Stochastic modelling in molecular biology: a probabilistic analysis of protein polymerisation and telomere shortening

Sarah Eugene

STOCHASTIC MODELLING IN MOLECULAR BIOLOGY: A PROBABILISTIC ANALYSIS OF PROTEIN POLYMERISATION AND TELOMERE SHORTENING

THÈSE

présentée pour obtenir le titre de

DOCTEUR

de l'Université Pierre et Marie Curie, Paris VI

École doctorale : Sciences Mathématiques de Paris Centre

Spécialité : Mathématiques Appliquées

par

Sarah Eugène

soutenue le 30 Septembre 2016 devant le jury composé de :

Directeurs:		
Marie Doumic	Ingénieur en Chef des Ponts	Inria Paris Rocquencourt
Philippe Robert	Directeur de recherche	Inria Paris Rocquencourt
Rapporteurs:		
Loïc Chaumont	Professeur	Université d'Angers
Peter Olofsson	Professeur	Trinity University
Examinateurs:		
Amaury Lambert	Professeur	Collège de France & UPMC
Marek Kimmel	Professeur	Rice University
Human Rezaei	Directeur de Recherche	INRA

Remerciements

Je remercie d'abord mes directeurs de thèse, Philippe Robert et Marie Doumic, sans qui cette thèse n'aurait pas existé. Merci pour votre encadrement, mais aussi pour vos conseils et votre patience. Je remercie aussi nos collaborateurs biologistes, Human Rezaei, Wei-Feng Xu et Zhou Xu, pour vos précieuses données.

Merci aussi à Loïc Chaumont et Peter Olofsson, d'avoir accepté de rapporter ma thèse, et ce, dans les délais. Merci aux examinateurs, Amaury Lambert, Marek Kimmel et Human Rezaei, d'avoir fait le déplacement pour être là aujourd'hui.

Je remercie ensuite toutes les autres personnes que j'ai pu rencontrer dans des laboratoires de mathématiques ou de biologie durant ces trois années. Je ne me mouille pas trop en remerciant tout le monde, de peur d'en oublier, mais je dois citer quelques personnes en particulier, de peur de les vexer. Je pense d'abord à Aurora dont les pauses café ont égaillé mes journées à jussieu ces trois dernières années; mais aussi à Riiinnnoo pour nos pauses café l'INRIA. Merci beaucoup aux coureurs de l'INRIA, Pauline, Victorien, Jonathan, parfois Renaud, Jacques-Henri et Xavier, de m'avoir initiée aux plaisirs de la course à pieds. Merci à Mehdi et Ryadh, pour nos déjeuners en dehors de la cantine de Jussieu. Merci aussi à l'équipe de Palestine, Brofesseur Bierre, Maman Brigitte et Papa Nicolas, et Cécile. Merci infiniment Brigitte pour tes précieux conseils et ton soutien, et merci Cécile pour nos éclats de rire. Merci à Camille, d'avoir imprimé ma thèse à la dernière minute, à Nicolas pour avoir gardé toutes mes affaires dans son bureau. Enfin une petite pensée pour mes co-bureaux du LJLL, Thibault, Maxime, Pierre et Pierre, Camille et Antoine (les inséparables) pour avoir subi mes râleries.

Je remercie aussi mes (autres) amis non mathématiciens. Ils se reconnaitront s'ils lisent ce document un jour. Mais je voudrais surtout remercier ma très chère Paula, mon soutien indéfectible, mon Pierrot, dont j'ai usé toute la patience, et enfin Uesh, mon cher Alix, qu'est-ce que j'aurais fait sans toi ?

Enfin, et plus que tout, je remercie ma famille pour m'avoir supportée ces vingt-huit dernières années, en toute circonstance, même quand j'avais tort. Merci Papa, merci Maman, d'être là, toujours, même quand je suis loin. Merci mon petit Kevin et ma petite Ambre, mes petits diables, j'espère ne pas vous avoir découragés de faire une thèse un jour. Merci enfin à ceux qui ne peuvent pas être là mais dont la pensée m'accompagne tous les jours.

Résumé

Dans cette thèse, nous proposons une analyse probabiliste de deux problèmes de biologie moléculaire dans lesquels la stochasticité joue un rôle essentiel : la polymérisation des protéines dans les maladies neurodégénératives ainsi que le raccourcissement des télomères.

L'agrégation des protéines en fibrilles amyloïdes est un important phénomène biologique associé à plusieurs maladies humaines telles que les maladies d'Alzheimer, de Huntington ou de Parkinson, ou encore l'amylose ou bien le diabète de type 2. Comme observé au cours des expériences reproduisant les petits volumes des cellules, les courbes d'évolution cinétique de l'agrégation des protéines présentent une phase de croissance exponentielle précédée d'une phase de latence extrêmement fluctuante, liée au temps de nucléation.
Après une introduction au problème de polymérisation des protéines dans le chapitre I, nous étudions dans le chapitre II les origines et les propriétés de la variabilité de ladite phase de latence ; pour ce faire, nous proposons un modèle stochastique minimal qui permet de décrire les caractéristiques principales des courbes expérimentales d'agrégation de protéines. On considère alors deux composants chimiques : les monomères et les monomères polymérisés. Au départ, seuls sont présents les monomères ; par suite, ils peuvent polymériser de deux manières différentes : soit deux monomères se rencontrent et forment deux monomères polymérisés, soit un monomère se polymérise à la suite d'une collision avec un autre monomère déjà polymérisé. Malgré son efficacité, la simplicité des hypothèses de ce modèle ne lui permet pas de rendre compte de la variabilité observée au cours des expériences. C'est pourquoi dans un second temps, au cours du chapitre III, nous complexifions ce modèle afin de prendre en compte d'autres mécanismes impliqués dans la polymérisation et qui sont susceptibles d'augmenter la variabilité du temps de nucléation. Lors de ces deux chapitres, des résultats asymptotiques incluant diverses échelles de temps sont obtenus pour les processus de Markov correspondants. Une approximation au premier et au second ordre du temps de nucléation sont obtenus à partir de ces théorèmes limites. Ces résultats reposent sur une renormalisation en temps et en espace du modèle de population, ainsi que sur un principe d'homogénéisation stochastique lié à une version modifiée d'urne d'Ehrenfest.

Dans une seconde partie, un modèle stochastique décrivant le raccourcissement des télomères est proposé. Les chromosomes des cellules eucaryotes sont raccourcis à chaque mitose à cause des mécanismes de réplication de l'ADN incapables de répliquer les extrémités du chromosome parental. Afin d'éviter une perte de l'information génétique, ces chromosomes possèdent à chaque extrémité des télomères qui n'encodent pas d'information génétique. Au fil des cycles de réplication, ces télomères sont raccourcis

jusqu'à rendre la division cellulaire impossible : la cellule entre alors en sénescence réplicative. L'objectif de ce modèle est de remonter aux caractéristiques de la distribution initiale de la taille des télomères à partir de mesures de temps de sénescence.

Abstract

This PhD dissertation proposes a stochastic analysis of two questions of molecular biology in which randomness is a key feature of the processes involved: protein polymerisation in neurodegenerative diseases on the one hand, and telomere shortening on the other hand.

Self-assembly of proteins into amyloid aggregates is an important biological phenomenon associated with human diseases such as prion diseases, Alzheimer's, Huntington's and Parkinson's disease, amyloidosis and type-2 diabetes. The kinetics of amyloid assembly show an exponential growth phase preceded by a lag phase, variable in duration, as seen in bulk experiments and experiments that mimic the small volume of the concerned cells. After an introduction to protein polymerisation in chapter I, we investigate in chapter II the origins and the properties of the observed variability in the lag phase of amyloid assembly. This variability is currently not accounted for by deterministic nucleation-dependent mechanisms. In order to tackle this issue, a stochastic minimal model is proposed, simple, but capable of describing the characteristics of amyloid growth curves. Two populations of chemical components are considered in this model: monomers and polymerised monomers. Initially, there are only monomers and from then, two possible ways of polymerising a monomer: either two monomers collide to combine into two polymerised monomers, or a monomer is polymerised by the encounter of an already polymerised monomer. However efficient, this simple model does not fully explain the variability observed in the experiments, and in chapter III, we extend it in order to take into account other relevant mechanisms of the polymerisation process that may have an impact on fluctuations. In both chapters, asymptotic results involving different time scales are obtained for the corresponding Markov processes. First and second order results for the starting instant of nucleation are derived from these limit theorems. These results rely on a scaling analysis of a population model and the proof of a stochastic averaging principle for a model related to an Ehrenfest urn model.

In the second part, a stochastic model for telomere shortening is proposed. In eukaryotic cells, chromosomes are shortened with each occurring mitosis, because the DNA polymerases are unable to replicate the chromosome down to the very end. To prevent potentially catastrophic loss of genetic information, these chromosomes are equipped with telomeres at both ends (repeated sequences that contain no genetic information). After many rounds of replication however, the telomeres are progressively nibbled to the point where the cell cannot divide anymore, a blocked state called replicative senescence. The aim of this model is to trace back to the initial distribution of telomeres from measurements of the time of senescence.

General introduction

Since the 19^{th} century, mathematical methods have been widely used in physics. The use of mathematical methods for biology is more recent. This delay is mainly due to the sheer complexity of biological systems and to the fact that fundamental discoveries in molecular biology occurred later, that is, in the second half of the 20^{th} century.

