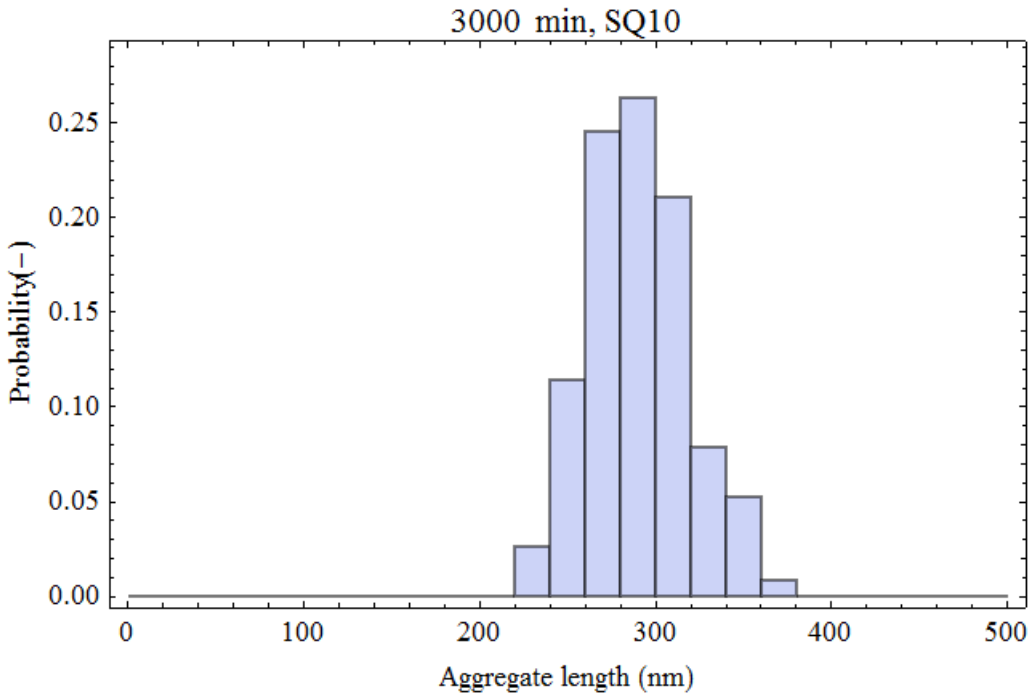

Figure 3.7: Visualisation of the measurement at 3000 minutes of the SQ10 data set. The measured aggregate lengths have been binned. The distribution is for increasing length first exponentially increasing. Afterwards, it exponentially decreases.

For SQ10 measurements are available at 10 different times: 2, 8, 15, 25, 85, 180, 360, 480, 1500 and 3000 minutes. On purpose there is a higher density of measurements at short times because we expect non-trivial behaviour to occur. A measurement at a given time consists of a list of the lengths of all measured capsids. The number of measured capsid aggregates per measurement is 92 - 130, so the uncertainty in each time measurement is quite large. To analyse the data we bin the lengths for every measurement. The uncertainty per bin is large for the number of aggregates per bin is of the order 10. The

binning enables us to plot the fraction of aggregates versus the length of the aggregate. This is shown in figure 3.7 for the measurement at 3000 minutes.

The first step in the fit is to determine s_{eq}. To do so we consider the measurement of 3000 minutes because we expect the system to be in equilibrium by then. In figure 3.7 we see first for increasing aggregate length exponentially increasing probability. Afterwards, there is exponential decay. By the law of mass action only exponential increase would be expected. Therefore, we assume that either the templates in the experiment are not mono disperse, or that some fully encapsulated templates have some sort of micelles structure attached to it such that they seem to have a larger length. The way to get around this ambiguity is to put all data points with a length longer than 300 nm - the length where exponential increase stops - in the bin which ends at 300 nm. In this way we enforce exponentially increasing behaviour.

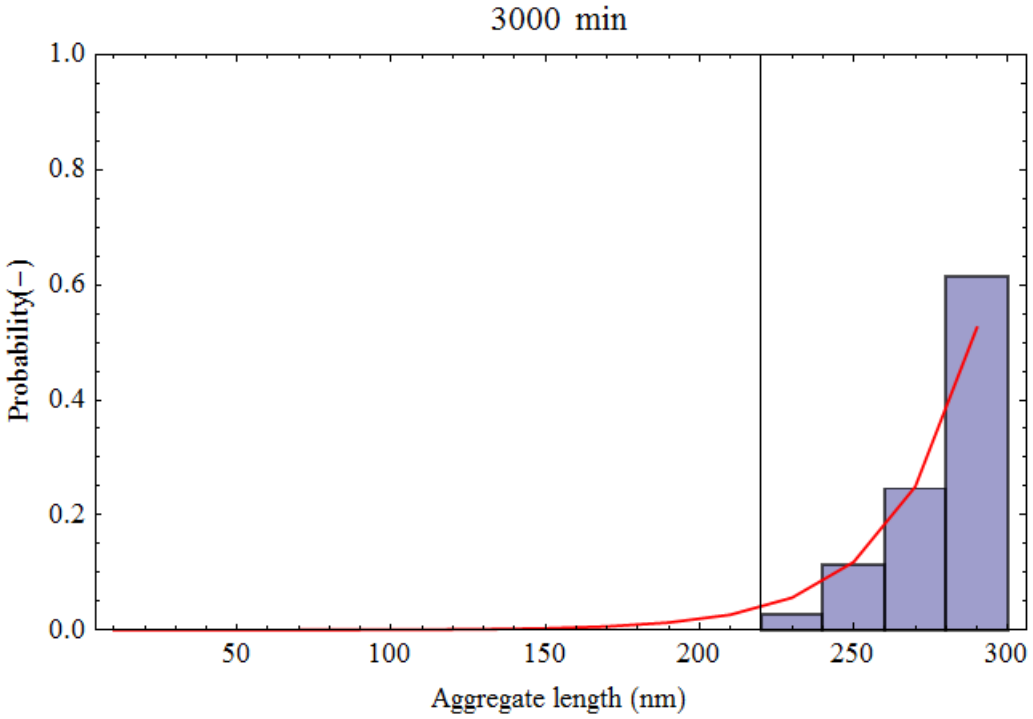

Figure 3.8: A visualisation of the exponentially enforced measurement at 3000 minutes together with an equilibrium distribution fit. The measurement has been made exponentially increasing by putting all data points of figure 3.7 which are larger than 300 nm in the bin which ends at 300 nm. The equilibrium distribution is fitted ath this time since the system is expected to be in equilibrium. The fit is focused on the two middle bins for they are assumed to be most reliable.

Next, we can fit the 3000 minutes measurement to the corrected distribution, $P_{eq,cor}(n)$, in the following way. We assume that a capsid of 300 nm is a fully formed capsid with $n = q = 417$. Then, to bin $P_{eq,cor}(n)$ we, for example, have for the probability of first bin, $P_{bin,1}$ that $P_{bin,1} = \sum_{n=1}^{\frac{b}{L}q} P_{eq,cor}(n)$. Here, $\mathrm{b} = 20\mathrm{nm}$ is the breadth of a bin and $\mathrm{L} = 300\mathrm{nm}$ is the maximum length of the aggregates. It poses no problem that $\frac{b}{L}q$ is not an integer since the sum then simply continues up to the last integer. In this way we find the fit shown in figure 3.8, which implies $s_{eq} = 1.027$.

With s_{eq} found through the experimental equilibrium distribution we are are ready to fit the dynamics. From the densities we have $\lambda = 0.1006$. To make the fit we choose $\sigma = 0.004, k_+ = 200$ and $\mathrm{t}_0 = 7\mathrm{min}$ to obtain the fit shown in figure 3.9. The way we chose this parameter requires some explanation though. Since we have three fitting parameters we focused on three features of the dynamical distribution to make the fit: the peak of the measurement of t = 15 minutes, the peak at t = 2 minutes, and the breadth of the distribution at 2,8 and 15 minutes. We will explain below how we controlled these three features. We optimised the fit to have the least discrepancy between the data and the model with regard to the three

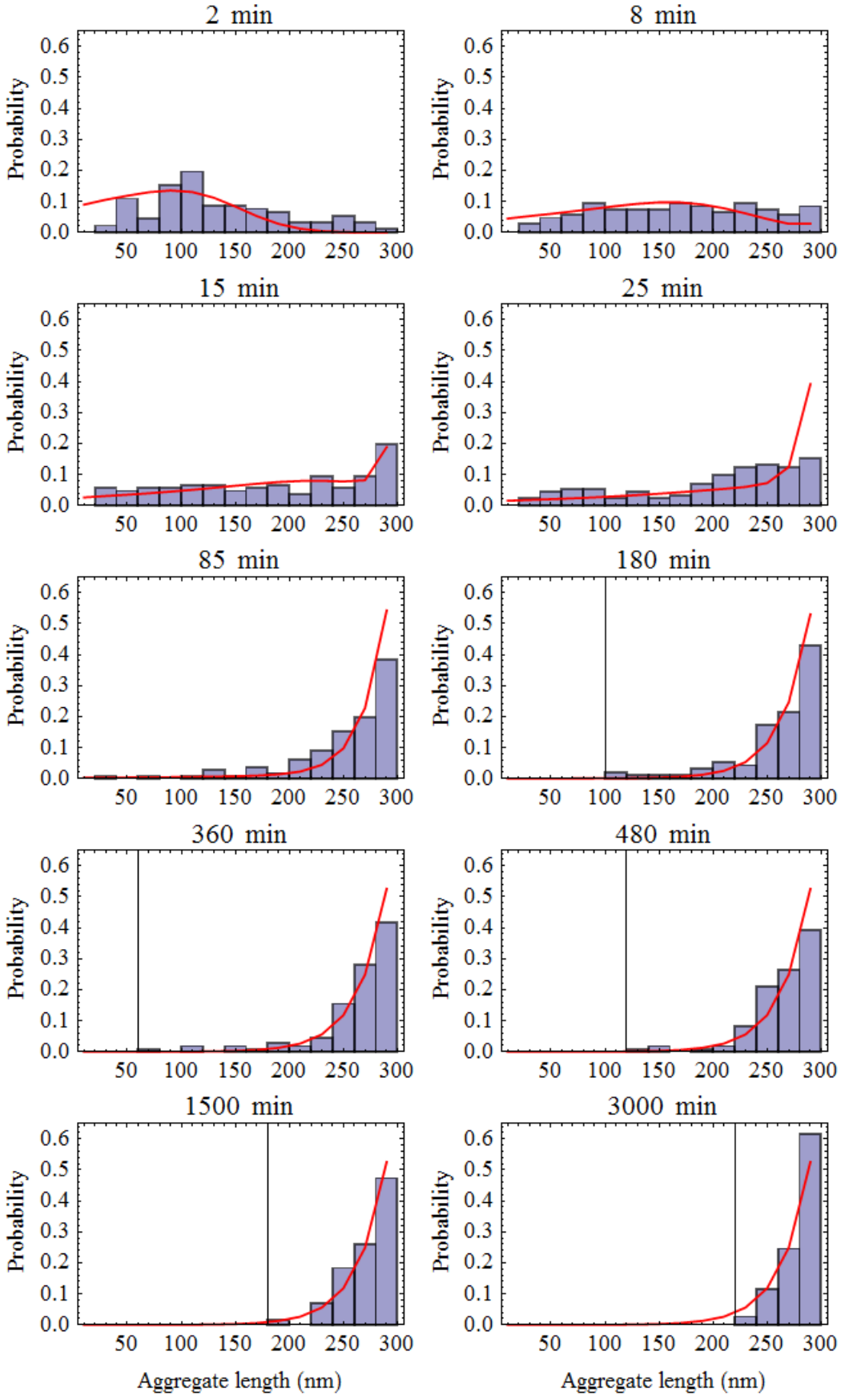

Figure 3.9: A fit of the SQ10 data set. The fitting parameters are $s_{eq} = 1.027$, $\lambda = 0.1006$, $\sigma = 0.004$, $k_+ = 200\text{min}^{-1}$, $\text{t}_0 = 7\text{min}$ and $S = 1.131$. The four fitting parameters have been determined by focusing on the position of the peak at 2 and 15 minutes, the breadth of the distribution and the form of the equilibrium distribution.

fitting features. We judged the optimal point in parameter space by eye, so we used no least square method or whatsoever. The reason for this is that we have a large uncertainty in the data for every bin. The reason for this is that for one time measurement we have between 92-130 lengths measured. This implies that per bin we have on average only 10 measured lengths. Therefore we have a very large uncertainty. Another reason is that we are dealing with a biological system described by a very crude model while in reality such a system is highly complex. Another aspect of the fitting is that in the measurements the uncovered templates - the aggregates of length zero - were not measured. Therefore, in the same way we defined $P_{eq,cor}(n)$ we have to correct the calculated probabilities by a factor $\frac{1}{1-P(0,\tau)}$. Now we will explain in detail how we pinned down σ, k_+ and t_0.

To start, we made a change of variables to $\sigma, \tau_{2\min}$ and t_0, where $\tau_{2\min} = (2 + t_0)k_+$, such that we could easily fit the peak at 2 minutes. We fitted the peak at 2 minutes by imposing that the fit showed the same asymmetry as the data does. This asymmetry comes from the fact that we see the assembly wave is at low aggregate length values.
In general, σ influences the breadth of the distribution and also the speed of the assembly wave travelling from 0 to 300 nm (at 2,8,15 minutes). By accounting for these two properties we could pin down σ.
Finally, we find the τ-values of the measurement at t = 8,15,etc. min in the following way

$$\tau_t = \frac{t + t_0}{2 + t_0}\tau_{2\min} = (t + t_0)k_+.$$

By tuning the value of t_0 we could fit the peak at 15 minutes.

To see how well this fit is, we evaluate the fits of all times we did not use to fit the parameters. For 8 minutes the agreement is very good. At 25 minutes it is good for aggregates which are not fully covered but bad for the last bin. With 85, 180, 360, 480 and 1500 min the agreement is good. Therefore, we would say that the fit does get the timescales correct and gives quite accurate predictions. of the distribution. With the fitting parameters found we obtain $S = 1.131$. We have that $\mathrm{c_P} = 2692\mathrm{nM}$ and thus, as explained in the paragraph on parameter determination, $\phi_P = 4.84 \times 10^{-8}$ and $\epsilon + g = -16.97$ in units of k_BT. Furthermore, from $\sigma = e^{-h+\epsilon}$ we find $-h + \epsilon = -5.521$.

SQ14

For the SQ14 data set are measurements made at t = 2, 8, 15, 30, 64, 373, 1440 and 2925 min. The fitting can be done in exactly the same way as for SQ10. The only difference is that as fitting features the peak of the measurements at 2 and 15 minutes was used. The peak at 15 minutes pinned down σ and t_0. With $s_{eq} = 1.038$, $\mathrm{k_+} = 129\mathrm{min}^{-1}$, $\sigma = 0.05$ and $\mathrm{t_0} = 12\mathrm{min}$ we obtain the fit from figure 3.10. These parameters give with $\lambda = 0.134$ that $S = 1.187$ and with $\mathrm{c_P} = 2016\mathrm{M}$ we have $\phi_P = 3.63 \times 10^{-5}$. This gives through the critical concentration $\epsilon + g = -10.40$. Furthermore, from $\sigma = e^{-h+\epsilon}$ we find $-h + \epsilon = -3.00$.

NP3

For the NP3 data set there are two aspects to consider. One is how the dynamical equations fit with $\lambda > 0$. The other is how with $\lambda = 0$ a fit can be made. The reason for the last aspect is that with the model of Kraft et al [7] this data set has ben fit, as noted in the paper of Hernandeze-Garcia et al. [21]. Therefore, it is interesting to see how the parameter values change by going fitting with $\lambda = 0$ instead of $\lambda > 0$.

The data set has six measurements at t = 10, 60, 350, 1485, 2880 and 7440 minutes. There were micelles structures in the solution which were also measured. Therefore, for the last four measurements all lengths shorter than 100 nm were excluded. At these times they disturbed the exponential increasing behaviour and were therefore identified as micelles. For $\lambda = 0.324$ we found the fit shown in figure 3.11 with $\sigma = 0.005$, $s_{eq} = 1.016$, $\mathrm{k_+} = 33\mathrm{min}^{-1}$ and $\mathrm{t_0} = 5\mathrm{min}$. From these values we obtain $S = 1.40$, $\epsilon + g = -18.36$ and $-h + \epsilon = -5.30$. These energy values are quite close to the values calculated by Hernandez-Garcia [21]. In making this fit we used as fitting features the form and position of the peak at 10 min and the absolute increase in the last three bins between 140 and 235 nm. That the fit lies under the

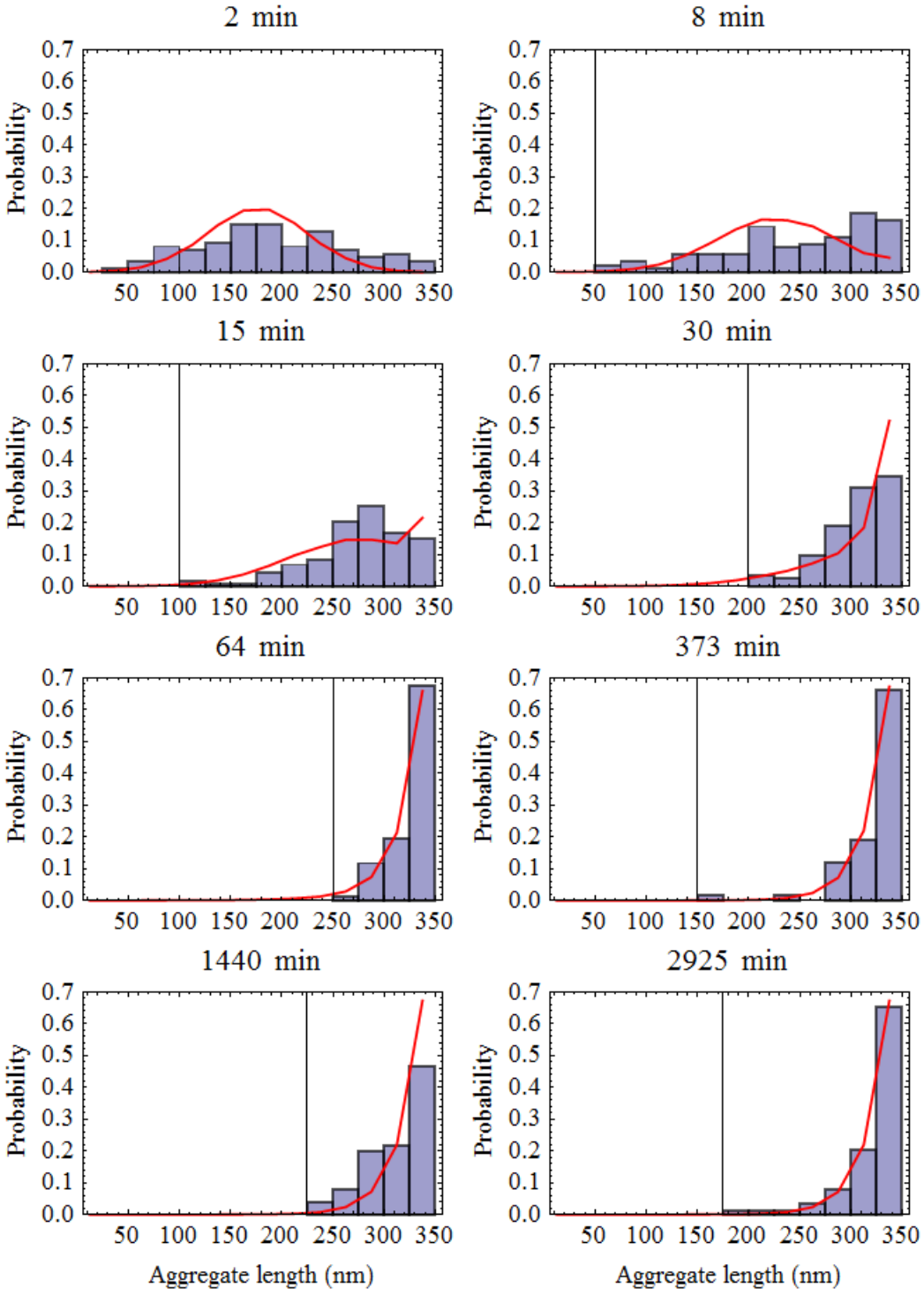

Figure 3.10: A fit of the SQ14 data set. The fitting parameters are $s_{eq} = 1.038$, $\lambda = 0.134$, $\sigma = 0.05$, $k_+ = 129\text{min}^{-1}$, $\text{t}_0 = 12\text{min}$ and $S = 1.187$. The four fitting parameters have been determined by focusing on the position of the peak at 2 and 15 minutes, the breadth of the distribution and the form of the equilibrium distribution.