In 1944, a seminal book by Erwin Schrödinger called 'What is life' [Schrödinger, 1944] paved the way for modelling in biology, introducing the concept of 'order-from-disorder' which was predominant in biology for almost fifty years [Symonds, 1986]. For Schrödinger, life being a highly ordered system, the randomness of interactions at the microscopic level disappears at the macroscopic level because of the large number of molecules involved in the mechanisms considered. This echoes the vision that prevails in statistical physics, where the order of magnitude of the number of interacting objects is very large—comparable, for instance in the kinetic theory of gases, to the Avogadro constant ($6.02 \cdot 10^{23}$). As a result, modelling biological phenomena followed a deterministic approach, although these orders of magnitude can be much lower in biology. In addition to this inadequacy, the discovery of DNA as the molecular basis of genetic information in 1962 by Watson and Crick, resulted in a mechanistic approach in which genetic information is unequivocally translated into proteins, which are in turn responsible for the functionalities of cells. However, many experiments have revealed the stochastic nature of the expression of genes, which is at the very core of molecular biology. For instance in [Elowitz et al., 2002], the author tracked the expression of a fluorescent protein in two separate cells in order to assess their individual production level—which turned out to be different, thus shedding light on the reality of cell-to-cell variability.

Although probabilistic methods are more systematically used in the fields of evolutionary biology and population dynamics (see for instance [Lambert, 2008, Méléard and Bansaye, 2015]), recent advances in cancer and neurodegenenerative diseases research, as well as extended possibilities to collect large sets of data at the micro level, have also allowed for rigorous mathematical analyses [Friedman, 2010], leading to stochastic modelling playing an increasingly important role in molecular biology. In this thesis, we use such probabilistic methods to address two questions of this field, examined in two separate parts: protein polymerisation in the framework of neurodegenerative diseases, and telomere shortening causing replicative senescence.

References

[Bharucha-Reid, 1960] Bharucha-Reid, A. (1960). *Elements of the theory of Markov processes and their applications*. New York, McGraw-Hill,.

[Elowitz et al., 2002] Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

[Friedman, 2010] Friedman, A. (2010). What is mathematical biology and how useful is it? *Notices of the AMS*, 57(7):851–857.

[Lambert, 2008] Lambert, A. (2008). Population dynamics and random genealogies. *Stochastic Models*, 24(S1):45–163.

[Méléard and Bansaye, 2015] Méléard, S. and Bansaye, V. (2015). *Stochastic Models for Structured Populations*. Springer.

[Mode et al., 2003] Mode, C. J., Kimmel, M., and Axelrod, D. (2003). Branching processes in biology.

[Schrödinger, 1944] Schrödinger, E. (1944). *What is life?*, volume 10. Cambridge University Press.

[Symonds, 1986] Symonds, N. (1986). What is life? : Schrodinger's influence on biology. *Q. Rev. Biol.*, 61(2):221–226.

Contents

Part I

Stochastic modelling of protein polymerisation

Chapter 1

An introduction to protein polymerisation

Contents

In this part, we propose a stochastic modelling of polymerisation of proteins, also called aggregation, involved in prion and prion-like diseases. We start by introducing the biological problem, and we then give some insights on the mathematical tools we use to model these diseases stochastically.

1.1 Prion diseases and prion-like diseases

Prion diseases, also called Transmissible Spongiform Encephalopathies (TSEs) are fatal neurodegenerative diseases characterised by the abnormal aggregation of prion protein. Early stages exhibit dementia, troubles in speech and coordination of movements, and visual problems. They include Bovine Spongiform Encephalopathy (BSE, or 'mad cow' disease) in cattle, Creutzfeldt-Jakob Disease (CJD), Fatal Familial Insomnia (FFI) or Gerstmann-Sträussler-Scheinker syndrome (GSS) in humans. First description of a prion disease dates back to scrapie, a disease of sheep and goats, in 1732, inducing among other many behavioural changes, excessive scraping sensations; hence the name. In humans, the first reported prion disease was Kuru, a disease of New Guinea, which diffusion was related to funerary cannibalism. Their transmission can be either genetic (GSS, FFI, CJD), infectious (CJD) or sporadic (CJD), which made them hard to identify as a single disorder. However, in the seventies, many similarities in the central nervous systems of people infected by kuru and scrapie were found: they presented the characteristic spongiform aspect. In 1967, Griffith [Griffith, 1967] proposed a new mechanism of infection due to a protein, rather than a virus, or a viroid-like pathogen: the concept of prion was born. Experiments in [Prusiner, 1998], revealed that the infectious agent was devoid of nucleic acid, and thus confirmed this hypothesis. Stanley Prusiner proposed the term 'prion' (contraction of '**pro**teinaceous','**in**fectious' and '**on**') and received the Nobel Prize in Physiology or Medicine in 1997 for his work. The prion agent is a protein, denoted PrP, that may change conformation and become infectious. Prion protein is normally present in many species, including all mammals, in the form called PrP^{C}, for 'cellular'. Its physiological role is not well understood, but it is believed to be involved in many cellular functions. In prion diseases, PrP^{C} is converted into PrP^{Sc}, for 'scrapie', an isomeric form of PrP rich in β-sheets. These β-sheets are sticky, and cause aggregation of prion into amyloids. Amyloid plaques have been proved to be pathogenic in the case of prion diseases [Prusiner, 1998].

The mechanism of protein misfolding followed by polymerisation is common to other diseases namely Alzheimer's, Huntington's and Parkinson's disease, amyloidosis and type-2 diabetes. They share with prion diseases the presence in tissues of amyloid fibrils, even if the pathogenic role of this plaques is not as clear as in prion diseases [Jarrett and Lansbury, 1993, G.Roberts, 2016]. However, each disease has its own protein. Among amyloidogenic proteins, one can cite β-amyloid aggregates, observed in the brains of patients infected by Alzheimer's [Glenner and Wong, 1984, Murphy and LeVine, 2010] as shown in figure 1.1. One also observes β_2-microglobulin (β_2m) amyloid fibrils in individuals infected by dialysis-related amyloidosis [Goto et al., 2005].

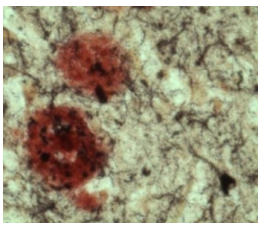


Figure 1.1: β-amyloid aggregates (red) *Credit: Sanford-Burnham Medical Research Institute, www.sciencedaily.com*

Hence, despite the variety of the diseases and the proteins involved, there seems to be a common mechanism underlying these disorders: protein polymerisation. Before introducing the mathematical details of the modelling, we start by describing the aggregation of proteins from a biophysical and kinetic point of view.

1.2 Polymerisation of proteins

As mentioned above, understanding the phenomenon of protein polymerisation is fundamental to study many neurodegenerative diseases and amyloid diseases. Nevertheless, it should be noted that protein aggregation is also involved in other areas. In biology, it is naturally and positively occurring in the cytoskeleton to maintain the shape and mobility of cells. Cells endure many changes in their environment and must adapt quickly. This is achieved through a rapid and controlled polymerisation of microtubules and actin filaments [J.M. Berg, 2002, Desai and Mitchison, 1997].
In the industry, it may happen that unwanted amorphous aggregates of proteins are produced. Thus, controlling the polymerisation also has significant impacts in the biotechnology industry [Roberts, 2003]. Hence, polymerisation mechanisms that will be introduced are actually general enough to take into account all these modes of polymerisation.

1.2.1 Collecting data

Collecting data of protein aggregation is not straightforward. From an epidemic point of view, it is not easy to determine the number of infected people since the incubation time is very long.

In vivo robust data are also complicated to get. Indeed, when following aggregation of proteins in a culture of cells, it is hard to isolate the kinetics of the polymerisation from the complex dynamics of the cell. Moreover, there are many unsolved questions concerning the interaction of polymers and the cell environment. For instance, it is not clear whether the amyloid fibrils are formed in the extra or intracellular medium, or both [Ma and Lindquist, 2002].

Recent techniques have allowed the study of the assembly of monomers into polymers *in vitro*. These experiments allow the monitoring of the mass of polymers formed during the polymerisation. However, the presumably intermediate species (details about these species will be given later) are not tractable using these approaches. For the purpose of our study we will look at data of aggregation of β_2m amyloid fibrils. These data have been published in [Xue et al., 2008]. Because of its quick polymerisation without formation of amorphous aggregates, β_2m is a good choice for the study of self-assembly of proteins. β_2m is, as previously stated, involved in dialysis-related amyloidosis. The native β_2 m structure is converted into a cross-β structure, typical of amyloid diseases, and becomes prone to formation of amyloid-like fibrils. During the experiment, the mass of polymers formed is measured via fluorescence of thioflavin (ThT). ThT is a dye that is amyloid-specific, and does not interfere with the process of fibril formation. ThT fluorescence is linearly correlated to the mass of polymers formed.

In our set of data, the mass of polymers formed is monitored until the total consumption of monomers for 20 different initial concentrations of monomers introduced in the test tube. For a given initial concentration of monomers, denoted m throughout this thesis, the experiment of polymerisation is repeated between 9 and 12 times. This gives a set of 235 traces shown in figure 1.2.

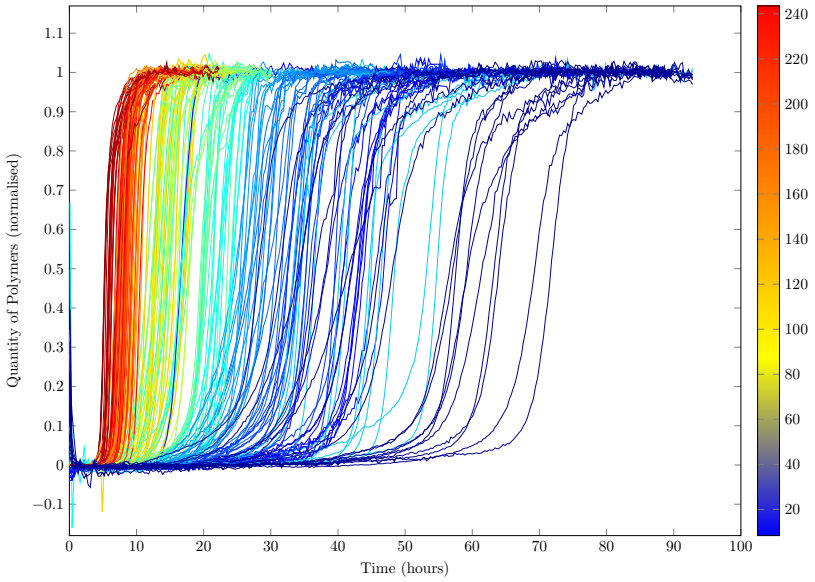


Figure 1.2: Experimental progress curves of polymerised mass for initial concentration of monomers ranging from $8.4\mu M$ to $243.5\mu M$ (color bar), curves obtained from the same initial monomer concentration have the same color, published in [Xue et al., 2008].

Observations. Looking at figure 1.2, one notes that:

— Each curve is characterised by a slow beginning, and then, as soon as a small fraction δ of polymers is formed, all monomers are very quickly polymerised, until total consumption of monomers. All curves have a sigmoid shape, typical of protein polymerisation experiments. Hence, the time to form a small fraction δ of polymers in a volume V of solution, the take-off of the curve, is an interesting feature that we will later be interested in. In the polymerisation literature, δ is often chosen to be between 10 and 20% [Ferrone, 1999], and is then called *lag time*. For the purpose of the subsequent study, we introduced the notation $T_V(\delta)$ for the time for δ reaction completion, that will be used throughout this thesis.

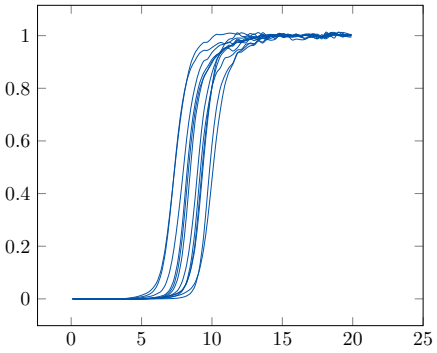


Figure 1.3: $m = 122\mu M$

— For a given initial concentration m of monomers, in the same volume V of solution, repeating the same experiments 12 times, the lag-time varies in a range within hours. For instance, if we look at the initial concentration 122μM (figure 1.3), the lag-time varies from approximately 7 hours, to approximately 10 hours [Xue et al., 2008]. This is very surprising regarding the large volumes considered ($15\mu L$). This suggests a high **stochasticity** for $T_V(\delta)$.

This randomness is sometimes attributed to experimental imprecisions. However, it is more likely that stochasticity is inherent to the mechanism of polymerisation since it obeys some rules:

- ◇ the smaller the initial concentration m is, the longer the lag-time is. It is intuitive in a stochastic framework since putting fewer monomers gives rise to a smaller probability for the monomers to meet, and so to aggregate.
- ◇ the smaller the initial concentration m is, the higher the variance of $T_V(\delta)$ is. Here again, when fewer monomers interact, it is intuitive to get a higher variance.

From this observation, we can define the random variable $T_V(\delta)$ mathematically, as the time when a fraction δ of the initial number of monomers is polymerised. It is the stopping time:

$$T_V(\delta) = \inf\{t > 0, X_2(t) \geq \delta M_V\}$$

where,

- ◇ M_V is the **number** of monomers introduced initially. For example, if m is given in μM (microMolar), V in litres (L), then

$$M_V = m \cdot 10^{-6} \cdot V \cdot N_A \tag{1.1}$$

  where N_A is the Avogadro constant, $N_A = 6.02 \cdot 10^{23}$.
- ◇ $X_2(t)$ is he number of polymerised monomers at time t. We will get into details of this concept later in this thesis.

It is important to note that the natural definition of the lag time here, according to our set of data, is the time when a **fraction** δ of the initial number of monomers introduced is polymerised, and not only the time of formation of one polymer as in [Yvinec, 2012, Szavits-Nossan et al., 2014].

— Nevertheless, for a given initial concentration m of monomers again, all the curves have the same shape and can be superimposed: while the lag-time is certainly stochastic, the rest of the process seems to be deterministic.

1.2.2 Intermediate states

From its native state to fibrils, a protein involved in a polymerisation process exists under different intermediary forms. We briefly describe these forms before getting into the details of the kinetic aspects.

Monomers: In the beginning, native proteins are in the state of soluble monomers. We denote the species 'native monomers' by $\mathcal{X}_0$, and their number X_0 (throughout this thesis, we will make the same distinction in the notation between the chemical species and their number.)

Misfolded monomers: It is widely admitted [Chiti and Dobson, 2006, Uversky et al., 2001, Uversky and Fink, 2004, Prusiner, 1998, Griffith, 1967] that before being able to polymerise, monomers undergo a conformational change that makes them 'active'. We call this form 'misfolded monomer' and denote it by $\mathcal{X}_1$.

Nucleus: Before a critical size, called the 'nucleus size' and denoted i_0 here, the aggregation is chemically unfavorable, whereas dissociation is favourable. When it reaches this critical size, the monomers form a nucleus that is thermodynamically stable and allow further growth. In 1999, Ferrone defined the critical nucleus as the aggregate of size after which 'the association rate exceeds the dissociation rate for the first time' [Ferrone, 1999]. The nucleus is the first aggregate thermodynamically stable on the pathway from monomers to polymers. Details about the mechanism of nucleation will be given in the next section.

Oligomers: Oligomers are intermediate states of polymerised mass that are not easy to track experimentally. There are two types of oligomers: soluble oligomers, which are intermediate species between the monomers and the nucleus, and insoluble oligomers, between nucleus and fibrils. The latter are also called protofibrils (as often in the polymerisation literature, many different words are used with the same meaning). Protofibrils of amyloid-β (related to Alzheimer's) have been experimentally observed by Lansbury [Caughey and Lansbury, 2003]. It has even been suggested that these intermediaries are actually the toxic pathogens of Alzheimer's (rather than the plaques) [Shankar et al., 2008]. Indeed, there is still a controversy on whether protofibrils are 'on' the pathway to fibril formation, or 'off', as shown in figure 1.5; their toxicity however seems to be quite likely [Chiti and Dobson, 2006, Kayed and Lasagna-Reeves, 2013].

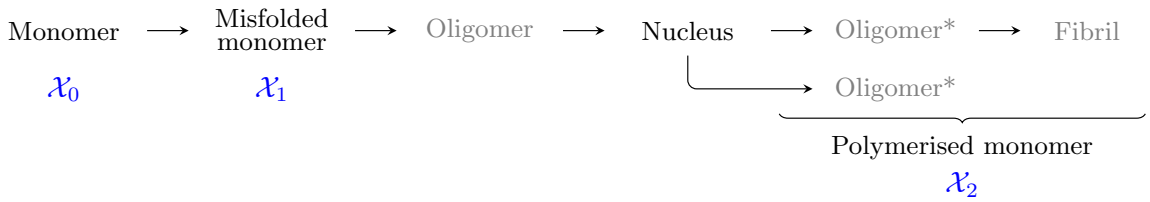


Figure 1.5: Intermediate states of a protein during polymerisation, from monomers to polymers. The species that will be considered in this thesis are in black. Inspired from [Morris et al., 2009].
*: oligomers are found 'on' and 'off' pathway to fibrils, depending on the protein involved.

Fibrils: The final product of polymerisation is amyloid fibrils. They are insoluble fibres detected in the extra or intracellular medium. Despite the diversity of amyloidogenic proteins, the amyloid fibrils have an incredibly common structure [Sunde et al., 1997] consisting of protofilaments twisted along the same axis, as shown in figure 1.4.

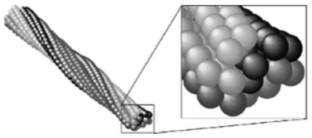


Figure 1.4: Fibril of sickle cell hemoglobin. Retrieved from [Ferrone, 2006]

Polymerised Monomers: Concerning the products of the polymerisation, in order to avoid the debate about fibrils, amorphous aggregates, oligomers, the concept of polymerised monomers is often used [Wegner and Engel, 1975, Ferrone, 1999] as it allows many simplifications. It has been first introduced by Oosawa in 1962 [Oosawa and Kasai, 1962]. A polymerised monomer is a monomer involved in a polymer with a size greater than the nucleus size. We will use this simplification in our modelling approach. Polymerised monomers will be denoted by $\mathcal{X}_2$.