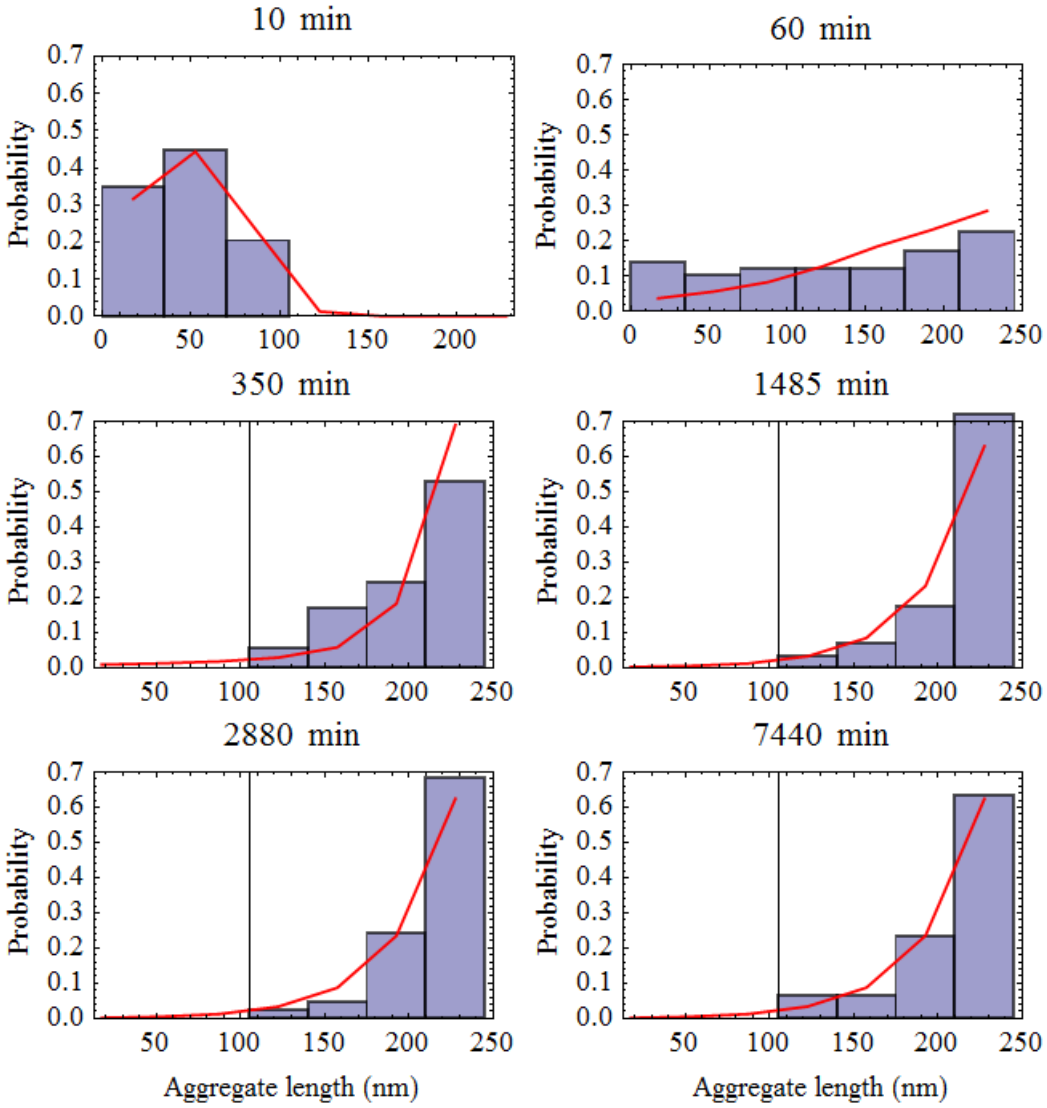

Figure 3.11: A fit of the NP3 data set. The fitting parameters are $s_{eq} = 1.016$, $\lambda = 0.324$, $\sigma = 0.005$, $k_+ = 33\text{min}^{-1}$, $\text{t}_0 = 5\text{min}$ and $S = 1.40$. The four fitting parameters have been determined by focusing on the position of the peak at 10 and 60 minutes, the increase of the distribution at 60 minutes for the last three bins and the form of the equilibrium distribution.

data of first bin(s) at 10 and 60 minutes is not a problem for there are probably already some micelles structures present.

If we take $\lambda = 0$ instead, we obtain the fit shown in figure 3.12 which has been constructed after the same features as for $\lambda = 0.324$. It satisfies $s_{eq} = S = 1.016$, $\text{k}_+ = 220\text{min}^{-1}$, $\text{t}_0 = 12\text{min}$, $\sigma = 0.03$ and gives $\epsilon + g = -18.04$ and $-h + \epsilon = -3.51$.

The calculated energy values for $\lambda = 0$ and $\lambda = 0.324$ are close to each other. There is only a few k_BT difference in $-h + \epsilon$. Furthermore, the $\lambda = 0$ fit has a much greater rate constant than the $\lambda > 0$ fit. In the former case an infinite supply of proteins is assumed. Therefore, intuitively it is plausible that this fit gives rise to a larger k_+ since the effect of a larger concentration can be understood as effectively heightening the rate constant for the forward reaction to happen.

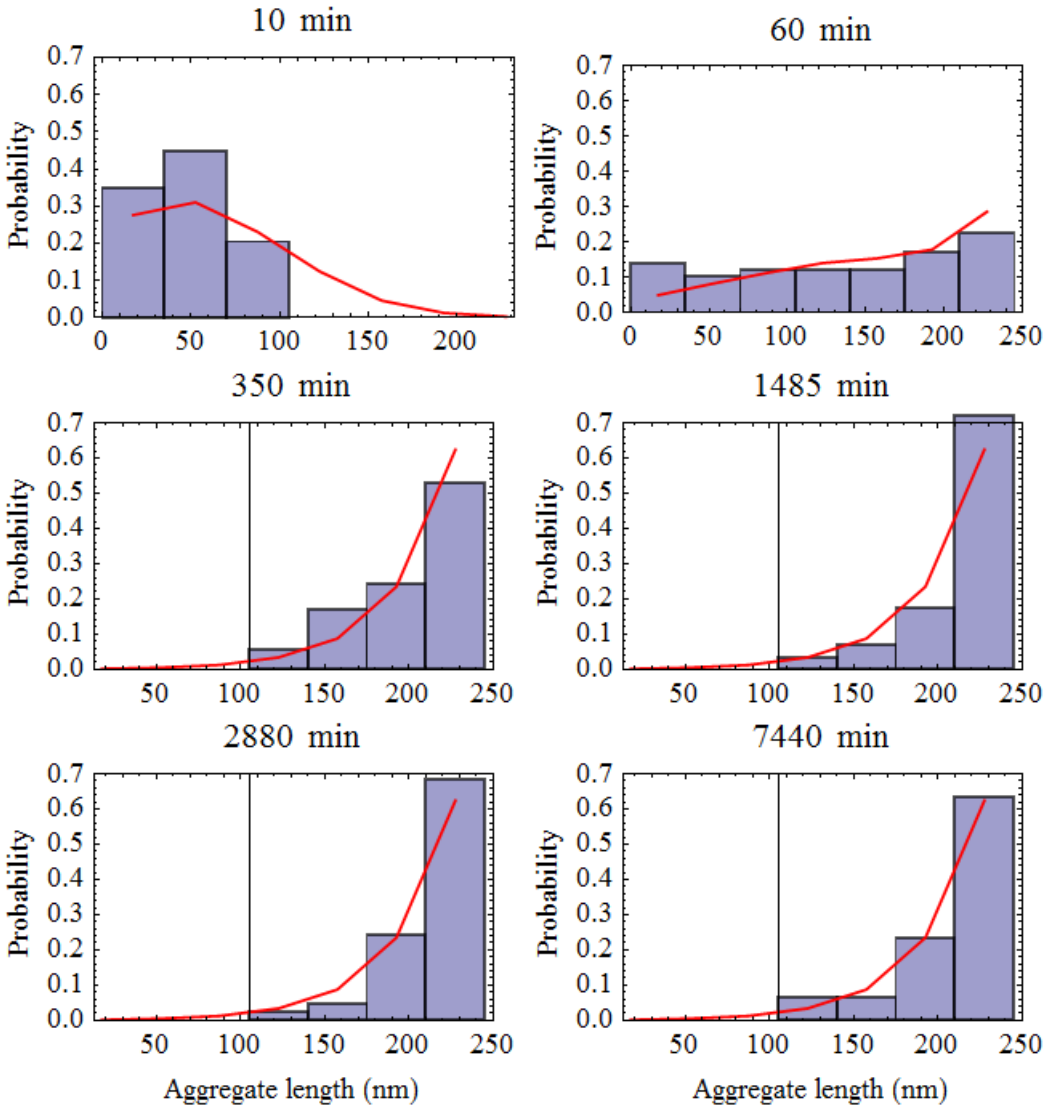

Figure 3.12: A fit of the NP3 data set with $\lambda = 0$. The fitting parameters are $s_{eq} = S = 1.016$, $\sigma = 0.03$, $k_+ = 220\text{min}^{-1}$ and $\text{t}_0 = 12\text{min}$. The four fitting parameters have been determined by focusing on the position of the peak at 10 and 60 minutes, the increase of the distribution at 60 minutes for the last three bins and the form of the equilibrium distribution.

Chapter 4

Ising-S model

As outlined in the introduction and theory we introduce the Ising-S model to account for multiple assembly signals and the effect of entropy in self-assembly. One could say that this model is in essence a one-dimensional Ising model with the assembly signals causing impurities. We will solely focus on the equilibrium properties of the assembly system within this model, the dynamics are out of the scope of this thesis. Below we will first calculate the general semi-grand partition function. Afterwards, the reduction of the Ising-$\{1\}$ model to the zipper model is shown. Next, the equilibrium distribution is calculated exactly and thereafter approximated in the high energy barrier limit to show it comprises zipper like behaviour.

4.1 General partition function

To find the general partition function of the Ising-S model, we first consider the partition function of the Ising-$\emptyset$ model, that is, a one-dimensional Ising model without impurities. This will be done in such a way that the one can, in principle, straightforwardly calculate the partition function in the general case. From the theory of section 2.4 we have for the Ising-$\emptyset$ model with free boundary conditions

$$\Xi^{\emptyset} = \sum_{\{n_i\}} \exp\Big[(-h'+\epsilon)\sum_{i=0}^{q-1}(1-n_i)n_{i+1} + (\mu_P - g - \epsilon)n\Big], \tag{4.1}$$

where we did a little bit of rewriting, defined $n_0 \equiv 0$ and recall that $n \equiv \sum_{i=1}^{q} n_i$. Furthermore, $h' > 0$ is the free energy barrier for nucleation, $\epsilon < 0$ is the free energy protein-protein interaction, $g < 0$ is the free energy template-protein interaction and μ_P is the chemical potential of the unbdound proteins. To proceed we introduce the transfer matrix $T_{n_i n_{i+1}} = e^{(-h+\epsilon)n_i n_{i+1} + (\mu_P - g - \epsilon)n_i}$ such that

$$\begin{aligned} \Xi^{\emptyset} &= \sum_{n_q=0,1} (T^q)_{n_0 n_q}, \\ \mathbf{T} &= \begin{pmatrix} 1 & \sigma' s_{eq} \\ 1 & s_{eq} \end{pmatrix}, \end{aligned} \tag{4.2}$$

where $s_{eq} \equiv e^{\mu_P - g - \epsilon}$ is the Boltzmann factor for having a protein cooperatively bound and $\sigma' \equiv e^{-h'+\epsilon}$ is the Boltzmann factor for a nucleation. Equation (4.2) shows that the semi-grand partition function consists of only two distinct terms. This reflects that there are only two distinct boundary conditions. Namely, at one of the ends we have a protein or we do not have a protein. In the matrix corresponds $n_i = 0$ to the first row or column and $n_i = 1$ to the second row or column. Next, we can calculate $\mathbf{T}^q$ as

$$\mathbf{T}^q = c_3 \begin{pmatrix} c_2\lambda_+^q + c_1\lambda_-^q & \sigma' s_{eq}(\lambda_+^q - \lambda_-^q) \\ \lambda_+^q - \lambda_-^q & c_1\lambda_+^q + c_2\lambda_-^q \end{pmatrix}, \tag{4.3}$$

where $c_1 \equiv \lambda_+ - 1$, $c_2 \equiv 1 - \lambda_-$ and $c_3 \equiv \frac{1}{\lambda_+ - \lambda_-}$ with $\lambda_\pm = \frac{1}{2}(1 + s_{eq} \pm \zeta)$ the eigenvalues of the transfer matrix and $\zeta \equiv \Big((1 - s_{eq})^2 + 4\sigma' s_{eq}\Big)^{\frac{1}{2}}$. By using the defining relation of the eigenvalues we find

$$\Xi^{\emptyset} = \frac{1}{\lambda_+ - \lambda_-}\Big((1 - \lambda_-)\lambda_+^{q+1} + (\lambda_+ - 1)\lambda_-^{q+1}\Big). \tag{4.4}$$

Next, we show that the matrix $\mathbf{T}^q$ can be used to find the partition function in general. To do so, it is instructive to first find the partition function for the Ising-$\{1\}$ model. This model has the following expression for $\Xi^{\{1\}}$

$$\begin{aligned}\Xi^{\{1\}} &= \sum_{\{n_i\}} \exp\Big[(-h' + \epsilon)\sum_{i=0}^{q-1}(1 - n_i)n_{i+1} + (\mu_P - g - \epsilon)n - (h_i - h')n_1\Big], \\ &= \sum_{n_1, n_q} (TC^{(1)})_{n_0 n_1}(T^{q-1})_{n_1 n_q},\end{aligned}$$

where we introduced the matrix $C^{(p_i)}_{n_i n_j} = \delta_{n_i n_j} e^{-(h_{p_i} - h')n_i} \equiv \delta_{n_i n_j}\chi_{p_i}^{n_i}$ for the i-th special site at position p_i. The partition sum can be written down in an expression consisting of four terms. This is to be expected for the symmetry is broken and therefore we now have four relevant boundary conditions

$$\Xi^{\{1\}} = T_{00}^{q-1} + T_{01}^{q-1} + \sigma' s_{eq}\chi_1(T_{10}^{q-1} + T_{11}^{q-1}). \tag{4.5}$$

The expression shows that the template can be thought of as a template with $q - 1$ binding sites connected with a single assembly site. The first two terms correspond to the states where the assembly site is not occupied and the last two to the states where the assembly signal is occupied. If the energetic advantage of binding at an assembly signal goes to infinity the last two terms dominate. This expression can be generalized for one special site at any position $p_1 \in P \equiv \{1, 2, ..., q\}$ in the following way

$$\Xi^{\{p_1\}} = \sum_{n_{p_1}, n_q} (T^{p_1}C^{(p_1)})_{n_0 n_{p_1}}(T^{q-p_1})_{n_{p_1} n_q}. \tag{4.6}$$

Subsequently, the general semi-grand partition function with $p_i, m \in P$ can be written as

$$\Xi^{S} = \sum_{\{n_i\}_{i\in S} \cup \{n_q\}} (T^{p_1}C^{(p_1)})_{n_0 n_{p_1}}(T^{p_2 - p_1}C^{(p_2)})_{n_{p_1} n_{p_2}} \cdot ... \cdot (T^{q-p_m})_{n_{p_m} n_q}, \tag{4.7}$$

where $S \equiv \{p_1, p_2, ..., p_m\}$ is the set of assembly signal positions.

Below we show how one can calculate the partition function with the cluster expansion such that the equilibrium distribution can be calculated. However, this does not give the partition function in closed form. Therefore is the general expression of the partition function useful. Also, one can easily calculate $\langle\theta\rangle$ from the partition function. In fact, one should be able to find the equilibrium distribution by expanding the partition function in a power series of s_{eq}. This should give a polynomial of order q where a term of order n gives the probability of having n proteins bound. Nevertheless, this will be left for future research.

In the next section the reduction of $\Xi^{\{1\}}$ to the partition function of the zipper model will be shown. This shows that the zipper model is contained within the Ising-S model.

4.2 Reduction to Zipper model

The Ising-S model contains the same parameters and the same physical processes as the zipper model. Therefore, one could suspect some connection between the two models. Furthermore, the one-dimensional Ising model exhibits a certain correlation length which gives the approximate size of a cluster with the same sign. Since the Ising-S model is a Ising like model it should have that characteristic either. In addition

to that, the zipper model is in essence protein cluster of variable length. This points to a correspondence between the Ising-$\{1\}$ and the zipper model in the limit of large correlation length. This only holds for the Ising-$\{1\}$ model for the zipper model has only one assembly signal at the end of the template. The parameter which governs the correlation is $\sigma' = e^{-h'+\epsilon}$. If it goes to zero the energy barrier is high, so no nucleation at normal sites is to be expected, and/or the protein-protein interaction is very strong which favours cluster forming, i.e., a large correlation length. Below we will show that in the $\sigma' \to 0$ limit and under further appropriate conditions the Ising-$\{1\}$ indeed reduces to the zipper model.