A first global picture of how these forms interact is given in figure 1.5. Our approach consists in focusing on:

— the reversible conversion of $\mathcal{X}_0$ into $\mathcal{X}_1$ (Conformation step)

— the formation of the nucleus (Nucleation)

— the formation of $\mathcal{X}_2$ (Growth and secondary pathway)

1.2.3 Steps of protein polymerisation

In this section, we review the mechanisms presumably involved in protein polymerisation. Subsequent models will be a combination of these possibilities.

Conformation step

Physico-chemical properties of a protein are not only inherited from its sequence of amino acids, but also from its 3d-structure. Hence, having the correct conformation is necessary for proteins to be functional. They are able to adopt a minimal-energy conformation surprisingly quickly. However, many other

minimal-energy conformations than the native one are possible so that it happens that proteins do not fold correctly. In this case, other proteins called chaperones come into play: chaperones [Ellis, 1987] 'help' the proteins to find their native structure, or even refold a non-native protein [Wright et al., 2015]. Usually in the native state, hydrophobic amino acids stick together in the core of the protein to avoid water molecules, and hydrophilic ones are at the surface, on contact with cellular medium.
Unfortunately, it happens that proteins misfold despite the help of chaperones. They convert into a structure rich in β-sheets, that are highly hydrophobic. Thus, while the native form is often soluble in the cellular environment, the misfolded one is frequently insoluble and has a propencity to aggregate.

Misfolding in prion and prion-like diseases. Already in 1967, Griffith suggested that the agent responsible of scrapie was becoming infectious because of a change of conformation. Prusiner confirmed this hypothesis with experiments revealing that PrP^{Sc} had a secondary structure rich in β-sheets whereas PrP^{C} was of α-helix secondary structure, but still, they both have the same sequence of amino acids. Hence, PrP^{C} and PrP^{Sc} are isomers that do not have the same physico-chemical properties, especially regarding aggregation: the higher proportion of β-sheet makes PrP^{Sc} aggregate. Formation of β-sheets has also been reported in various amyloid diseases [Knowles et al., 2014a]. This hypothesis also explains the diversity of prions. Indeed, varieties of prions were brought to light by experiments revealing different features, like incubation times for instance, each of them being induced by a certain conformation of PrP^{Sc} [Prusiner, 1998]. In conclusion, some proteins have an intrisic, i.e encoded in its amino acid sequence, capacity to convert into another 3d-structure [Knowles et al., 2014a]. Normally, chaperones prevent this from happening, but, in the case of prion and prion-like diseases, misfolding has been experimentally observed as being the first step before aggregation of proteins [Collins et al., 2004]. Misfolding is now often stated as a pre-step for polymerisation [Prigent et al., 2012, Serio et al., 2000, Chiti and Dobson, 2006]. It explains:

— **sporadic appearance of prion and prion-like diseases:** a protein spontaneously changes its conformation and becomes able to polymerise.

— **genetic transmission:** a genetic mutation in the gene coding for the protein will induce a higher propensity to change conformation.

— **infectious transmission:** transmission of the protein in the misfolded state, ready to polymerise.

Misfolding is chemically modelled by equation (1.2).

$$\mathcal{X}_0 \underset{\gamma^*}{\overset{\gamma}{\rightleftharpoons}} \mathcal{X}_1 \tag{1.2}$$

where,

— γ is the chemical rate of misfolding of protein from its native state to its 'active' state.

— γ^* is the chemical rate of conversion of a misfolded protein back to its native state.

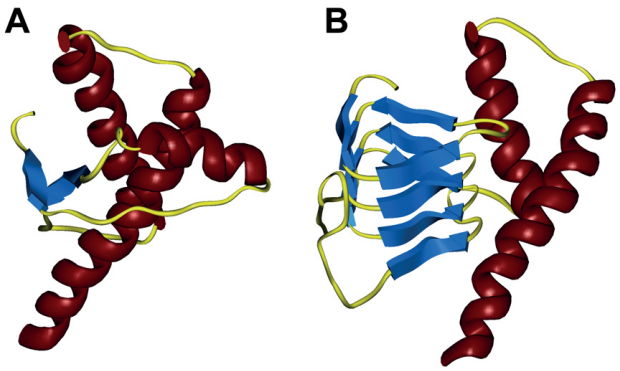


Figure 1.6: A: Functional PrP^C. B: PrP^C is converted into PrP^{Sc}, rich in sticky β-sheets (the blue structure). β-sheet are composed of β-strands. In the amyloid case, β-strands are perpendicular to the fibril axis [Knowles et al., 2014a]. Retrieved from: www.ucsf.edu/news/2001/08/4709/ucsf-study-finds-two-old-drugs-may-help-fight-prion-diseases

— $\gamma << \gamma^*$, misfolding is fortunately thermodynamically unfavourable.

Misfolding is a slow reaction that might by itself explain the slow beginning of the reaction. However, this initial lag phase is also commonly modelled by a nucleation phase, or by both conformation and nucleation.

Nucleation

The concept of nucleation was initially introduced in the framework of crystallisation [Volmer and Weber, 1926]. It is a phase transition after which growth can occur. Before, aggregation is thermodynamically unfavourable. Protein polymerisation does present some distinctive features of a nucleated growth:

— First, experimental curves reveal a long lag phase (figure 1.2), attributed to the formation of nuclei [Chiti and Dobson, 2006].

— Second, this lag phase strongly depends on the initial concentration of monomers introduced (figure 1.2).

— Finally, introduction of preformed fibrils drastically reduces the lag phase [Jarrett and Lansbury, 1993]. Indeed, suppose nucleation is happening, it acts like a kinetic barrier to overcome for amyloid formation. When introducing preformed fibrils, the growth is not limited anymore by the nucleation step [Chiti and Dobson, 2006]. This 'seeding' effect also suggests that the initiation of the polymerisation is very slow compared to the growth. This interplay will be important for our modelling.

Nucleation had already been introduced as a first step for actin polymerisation in 1959 [Oosawa and

Kasai, 1962]. Later, in 1993, Jarrett and Lansbury proposed it as a key step in Alzheimer's disease and scrapie [Jarrett and Lansbury, 1993]. In fact, nucleated polymerisation explains the infectivity of prion, the pathogenic agent being the aggregated PrP^{Sc} acting as seeds [Jarrett and Lansbury, 1993]. The general chemical scheme of nucleation is given by equation (1.3). We consider here that the aggregation happens by addition of misfolded monomers.

$$\begin{aligned} \mathcal{X}_1 + \mathcal{X}_1 &\underset{k_{\text{off}}^2}{\overset{k_{\text{on}}^1}{\rightleftharpoons}} \mathcal{X}_2 \\ &\vdots \\ \mathcal{X}_{i_0-1} + \mathcal{X}_1 &\underset{k_{\text{off}}^{i_0}}{\overset{k_{\text{on}}^{i_0-1}}{\rightleftharpoons}} \mathcal{X}_{i_0} \end{aligned} \tag{1.3}$$

where:

— i_0 is the size of the nucleus,

— k_{on}^i is the rate of polymerisation of an oligomer of size $i \leq i_0$, before nucleation,

— k_{off}^{i+1} is the rate of depolymerisation of an oligomer of size $i+1$, $i \leq i_0$, before nucleation,

— $k_{\text{on}}^i \cdot c_1 < k_{\text{off}}^{i+1}$: aggregation is unfavourable before nucleation, with c_1 the concentration of $\mathcal{X}_1$. Note that multiplying by c_1 on the left hand side is necessary for the comparison between the rates because of homogeneity.

However, since intermediate species before nucleus are very unstable, and so never observed, it is also considered that the formation of the nucleus is not an addition of $\mathcal{X}_1$ until i_0, but obeys an i_0 kinetic order, that is equation (1.4) [Oosawa and Asakura, 1975].

$$\underbrace{\mathcal{X}_1 + ... + \mathcal{X}_1}_{i_0} \quad \underset{k_{\text{off}}^n}{\overset{k_{\text{on}}^n}{\rightleftharpoons}} \mathcal{X}_{i_0} \tag{1.4}$$

where k_{on}^n is the chemical rate of formation of the nucleus, k_{off}^n its rate of breakage.

Nucleation is a highly stochastic phenomenon. The key feature of the nucleation step is its stochasticity. It corresponds to the very unlikely encounter of i_0 monomers, where i_0 is the size of the nucleus. First, as said above, the nucleation time strongly depends on the initial concentration of monomers. The lower the initial concentration is, the longer the lag phase is. It could be explained by the decrease of probability to meet for monomers, and thus to form nuclei. Second, one notes on the data figure 1.2 that for small initial concentrations, the variance of the lag time is higher than for high concentrations, so that the stochasticity seems inherent to the polymerisation mechanism. This randomness of the initial phase was emphasised in [Szabo, 1998, Hofrichter, 1986b].

Choice $i_0 = 2$. A nucleation step involving dimer formation is often made in the literature [Wegner and Engel, 1975, Knowles et al., 2009]. Indeed, the first attachment step towards nucleation has the biggest

energy penalty in the classical nucleation theory. Back in 1967, Griffith suggested that the dimerisation was initiating the polymerisation. For the sake of simplification, in order to capture the main features of a nucleated polymerisation, choosing $i_0 = 2$ is then reasonable. In our models, we will consider that the size of the nucleus is equal to 2.

As a conclusion, one must keep in mind that the nucleation step is stochastic and slow.