From equation (4.3) and (4.5) we find, using the defining relation of the eigenvalues, that the semi-grand partition sum of the Ising-$\{1\}$ model is

$$\Xi^{\{1\}}(\sigma') = \frac{1}{\lambda_+ - \lambda_-}\Big([s_{eq}\sigma'\chi_1 + 1 - \lambda_-]\lambda_+^q + [-s_{eq}\sigma'\chi_1 + \lambda_+ - 1]\lambda_-^q\Big), \tag{4.8}$$

where $\lambda_\pm = \frac{1}{2}(1+s_{eq} \pm \zeta)$, $\zeta \equiv \Big((1-s_{eq})^2 + 4\sigma' s_{eq}\Big)^{\frac{1}{2}}$, $\sigma' \equiv e^{-h'+\epsilon}$, $\chi_1 \equiv e^{-(h_1-h')}$, $s_{eq} \equiv e^{-\epsilon-g+\mu_p}$. To take the limit of $\sigma' \to 0$ we Taylor expand around $\sigma' = 0$. For $\sigma' = 0$ we have $\zeta \to \pm(1 - s_{eq})$ and one may pick any of the two signs because $\Xi^{\{1\}}$ is invariant under $\zeta \to -\zeta$. We arbitrarily pick the plus sign, such $\lambda_+ \to 1$ and $\lambda_- \to s_{eq}$, and obtain for the zeroth order

$$\Xi^{\{1\}}(0) \quad = \quad 1. \tag{4.9}$$

This is to be expected because for infinite nucleation cost one would expect that the probability of having a nucleation is zero and thus is the only contribution to the partition function that of an uncovered template. Including the first order term we obtain

$$\begin{aligned}\Xi^{\{1\}}(\sigma') \quad &\approx \quad 1 + \chi_1\sigma' s_{eq}\frac{\frac{q}{\chi_1} + (1-s_{eq}^q)(1-\frac{1}{\chi_1(1-s_{eq})})}{1-s_{eq}} = 1 + \chi_1\sigma' s_{eq}\frac{\frac{q-1}{\chi_1} + 1 - s_{eq}^q + \frac{s_{eq}}{\chi_1}\frac{s_{eq}^{q-1}-1}{1-s_{eq}}}{1-s_{eq}}, \\ &= \quad 1 + \chi_1\sigma' s_{eq}\frac{1-s_{eq}^q}{1-s_{eq}} + \sigma' s_{eq}\Big(\frac{q-1}{1-s_{eq}} + s_{eq}\frac{s_{eq}^{q-1}-1}{(1-s_{eq})^2}\Big).\end{aligned} \tag{4.10}$$

The last form of the expression shows that the partition sum is composed of three terms. The first term gives the empty template contribution and is therefore equal to unity. The next gives the zipper states contribution because $\chi_1\sigma' = \sigma$ and therefore exactly corresponds to the zipper partition function from section 3.1. The third term corresponds to competitor states. In fact, the first order approximation is equal to the partition function of a self-competing system, Ξ_{sc}, which was introduced in section 2.3.1 and will be calculated in chapter five. This is to be expected for the following reason. The self-competition partition function contains all states with one nucleation: the zipper and the competitor states. Furthermore, every nucleation generates a factor σ'. Therefore, the first order approximation in σ' should give the zipper as well as the competitor states.

To see when the competitor states are negligible we can put several conditions on the parameters. By doing so we find

$$\Xi^{\{1\}}(\sigma') \approx 1 + \chi_1\sigma' s_{eq}\frac{1-s_{eq}^q}{1-s_{eq}}, \tag{4.11}$$

for the conditions $s_{eq} \neq 1$, $|1 - s_{eq}|\chi_1 \gg 1$ and $\chi_1 \gg q$. The second condition prevents the competitor states to dominate if s_{eq} is close to unity. It ensures that the assembly signal is strong enough to withstand the entropic attraction of the competitor states. The third condition ensures that a nucleation at a normal site is not favourable. This is the proof of the condition for the zipper model which was quoted in section 2.3: $\ln q \ll h' - h$, identifying $h = h_1$. For completeness, the condition $\chi_1 \gg q$ guarantees that the competitor states are negligible for $s_{eq} \to 1$.

This analysis shows that the Ising-$\{1\}$ model indeed reduces to the zipper model under appropriate conditions. Also, it shows that the first order approximation of $\Xi^{\{1\}}$ comprises the zipper model together with competitor states. That is, the first order approximation contains all states with one nucleation.

4.3 Probability distribution

In this section we calculate the equilibrium probability distribution, $P_{eq}(n)$, of the Ising-S model. To do so we introduce the cluster expansion in the footsteps of [24, 25]. The advantage of this method is that it gives for a state with a fixed number of protein clusters k and proteins n what its multiplicity is. That is, how many configurations exist with the same value of k and n. The Boltzmann weight of states we consider in the Ising-S model can be expressed in terms of k and n. Therefore, one can in principle calculate the partition function with the cluster expansion by adding up the Boltzmann factor of all states multiplied with their respective multiplicity.
In section 4.1 we already found the partition function. However, we could not write the partition function as a sum of the $n = 0$, $n = 1$, etc. contributions. This is possible with the cluster expansion because by fixing n and summing over k one can find the contribution of any n. Subsequently, one can write $\Xi^S = \sum_{n=0}^{q} W(n)$, with $W(n)$ the portion of the partition function having n proteins, and $P_{eq}(n) = \frac{W(n)}{\Xi^S}$. In the following we will first give a recipe to calculate the partition function with the cluster expansion in general. Afterwards, we give for some special cases a detailed form of $P_{eq}(n)$. Nevertheless, we do not succeed in obtaining a closed form expression of $P_{eq}(n)$.

4.3.1 Exact probability distribution

To find the probability distribution of the Ising-S model we will first determine the multiplicity factors for different boundary conditions. These multiplicity factors give for a fixed number of protein clusters k and number of proteins n how many configurations are possible. One can picture this possibility by imagining one protein cluster of size $n = q - 1$ and imagining it has two configurations which both have $k = 1$ and $n = q - 1$. We denote the multiplicity as $\Omega^{n_1 n_q}$ where n_1 and n_q are the occupation numbers of the boundaries and therefore specify the boundary conditions.

First, consider a protein at both boundaries. The multiplicity of a chain of q sites can be written as

$$\Omega^{11}(n,q) = \begin{cases} 0, & n = 0, n = 1 \\ \binom{n-1}{k-1}\binom{q-n-1}{k-2}, & 2 \leq n \leq q-1 \\ 1, & n = q \end{cases}, \tag{4.12}$$

where $q \geq 3$. Furthermore, the number of protein clusters satisfies $2 \leq k \leq k^{11}_{max} = \mathrm{Max}(\mathrm{Min}(n, q - n + 1), 2)$. For the boundary conditions $n_1 = 0, n_q = 1$ or $n_1 = 1, n_q = 0$ we find

$$\Omega^{10}(n,q) = \Omega^{01}(n,q) = \begin{cases} 0, & n = 0 \\ \binom{n-1}{k-1}\binom{q-n-1}{k-1}, & 1 \leq n \leq q-1 \\ 0, & n = q \end{cases}, \tag{4.13}$$

where the number of protein clusters is given by $1 \leq k \leq k^{10}_{max} = \mathrm{Max}(\mathrm{Min}(n, q - n), 1)$. Finally, $n_1 = n_q = 0$ gives

$$\Omega^{00}(n,q) = \begin{cases} 1, & n = 0 \\ \binom{n-1}{k-1}\binom{q-n-1}{k}, & 1 \leq n \leq q-2 \\ 0, & n = q-1, n = q \end{cases}, \tag{4.14}$$

where the number of spin up clusters satisfies $1 \leq k \leq k^{00}_{max} = \mathrm{Max}(\mathrm{Min}(n, q - n - 1), 1)$. The sum of the different multiplicities adds up to 2^q, provided $q \geq 3$. Of course, this is what to be expected for an Ising like model.

The next step is to rewrite the general partition function as derived in section 2.4 such that we can put it in the right form. With a little rewriting, as we did for the general partition function calculation, we obtain

$$\Xi^S(\mu_P) = \sum_{\{s_i\}} \exp\Big[(-h' + \epsilon)\sum_{i=0}^{q-1}(1 - n_i)n_{i+1} + (\mu_P - g - \epsilon)n - \sum_{i \in S}(h_i - h')n_i\Big]. \tag{4.15}$$

To handle the special sites we now imagine breaking up the template into parts which have at the ends special sites. Here we assume site 1 and q to be special no matter the value of h_1 and h_q for notational convenience. This implies that for $m \geq 2$ special sites we have $m-1$ parts. Therefore, we define the bare partition function $\Xi_{bare}^{n_{p_x},n_{p_y}}(q_{xy})$ to be the partition function over all sites between special site p_x and p_y. This partition function only depends on whether or not we have proteins occupying the special sites at the ends of the part. By using this bare partition function we can rewrite Ξ^S as

$$\begin{aligned}\Xi_{bare}^{n_{p_x},n_{p_y}}(q_{xy}) &\equiv \sum_{n_{p_x+1},\ldots,n_{p_y-1}} \exp\Big[(h'-\epsilon)\sum_{i=p_x}^{p_y-1} n_i n_{i+1} + (\mu_P - g - h')\sum_{i=p_x+1}^{p_y-1} n_i \\ &\quad -\frac{b_{p_x}}{2}n_{p_x} - \frac{b_{p_y}}{2}n_{p_y}\Big], \qquad (4.16)\\ \Xi^S(\mu_P) &= \sum_{\{n_i\}_{i\in S}} \Xi_{bare}^{n_1,n_{p_2}}(q_{12})\Xi_{bare}^{n_{p_2},n_{p_3}}(q_{23})\cdot\ldots\cdot\Xi_{bare}^{n_{p_{m-1}},n_q}(q_{m-1,m}), \qquad (4.17)\end{aligned}$$

where we defined $p_y - p_x + 1 \equiv q_{xy}$, $b_{p_1} \equiv b_1 = 2(h_1 - \mu_P + g)$, $b_{p_m} \equiv b_q = 2(h_q - \mu_P + g)$ and $b_i = h_i - \mu_P + g$ for $i \in S/\{1,q\}$ and $p_x, p_y \in P \equiv \{1,2,...,q\}$. In order to find the general partition function we first calculate the bare partition function with the cluster expansion such that we can calculate the weight functions. We have to calculate the bare partition function for all values of n_{p_x} and n_{p_y}, that is, for all boundary conditions. To do this we may use equation (4.12) up to and including (4.14). This gives for the part of the template, having n proteins on it, from p_x to p_y

$$\begin{aligned}\Xi_{bare}^{n_{p_x}n_{p_y}}(q_{xy}) &= e^{-\frac{b_{p_x}n_{p_x}+b_{p_y}n_{p_y}}{2}}\sum_{n=0}^{q_{xy}} W^{n_{p_x}n_{p_y}}(n,q_{xy}), \qquad (4.18)\\ W^{11}(n,q_{xy}) &= \begin{cases}\sum_{k=2}^{k_{max}^{11}(n,q_{xy})}\Omega^{11}(n)\sigma'^{k-2}s_{eq}^{n-2}, & 2\leq n\leq q_{xy}-1\\ \frac{s_{eq}^{q_{xy}-2}}{\sigma'}, & n = q_{xy}\\ 0, & \text{otherwise}\end{cases}, \qquad (4.19)\\ W^{10}(n,q_{xy}) &= W^{01}(n,q_{xy}) = \begin{cases}\sum_{k=1}^{k_{max}^{10}(n,q_{xy})}\Omega^{10}(n)\sigma'^{k-1}s_{eq}^{n-1}, & 1\leq n\leq q_{xy}-1\\ 0, & \text{otherwise}\end{cases}, \qquad (4.20)\\ W^{00}(n,q_{xy}) &= \begin{cases}\sum_{k=1}^{k_{max}^{00}(n,q_{xy})}\Omega^{00}(n)\sigma'^{k}s_{eq}^{n}, & 1\leq n\leq q_{xy}-2\\ 1, & n=0\\ 0, & \text{otherwise}\end{cases}. \qquad (4.21)\end{aligned}$$

With the expression for the bare partition function found we can calculate the partition function Ξ^S for a few special cases. We will focus on the case $S = \{1,q\}$ and $S = \{1,q_*,q\}$, where in the latter case q is odd and $q_* = \frac{q+1}{2}$. Afterwards, we will discuss the influence of the assembly signals on the equilibrium distribution.

4.3.2 Ising-$\{1,q\}$ model

Suppose we have two special sites which are located at the ends of the template. Then we have $S = \{1,q\} = \{p_1,p_2\}$ and we obtain for the partition function the following

$$\begin{aligned}\Xi^{\{1,q\}}(q) &= \sum_{n=0}^{q}\Big(\chi_1\chi_q s_{eq}^2\sigma'^2 W^{11}(n,q) + \chi_1 s_{eq}\sigma' W^{10}(n,q) + \\ &\quad \chi_q s_{eq}\sigma' W^{01}(n,q) + W^{00}(n,q)\Big),\\ &\equiv \sum_{n=0}^{q} W^{\{1,q\}}(n), \qquad (4.22)\end{aligned}$$

where $s_{eq} = e^{\mu_P - \epsilon - g}$, $\sigma' = e^{-h'+\epsilon}$ and $\chi_i = e^{-(h_i - h')}$. Now we may easily find the equilibrium distribution for this case

$$\begin{aligned} P_{eq}^{\{1,q\}}(n) &= \frac{W^{\{1,q\}}(n)}{\Xi^{\{1,q\}}}, \\ &= \frac{1}{\Xi^{\{1,q\}}}\Big(\chi_1\chi_q(1+\delta_{n,q}(\frac{1}{\sigma'}-1))\sum_{k=2}^{k_{max}^{11}(n)}\Omega^{11}(n,q)\sigma'^k s_{eq}^n + \\ &\quad (\chi_1+\chi_q)\sum_{k=1}^{k_{max}^{10}(n)}\Omega^{10}(n,q)\sigma'^k s_{eq}^n + \sum_{k=1}^{k_{max}^{00}(n)}\Omega^{00}(n,q)\sigma'^k s_{eq}^n + \delta_{n,0}\Big). \end{aligned} \tag{4.23}$$

From this expression we can see immediately that the weight of $n = q$ is the same as for the zipper model if $\chi_q = 1$ and one identifies $\chi_1\sigma' = \sigma$. This is to be expected because for $n = q$ there is only one possible configuration in the Ising-S model as well as in the zipper model.

4.3.3 Ising-$\{1, q_*, q\}$ model

Suppose q is odd, we have two special sites at the ends and one exactly in the middle at $q_* \equiv \frac{q+1}{2}$. Then we obtain from equation (4.17) with $S = \{1, q_*, q\}$

$$\Xi^S(\mu_P) = \sum_{n_1,n_*,n_q=0,1}\Xi_{bare}^{s_l,s_{q_*}}(q_*)\Xi_{bare}^{s_{q_*},s_r}(q_*), \tag{4.24}$$

where we defined $n_{q_*} \equiv n_*$. From equation (4.18) we obtain

$$\begin{aligned} \Xi^S(\mu_P) &= \sum_{\{n_i\}_{i\in S}} e^{-\frac{b_1}{2}n_1 - b_*n_* - \frac{b_q}{2}n_q}\sum_{m_l=n_{l,min}}^{n_{l,max}}\sum_{m_r=n_{r,min}}^{n_{r,max}} W^{n_1n_*}(m_l,q_*)W^{n_*n_q}(m_r,q_*), \\ &\equiv \sum_{\{s_i\}_{i\in S}} e^{-\frac{b_1}{2}n_1 - b_*n_* - \frac{b_q}{2}n_q}\Gamma, \end{aligned} \tag{4.25}$$

where l stands for the left part of the template between site 1 and q_* and r stands for the right part. Also, we defined $n_{l,min} = n_1 + n_*$, $n_{l,max} = q_* - 2 + n_{l,min}$, $n_{r,min} = n_q + n_*$ and $n_{r,max} = q_* - 2 + n_{r,min}$. To calculate the equilibrium distribution we wish to rewrite this double sum as a double sum over the total number of proteins on the entire template and over the number of proteins on either the left or the right part. It turns out that this requires to sum over the part which has the least possible proteins given the boundary conditions. Therefore we introduce α

$$\alpha = \alpha(n_1, n_q) = \begin{cases} l, & n_q \geq n_1 \\ r, & n_1 > n_q \end{cases}, \bar{\alpha} = \begin{cases} r, & \alpha = l \\ l, & \alpha = r \end{cases}.$$

This allows us to write

$$\begin{aligned} \Gamma &= \Big[\sum_{n=n_{min}}^{n_{1,b,f}-1}\sum_{m_\alpha=n_{\alpha,min}}^{n-n_{\bar{\alpha}}} + \sum_{n=n_{1,b,f}}^{n_{max}}\sum_{m_\alpha=n-n_{1,b,f}+n_{\alpha,min}}^{n_{\alpha,max}}\Big]\cdot \\ &\quad \Big[\delta_{\alpha,l}W^{n_1n_*}(m_\alpha,q_*)W^{n_*n_q}(n-m_\alpha+n_*,q_*) + \\ &\quad \delta_{\alpha,r}W^{n_1n_*}(n-m_\alpha+n_*,q_*)W^{n_*n_q}(m_\alpha,q_*)\Big], \\ &\equiv \Big[\sum_{n=n_{min}}^{n_{1,b,f}-1}\sum_{m_\alpha=n_{\alpha,min}}^{n-n_{\bar{\alpha}}} + \sum_{n=n_{1,b,f}}^{n_{max}}\sum_{m_\alpha=n-n_{1,b,f}+n_{\alpha,min}}^{n_{\alpha,max}}\Big]W(n_1,n_*,n_q,n,m_\alpha,q_*) \end{aligned} \tag{4.26}$$

where $n = m_l + m_r$, $n_{min} = n_l + n_* + n_r$, $n_{max} = q - 3 + n_{min}$ and the number of proteins needed to fill the α part is $n_{1,b,f} = q_* - 2 + n_{min}$. The idea of this rewriting is that the first double sum holds for all n at which the α part can not be completely filled. The second sum holds for all n at which the α part can be completely filled. Now we introduce the bare weight function

$$W_{bare}^{\{1,q_*,q\}}(n_1, n_*, n_q, n, q_*) \equiv \begin{cases} 0, & n < n_{min} \\ \sum_{m_\alpha=n_{\alpha,min}}^{n-n_{\bar{\alpha}}} W(n_1, n_*, n_q, n, m_\alpha, q_*), & n_{min} \leq n < n_{1,b,f} \\ \sum_{m_\alpha=n-n_{1,b,f}+n_{\alpha,min}}^{n_{\alpha,max}} W(n_1, n_*, n_q, n, m_\alpha, q_*), & n_{1,b,f} \leq n \leq n_{max} \\ 0, & n > n_{max} \end{cases},$$

such that

$$\begin{aligned} \Xi^S(\mu_P) &= \sum_{\{s_i\}_{i\in S}} e^{-\frac{b_1}{2}n_1 - b_* n_* - \frac{b_q}{2}n_q} \sum_{n=0}^{q} W_{bare}^{\{1,q_*,q\}}(n_1, n_*, n_q, n, q_*), \\ &\equiv \sum_{n=0}^{q} W^{\{1,q_*,q\}}(n). \end{aligned} \tag{4.27}$$

Now one may find the equilibrium distribution exactly by calculating: $P_{eq}^{\{1,q_*,q\}}(n) = \frac{W^{\{1,q_*,q\}}(n)}{\Xi^S(\mu_P)}$. Because the expression is not closed it is not easy to calculate by hand. However, a program like Mathematica would suffice.

In the next paragraph we will consider what the influence of the special sites is on the equilibrium distribution.

4.4 Analysis

In the previous section expressions were obtained for the equilibrium distribution of the Ising-$\{1, q\}$ and the Ising-$\{1, q_*, q\}$ model. To analyse its behaviour we take for simplicity $\chi_1 = \chi_2 = \chi_3 \equiv \chi$. Therefore, q, s_{eq}, σ and χ are the relevant parameters. In the following will first the interchangeability of σ and q be shown, with regard to their effect on having strong correlation or not. Subsequently, to see the effect of the assembly signals, nucleation entropy and positional entropy only, we put $s_{eq} = 1$ such that their is no energetic advantage/disadvantage of having a protein cooperatively bound at a 'normal' site. The influence of assembly signals is opposed to that of positional entropy. For if the assembly signal is strong there will be a single posible position for a protein cluster of given size. Therefore, solely the effect of the assembly signals and entropy can be studied.

First, we consider the effect of σ' and q. For larger q one would expect more nucleation entropy driven behaviour for there are more possibilities for nucleation. Next, for large $\sigma' = e^{-h'+\epsilon}$ - so low nucleation cost - one would also expect more nucleation entropy driven behaviour. To make an estimate of the parameter values of q and σ' at which entropic or energetic dominated behaviour occurs we consider the correlation length of the partition functions which were calculated in section 4.1. There we see that the partition function comprises of a linear combination of λ_+^q and λ_-^q over a chain of length q. Furthermore, if $s_{eq} \to 1$ is $\lambda_+ - \lambda_- \approx 2\sigma'^{\frac{1}{2}}$. So for low values of σ' - which implies high nucleation cost, long clusters and thus a high correlation - we expect $\lambda_+ \approx \lambda_-$. Therefore, the relative sizes of the eigenvalues signify in which regime the system is. To obtain an expression of the correlation length we rewrite the quantity $\left(\frac{\lambda_-}{\lambda_+}\right)^q$, for $\sigma' < 1$, as

$$\left(\frac{\lambda_-}{\lambda_+}\right)^q = e^{q \ln \frac{\lambda_-}{\lambda_+}} \equiv e^{-\frac{q}{\xi}}, \tag{4.28}$$

with $\xi \equiv \frac{-1}{\ln \frac{\lambda_-}{\lambda_+}}$ the correlation length. From this expression one can see that if $\lambda_+ \approx \lambda_-$ there is a long correlation length because then $\xi \gg q$ is required. On the other hand, for $\lambda_+ \gg \lambda_-$ the correlation length

is much smaller than q. The value of ξ at which the crossover between the two regimes lies is q. So, if we consider $s_{eq} = 1$ for the above mentioned reason, we obtain the following relation $q \ln\left(1 + \frac{2\sqrt{\sigma_c'}}{1-\sqrt{\sigma_c'}}\right) = 1$, where σ_c' is the crossover value at which $\xi = q$. Since typically $q \gg 1$ it is from this relation safe to assume that $\sqrt{\sigma_c'} \ll 1$. Therefore, we have $2q\sqrt{\sigma_c'} = 1$. This shows clearly that for larger q the crossover lies at a lower value of σ'. This is to be expected because for larger q there should be a greater nucleation cost in order to suppress entropic nucleation effects. From this analysis we conclude that it is only necessary to analyse one value of q at different values of σ' around σ_c' in order to see nucleation entropic effects or not.

Furthermore, in section 4.2 it was shown that for $s_{eq} \to 1$ the competitor states - the position entropy favourable states - are suppressed for $\chi \gg q$. To see the influence of positional entropy we will cover the cases where this condition is either satisfied or not. With these considerations we choose two parameter sets which should give all interesting behaviour of the system. For the position entropy dominating regime we choose $q = 75, s_{eq} = 1, \chi = e^3$ and for the assembly-signal-dominated regime we choose $q = 75, s_{eq} = 1, \chi = e^9$. We plot the logarithm of the equilibrium probability, $P_{eq}(n)$ versus n, which is scaled to the probability of having an empty template for graphical convenience. See for example figure 4.1. In the following paragraphs the Ising-$\{1, q\}$ model will first be analysed. Afterwards, will the Ising-$\{1, q_*, q\}$ model be covered.

4.4.1 Ising-$\{1, q\}$ model

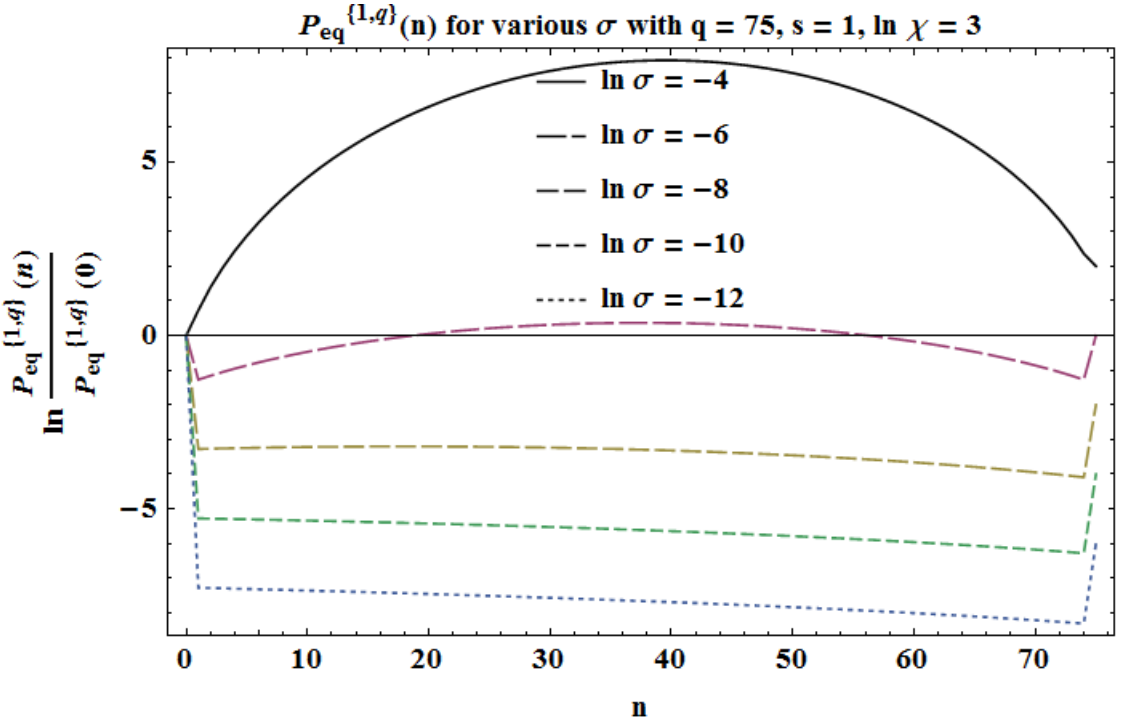

Figure 4.1: The equilibrium probability distribution of the Ising-$\{1, q\}$ model for $q = 75$, various σ', $\chi = e^3$ and $s_{eq} = 1$. It shows the logarithmic probability of having a template with n proteins attached, relative to the the probability of an empty template, versus the number of proteins attached. The value of n does not specify a position on the template but the number of bound proteins which can be distributed in many different configurations. The crossover value is $\sigma_c' = e^{-10}$. For lower σ' a $\ln q - n$ dependence is visible because of the multiplicity of competitor states. For higher values the distribution becomes Gaussian.

To analyse the probability distribution we first consider the position entropy dominated regime. The resulting distribution is given in figure 4.1 for multiple values of σ'. In this regime we have $\sigma_c' = e^{-10}$.

Therefore, if we consider $\ln \sigma' = -12$, we expect strongly correlated behaviour. For this σ' value there could, in principle, be a nucleation at both assembly signals. However, since we are in the position entropy dominated regime, these states gain a factor $\sigma'\chi = e^{-7} \ll 1$ relative to a single assembly signal nucleated state. Therefore, the states with two nucleations are negligible and we expect the Ising-$\{1, q\}$ model to behave as the Ising-$\{1\}$ model for $n < q$. For $n = q$ both assembly signals are necessarily occupied. From section 4.2 we have that the Ising-$\{1\}$ model contains up to first order in σ' - so for one nucleation - zipper as well as competitor states. That is, assembly-signal-dominated and position-entropy dominated states. This is what is observed in figure 4.1 because, unlike figure 3.1, the distribution is not horizontal. Instead, it shows a $\ln(q - n)$ behaviour which is to be expected because from the theory of section 2.3.1 we know that the competitor states have a multiplicity factor of $q - n$.

For higher values of σ the nucleation entropy slowly starts dominating for $\sigma' > \sigma'_c$. At $\ln \sigma' = -8$ the distribution has a local maximum at $n = 20$. Next, for $\sigma' = e^{-6}$ a parabolic shape of the distribution arises for $1 < n < q$. Finally, at $\sigma' = e^{-4}$ a very clear parabolic behaviour is visible. This parabolic behaviour implies that the distribution is is Gaussian for we plotted the logarithm of the distribution. This Gaussian behaviour is to be expected because the binomial distribution - which determines the multiplicity as given by equation (4.12) up to and including (4.14) - can be approximated for large q by a normal distribution.

Another aspect is the energetic advantage of binding at an assembly signal. As noted above, for $0 < n < q$ there is only one nucleation, thus only one assembly signal is occupied. Therefore, the second assembly signal only favours the state $n = q$ where both signals are occupied by the same cluster. This is a significant effect for $\sigma' \leq \sigma'_c$. For higher values its effect diminishes.

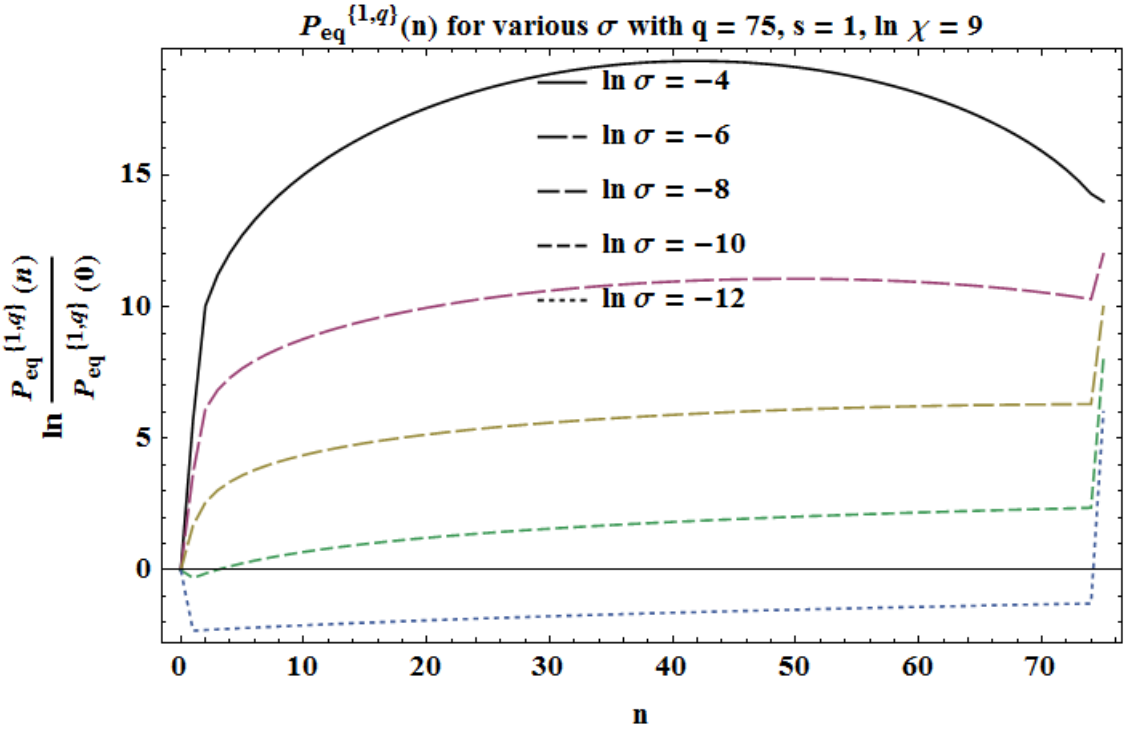

Figure 4.2: The equilibrium probability of the Ising-$\{1, q\}$ model for $q = 75$, various σ', $\chi = e^9$ and $s_{eq} = 1$. It shows the logarithmic probability of having a template with n proteins attached, relative to the the probability of an empty template, versus the number of proteins attached. The crossover value is $\sigma'_c = e^{-10}$. For lower σ' a $\ln n$ dependence is visible due to two-cluster states. For higher values the distribution becomes Gaussian.

For the assembly signal dominated regime is the distribution given in figure 4.2. It shows in the strongly correlated regime, $\sigma' = e^{-12}$, a $\ln n - 1$ behaviour. This is new compared to the position energy dominated regime. It originates from two nucleations at the ends, such that both assembly signals

are occupied. This gives a multiplicity factor of $n-1$ because the proteins can be arranged in $n-1$ equivalent ways. For $\sigma' = e^{-8}$ this two nucleation behaviour is even more apparent. At larger values of σ' the nucleation entropy takes over such that the distribution becomes nearly Gaussian for $\sigma' = e^{-4}$ at $2 < n < q$. At $n = q$ the probability exhibits a jump. This is because the two clusters merge into one cluster. Therefore, the system gains a factor $\frac{1}{\sigma}$.

4.4.2 Ising-$\{1, q_*, q\}$ model

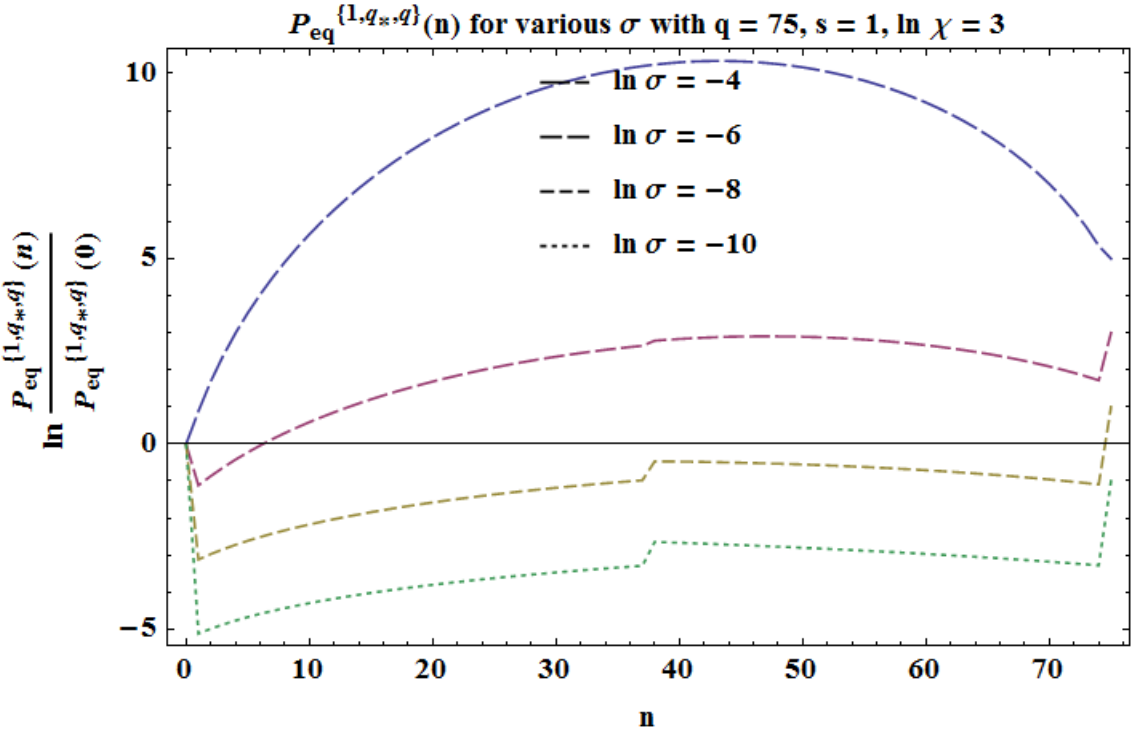

Figure 4.3: The equilibrium probability of the Ising-$\{1, q_*, q\}$ model for $q = 75$, various σ', $\chi = e^3$ and $s_{eq} = 1$. It shows the logarithmic probability of having a template with n proteins attached, relative to the the probability of an empty template, versus the number of proteins attached. The crossover value is $\sigma'_c = e^{-8.66}$. For lower σ' a $\ln n$ dependence is visible at $n < q_*$ due to multiplicity of one nucleation at position q_*. A $\ln q - n$ dependence occurs for $n > q_*$ due to competitor states. For higher values of σ' the distribution becomes Gaussian.

To analyse $P_{eq}^{\{1,q_*,q\}}$ in the position entropy dominated regime, which is given in figure 4.3, we first note that with $q = 75$ we have $q_* = 38$ wherefore σ'_c is larger than for the Ising-$\{1, q\}$ model. It gives $\ln \sigma'_c = -8.66$. Therefore, with $\sigma' = e^{-10}$ we expect strongly correlated behaviour. In this regime we see more complicated behaviour than for the Ising-$\{1, q_*, q\}$ model. There is $\ln q - n$ behaviour for $n \geq q_*$ and some other entropic behaviour for $n < q_*$. This behaviour stems from the fact that with three assembly signals a new entropic factor comes into play. The states with one nucleation, namely that of the assembly signal at q_*, carry a multiplicity factor of n since the cluster can have n equivalent positions. Therefore, the distribution shows $\ln n$ behaviour for these values of n. Furthermore, for higher values of σ' the Gaussian behaviour sets in like for the Ising-$\{1, q\}$ model.

The energetic advantage of binding at an assembly signal is interesting. For $\sigma' \leq \sigma'_c$ a jump in the probability is observed at $n = q_*$. This comes from the fact that the cluster which occupied the central assembly signal 'reached' an other assembly signal. Therefore, for $n \geq n_*$ the states where one cluster covers two assembly signals while the third is unoccupied dominate. Finally, for $n = q$, the third assembly signal is reached which gives another probability jump.

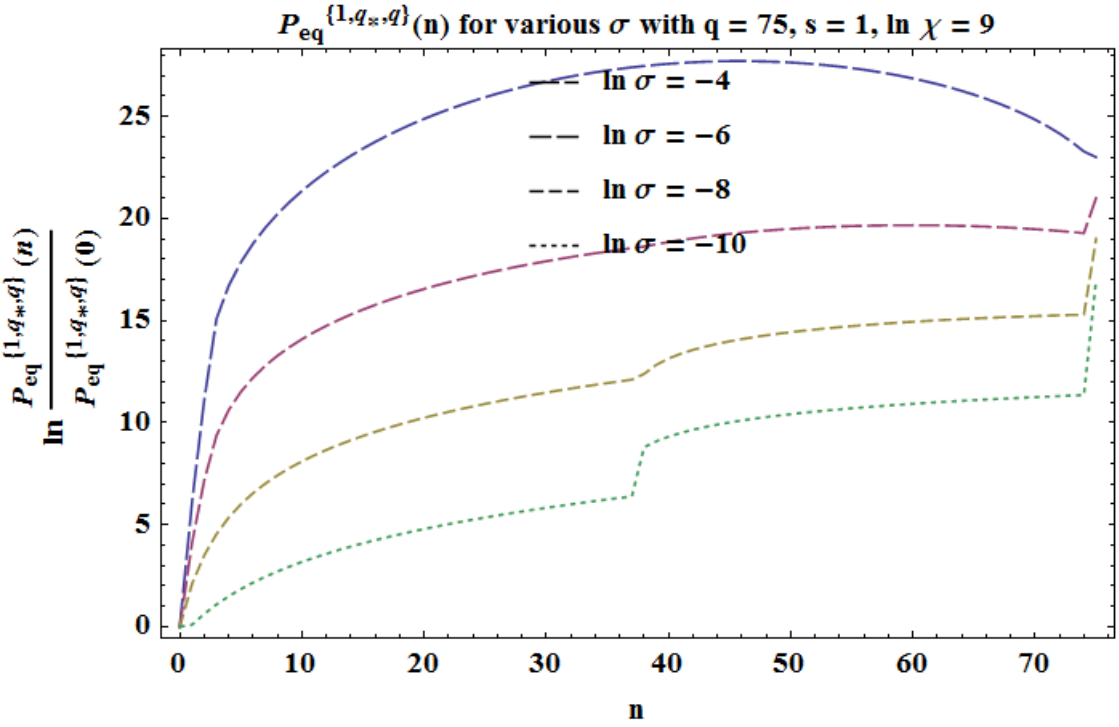

Figure 4.4: The equilibrium probability of the Ising-$\{1, q_*, q\}$ model for $q = 75$, various σ', $\chi = e^9$ and $s_{eq} = 1$. It shows the logarithmic probability of having a template with n proteins attached, relative to the the probability of an empty template, versus the number of proteins attached. The crossover value is $\sigma'_c = e^{-8.66}$. For lower σ' a $3 \ln n$ dependence is visible at $n < q_*$ due to multiplicity of the three nucleation state. For $n > q_*$ this state give rise to a dominant $\ln n$ dependence. For higher values of σ' the distribution becomes Gaussian.