Growth and secondary pathway

Once the nucleus is formed, polymerisation becomes more favourable than depolymerisation, and polymers start to elongate. Polymerisation is often considered as occurring by monomers addition. However, when a conformation step occurs, it is not clear whether the additional monomer is misfolded or not. Here, we will consider that only a misfolded monomer can polymerise, as suggested by [Jarrett and Lansbury, 1993, Frieden and Goddette, 1983]. To be fully general, we introduce the possibility to depolymerise, even if in this dissertation, regarding our data, we will consider that a polymer can only grow. If we denote i_0 the size of the nucleus, $\mathcal{X}_i$ a polymer of size i, then, for $i \geq i_0$ the growth mechanism follows the scheme:

$$\mathcal{X}_{\mathrm{i}} + \mathcal{X}_1 \quad \underset{k^{i+1}_{\mathrm{dep}}}{\overset{k^{i}_{\mathrm{pol}}}{\rightleftharpoons}} \quad \mathcal{X}_{\mathrm{i+1}} \tag{1.5}$$

where,

— k^i_{pol} is the chemical rate (not to be confused with the rate of the associated Markov process that will be defined later) of the reaction of polymerisation for a polymer of size $i \geq i_0$,

— k^{i+1}_{dep} the chemical rate of the reaction of depolymerisation of a polymer of size $i+1$, $i \geq i_0$,

— $\mathcal{X}_1$ a monomer, either native or misfolded, depending on the model chosen,

— $k^{i+1}_{\mathrm{dep}} < k^i_{\mathrm{pol}} \cdot c_1$: in the growth phase, polymerisation is favourable.

However, this mechanism of formation of polymers is not sufficient to explain by itself experimental curves presenting a very quick consumption of monomers as soon as the process has started. There are also autocatalytic mechanisms involved that enhance the polymerisation called 'secondary pathways'. These are for instance:

— **Heterogenous nucleation:** Heterogenous nucleation was shown to take part in sickle hemoglobin polymerisation by Ferrone and his coworkers in 1985 [Ferrone et al., 1985]. In this framework, the nucleation initiating polymerisation in bulk is referred as homogenous nucleation. Once a polymer is formed, another nucleation happens on the surface of the polymer, initiating a new fibril assembly. Hence, the longer a fibril is, the higher the possibilities of heterogenous nuclei formation are. This explains the autocatalytic part observed in experiments. As soon as homogenous nucleation has happened, there is an exponential growth of polymers [Ferrone et al., 1980].

— **Fragmentation:** Another autocatalytic mechanism usually proposed is fragmentation. It comes

from the same idea as heterogenous nucleation. What takes time in polymerisation is the initiation of the process. Secondary processes are mechanisms that shortens, or removes the lag phase to initiate new polymers via existing polymers. Fragmentation plays this role. Suppose a (rather long) polymer is formed. Then, if it breaks into two shorter polymers, this allows to initiate the formation of two new polymers from the initial long one, and so on. At the end, we get an exponential growth. This mechanism was suggested in [Xue et al., 2008, Collins et al., 2004] for instance.
Fragmentation hypothesis was experimentally supported by the fact that mechanical agitation was drastically increasing the rates of fibrillation [Bishop and Ferrone, 1984]. This indicates that the breakages induced by agitation were accelerating the polymerisation. We consider that fragmentation is irreversible. The chemical reaction modelling the breakage of a polymer of size $i+j$ into two polymers of size i and j is:

$$\mathcal{X}_{\mathrm{i+j}} \xrightarrow{k_{\mathrm{fr}}^{i+j}} \mathcal{X}_{\mathrm{i}} + \mathcal{X}_{\mathrm{j}} \tag{1.6}$$

where k_{Fr}^{i+j} is the rate of fragmentation of a polymer of size $i+j$.

Now that we know most of the mechanisms suspected to be involved in the polymerisation so far, we can introduce the classical models used in the literature of polymerisation.

1.2.4 Review of polymerisation models

The broad polymerisation literature contains many models of self-assembly of proteins relying on the steps described in the previous section. However, depending on the protein involved and on the purpose of the study (estimating parameters, understanding a precise mechanism not directly observed in the experiments etc.), different approaches can be used. In this section, we will try to review the main models, in our sense, introduced since the last fifty years concerning protein polymerisation in general, not only prion aggregation.
Before starting, we should note that kinetic equations derived from these models are up to now essentially deterministic rather than stochastic and rely on the law of mass action, which will be described later in this thesis.

For the sake of clarity, we summarise the notations that will be used throughout this section.

— $\mathcal{X}_0$: monomers in their native state.

— $\mathcal{X}_1$: monomers in their 'active' state, able to polymerise. When conformation step is taken into account, it corresponds to the misfolded monomer.

— $\mathcal{X}_i$: polymers of size i. This notation will be useful for mechanistic models taking into account the whole polymer size distribution, instead of the concept of polymerised monomers previously introduced.

— i_0: size of the nucleus.

— c_i: concentration of polymers of size i.

— k^i_{on}: chemical rate of association for a polymer of size $i < i_0$.

— k^i_{off}: chemical rate of dissociation for a polymer of size $i < i_0$.

— k^i_{pol}: chemical rate of polymerisation for a polymer of size $i \geq i_0$.

— k^i_{dep}: chemical rate of depolymerisation for a polymer of size $i \geq i_0$.

— k^{i+j}_{fr}: chemical rate of fragmentation of a polymer of size $i + j$.

— γ: chemical rate of misfolding of protein from its native state to its 'active' state.

— γ^*: chemical rate of conversion of a misfolded protein nack to its native state.

Towards a complete model of protein self-assembly

Models of polymerisation have progressively evolved to a complete picture of the mechanism of polymerisation.

A basic approach, nucleation and growth. First contributions to the modelling of protein aggregation are due to Oosawa and his coworkers [Oosawa et al., 1959]. They studied actin protein and modelled the conversion of G-actin into F-actin. Some key features of polymerisation were already identified:

— The beginning is very slow, and is followed by a steep slope, suggesting an autocatalytic mechanism.

— The lag phase is shortened by addition of native G-actin, suggesting a stochastic effect due to the encounter of initial monomers.

— Introduction of F-actin removes the lag-phase, impling a seeding effect suggesting nucleation.

In 1974, Hofrichter et al. proposed a mechanism based on these observations with two steps: nucleation followed by polymerisation via subsequent addition of monomers. The constants of reaction chosen do not depend on the size of the polymer, so that we forget about the superscript i. The sigmoidal shape of experimental curves suggests that:

— Until the nucleus size, polymerisation is thermodynamically unfavourable, inducing a lag phase. This is modelled in [Hofrichter et al., 1974] by $k_{\text{on}} \cdot c_1 < k_{\text{off}}$. (See picture 1.7)

— For $i \geq i_0$, polymerisation becomes favourable, inducing a steep take-off. They [Hofrichter et al., 1974] thus chose $k_{\text{pol}} \cdot c_1 > k_{\text{dep}}$.

Note that there is neither a conformation step nor a secondary pathway. Chemical reactions associated to [Hofrichter et al., 1974] are shown in figure 1.7.

$$\begin{aligned}
&\mathcal{X}_1 + \mathcal{X}_1 \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} \mathcal{X}_2 \\
&\qquad\vdots \\
&\mathcal{X}_1 + \mathcal{X}_{i_0-1} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} \mathcal{X}_{i_0}
\end{aligned}
\quad \Big\} \text{ Nucleation: } k_{\text{on}} \cdot c_1 < k_{\text{off}}$$

$$\begin{aligned}
&\mathcal{X}_1 + \mathcal{X}_{i_0} \underset{k_{\text{dep}}}{\overset{k_{\text{pol}}}{\rightleftharpoons}} \mathcal{X}_{i_0+1} \\
&\qquad\vdots \\
&\mathcal{X}_1 + \mathcal{X}_i \underset{k_{\text{dep}}}{\overset{k_{\text{pol}}}{\rightleftharpoons}} \mathcal{X}_{i+1} \\
&\qquad\vdots
\end{aligned}
\quad \Big\} \text{ Growth: } k_{\text{pol}} \cdot c_1 > k_{\text{dep}}$$

Figure 1.7: Polymerisation mechanism suggested in [Hofrichter et al., 1974]. Before nucleation, depolymerisation is favourable. Once the nucleus is formed, polymerisation becomes favourable.

Addition of a secondary pathway. Later, Ferrone [Ferrone et al., 1985] argued that this model wasn't able to take into account the 'extreme autocatalysis'. He introduced a secondary pathway called 'heterogenous nucleation', previously described. In 1982, Wegner and Savko [Wegner and Savko, 1982] added another secondary process for polymerisation of actin: fragmentation (equation (1.6)). However, actin kinetics do not present a lag phase contrary to our experiments. Hence, they didn't include nucleation, but only growth and fragmentation, what allows to predict the steep slope of the polymerisation as shown in figure 1.8.

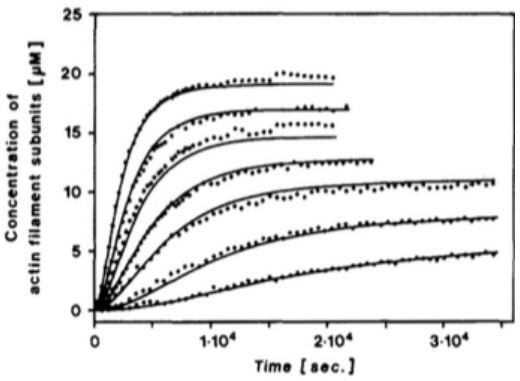


Figure 1.8: Actin polymerisation curves compared to predictions from Wegner and Savko's model. Retrieved from [Wegner and Savko, 1982]

.