For the assembly signal dominated regime is the distribution given in figure 4.4. In the strongly correlated regime, for $\sigma' = e^{-10}$, there is similar behaviour as for the position entropy dominated regime at $n < q_*$, but very different behaviour for $n \geq q_*$. This can be understood by considering three dominant states which mainly contribute in this regime. First, the states with one nucleation at position 1,q_* or q. Moreover, the states with two of the three assembly signals nucleated. Finally, the state where all assembly signals are nucleated. Since every nucleation at an assembly signal gives a factor $\chi\sigma'$ one could expect the three and two nucleation states to be unfavourable. However, this is not the case for the multiplicities are not to be neglected. As a rule of thumb does the multiplicity of a state scale as n^{x-1}, with x the number of loose ends of that state. A loose end being an end of a cluster which is not at an assembly signal. This rule of thumb only holds in the assembly signal dominated and strong correlated regime. Therefore, for $n < q_*$ does the three nucleation state multiplicity, for example, scale as n^3 while that of the two nucleation state at position 1 and q_* scales as n^2. From this two nucleation state one can understand this rule of thumb. It has two loose ends at the left part and one loose end at the right part. For a given number of proteins at the right part, proteins of the left part can be arranged in a number of ways which scales roughly as n. Independently, the right part can have n proteins such that the total multiplicity goes as n^2. One can calculate the multiplicity exactly and finds $\frac{n(n-1)}{2} \propto n^2$. Therefore, in this regime for $n < q_* = 38$ does n^3, for increasing n, quickly start to dominate over n^2. Also, the cost of nucleation is only e^{-1}. These two reasons combined make the three nucleation state dominant and explains the the $3 \ln n$ behaviour of the distribution for $n < q_*$. For $n < q_*$ this state is also dominant. The reason is that for these values the probability is increasing like $\ln n$ instead of decreasing as $\ln q - n$. The only possible state to give this behaviour is the three nucleation state since its multiplicity predominantly scales as $n - q_*$ while the two nucleation states have a multiplicity of either n^0 or $q - n$.

Furthermore, for higher values of σ' the Gaussian behaviour starts dominating. The jump of the

probability at $n = q_*$ is mainly caused by the merging of two clusters of the three nucleation state at either the left or the right part. This gives a factor $\frac{1}{\sigma'}$. The jump at $n = q$ is for the same reason.

In this chapter we encountered twice competitor states, more precisely, the self-competition states as introduced in section 2.3.1. In the first order approximation of the partition function of the Ising-$\{1\}$ model they occurred. Afterwards, they were found to be dominant in the equilibrium distribution of both the Ising-$\{1, q\}$ and the Ising-$\{1, q_*, q\}$ model for the position entropy dominated regime. The next chapter will focus in more detail on competition. It will cover both self-competition and species competition.

Chapter 5

Competition

As introduced in section 2.3.1 we consider two kinds of competition: self-competition and species competition. Both kinds of competition are between position entropy rich states - the competitor states - and binding energy rich states - the zipper states. In self-competition both kind of states are encountered on the same template. For species competition one species only has competitor states while the other only has zipper states. In the following we will first consider self-competition. Afterwards, we will have a short word on species competition.

5.1 Self-competition

The self-competition partition function, as introduced in section 2.3.1, can be calculated as

$$\Xi_{sc} = 1 + \chi\sigma' s_{eq}\frac{1-s_{eq}^q}{1-s_{eq}} + \sigma' s_{eq}\frac{-1+q(1-s_{eq})+s_{eq}^q}{(1-s_{eq})^2}, \tag{5.1}$$

with $\sigma' = e^{-h'+\epsilon}$ the Boltzmann factor of the nucleation cost, $\chi = e^{-(h-h')}$ the factor for binding at the assembly signal and $s_{eq} = e^{\mu_P - g - \epsilon}$ the factor for having a protein cooperatively bound. This partition function was encountered in section 4.2 as the first order approximation of the partition function of the Ising-$\{1\}$ model. To see the competition between the competitor and zipper states we define, respectively, the fraction of templates in a competitor and zipper state as

$$P_{zip} \equiv \frac{\Xi_{zip}(q)-1}{\Xi_{sc}},\ P_{comp} \equiv \frac{\Xi_{comp}(q-1)-1}{\Xi_{sc}}, \tag{5.2}$$

where the definitions of the zipper and competitor partition functions from section 2.3.1 was used. Moreover, the average occupation numbers of both kind of states are important measurable quantities

$$\langle\theta\rangle_{zip} \equiv \frac{\partial \ln\Big(\Xi_{zip}(q)-1\Big)}{\partial\mu_P},\ \langle\theta\rangle_{comp} \equiv \frac{\partial \ln\Big(\Xi_{comp}(q)-1\Big)}{\partial\mu_P}. \tag{5.3}$$

All these quantities depend on s_{eq} which is determined by mass conservation

$$S = s_{eq} + S\lambda\langle\theta\rangle_{sc},$$

with $S = \frac{\phi_P}{\phi_c}$ the total concentration of proteins relative to the critical concentration, $\lambda = \frac{q\rho_T}{\phi_P}$ the stoichiometric ratio of the number of available binding sites in the system to the total number of proteins and $\langle\theta\rangle_{sc}$ the occupation number of all templates. This equation determines $s_{eq} = s_{eq}(S,\lambda,\sigma',\chi)$. The expression of $\langle\theta\rangle_{sc}$ is naturally given by

$$\langle\theta\rangle_{sc} \equiv \frac{\partial \ln\Xi_{sc}}{\partial\mu_P} = P_{zip}\langle\theta\rangle_{zip} + P_{comp}\langle\theta\rangle_{comp}. \tag{5.4}$$

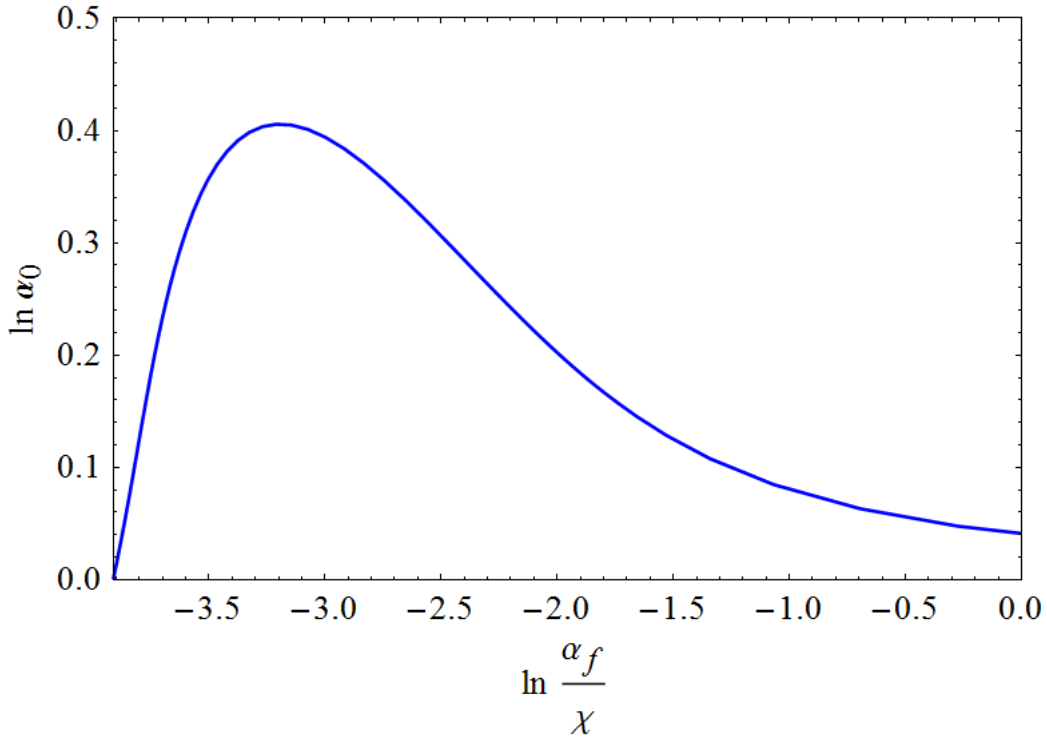

Figure 5.1: For any value of S, λ, σ' and χ the values of α_o and $\frac{\alpha_f}{\chi}$ appear to lie on this curve. The origin of this universality is not known. However, the curve can be used to determine the strenght of the zipper assembly signal: χ.

With these definitions we introduce the specificity ratios, $\alpha_f \equiv \frac{P_{zip}}{P_{comp}}$ and $\alpha_o \equiv \frac{\langle\theta\rangle_{zip}}{\langle\theta\rangle_{comp}}$, to characterize the competition. Once could ask whether the specificity ratios have some relation. It turns out that for given S, λ, σ' and χ the specifity ratios lie somewhere on the curve given in figure 5.1. For increasing values of $\frac{\alpha_f}{\chi}$ greater than unity, α_o slowly goes to unity. In this limit $\lambda \to 0$ wherefore the occupation ratio goes unity and α_f goes to infinity. Furthermore, it shows that α_f scales with χ. This scaling relation can easily be seen by computing α_f. The universality observed in figure 5.1 is in essence a characteristic of the mass conservation equation. Namely, it gives $s_{eq} = s_{eq}(S, \lambda, \sigma', \chi)$ and subsequently determines α_f and α_o. Therefore, in principle one should be able to show this universality arising from the mass conservation equation. however, this is left for future research.

The universality curve is useful for the following reason. If one measures both α_f and α_o, one can infer with this curve two possible values of χ. However, this does not specify χ uniquely. A way to get around this is by considering two different systems with the same χ and making measurements of both of them, so one with $\{\sigma'_1, \lambda_1, S_1\}$ and the other with $\{\sigma'_2, \lambda_2, S_2\}$. These parameters can have any value. For both systems α_f and α_o can be measured experimentally. So for system one $\alpha_{f,1}$ and $\alpha_{o,q}$ are measured, likewise for system two. Subsequently, from the universal curve does $\alpha_{o,1}$ gives two possible values of χ - one left of the maximum and one right of the maximum - as well as $\alpha_{o,2}$ does. The value of χ which is possible for both systems is the value which the system has. In this way the strength of an assembly signal can be readily measured.

To have a visualization for the point at the curve which a given combination of S, λ, σ' and χ gives, we show in figure 5.2 α_f as a function of λ for $S = 2$, $\sigma' = e^{-4}$ and various values of χ. It shows that α_f is largest for $\lambda \to 0$. In this limit there is no competition, for $s_{eq} = S$ and thus the competitor and zipper states are not influenced by each other. Therefore, the amount of fully covered templates is largest and thus α_f is largest since fully covered templates are per definition zipper like. For $0 < \lambda < 1$ the ratio is for increasing λ first relatively constant and subsequently drops sharply around $\lambda = 1$. For $\lambda > 1$ the ratio steadily decreases. This is to be expected since competition should become important if the number of

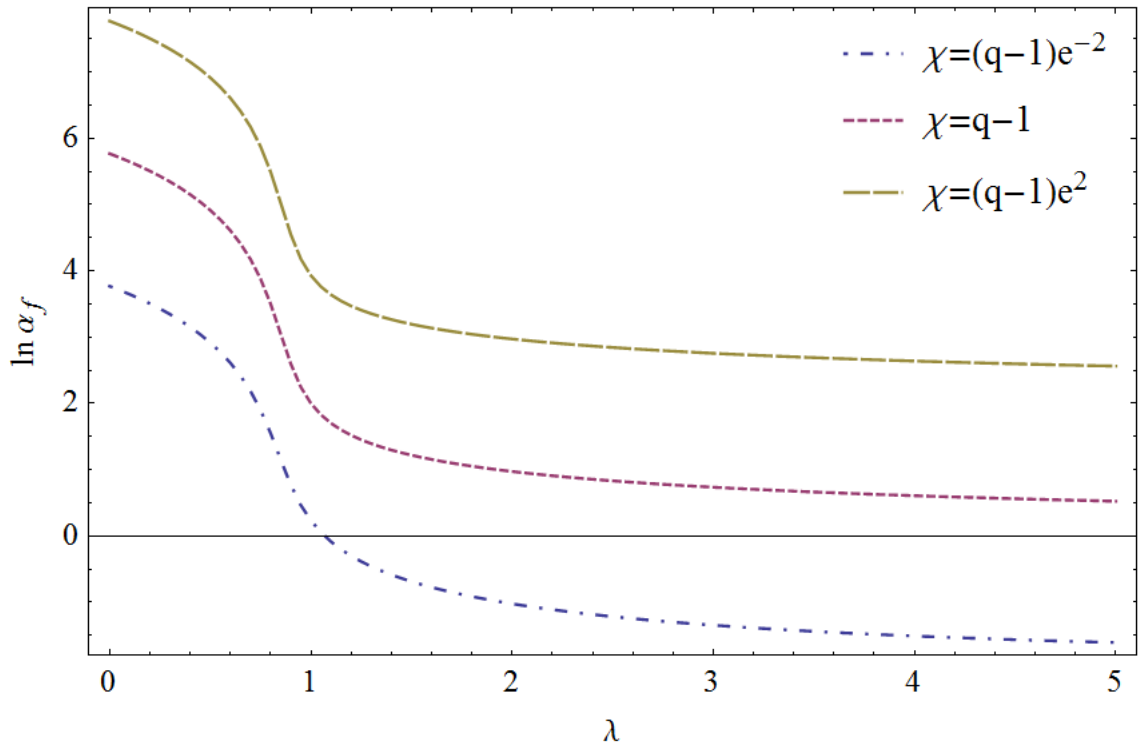

Figure 5.2: For $S = 2$ and $\sigma' = e^{-4}$ is α_f given as a function of λ. It shows that α_f indeed scales with χ. Also, if $\chi = q - 1$ we have $\alpha_f \to 1$ in the limit of infinite competition $\lambda \to \infty$. This is to be expected because under this condition are the entropic and energetic gain of binding the first protein at, respectively, the zipper and competitor states equal.

available proteins is of the order of the available number of binding sites. Furthermore, for $\chi = q - 1$ one observes that $\alpha_f \to 1$ for $\lambda \to \infty$. The reason is that in this limit the competition will mainly be among states with only one protein bound. For the zipper states the binding of a protein at the assembly signal gives χ while the binding of one protein in a competitor state has a multiplicity factor of $q - 1$. Therefore, if these two contributions are equal one would expect none to have an advantage wherefore $\alpha_f = 1$. In the next section, a very short note will be given on the essence of species competition.

5.2 Species competition

In species competition one species has a template with a (strong) assembly signal - therefore giving rise to zipper states only - while the other species has no signal and thus only competitor states. The system can be well characterized from figure 5.3, where $\sigma' = e^{-12}$, $\chi = e^9$ and $\lambda_{c,zip} = \lambda_{c,comp} = e^4$ where $\lambda_c \equiv \frac{q\rho_T}{\phi_c}$ is the number of available binding sites on a template species relative to the critical concentration. It shows that for $S > 1$ only the zipper templates are being filled up to $S \approx \lambda_{c,zip}$, the point where the system has enough proteins to, in principle, fill all zipper templates. Nevertheless, for increasing S the growth halts at $\langle\theta\rangle_{zip} \approx 0.96$. While it halts the competitor templates are filled and finally both occupancies rise to unity. The flatness of the plateau is dependent on σ' and the breadth on χ. This is the most important characteristic of species competition. Varying $\lambda_{c,zip}$ and $\lambda_{c,comp}$ gives rise to the same kind of behaviour. An interesting feature is the value of $\langle\theta\rangle_{zip}$ at which the filling halts. The investigation of this property is left for future research.

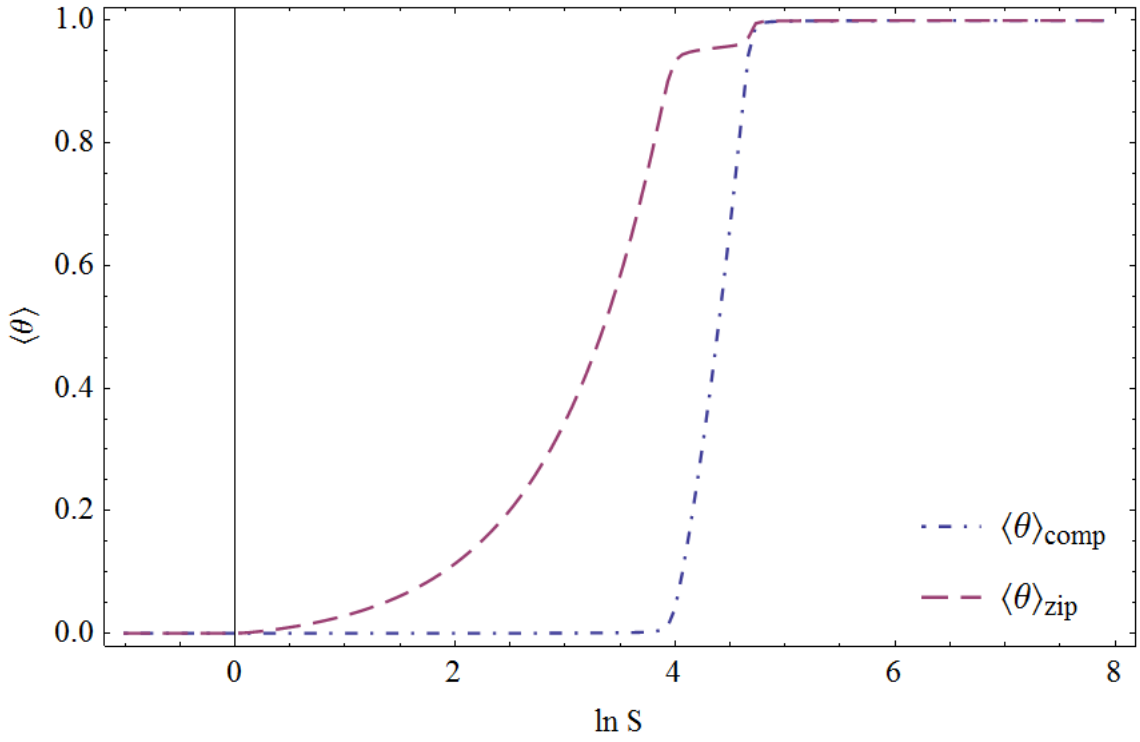

Figure 5.3: For both species is $\langle\theta\rangle$ given as a function of S for $\lambda_{c,zip} = \lambda_{c,comp} = e^4$, $\sigma' = e^{-12}$ and $\chi = e^9$. It shows for increasing S that the zipper templates are first filled up to soe treshold value $\langle\theta\rangle = 0.96$. Afterwards, the competitor templates are filled until both species are filled at $S \approx \lambda_{c,zip} + \lambda_{c,comp}$.

Chapter 6

Conclusion and discussion

In this thesis on linear self-assembly four main topics have been discussed. The dynamics of zipper self-assembly, the agreement of these calculated dynamics with experiment, the equilibrium distribution of the Ising-S model and the competition between energy versus entropy rich states. In the following a summary of the results of all topics will be given, their range of validity will be discussed and recommendations for future research will be given.

Summary

The dynamics of zipper self-assembly show surprising behaviour when probed with a finite protein concentration. It turns out that an overshoot in the concentration of fully covered templates can occur. Also, undershoots in the concentration of empty templates and proteins are found. The cause of this behaviour is not yet clear.