And the conformation step? Based on Wegner and Savko's model, Frieden and Goddette added a conformation step for polymerisation of actin [Frieden and Goddette, 1983], as modelled by reaction (1.2). Growth also takes into account different rates of polymerisation and depolymerisation for different sizes i of polymers. It corresponds to reaction (1.5).

Prion aggregation

Early models of polymerisation are essentially based on actin. However, the situation is different for prion and prion-like aggregation. The main difference lies in the long lag phase observed on experimental curves, highly variable, unlike actin filaments formation for instance (see figure 1.8). This phase is the stochastic step of polymerisation, and is therefore crucial for us. The models proposed for prion aggregation all include this slow beginning.

The mathematician Griffith was the first to propose models of prion infection for scrapie. He included

a conformation step followed by the formation of a dimer. The misfolding is unfavourable, but if the rate of polymerisation is large enough, it pushes the equilibrium towards the formation of misfolded monomers. Then, growth occurs by addition of native monomers, their integration being promoted by the presence of other polymers. In 1982, Prusiner proposed that the misfolded monomer $\mathcal{X}_1$ was catalyzing the conversion of $\mathrm{PrP}^{\mathrm{C}}$ into $\mathrm{PrP}^{\mathrm{Sc}}$. In these models, the slow take-off was supposed to be explained by the slow conformation step, and the autocatalysis part by the fact that presence of polymers and misfolded monomers is boosting the conformation step. Still, no kinetic analysis was performed, so that in 1996, Eigen showed that the rates for $\mathcal{X}_1$ to be the catalytic agent were numerically unrealistic [Eigen, 1996]. Later models added a secondary pathway to enhance the catalytic part.

Finally, Lansbury proposed in 1993 [Jarrett and Lansbury, 1993] a model including nucleation. It follows the chemical scheme of figure 1.7 with a prenucleation step, misfolding of protein, as described by reaction (1.2).

It seems clear from the vast literature of protein aggregation that a complete model should take into account both the initial phase, highly stochastic and very slow, and the following autocatalytic reaction leading to an explosion of the formation of polymers. Our work will essentially be based upon the complete model of nucleated polymerisation by subsequent monomer addition, proposed in [Prigent et al., 2012], and represented by equation (1.7).

$$\begin{aligned} \mathcal{X}_0 &\underset{\gamma^*}{\overset{\gamma}{\rightleftharpoons}} \mathcal{X}_1 \\ \underbrace{\mathcal{X}_1 + \ldots + \mathcal{X}_1}_{i_0} &\underset{k_{\mathrm{off}}^n}{\overset{k_{\mathrm{on}}^n}{\rightleftharpoons}} \mathcal{X}_{\mathrm{i}_0} \\ \mathcal{X}_{\mathrm{i}} + \mathcal{X}_1 &\underset{k_{\mathrm{dep}}^{i+1}}{\overset{k_{\mathrm{pol}}^i}{\rightleftharpoons}} \mathcal{X}_{\mathrm{i}+1} \end{aligned} \tag{1.7}$$

Polymerised monomers. The complete model of protein aggregation involves an infinite set of chemical reactions, which leads to mathematical complications for the derivation of the kinetics. Instead of taking into account the distribution size of the filaments, a common simplification consists in considering only two species: free monomers, $\mathcal{X}_1$, and monomers polymerised in aggregates of sizes greater than the nucleus size, $\mathcal{X}_2$ (we make here a small abuse of notations: in the previous complete models, $\mathcal{X}_2$ was referring to dimers). Our work relies on this approach. Hence, forgetting about the conformation step, this simplification transforms (1.7) into:

$$\begin{aligned} \underbrace{\mathcal{X}_1 + \ldots + \mathcal{X}_1}_{i_0} &\rightharpoonup i_0\mathcal{X}_2 \\ \mathcal{X}_1 + \mathcal{X}_2 &\rightleftharpoons 2\mathcal{X}_2 \end{aligned} \tag{1.8}$$

The first reaction captures the slow nucleation step, while the second the fast autocatalytic growth.

Finally, $\mathcal{X}_2$ represents the mass polymerised, so that, for a volume of solution V:

$$\frac{X_2(t)}{V} = \sum_{i \geq i_0} i c_i(t)$$

which is precisely the quantity usually measured in experiments, allowing curve-fitting.

This minimalistic 2-step model was initially used in [Watzky and Finke, 1997] for transition-metal nanocluster formation, and after for protein polymerisation [Morris et al., 2008] in a slightly different form [Morris et al., 2009]:

$$\begin{aligned} \mathcal{X}_1 &\xrightarrow{k_1} \mathcal{X}_2 \\ \mathcal{X}_1 + \mathcal{X}_2 &\xrightarrow{k_2} 2\mathcal{X}_2 \end{aligned} \tag{1.9}$$

It is referred as the Finke-Watzky mechanism (F–W mechanism), and is said to be equivalent to (1.8) in [Morris et al., 2008]. However, from our stochastic point of view, models (1.9) and (1.8) differ on the rates of reaction, since the F–W mechanism does not take into account the low probability of the encounter of monomers to initiate the polymerisation. This will lead to different dynamics for the stochastic processes $(X_1)_t$ and $(X_2)_t$. The crucial choice of rates of reactions in the stochastic framework will be discussed deeper later.

Now that we have a general picture on the models of protein polymerisation, we get into the details on the way to derive a kinetic analysis, and ultimately, to get information about the parameters involved both in the deterministic and in the stochastic settings.

1.3 Mathematical modelling of protein polymerisation

It is the role of biologists and physicists to collect data and interpret them to propose mechanistic models. In the modelling process, mathematicians have then to understand what the biologists made, or at least try to, often simplify their models, and mathematically formalise the mechanisms in order to derive an analytical solution to the model capable to predict the behaviour of the quantities involved. Then, the dream comes true when the model, with parameters estimated from biological data, reproduces the experiments.

Here, we explain how to derive a mathematical analysis from the models presented before. We also review some methods used in the literature to fit the kinetic parameters of the polymerisation mechanism. As it will be systematically be done in the subsequent chapters, we will, from now on, always assume a nucleus size of two and that no depolymerisation occurs:

$$\begin{cases} i_0 &= 2, \\ k_{\text{off}}^N &= 0, \\ k_{\text{dep}}^i &= 0 \quad \forall i \geq i_0. \end{cases} \tag{1.10}$$

1.3.1 Deterministic modelling

The mechanisms involved in the polymerisation are essentially written as a set of chemical reactions. The most famous way to derive kinetics of a reaction network (*i.e* a set of chemical reactions) is the law of mass action, due to Guldberg and Waage in 1867, which leads to a set of ordinary differential equations (ODEs) (see [Rubinov, 1975] for instance and references therein). The key idea underlying this modelling is the same in the stochastic framework.
More precisely, let's consider two species, $\mathcal{A}$ and $\mathcal{B}$ that tend to form $\mathcal{C}$ in a volume V of solution:

$$\mathcal{A} + \mathcal{B} \xrightarrow{k} \mathcal{C} \tag{1.11}$$

We consider that the content of the solution is homogeneous so that only the temporal evolution of the concentrations of chemical species interests us. The concentration in $\mathcal{A}$ (resp. $\mathcal{B}$ and $\mathcal{C}$) at time t is denoted by $c_A(t)$ (resp. $c_B(t)$ and $c_C(t)$), the total composition of the solution by $c(t) = (c_A(t), c_B(t), c_C(t))$. Then, in order to write a differential equation representing the temporal evolution of this reaction, we have to define $\mathcal{K}_{\mathcal{A}+\mathcal{B}\to\mathcal{C}}(c)$, the instantaneous rate at which reaction (1.11) occurs. Deriving a kinetics for a reaction network means defining a rate function for each reaction in the network. Once this assignment is done, it is easy to describe the time evolution of the composition of the solution: each time reaction (1.11) occurs, we loose one molecule of $\mathcal{A}$, one molecule of $\mathcal{B}$, and we gain one molecule of $\mathcal{C}$ so that:

$$\begin{aligned} \dot{c_A}(t) &= \dot{c_B}(t) = -\mathcal{K}_{\mathcal{A}+\mathcal{B}\to\mathcal{C}}(c(t)) \\ \dot{c_C}(t) &= \mathcal{K}_{\mathcal{A}+\mathcal{B}\to\mathcal{C}}(c(t)) \end{aligned}$$

The law of mass action states that $\mathcal{K}_{\mathcal{A}+\mathcal{B}\to\mathcal{C}}(c)$ is proportional to the probability for $\mathcal{A}$ and $\mathcal{B}$, that is proportional to $c_A(t) \cdot c_B(t)$. The usual kinetics follows [Rubinov, 1975]:

$$\dot{c_A}(t) = -k c_A(t) \cdot c_B(t) \tag{1.12}$$

where k is the constant of proportionality. In the chemical jargon, this quantity is called 'rate of reaction' (which is different from the rate of occurence of the reaction).