Experimental data which shows zipper like behaviour can be fitted reasonably well with the zipper dynamics. For the SQ10 data set is found that $\epsilon + g = -16.97$, $-h + \epsilon = -5.521$ and for SQ14 $\epsilon + g = -10.40$, $-h + \epsilon = -3.00$, with all energies in units of $k_B T$. All these values are of the expected orders of magnitude for the binding energy scales of self-assembly are of the order of 10^0. Furthermore, for NP3 the fits of $\lambda = 0$ and $\lambda = 0.324$ give similar results which is to be expected because both have excess protein concentration. For $\lambda = 0.324$ we find $\epsilon + g = -18.36$, $-h + \epsilon = -5.30$. With $\lambda = 0$ is $\epsilon + g = -18.04$, $-h + \epsilon = -3.51$ found.

For the Ising-S model the general partition function can be calculated. The Ising-$\{1\}$ partition function reduces in first order of the nucleation cost to the self-competition partition of the zipper model. This is to be expected because for a single nucleation only zipper and competitor states are possible. For the Ising-$\{1, q\}$ and Ising-$\{1, q_*, q\}$ model the equilibrium distribution can be calculated, though not in a closed form. Upon analysis of this distribution it was found that three regimes occur. The first, where nucleation entropy dominates, has a Gaussian distribution due to low nucleation cost. The other regimes have high nucleation cost. The second regime is position-entropy-dominated with weak assembly signals and costly nucleation. The third is assembly-signal-dominated and the nucleation of all assembly signals gives rise to new entropic behaviour.

Finally, in self-competition it turns out that the specificity ratio of the fractions of zipper and competitor states obey a universal curve. This curve can be used to determine the strength of an assembly signal through measurements. Furthermore, filling of templates when increasing protein concentration shows a plateau for species competition. The zipper states are filled up to some particular value which defines the height of the plateau. Afterwards, the competitor states start to be filled also. Finally, they are both completely filled simultaneously.

Discussion

With these main results given we now discuss their validity.

The peculiarities in the zipper assembly dynamics are criterion dependent. The reason is that for some assemblies the overshoot can be very small. The question is whether one defines this to be an overshoot or not. At a certain point uncertainty in the numerical simulation comes into play. This makes boundaries in the peculiarities phase diagram uncertain.

Next, for the comparison of data to the zipper assembly a great uncertainty arises because the data is obtained by counting. This obeys Poisson statistics and since the counts are of the order 10^1 per bin, the uncertainty is very large. This translates into a large uncertainty in the the calculated energies. Furthermore, due to the availability of four fitting parameters the predictive value of the fit is questionable. The main features of the data could be fitted by pinning down the fitting parameters. Discrepancy with other features can be attributed to uncertainty in the data.

Moreover, the calculated equilibrium distributions for the Ising-$\{1, q\}$ and Ising-$\{1, q_*, q\}$ model are exact but for more assembly signals the calculations become ever more complex. This makes them unsuitable for investigating the effect of more assembly signals.

Recommendations

For future research we provide the following recommendations.

To obtain a better understanding of the dynamical equations we propose that approximations, for example those of appendix C, are worked out and that a more mathematical approach is used to analyse the system of non-linear, first order couple differential equations. Possibly, more general theorems can shed light on the occurrence of over- and undershoots.

Furthermore, to be able to judge better the validity of the fit of experimental data we propose data with more length measurements, of the order of 10^3 per bin, such that the model can possibly be refuted. Also, an accurate estimation of the offset time would be helpful because it gives one fitting parameter less. This would also increase the possibilities of refutation.

Next, for the Ising-S model we recommend to expand the partition function in a power series of s_{eq}. This should give rise to a polynomial since the cluster expansion is essentially such a polynomial expansion and it is exact. By expanding the partition function, however, one does not need to sum over the number of clusters. Therefore, starting from a closed form expression, one should be able to find the terms of the polynomial expansion in a more explicit form than the result obtained with the cluster expansion. In this way the equilibrium distribution can possibly be found in closed form. This possibility can be seen by using Mathematica to make such an expansion of the partition function for definite parameter values.

Finally, for self-competition should it be possible to derive the universal curve from mass conservation. However, this implies that $s_{eq} = s_{eq}(S, \lambda, \sigma', \chi)$ needs to be found which is non-trivial. Therefore, some novel method of attacking this problem should be used.

Appendix A

Canonical multi-component derivation

In section 2.1 we derived the sub-grand canonical partition function describing our system in the grand canonical ensemble. One might object that this is a strange description because in any self-assembly experiment one does not have an infinite bath of particles which determines the chemical potential, but one has a given V, T and N. Therefore we will give in this appendix a derivation in the canonical ensemble. Of course, in the thermodynamic limit a description in the grand canonical ensemble and in the canonical ensemble should be give the same results. Nonetheless, it is instructive to make the derivation in the canonical ensemble also. We will do this both for one kind of templates and for two kinds of templates.

A.1 Lagrange formalism

To calculate the equilibrium quantities in the canonical ensemble we write down the Helmholtz free energy in the same way as in section 2.1 with exactly the same definitions. This gives

$$F = \rho_P\Big[\ln\rho_P - 1\Big] + \sum_{\{n_i\}}\rho_T(\{n_i\})\Big[\ln\Big(\rho_T(\{n_i\})\frac{V_{mol,T(\{n_i\})}}{V_{mol,P}}\Big) - 1 + E_{int}(\{n_i\})\Big],$$

To find the equilibrium values of ρ_P and $\rho_T(\{n_i\})$ we should minimize F with regard to the densities. Nevertheless, to do so we should take into account that we have a finite amount of both templates and proteins. The fixed total concentration of templates is $\rho_T = \sum_{\{n_i\}}\rho_T(\{n_i\})$ and the fixed total concentration of protein is $\phi_P = \rho_P + \sum_{\{n_i\}} n\rho_T(\{n_i\})$. To take this into account we use the Lagrange multiplier formalism to define F_L. We obtain

$$F_L = F + \lambda_0\Big(\phi_P - \rho_P - \sum_{\{n_i\}} n\rho_T(\{n_i\})\Big) + \lambda_1\Big(\rho_T - \sum_{\{n_i\}}\rho_T(\{n_i\})\Big),$$

To find the equilibrium densities we minimize F_L

$$\rho_P = \exp[\lambda_0], \tag{A.1}$$

$$\rho_T(\{n_i\})\frac{V_{mol,T(\{n_i\})}}{V_{mol,P}} = \exp[-E_{int}(\{n_i\}) + n\lambda_0 + \lambda_1], \tag{A.2}$$

from which we see that the λ_0 simply takes the role of μ_P and $\lambda_1 = \mu_T$.

A.2 Multi-component

One might conduct an experiment with multiple templates species. This can be, for example, the case if one uses zipper model templates together with competitor zipper model templates. If one mixes these two

kinds of templates the difference between the two with regard to occupation should be observable. In this case we have conservation of the two kinds of templates and of the proteins. This gives for the Helmholtz free energy

$$\begin{aligned} F &= \rho_P\Big[\ln\rho_P - 1\Big] + \sum_{\{n_i\}}\rho_{T,1}(\{n_i\})\Big[\ln\Big(\rho_{T,1}(\{n_i\})\frac{V_{mol,T_1(\{n_i\})}}{V_{mol,P}}\Big) - 1 + E_{int,1}(\{n_i\})\Big] + \\ &\quad + \sum_{\{n_i\}}\rho_{T,2}(\{n_i\})\Big[\ln\Big(\rho_{T,2}(\{n_i\})\frac{V_{mol,T_2(\{n_i\})}}{V_{mol,P}}\Big) - 1 + E_{int,2}(\{n_i\})\Big], \end{aligned}$$

where the sums runs over the allowed configurations of the respective model. To take into account the finite amount of templates and protein we write

$$\begin{aligned} F_L &= F + \lambda_0\Big(\phi_P - \rho_P - \sum_{\{n_i\}} n\rho_{T,1}(\{n_i\}) - \sum_{\{n_i\}} n\rho_{T,2}(\{n_i\})\Big) + \lambda_1\Big(\rho_{T,1} - \sum_{\{n_i\}}\rho_T(\{n_i\})\Big) + \\ &\quad \lambda_2\Big(\rho_{T,2} - \sum_{\{n_i\}}\rho_{T,2}(\{n_i\})\Big), \end{aligned}$$

To find the equilibrium densities we minimize F_L

$$\rho_P = \exp[\lambda_0], \tag{A.3}$$

$$\rho_{T,1}(\{n_i\})\frac{V_{mol,T_1(\{n_i\})}}{V_{mol,P}} = \exp[-E_{int,1}(\{n_i\}) + n\lambda_0 + \lambda_1], \tag{A.4}$$

$$\rho_{T,2}(\{n_i\})\frac{V_{mol,T_2(\{n_i\})}}{V_{mol,P}} = \exp[-E_{int,2}(\{n_i\}) + n\lambda_0 + \lambda_2], \tag{A.5}$$

from which we see that $\lambda_0 = \mu_P$, $\lambda_1 = \mu_{T,1}$ and $\lambda_2 = \mu_{T,2}$ in the same way as for the one-component case. These results show that the distribution of the two different templates species is connected through λ_0, or ρ_P. The density of unbound proteins determines of both templates species the distribution.

Appendix B

Mass conservation analysis

In section 3.1 we found $\langle\theta\rangle$ and P_{eq} in equilibrium as a function of s_{eq}, that is, s in equilibrium. From the definition of $s = \frac{\rho_P}{\phi_c}$ we can see that its the density of unbound proteins relative to the critical density $\phi_c = e^{\epsilon+g}$. Nevertheless, experimentally we do not impose the density of unbound proteins but the total density of proteins $S \equiv \frac{\phi_P}{\phi_c}$. This total density must constant - we do not consider protein degradation - and the sum of the unbound and bound proteins. This gives

$$\begin{aligned} \phi_P &= \rho_P(t) + \sum_{n=0}^{q} n\rho_T(n,t), \\ S &= s(t) + \lambda_c\langle\theta\rangle(t), \end{aligned}$$

where $\lambda_c = \frac{\rho_T q}{\phi_c}$, $\phi_c = e^{\epsilon+g}$, $\rho_T = \sum_{n=0}^{q}\rho_T(n)$ and $\langle\theta\rangle = \sum_{n=0}^{q}\frac{n}{q}\frac{\rho_T(n)}{\rho_T}$. This expression holds in and out of equilibrium. In particular, when the system has reached equilibrium we have that $s = s_{eq}$ and from equation (3.2)

$$\langle\theta\rangle(s_{eq},\sigma) = \frac{\sigma}{q}\frac{s_{eq}}{(1-s_{eq})}\Big(\frac{1-(q+1)s_{eq}^q + qs_{eq}^{q+1}}{1-s_{eq}+\sigma s_{eq}(1-s_{eq}^q)}\Big). \tag{B.1}$$

With this expression we have the following equation for $s_{eq} = s_{eq}(S,\lambda_c,\sigma)$

$$S = s_{eq} + \lambda_c\langle\theta\rangle(s_{eq},\sigma). \tag{B.2}$$

This is an implicit equation for $s_{eq}(S,\lambda_c,\sigma)$. One would like to know $s_{eq}(S,\lambda_c,\sigma)$ exactly because it gives the concentration of unbound proteins and with that also the equilibrium properties of the system. To obtain a more explicit equation we rewrite the mass conservation equation to

$$\begin{aligned} &S + s_{eq}\Big(S(\sigma-1) - (1+S) - \frac{\lambda_c\sigma}{q}\Big) + s_{eq}^2\Big(1-(1+S)(\sigma-1)\Big) + s_{eq}^3(\sigma-1) + \\ &\sigma s_{eq}^{q+1}\Big(\lambda_c\frac{q+1}{q} - S + s_{eq}(1+S-\lambda_c) - s_{eq}^2\Big) = 0, \end{aligned} \tag{B.3}$$

which we can use to obtain approximations for $s_{eq}(S,\lambda_c,\sigma)$, or $s_{eq}(S)$ for simplicity.
However, first we take a closer look at equation (B.2). First thing to note is that for all S, $s_{eq}(S) < S$, since $\lambda_c\langle\theta\rangle(s_{eq}(S),\sigma) > 0$ for all values of s_{eq}. Second, if $\lambda_c \ll 1$ then we have that $s_{eq}(S) \approx S$ because $\langle\theta\rangle \in [0,1]$ for all values of s_{eq}. This implies that for $\lambda_c \ll 1$ we know $\langle\theta\rangle(S)$ because we know $\langle\theta\rangle(s_{eq})$. Furthermore, if $\lambda_c \gg 1$ we have three seperate S regions depending on the relative size of s_{eq} compared to $\langle\theta\rangle$. The first region defined by $\lambda_c\langle\theta\rangle \gg s_{eq}$ has from mass conservation that $\langle\theta\rangle(s_{eq}(S),\sigma) = \frac{S}{\lambda_c}$, so $\langle\theta\rangle$ grows linearly. The second and third have $\lambda_c\langle\theta\rangle \ll s_{eq}$. In the second region we have $\frac{s_{eq}}{S} = 1 - \frac{\lambda_c}{S}\langle\theta\rangle \approx 1$ because $\langle\theta\rangle \in [0,1]$ and $S > s_{eq} \gg \lambda_c$. So we have $s_{eq}(S) \approx S$. In the third region we have $S \ll 1$. From equation (B.1) we see that if $s_{eq} \ll 1$, we have $\langle\theta\rangle \approx s_{eq}\frac{\sigma}{q}$. This

implies that $s_{eq}(S) \approx \frac{S}{1+\frac{\lambda_c \sigma}{q}}$, because $s_{eq} < S \ll 1$.

The above considerations show that a transition from one region to another happens when $\lambda_c \langle \theta \rangle = s_{eq}$. This defines two crossover points since we have three different regions. Furthermore, from mass conservation we immediately obtain $s_{eq} = \frac{S}{2}$ at the crossover points. We may find what this value of S is by putting $s_{eq} = \frac{S}{2}$ in the defining equation for s_{eq}. The result is shown in B.1. This figure shows for $\sigma = e^{-8}$ what the two crossover values for S, $S_{c,1}$ and $S_{c,2}$, are and also what S_n is as a function of λ_c. Below we will explain what S_n is. As explained above we expect for $S \gg S_{c,2}$ and for $S \ll S_{c,1}$ that $s_{eq} \approx S$. For $S_{c,1} \ll S \ll S_{c,2}$ we expect to find linear growth of $\langle \theta \rangle$. From this figure we see that it appears that $S_{c,2} \approx 2\lambda_c$. The validity of this relation may be shown by putting $s_{eq} = \frac{S}{2}$ in equation (B.3) and by considering only the leading terms in S for $S \gg 1$ and $\lambda_c \gg 1$. This gives:

$$\frac{S}{2}(S - \lambda_c) - \frac{S^2}{4} = 0,$$

from which we find indeed $S = 2\lambda_c$. Also, we see that $S_{c,1}$ disappears if λ_c becomes very large. This is to be expected because if $\frac{\lambda_c \sigma}{q} \gg 1$, then we have in the third region that $s_{eq}(S) = \frac{S}{\frac{\lambda_c \sigma}{q}} \ll S$. This implies that the third region becomes practically non-existent and therefore does also $S_{c,1}$ go to zero.

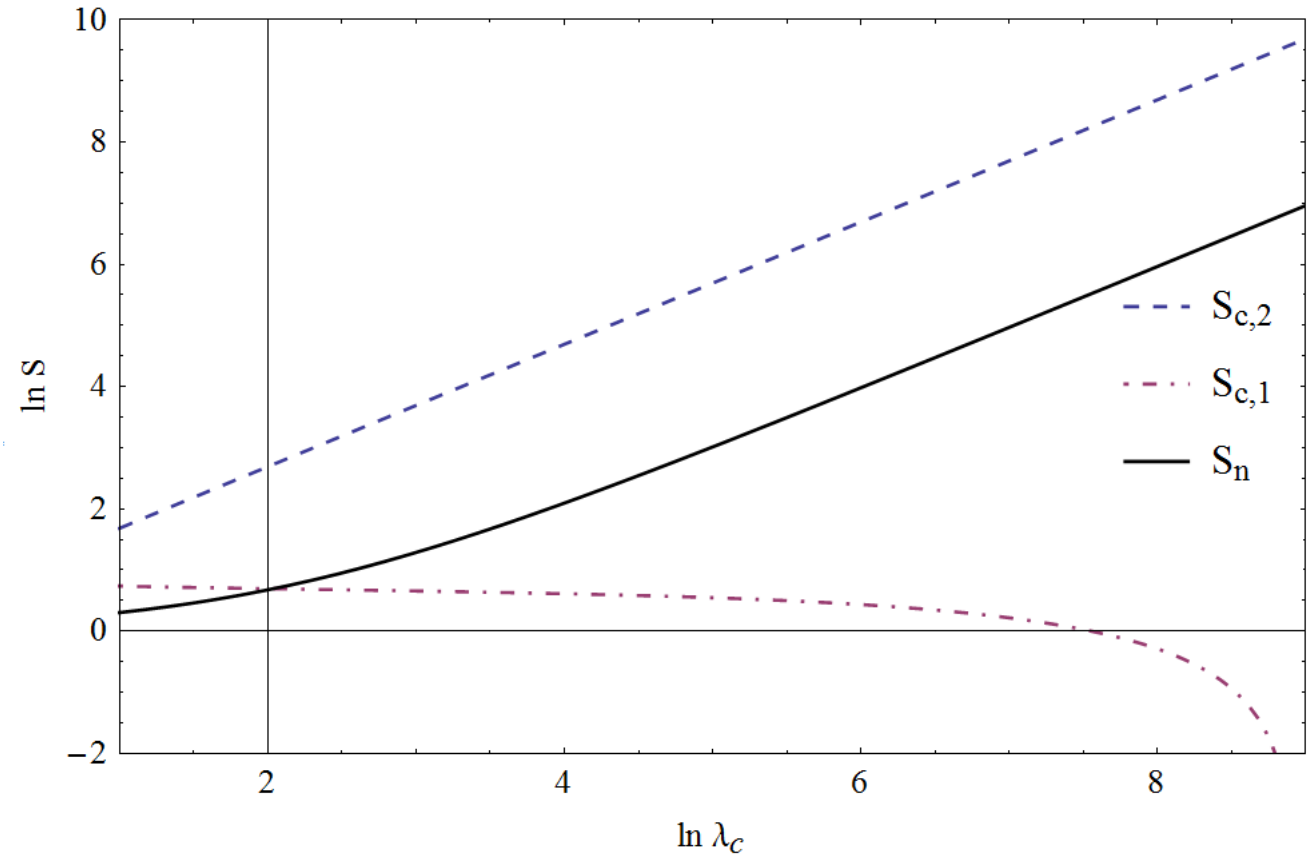

Figure B.1: The different regimes where $s_{eq}(S)$ can be in for $\sigma = e^{-8}$ and $q = 51$. For a given value of λ_c two crossover points, $S_{c,1}$ and $S_{c,2}$, can be distinguished. For $S > S_{c,2}$ and $S < S_{c,1}$ we have respectively $s_{eq}(S) \approx S$ and $s_{eq}(S) \approx \frac{S}{1+\frac{\lambda_c \sigma}{q}}$. For $S_{c,1} < S < S_{c,2}$ does $\langle \theta \rangle$ depend linearly on S. The value of S_n gives the value of S around which $\langle \theta \rangle$ can be approximated. If S_n is in the linear regime the approximation of $\langle \theta \rangle$ in section B.3 is expected to be accurate.

B.1 Large $s_{eq}(S)$

For large s_{eq}, or s_{eq}, we assume that $\sigma s_{eq}^{q+1} \gg s_{eq}^3$ such that we have

$$\lambda_c \frac{q+1}{q} - S + s_{eq}(1 + S - \lambda_c) - s_{eq}^2 = 0. \tag{B.4}$$

The solutions are given by

$$s_{eq,\pm} = \frac{1 + S - \lambda_c}{2}\Big(1 \pm \sqrt{1 - 2\frac{S - \lambda_c}{(1 + S - \lambda_c)^2}}\Big), \tag{B.5}$$

where we assumed that $q \gg 1$. The minus solution is unphysical. In figure B.2 we show the validity of the approximation. This figure shows that the approximation breaks down around the point $S = S_{c,2}$.

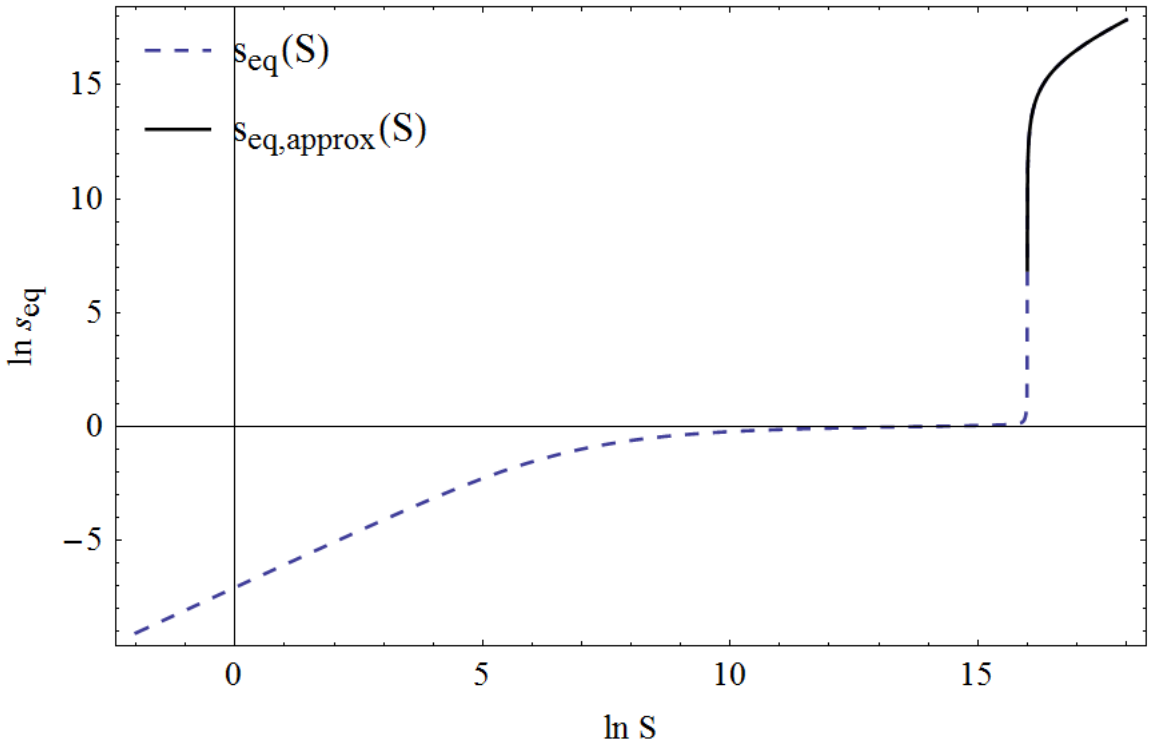

Figure B.2: For $S > S_{c,2}$ we can solve for $s_{eq}(S)$ by assuming $\sigma s_{eq}^{q+1} \gg s_{eq}^3$ and considering the dominant terms in the mass conservation equation. The parameter values are $\lambda_c = e^{16}$, $\sigma = e^{-5}$ and $q = 51$. After the steep descent does the approximation break down.

B.2 Small $s_{eq}(S)$

For small s_{eq} we assume that $\sigma s_{eq}^{q+1} \ll \sigma s_{eq}^3$ and we are left with a cubic polynomial which may be solved exactly. The result is shown in figure B.3. This figure shows that the approximation breaks down somewhere around $S = S_{c,1}$.

B.3 Around $s_{eq}(S) = 1$

For high and low values of $s_{eq}(S)$ the approximation breaks down around the crossover points. This implies we have a good approximation for region two and three but not for region one. To get a grib on

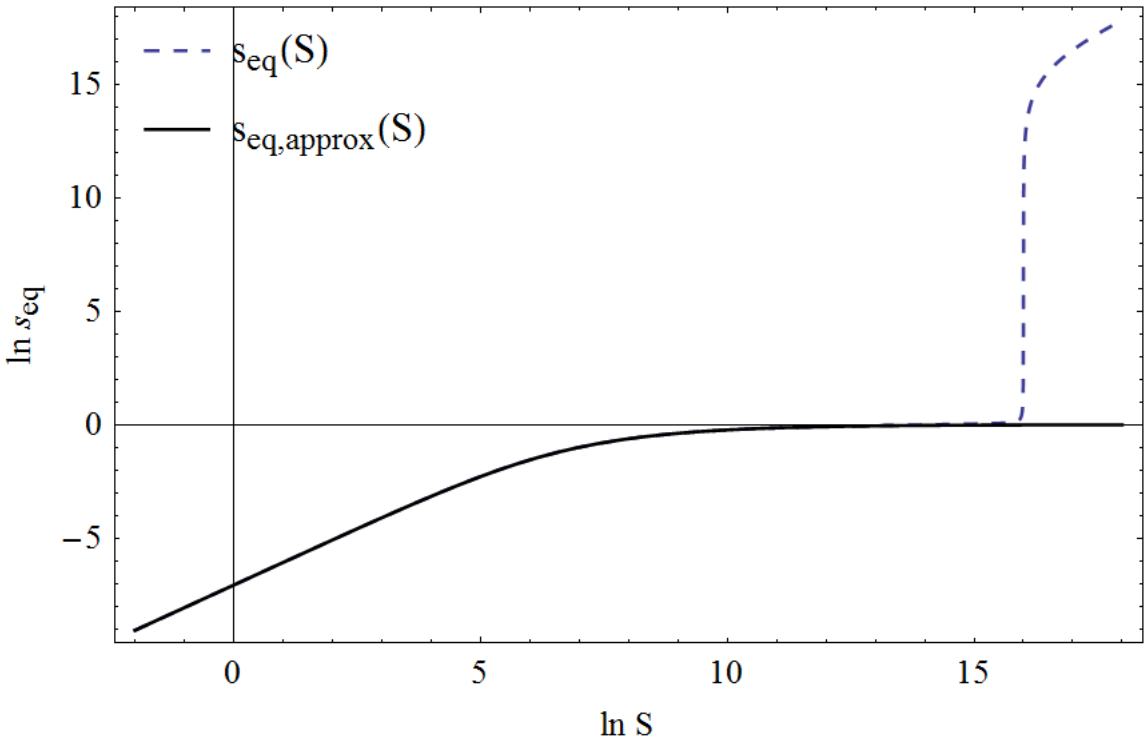

Figure B.3: For $S < S_{c,1}$ we can solve for $s_{eq}(S)$ by assuming $\sigma s_{eq}^{q+1} \ll s_{eq}^3$ and considering the dominant terms in the mass conservation equation. The parameter values are $\lambda_c = e^1 6$, $\sigma = e^{-5}$ and $q = 51$. The approximation holds up to $s_{eq} \approx 1$.

region one we use the fact that $s_{eq} = 1$ is a special point and in fact a critical point if $q \to \infty$. In the large q limit we have a phasetransition at $s_{eq} = 1$. We call this point the nuclation point, because then $s_{eq} = s_{eq} = e^{\mu_P - \epsilon - g} = 1$. This means that the energy cost to get a protein out of the solution is equal to the energy gain a protein gets upon binding. Therefore, we write $s_{eq}(S) = 1 + \delta(S)$, with $\delta(S)$ some function yet to be determined. Putting this in equation (B.3) we find

$$\begin{aligned} 0 &= \left(-\frac{\lambda_c}{2}\sigma(1+q) - q\sigma - 1 + (q\sigma+1)S\right) + \\ &\quad \delta\left(\sigma\left[\frac{1}{6}\lambda_c(q-1)(q+1) + \frac{1}{2}(q+1)q(-\lambda_c + S - 1) - q\right] - 1\right) + O(\delta^2), \end{aligned} \tag{B.6}$$

from which we obtain by neglecting all orders higher than one

$$s_{eq}(S) = 1 + \frac{\frac{\lambda_c}{2}\sigma(1+q) + q\sigma + 1 - (q\sigma+1)S}{\sigma\left[\frac{1}{6}\lambda_c(q-1)(q+1) + \frac{1}{2}(q+1)q(-\lambda_c + S - 1) - q\right] - 1}. \tag{B.7}$$

With this expression we may calculate approximatly the value of S_n at which we have $s_{eq}(S_n) = 1$. The result is

$$S_n = 1 + \lambda_c \langle\theta\rangle(1, \sigma),$$

where we used that $\langle\theta\rangle(1,\sigma) = \frac{\sigma}{2}\frac{1+q}{1+\sigma q}$. Remarkably, this result is exact because we can use mass conservation to obtain S exactly if $s_{eq} = 1$. This approximation only holds for values of s_{eq} close to one and thus for S close to $S = S_n$. In figure B.4 we see that this is indeed the case. Nevertheless, this approximation may be used to approximate $\langle\theta\rangle(s_{eq}(S), \sigma)$ in some parameter region. If we write

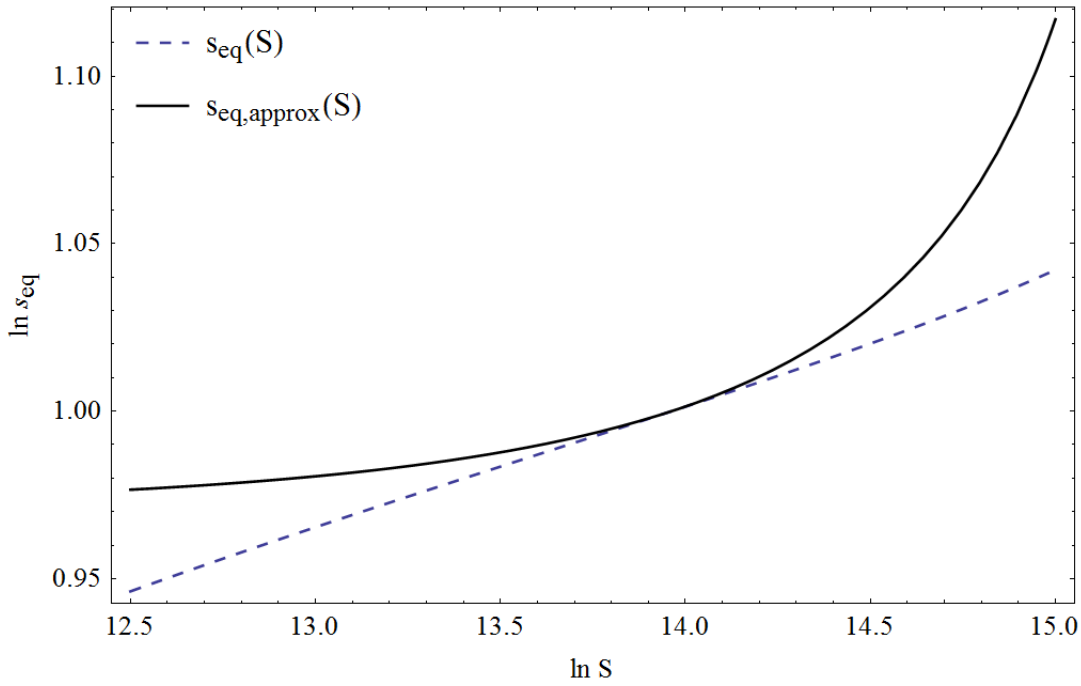

Figure B.4: In first order approximation we can find $s_{eq}(S)$ for S around S_n. This figure shows that the approximation is tangent to the exact curve. If we use $s_{eq,approx}(S)$ to find $\langle\theta\rangle$ up to first order in S and if $\langle\theta\rangle$ behaves linearly at $S = S_n$, we have a valid approximation of the linear behaviour of $\langle\theta\rangle$.

$\langle\theta\rangle(s_{eq}(S_n + (S - S_n)), \sigma)$ and taylor expand for small $S - S_n$ we obtain up to first order

$$\langle\theta\rangle(s_{eq}(S),\sigma) = \langle\theta\rangle(1,\sigma) + (S - S_n)\frac{1}{\lambda_c + a(\sigma,q)} + O((S - S_n)^2), \tag{B.8}$$

$$= \frac{1}{\lambda_c}\left(S_n(1 - \frac{1}{1 + \frac{a(\sigma,q)}{\lambda_c}}) - 1\right) + \frac{S}{\lambda_c}\frac{1}{1 + \frac{a(\sigma,q)}{\lambda_c}} + O((S - S_n)^2), \tag{B.9}$$

where $a(\sigma, q) \equiv \frac{12+24q\sigma+12q^2\sigma^2}{\sigma(q^3\sigma+4q^2-q\sigma+6q+2)}$. This expression may be used because the exact $\langle\theta\rangle(s_{eq}(S), \sigma)$ has a range of S values where it grows linearly with S. This implies that if $S_{c,1} \ll S_n \ll S_{c,2}$ we expect our approximation of $\langle\theta\rangle(s_{eq}(S), \sigma)$ to hold for the whole of region one, that is, $S_{c,1} \ll S \ll S_{c,2}$. This is because we already derived that in region one we have approximately $\langle\theta\rangle(s_{eq}(S), \sigma) \approx \frac{S}{\lambda_c}$. From our approximation we see that indeed if $\lambda_c \gg a(\sigma, q)$, we have the approximate result derived earlier.
The validity of this approximation is wholly dependent on whether S_n is indeed within region one. Because if it is not then our linear approximation of $\langle\theta\rangle(s_{eq}(S), \sigma)$ will not be very good. To know whether S_n is within the region one we refer to figure B.1. There is a value for λ_c where $S_n = S_{c,1}$. Since S_n is determined for $s_{eq} = 1$ and we also have that $S_{c,1} = 2s_{eq}$ we must conclude that $S_n = S_{c,1}$ if $S_n = S_{c,1} = 2$. When $S_n = 2$ we say that $\lambda_c = \lambda_{c,min}$ and this minimum value may be calculated from the expression for S_n to be

$$\lambda_{c,min} = \frac{1}{\langle\theta\rangle(1,\sigma)} = \frac{2}{\sigma}\frac{1+q}{1+\sigma q}.$$

So for any combination of (σ, q) we know the minimum value of λ_c for which our linear approximation for $\langle\theta\rangle(s_{eq}(S), \sigma)$ can be expected to hold.

Appendix C

Analytical approximations

In section 3.2 we derived the dynamical equations for zipper assembly and in section 3.3 we showed numerical solutions of these equations. To obtain a better grip on the properties of these solutions we propose a number of approximations. These are mere propositions and in no way fully worked out approximations with well defined ranges of validity and accuracy.

C.1 Coupled equations

To find how the system behaves at early times, especially during the assembly wave as seen in figure 3.4 we use the coupled reactions approximation (CRA). This comprises that we consider all forward rates to be much greater than the backward rates such that we obtain the following set of equations

$$\frac{\partial f(0,\tau)}{\partial \tau} = -\frac{\kappa s_{eq}}{S} y(\tau) f(0,\tau), \tag{C.1}$$

$$\frac{\partial f(1,\tau)}{\partial \tau} = -\frac{s_{eq}}{S} y(\tau) f(1,\tau) + \frac{\kappa}{\sigma S} y(\tau) f(0,\tau), \tag{C.2}$$

$$\frac{\partial f(n,\tau)}{\partial \tau} = -\frac{s_{eq}}{S} y(\tau) f(n,\tau) + \frac{1}{S} y(\tau) f(n-1,\tau), \tag{C.3}$$

$$\frac{\partial f(q,\tau)}{\partial \tau} = \frac{1}{S} y(\tau) f(q-1,\tau), \tag{C.4}$$

$$\frac{dy(\tau)}{d\tau} = -\frac{\lambda}{q\Xi_{eq}} \Big[\kappa y(\tau) f(0,\tau) + \sigma \sum_{n=1}^{q-1} s_{eq}^n y(\tau) f(n,\tau) \Big]. \tag{C.5}$$

If we consider $\kappa = 1$, we can write equation (C.5) to be

$$\frac{dy(\tau)}{d\tau} = -\frac{\lambda}{q} y(\tau) \Big[1 - P(q,\tau) \Big], \tag{C.6}$$

where we used that $\sum_{n=0}^{q} P(n,\tau) = 1$. From this we see that if $P(q,\tau) << 1$, we can solve this equation. This condition simply means that the assembly wave has not yet reached $n = q$, so we specialize to the regime where the fraction of fully covered templates is negligible. The solution is

$$y(\tau) = y_0 e^{-\frac{\tau}{T}}, \tag{C.7}$$

where we defined $T \equiv \frac{q}{\lambda}$ to be the typical timescale. from which we can immediately make an estimate for the time after which this approximation does not hold anymore, τ_m. Namely, in equilibrium, when $\tau \to \infty$, we have $y(\tau) = 1$, so the above expression is certainly invalidated when it is equal to 1. This gives

$$\tau_m = T \ln y_0. \tag{C.8}$$

So we have that the validity of the approximation is dependent on the starting conditions. Below we will assume that as a starting condition we have $P(0,0) = 1$ and thus that $s(0) = S$. This implies that $y_0 = \frac{S}{s_{eq}}$. As can be seen in figure 3.2 we have that $y_0 \equiv \frac{S}{s_{eq}}$ is large when S is much larger than unity and $\lambda = 1.5$. So neither in excess nor shortage of proteins we expect this approximation to hold well. Now we can solve equation (C.1) to be

$$f(0,\tau) = f(0,0)\exp[-\frac{s_{eq}}{S}y_0 T(1-e^{-\frac{\tau}{T}})].$$

To solve the other equations we assume that initially we have only empty templates, which is quite a reasonable assumption with regard to a real experiment. Therefore we have $f(n,0) = 0$ for $n > 0$ and $y_0\frac{s_{eq}}{S} = 1$. If we define an auxilary function $g(\tau)$ as

$$g(\tau) = T(1-e^{-\frac{\tau}{T}}), \tag{C.9}$$

we find

$$f(n,\tau) = \frac{1}{n!}\Big(\frac{y_0}{S}\Big)^n \frac{1}{\sigma} f(0,0) g(\tau)^n e^{-g(\tau)},$$

for $0 < n < q$ and we can rewrite it to obtain

$$P(n,\tau) = \frac{1}{n!} g(\tau)^n e^{-g(\tau)}, \tag{C.10}$$

for $0 \leq n < q$. Obviously, if $q \to \infty$ we can check that $\sum_{n=0}^{q-1} P(n,\tau) = 1$. To see what the maximum normalisation error is when q is finite we calculate

$$1 = e^g e^{-g} = e^{-g}\sum_{n=0}^{\infty}\frac{1}{n!}g^n = e^{-g}\Big(\sum_{n=0}^{q-1}\frac{1}{n!}g^n + R_{q-1}\Big),$$

where R_{q-1} is the remainder. The remainder is given by $R_{q-1} = \frac{e^x}{q!}g^q$ with $x \in [0,g]$. If we allow the error to be maximally α, so $\alpha = e^{-g}R_{q-1}$, and we take $x = g$ such that the remainder estimation is maximum, then we can find a very safe upper limit on the time up to which the approximation is valid. This limit is given by

$$\tau_s = -T\ln\Big[1 - \frac{(\alpha q!)^{\frac{1}{q}}}{T}\Big]. \tag{C.11}$$

Another interesting thing is how the top of the wave behaves in time. The top can be found by considering $\partial_\tau P(n,\tau)|_{\tau=\tau^*} = 0$ and gives

$$n = g(\tau^*). \tag{C.12}$$

So for a given n this equation gives the time at which this $P(n,\tau)$ is maximum, that is, when the wave passes this value of n. For small times $\tau \ll T$ we have that $n = \tau^*$, so we find classic wavelike behaviour. Furthermore, the wave actually slows down

$$\frac{dn}{d\tau^*} = e^{-\frac{\tau}{T}}.$$

Nevertheless, this will only be significant is T is sufficiently small.
The above considerations show that, at least, qualitatively the CRA can account for the wavelike behaviour which is seen figure 3.4. Nevertheless, it remains a question for which parameter values the approximation holds well. What seems to be the case is that for large values of S and values of λ around unity the approximation holds best. This is to be expected because, as noted above, y_0 is largest in this parameter regime.
Another interesting feature of our result of the CRA is that σ does not enter wherefore the dynamics are universal in τ.