Law of Mass Action For Polymerisation. We now apply the law of mass action, in the deterministic framework to the general model of nucleated polymerisation (5.5) with assumptions (1.10). We obtained the infinite set of ODEs shown in equation (1.13).

$$\begin{cases} \dot{c_0}(t) = \gamma c_0(t) - \gamma^* c_1(t) \\ \dot{c_1}(t) = -\gamma c_0(t) + \gamma^* c_1(t) - k_{\text{on}}^N c_1(t)^2 - \sum_{i=2}^{\infty} k_{\text{on}}^i c_1(t) \cdot c_i(t) \\ \dot{c_2}(t) = k_{\text{on}}^N c_1(t)^2 - k_{\text{pol}}^2 c_1(t) c_2(t) \\ \dot{c_i}(t) = k_{\text{pol}}^{i-1} c_1(t) \cdot c_{i-1}(t) - k_{\text{pol}}^i c_1(t) \cdot c_i(t), \text{ for } i \geq 3. \end{cases} \tag{1.13}$$

Becker-Döring. In this paragraph, we do not take into account the conformation step. We recall that this infinite system is in fact well-known. Indeed, equation (1.13) can be rewritten as a Becker-Döring system as follows:

$$\begin{aligned} \dot{c_i}(t) &= J_{i-1} - J_i \text{ for } i \geq 2 \\ \dot{c_1}(t) &= -\sum_{i=2}^{\infty} J_i - J_1 \end{aligned} \tag{1.14}$$

where $J_i = k^i_{\text{pol}} c_i \cdot c_1$ for $i \geq 2$, $J_1 = k^N_{\text{on}} c_1^2$.

In 1935, Becker and Döring introduced this model for deriving kinetics of many phenomena involving phase transition such as metastability of a ferromagnet for instance [**?**]. Their model applies to systems invoking two components: monomers and i-clusters of size $i \geq 2$. In our case, the two components are free monomers and fibrils of different sizes. The resulting system (1.14) is infinite but, under certain assumptions, has a unique solution.

Theorem 1.3.1. *[Ball et al., 1986] Suppose that:*

(i) for all $i \geq 2$, $k^i_{on} = O(i)$ when i tends to infinity,

(ii) for all $i \geq 1$, $c_i(0) \geq 0$,

(iii) $\sum_{i=1}^{\infty} i c_i(0) < \infty$,

then, the system defined by (1.14) does have a solution for all positive times t. Moreover, if we also have:

$$\sum_{i=1}^{\infty} i^2 c_i(0) < \infty,$$

then the solution is unique for a given initial condition.

In addition, we also have a density conservation if all the conditions for the existence of a solution are satisfied. We define:

$$\rho(t) := \sum_{i=1}^{\infty} i c_i(t)$$

which represents the density of particles at time t (*i.e* the number of particles per unit of volume).

Theorem 1.3.2. *[Ball et al., 1986] Let $c(t) = (c_i(t))_{i \geq 1}$ be a solution of (1.14). Then for all $t \geq 0$:*

$$\rho(t) = \rho(0).$$

In practice, in our experiments, at time $t = 0$, we introduce only monomers *i.e* $\sum_{i=1}^{\infty} i c_i(0) = m$, the initial concentration of monomers introduced, so that these theorems provide the existence of a solution of

the infinite deterministic system (1.14) for an appropriate choice of polymerisation rates. Theorem 1.3.2 corresponds to the principle of mass conservation.

We want, in this PhD dissertation, to adapt these models to the stochastic framework.

1.3.2 Stochastic modelling of chemical reactions

We go back to a general simple chemical reaction (1.11) of type

$$\mathcal{A} + \mathcal{B} \xrightarrow{k} \mathcal{C}.$$

In the stochastic framework, we want to count the number of molecules of type $\mathcal{A}$ (resp. $\mathcal{B}$ and $\mathcal{C}$) at time t in a volume N, that we denote by $A_N(t)$ (resp. $B_N(t)$ and $C_N(t)$) and derive its temporal evolution. We consider here that the solution is homogenous.

We shall first introduce the concept of 'volume' before writing the stochastic evolution equations. So far, we referred to V as the physical volume of the solution, *i.e* in litres, but since we are dealing with number of molecules here we need to introduce a molecular volume, denoted N, such that

$$N = \lfloor N_A \cdot V \rfloor,$$

where N_A is the Avogadro constant, $N_A = 6.02 \cdot 10^{23}$. Hence, the quantity $A_N(t)/N$ is the concentration of the species $\mathcal{A}$ in the solution per unit volume at time t; it is the stochastic equivalent of the previous quantity $c_A(t)$. It is also coherent with the definition (1.1) of M_V that becomes

$$M_V = \lfloor m \cdot N \rfloor$$

where m is the initial concentration of monomers given in M (molar). N is very large (the Avogadro constant is huge) and will therefore be our scaling parameter. Note that the introduction of a molecular volume is very convenient since N can be considered as large even for very small volumes of order $1\mu L$ (microlitre). From now on, we will always refer to N when we mention the volume, and all the subscripts and superscripts of our variables will be changed into N.

Before getting into the mathematical details of the stochastic modelling, the first question to address is: what is the added value of a stochastic modelling as compared to a deterministic one? In fact, a chemical reaction is by nature probabilist but in most cases, the law of mass action captures exactly what happens. As we will develop it in this section, the stochastic treatment of (1.11) leads to:

$$\lim_{N\to\infty} \frac{A_N(Nt)}{N} = c_A(t).$$

where c_A is the solution of the law of mass equation (1.12), so that for large volumes we do not see the

stochastic fluctuations around the deterministic (and continuous) mean. However, when the number of reactions is small, as it is sometimes the case in molecular biology, a stochastic approach is necessary. In our case, we might a priori think that the quantities involved are large enough to use directly the law of mass action, and this is why the fluctuations observed on the data presented figure 1.2 are quite surprising. Hence, to explain this variance observed, we do not have any other choice than going back to the microscopic scale and study precisely the statistical fluctuations in a stochastic manner.

In this PhD dissertation, we treat the chemical network as a continuous-time Markov process, as usually done in the literature in this framework [Anderson and Kurtz, 2011, McQuarrie, 1967, Ball et al., 2006]. For this purpose, we introduce here the Markov process $(X(t)) := (A(t), B(t), C(t))$.

Transition rates of the corresponding Markov process

We start by the easy general case (1.11). In order to translate this reaction into a Markov process, we need to define its transition rates. We choose these transition rates according to the law of mass action. More precisely, for $(a, b, c) \in \mathbb{N}^3$:

(i) the probability of the transition $(a, b, c) \to (a, b, c) + (-1, -1, 1)$ in the interval $(t, t + \Delta t)$ is proportional to the product of the concentrations in $\mathcal{A}$ and $\mathcal{B}$ at time t, *i.e* of the form $ka \times b/N^2 \Delta t + o(\Delta t)$, where k is the chemical constant of reaction and $o(\Delta t)/\Delta t \to 0$ for $\Delta t \to 0$. Since we are looking at reaction (1.11) that involves an **encounter** of molecules, what matters is the **concentration** of reactants.

(ii) the probability of the transition $(a, b, c) \to (a, b, c)$ in the interval $(t, t+\Delta t)$ is $1-ka \times b/N^2 \Delta t + o(\Delta t)$.

These assumptions define completely the Markov process $(X(t))$. These are usual assumptions for chemical reactions [Anderson and Kurtz, 2011].

Case of the reaction $\mathcal{A} + \mathcal{A} \to \mathcal{C}$. The previous set-up directly derived from the law of mass action works for $\mathcal{A}$ different from the species $\mathcal{B}$. However, the stochastic framework allows us to be more precise than the brutal law of mass action. Indeed, for the reaction

$$\mathcal{A} + \mathcal{A} \to \mathcal{C}$$

seen as the $\mathbb{N}^2$-valued Markov process $(A(t), C(t))$, the rate of the transition from the state (a, c) to $(a, c) + (-2, 1)$ is proportional to the number of couples of molecules of type $\mathcal{A}$, that is $a(a-1)/2$, and not a^2 as would say the law of mass action. For the same reason as before, this rate is inversely proportional to the square of the volume. As a result, for $(a, c) \in \mathbb{N}^2$:

$$(a, c) \longrightarrow (a, c) + (-2, 1) \text{ at rate } k\, a(a-1)/(2N^2).$$

This kind of reaction involving the meeting of two molecules of the same type will be used in this thesis for the modelling of the nucleation step (chosen here to be a dimerisation according to (1.10)).

Case of the reaction $\mathcal{A} \to \mathcal{B}$. In the case of a unimolecular reaction, there is no reason why the transition rates should depend on the volume. They only depend on the **number** of molecules of type $\mathcal{A}$ at time t. Hence, for $(a, b) \in \mathbb{N}^2$:

$$(a, b) \longrightarrow (a, b) + (-1, 1) \text{ at rate } k\, a.$$

This type of reaction will be used in this thesis to model the conformation step.

Derivation of the stochastic evolution equations

In this section, we use the formalism of [Robert, 2003].

Poisson counting process. In order to count the number of occurrences of reaction (1.11) in an interval of time, we will use point measures. We start by recalling some useful mathematical definitions, most of them being given in [Robert, 2003], for the non probabilist reader.

The space $\mathbb{R}$ is endowed with the Borelian σ-field. $\mathcal{M}(\mathbb{R})$ denotes the set of non-negative Radon measures on $\mathbb{R}$.