We have seen that the CRA does not hold for the entire relaxation process. Namely, it breaks down when $P(q,\tau)$ becomes significant. This brings us to the interesting question of how $P(q,\tau)$ behaves at later times, so how the fraction of fully covered templates behaves as a function of time. This question will be covered as much as possible in the following two sections.

C.2 Steady state

There are times at which the $P(n,\tau)$ of the intermediate states, so $0 < n < q$, do not vary strongly while $P(q,\tau)$ grows strongly. At these times one may invoke the steady state approximation (SSA) to find how $P(q,\tau)$ behaves. This approximation asserts that $\partial_\tau f(n,\tau) = 0$ for $0 < n < q$. Furthermore, for convenience we choose for these values of n: $f(n,\tau) = 1$. From equation (3.23) and (3.26) we then obtain

$$\partial_\tau f(q,\tau) = \frac{1}{S}\Big(y(\tau) - f(q,\tau)\Big), \tag{C.13}$$

$$s(\tau) + \lambda_c\langle\theta\rangle(\tau) = s_{eq} + \lambda_c\langle\theta\rangle_{eq}, \tag{C.14}$$

where we used the fact that equation (3.26) holds for any time τ and thus also in equilibrium. Using that $y(\tau) \equiv \frac{s(\tau)}{s_{eq}}$ we obtain from equation (C.14)

$$y(\tau) = 1 + \frac{\lambda_c}{s_{eq}}(\langle\theta\rangle_{eq} - \langle\theta\rangle(\tau)) = 1 + \lambda_c P_{eq}(q-1)(1 - f(q,\tau)), \tag{C.15}$$

with which we find for equation (C.13)

$$\partial_\tau f(q,\tau) = \frac{K_1}{S}\Big(1 - f(q,\tau)\Big).$$

Here we defined $K_1 \equiv 1 + \lambda_c P_{eq}(q-1)$. The solution, while assuming that $f(q,0) = 0$, is

$$f(q,\tau) = 1 - \exp[-\frac{K_1}{S}\tau],$$

from which we find using $\sum_{n=0}^{q} P(n,\tau) = 1$ and equation (C.15)

$$P(q,\tau) = P_{eq}(q)\Big(1 - \exp[-\frac{K_1}{S}\tau]\Big), \tag{C.16}$$

$$P(0,\tau) = P_{eq}(0) + P_{eq}(q)\exp[-\frac{K_1}{S}\tau], \tag{C.17}$$

$$y(\tau) = 1 + (K_1 - 1)\exp[-\frac{K_1}{S}\tau]. \tag{C.18}$$

From these results we see that greater λ_c, so more templates, enhances the growth of fully covered templates. This is to be expected because a higher template concentration gives for fixed S a lower value of $P_{eq}(q)$. Thus, the covering will take less time. Also, for higher K_1 we have that $y(\tau)$ is higher at $\tau = 0$ and thus that there is a greater assembly 'force' from the protein concentration. On the other hand, for fixed λ_c and increasing S we have that $P_{eq}(q)$ is increased. Thus it is to be expected that it takes longer to fill the templates. Nevertheless, the exact range of validity of this approximation is still to be examined and this will not be covered here.

C.3 Pre-equilibrium

In the previous section we found how $P(q,\tau)$ shows exponential behaviour for late times. Nevertheless, it is observed in numerical simulations that $P(q,\tau)$ typically has a sigmoid-like shape. Moreover, often $f(q-1,\tau)$ and $f(q,\tau)$ are almost equal while they increase significantly as opposed to the other probabilities. These considerations give rise to the pre-equilibrium approximation (PEA) which asserts the following approximations: $\partial_\tau f(n,\tau) = 0$ for $0 < n < q-1$ and $f(q-1,\tau) = f(q,\tau)$. Also, we assume that $f(n,\tau) = 1$ for $0 < n < q-1$. With these approximations we obtain from equation (3.23) and (3.26)

$$\partial_\tau f(q,\tau) = \frac{1}{S}\Big(y(\tau) - 1\Big)f(q,\tau), \tag{C.19}$$

$$s(\tau) + \lambda_c\langle\theta\rangle(\tau) = s_{eq} + \lambda_c\langle\theta\rangle_{eq}. \tag{C.20}$$

From equation (C.20) we obtain

$$y(\tau) - 1 = \frac{\lambda_c}{s_{eq}}(\langle\theta\rangle_{eq} - \langle\theta\rangle(\tau)) = K_2(1 - f(q,\tau)), \tag{C.21}$$

where we defined $K_2 \equiv \lambda_c P_{eq}(q-2)(1+s_{eq})$. This gives for equation (C.19)

$$\partial_\tau f(q,\tau) = \frac{K_2}{S}\Big(1 - f(q,\tau)\Big) f(q,\tau),$$

which can be solved, using $A \equiv \frac{1}{f(q,0)} - 1$, to find

$$f(q,\tau) = \frac{1}{1 + A\exp[-\frac{K_2}{S}\tau]}.$$

When using probability conservation and equation (C.21) one obtains as results

$$P(q,\tau) = \frac{P_{eq}(q)}{1 + A\exp[-\frac{K_2}{S}\tau]}, \tag{C.22}$$

$$P(q-1,\tau) = \frac{P_{eq}(q-1)}{1 + A\exp[-\frac{K_2}{S}\tau]}, \tag{C.23}$$

$$P(0,\tau) = P_{eq}(0) + \Big(P_{eq}(q-1) + P_{eq}(q)\Big)\frac{A\exp[-\frac{K_2}{S}\tau]}{1 + A\exp[-\frac{K_2}{S}\tau]}, \tag{C.24}$$

$$y(\tau) = 1 + K_2\frac{A\exp[-\frac{K_2}{S}\tau]}{1 + A\exp[-\frac{K_2}{S}\tau]}. \tag{C.25}$$

From these equations we see qualitatively the same behaviour as for the SSA: again for higher λ_c the relaxation is faster while for higher S it is slower. Also, we see that $P(q,\tau)$ is an S-shaped curve. For late times, when $e^{-\frac{K}{S}\tau} \ll 1$, we have the same exponential behaviour as for the SSA but with a different typical timescale since in general $K_1 \neq K_2$. But if $P_{eq}(q-1)\lambda_c \gg 1$ and $s_{eq} \gg 1$ then $K_1 \approx K_2$. So for large template concentration and excess protein concentration - such that $\lambda \ll 1$ and $s_{eq} \approx S$ - we have that the SSA and PEA coincide for late times. Nevertheless, to have a sound understanding of the range of validity of this approximation, more research is required.

C.4 Transition state theory

In section 3.2 we have seen how the dynamical equations could be made dimensionless by scaling the time to k_+. This allowed to calculated all dynamical quantities as a function of dimensionless time $\tau = k_+ t$. Nevertheless, real experiments are a function of real time t and thus an estimation for k_+ would be helpful for knowing the timescale of an experiment. In this section we use transition state theory (TST) to find an expression for k_+ in case of the zipper model. This derivation is by no means complete and is, like the preceding approximation sections, meant to provide inspiration for future research. First we will sketch the picture of a protein and a template forming a transition state (TS) and then we will give two routes to calculate k_+.

An unbound protein has three translational degrees of motion - for we assume it to have ideal gas behaviour in the solution - and vibrational plus rotational internal degrees of freedom. When the TS if formed we picture the protein to move in a space close to the template. This gives rise to new, more restricted, translational degrees of freedom for the protein. Also, both the internal degrees of freedom of the protein and the template may change in the TS. With this picture in mind we can write the forward rate of the n-th reaction of equation (3.4) as

$$v_n P^{\#}(n),$$

with v_n the attempt frequency of a protein in the TS to bind to the template and $P^{\#}(n)$ the (non-normalised) probability that a protein and a template with n proteins already attached to it form a TS. To proceed we assume that the transition reaction to bring about the TS occurs much faster than the binding of a protein from the TS to the template. This implies that the transition reaction is in equilibrium and thus that

$$k^{\#}_{+,n}\rho_P P(n) = k^{\#}_{+,n} P^{\#}(n),$$

which gives

$$P^{\#}(n) = K^{\#}_n \rho_P P(n).$$

Furthermore, we have for the protein, the template and the TS that their chemical potential in a certain state is given by

$$\mu = -k_B T \ln \frac{Z}{N},$$

with Z the partition function of a given reactant in a certain state and N the number of that reactant in this state present in the solution. In chemical equilibrium the chemical potential of the constituents in the unbound and in the TS should be equal and thus we have

$$K^{\#}_n = \frac{k^{\#}_{+,n}}{k^{\#}_{-,n}} = \frac{P^{\#}(n)}{\rho_P P(n)} = \frac{Z^{\#}(n)}{Z(n)}, \tag{C.26}$$

with $Z^{\#}(n)$ the combined partition function of the protein and the template in the TS and $Z(n)$ being the combined partition function in the unbound state. So by knowing how the partition function of the protein-template complex changes by going from the unbound state to the TS, we know what the TS equilibrium constant is. Once we know $K^{\#}_n$ we can also know $k^{+}_n = v_n K^{\#}_n$. Now we will propose two different ways of calculating k^{+}_n which, remarkably, give almost the same result.

C.4.1 Route 1

On the first route we assume that the protein moves on a two dimensional cylindrical surface of which the axis is given by the template. This gives rise to two translational degrees of freedom of the protein in the TS. One degree is parallel to the template and has length $a(q-n)$ with a the distance between two neighbouring binding sites. The other degree is on a circle with circumference L_{cir} of which the diameter is determined by the thickness of the template. This gives the following for the TS equilibrium constant

$$K^{\#}_n = \frac{\frac{a(q-n)}{\lambda_{th}}\frac{L_{cir}}{\lambda_{th}}}{\frac{V}{\lambda^3_{th}}} S_0(T) = \lambda_{th}\frac{a(q-n)L_{cir}}{V} S_0(T),$$

with V the volume of the solution, $\lambda_{th} \equiv \frac{h}{\sqrt{2\pi m_P k_B T}}$ the thermal wavelength and $S_0(T)$ the sticking coefficient which accounts for the change in internal degrees of freedom for both the protein and the template by going from the unbound state into the TS. Finally, since we are considering the zipper model the protein can only go from the TS to the bound state on the template when it is at the binding site next to the last bound protein. This implies a characteristic time equal to v_n^{-1} in which the protein gains one attempt to bind. This should approximately be given by

$$\frac{1}{v_n} = \frac{a(q-n)}{\sqrt{\langle v^2 \rangle}},$$

with $\langle v^2 \rangle$ the average squared velocity in the transition space. Since this space is two dimensional we have from the equipartition theorem that

$$2\frac{1}{2}k_B T = \frac{m_P}{2}\langle v^2 \rangle,$$

with m_P the mass of the protein. So we find

$$v_n = \frac{h}{\lambda_{th}\sqrt{\pi} m_P a(q-n)},$$

and finally

$$v_n P^{\#}(n) = \frac{h}{\sqrt{\pi} m_P} \frac{L_{cir}}{V} S_0(T) \rho_P P(n) \equiv k_n^+ \rho_P P(n). \tag{C.27}$$

This forward rate turns out to be independent of n which means that the number of already bound proteins does not influence the adsorption rate of unbound proteins. This is exactly what we already assumed in the derivation of the dynamical equations.

C.4.2 Route 2

On the second route we assume the proteins not to be moving in a two dimensional cylindrical space but in a one dimensional circular space around the binding site of the template where the protein can bind. Also, we assume that the restriction of the movement of the protein implies a vibrational mode perpendicular to the template and the circle with a frequency ν_n. The reason for this is that we assume the protein to be in a local energy minimum which causes the protein to vibrate perpendicular to a point on the circle. Furthermore, this vibration will bring the protein closest to the template with a frequency ν_n wherefore we assume that the attempt frequency is equal to ν_n, so $v_n = \nu_n$. This then gives

$$K_n^{\#} = \frac{k_B T}{h\nu_n} \frac{\frac{L_{cir}}{\lambda_{th}}}{\frac{V}{\lambda_{th}^3}} S_0(T) = \lambda_{th}^2 \frac{L_{cir}}{V} \frac{k_B T}{h\nu_n} S_0(T),$$

where $\frac{k_B T}{h\nu_n}$ is the partition function of the vibrational mode. Finally, we find

$$v_n P^{\#}(n) = v_n K_n^{\#} \rho_P P(n) = \frac{k_B T}{h} \lambda_{th}^2 \frac{L_{cir}}{V} \rho_P P(n) = \frac{h}{2\pi m_P} \frac{L_{cir}}{V} \rho_P P(n) \equiv k_n^+ \rho_P P(n). \tag{C.28}$$

So up to a factor $\frac{1}{2\sqrt{\pi}}$ we find the same result for k_n^+ on both routes. This factor could possibly be caused by, on route 1, simply taking $a(q-n)$ to be the typical length which the protein should travel upon binding. Since the protein moves on a cylindrical surface the typical distance could well be different.

Bibliography

[1] S. Modrow, D. Falke, U. Truyen, and H. Schätzl, *Molecular Virology*. Berlin Heidelberg: Springer-Verlag, 2013.

[2] M. Beijerinck, "About a contagium vivum fluidum as the cause of the mottling of tobacco leaves [Over een contagium vivum fluidum als oorzaak van de vlekziekte der tabaksbladen]," *Verhandelingen der Koninklijke Nederlandse Akademie van Wetenschappen*, vol. 65, pp. 3–21, 1898.

[3] G. Stubbs and A. Kendall, "Helical viruses," in *Viral Molecular Machines* (M. Rossmann and V. Rao, eds.), New York Dordrecht Heidelberg London: Springer, 2012.

[4] E. Norrby, "Nobel prizes and the emerging virus concept," *Archives of Virology*, vol. 153, pp. 1109–1123, 2008.

[5] O. U. Press, "Oxford dictionaries," 2015. [Online; accessed 9-June-2015].

[6] M. Mateu, "Introduction: The structural basis of virus function," in *Structure and Physics of Viruses* (M. Mateu, ed.), New York Dordrecht Heidelberg London: Springer, 2013.

[7] D. J. Kraft, W. K. Kegel, and P. van der Schoot, "A Kinetic Zipper Model and the Assembly of Tobacco Mosaic Virus," *Biophysical Journal*, vol. 102, pp. 2845–2855, June 20 2012.

[8] J. Castón and J. Carrascosa, "The basic architecture of viruses," in *Structure and Physics of Viruses* (M. Mateu, ed.), New York Dordrecht Heidelberg London: Springer, 2013.

[9] A. Bittner, M. G. J.M. Alonso, and C. Wege, "Nanoscale science and technology with plant viruses and bacteriophages," in *Structure and Physics of Viruses* (M. Mateu, ed.), New York Dordrecht Heidelberg London: Springer, 2013.

[10] T. Doll, S. Raman, R. Dey, and P. Burkhard, "Nanoscale assemblies and their biomedical applications," *Journal of the Royal Society: Interface*, vol. 10, p. 20120740, 2012.

[11] Q. Chenn and H. Lai, "Plant-derived virus-like particles as vaccines," *Human Vaccines & Immunotherapeutics*, vol. 9, no. 1, pp. 26–49, 2013.

[12] J. Chroboczek, I. Szurgot, and E. Szolajska, "Virus-like particles as vaccine," *Acta Biochemica Polonica*, vol. 61, no. 3, pp. 531–539, 2014.

[13] A. Hernandez-Garcia, M. Werten, M. Stuart, F. de Wolf, and R. de Vries, "Coating of single dna molecules by genetically engineered protein diblock copolymers," *Small*, vol. 8, no. 22, pp. 3491–3501, 2012.

[14] P. Nestola, C. Peixoto, R. Silva, P. Alves, J. Mota, and M. Carrondo, "Improved virus purification processes for vaccines and gene therapy," *Biotechnology & Bioengineering*, vol. 112, pp. 843–857, 2015.

[15] U. Unzueta, P. Saccardo, J. Domingo-Espín, J. Cedano, O. Conchillo-Solé, E. García-Fruitós, M. Céspedes, J. Corchero, X. Daura, R. Mangues, N. Ferrer-Miralles, A. Villaverde, and E. Vázquez, "Sheltering dna in self-organizing, protein-only nano-shells as artificial viruses for gene delivery," *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 10, no. 3, pp. 535 – 541, 2014.

[16] A. Naskalska and K. Pyrć, "Virus like particles as immunogens and universal nanocarriers," *Polish Journal of Microbiology*, vol. 64, no. 1, pp. 3–13, 2015.

[17] G. Destito, A. Schneemann, and M. Manchester, "Biomedical nanotechnology using virus-based nanoparticles," in *Viruses and Nanotechnology* (M. Manchester and N. F. Steinmetz, eds.), Berlin Heidelberg: Springer-Verlag, 2009.

[18] A. Bittner, "Biomolecular rods and tubes in nanotechnology," *Naturwissenschaften*, vol. 92, pp. 51–64, 2005.

[19] J. Zhou, C. Soto, M. Chen, M. Bruckman, M. Moore, E. Barry, B. Ratna, P. Pehrsson, B. Spies, and T. Confer, "Biotemplating rod-like viruses for the synthesis of copper nanorods and nanowires," *Journal of Nanobiotechnology*, vol. 10, no. 1, p. 18, 2012.

[20] M. Knez, M. Sumser, A. Bittner, C. Wege, H. Jeske, T. Martin, and K. Kern, "Spatially selective nucleation of metal clusters on the tobacco mosaic," *Advanced Functional Materials*, vol. 14, pp. 116–124, February 2004.

[21] A. Hernandez-Garcia, D. J. Kraft, A. F. Janssen, P. H. Bomans, N. A. Sommerdijk, D. M. Thies-Weesie, M. E. Favretto, R. Brock, F. A. de Wolf, M. W. Werten, P. van der Schoot, M. C. Stuart, and R. de Vries, "Design and self-assembly of simple coat proteins for artificial viruses," *Nature Nanotechnology letters*, vol. 9, pp. 698–702, August 24 2014.

[22] P. van der Schoot and R. Zandi, "Kinetic theory of virus capsid assembly," *Physical Biology*, vol. 4, pp. 296–304, 2007.

[23] C. Kittel, "Phase Transition of a Molecular Zipper," *American Journal of Physics*, vol. 37, no. 9, pp. 917–&, 1969.

[24] H. Shigematsu, "Asymptotic behavior of fluctuations for the 1d ising model in zero-temperature limit," *Journal of Statistical Physics*, vol. 71, pp. 5–6, 1993.

[25] T. Antal, M. Droz, and Z. Rácz, "Probability distribution of magnetization in the one-dimensional ising model: effects of boundary conditions," *Journal of Physics A: Mathematical and General*, vol. 37, pp. 1465–1478, 2004.