Definition 1.3.1 (Point measure)**.** *If m is an element of $\mathcal{M}(\mathbb{R})$, then m is a point measure if it can be represented as*

$$m = \sum_n \delta_{u_n}$$

where $(u_n)_{n\in\mathbb{Z}}$ is a element of $\mathbb{R}^{\mathbb{Z}}$, and δ_a is the Dirac measure at the point a.

We denote the set of point measures on $\mathbb{R}$ by $\mathcal{M}_p(\mathbb{R})$.

Definition 1.3.2 (Point process)**.** *A point process N is a random variable with values in $\mathcal{M}_p(\mathbb{R})$.*

For instance, if we have a set of random observations, encounters of molecules provoking the chemical reaction in our case, and we want to count the number of observations until time t, then we define a point process

$$\begin{aligned} N : (\Omega, \mathcal{F}, \mathbb{P}) &\to \mathcal{M}_p(\mathbb{R}) \\ \omega &\mapsto N(\omega, \mathrm{d}t) \end{aligned}$$

attached to these observations. In our case of chemical networks, as suggested in the previous section, and as it is done for example in [Anderson and Kurtz, 2011], we make the following reasonable hypothesis on our chemical processes:

(i) Encounters of molecules occur one at a time.

(ii) The number of encounters occurring in disjoint intervals of times are independent.

(iii) The number of encounters of molecules occurring in an interval of time depends only on the length of the interval.

This allows us to choose, for our counting process, a Poisson point process.

Definition 1.3.3 (Poisson point process)**.** *A Poisson process $\mathcal{N}_\mu$ of intensity μ is a point process on $\mathbb{R}$ such that:*

(i) If I is an interval of $\mathbb{R}$, the distribution of the random variable $\mathcal{N}(\omega, I)$ is Poisson with parameter $\mu(I)$, i.e for $k \geq 0$,

$$\mathbb{P}(\mathcal{N}(\omega, I) = k) = \frac{\mu(I)^k}{k!} e^{-\mu(I)}.$$

(ii) If $I_1, \ldots, I_n$ are disjoint intervals of $\mathbb{R}$, then $\mathcal{N}(I_1), \ldots, \mathcal{N}(I_n)$ are independent random variables.

We now apply this to our favourite reaction (1.11)

$$\mathcal{A} + \mathcal{B} \xrightarrow{k} \mathcal{C}$$

associated to the Markov process $(A(t), B(t))$ which transitions rates are, for $(a, b) \in \mathbb{R}^2$

$$(a, b) \longrightarrow (a, b) + (-1, -1) \text{ at rate } k \cdot ab/N^2$$

with the initial conditions $A_N(0) = M_N$ and $B_N(0) = 0$, and $\lim_{N \to +\infty} M_N/N = m$.

Throughout this thesis, $\mathcal{N}_\xi$ denotes a Poisson point process on $\mathbb{R}$ with parameter $\xi \in \mathbb{R}^+$. $\mathcal{N}_\xi([0, t])$ denotes then the number of points of the point process $\mathcal{N}_\xi$ in the interval of time $[0, t]$. We add a superscript $\mathcal{N}^i_\xi$, $i \in \mathbb{N}$, when we consider an i.i.d sequence of Poisson processes of parameter ξ.

We now derive the stochastic differential equation describing the temporal evolution of the number of molecules of type $\mathcal{A}$, $(A(t))$. Each encounter of a molecule of type $\mathcal{A}$ and a molecule of type $\mathcal{B}$ causes the disappearance of a molecule of type $\mathcal{A}$. According to the previous discussion, we attach to each couple of a molecule of type $\mathcal{A}$ and a molecule of type $\mathcal{B}$ a Poisson point process of parameter k/N^2. There are $A_N(t) \cdot B_N(t)$ such couples at time t. If the differential $\mathrm{d}A_N(t)$ is defined as

$$\mathrm{d}A_N(t) = A_N(t) - A_N(t-)$$

then, the stochastic differential equation ruling the evolution of $(A_N(t))$ is

$$\mathrm{d}A_N(t) = - \sum_{i=1}^{A_N(t-)B_N(t-)} \mathcal{N}^i_{k/N^2}(\mathrm{d}t) \tag{1.15}$$

with $A_N(0) = M_N$.
The existence and uniqueness of such an equation is given by proposition A.11 p.356 of [Robert, 2003].

From microscopic interactions to the deterministic law of mass action

We now draw the link between the stochastic modelling and the deterministic modelling of chemical reactions. We do not get into the details of the rigorous mathematical proofs since it will done in the following chapters for the reactions involved in polymerisation.

The idea is, from the equivalent integral form of equation (1.15)

$$A_N(t) = A_N(0) - \sum_{i=1}^{+\infty} \int_0^t \mathbb{1}_{\{A_N(s-)B_N(s-)\geq i\}} \mathcal{N}^i_{k/N^2}(\mathrm{d}s)$$

to rewrite it according to the martingales associated to the Poisson point measures $\mathcal{N}^i_{k/N^2}$

$$\mathcal{M}_N(t) = \sum_{i=1}^{+\infty} \int_0^t \mathbb{1}_{\{A_N(s-)B_N(s-)\geq i\}} \left(\mathcal{N}^i_{k/N^2}(\mathrm{d}s) - \frac{k}{N^2}\,\mathrm{d}s \right)$$

with quadratic variation

$$\langle \mathcal{M}_N \rangle(t) = k \int_0^t \frac{A_N(s)B_N(s)}{N^2}\,\mathrm{d}s$$

in order to obtain, at the first order, the deterministic ODE (1.12). The proof of the fact that $\mathcal{M}_N$ is actually a square integrable martingale with the corresponding quadratic variation is given for instance in proposition 6.2 p.143 of [Robert, 2003].
Hence,

$$A_N(t) = A_N(0) + \mathcal{M}_N(t) - \int_0^t \frac{A_N(s)B_N(s)}{N^2}\,\mathrm{d}s.$$

We consider the fluid limits

$$\overline{A}_N(t) = \frac{A_N(Nt)}{N}$$

and

$$\overline{B}_N(t) = \frac{B_N(Nt)}{N}$$

(see section 1.4 for the definition of a fluid limit) that satisfy

$$\overline{A}_N(t) = m + \frac{\mathcal{M}_N(Nt)}{N} - \int_0^t \overline{A}_N(s)\overline{B}_N(s)\,\mathrm{d}s. \tag{1.16}$$

In all the following chapters, we use the same methodology. We start by proving with Doob's inequality that the term martingale $\mathcal{M}_N(t)/N$ vanishes uniformly on finite intervals. Then, by writing the same stochastic differential equation for $(\overline{B}_N(t))_N$ and using the fact that $(\overline{A}_N(t))$ and $(\overline{B}_N(t))$ are bounded, we show that the sequences of processes $(\overline{A}_N(t))_N$ and $(\overline{B}_N(t))_N$ are tight. Let $(c_A(t))$ (resp. $(c_B(t))$) be one of the limiting points of $(\overline{A}_N(t))_N$ (resp. $(\overline{B}_N(t))_N$), they necessarily satisfy the following differential

equation:

$$\dot{c_A}(t) = -kc_A(t) \cdot c_B(t)$$

with $c_A(0) = m$, which is precisely the law of mass action (1.12). Hence, at the first order, with a change of time scale, when N tends to infinity, we go back to the deterministic modelling.

In the following chapters, we will also prove functional central limit theorems, *i.e* limits for the convergence in distribution of process of the quantity

$$\left(\frac{A_N(t) - Nc_A(t)}{\sqrt{N}}\right)_t.$$

The last step in the modelling process is to estimate the kinetic parameters of the chemical reaction, like the parameter k of equation (1.11) for instance.

1.3.3 Curve-fitting

In this section, we briefly discuss some methods of parameter estimation. Essentially, the kinetic behaviour of the models is obtained deterministically and is derived from the law of mass action. The Finke-Watzky mechanism (F–W mechanism) (1.9) is often considered as one of the best models able to fit a broad set of data, including amyloid, α-synuclein, and polyglumatine [Morris et al., 2008] and still be simple. The constant of reactions k_1 and k_2 are obtained deterministically by least-squares from the law of mass action derived from (1.9), with c_1 the concentration of $\mathcal{X}_1$, the free monomers, and c_2 the concentration of $\mathcal{X}_2$, the polymerised monomers:

$$\dot{c_2}(t) = k_1c_1(t) + k_2c_1(1) \cdot c_2(t).$$

A 'fit' is a set of parameters that allow the model to reproduce the experiments. An example of a fit obtained by Morris *et al.* by the F–W mechanism in [Morris et al., 2008] is shown in picture 1.9.

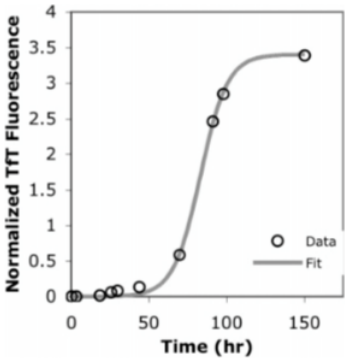


Figure 1.9: Retrieved from [Morris et al., 2008]. Fit of the F–W mechanism on data of amyloid β peptide aggregation published in [Bieschke et al., 2005].

In this case, data consist in one experiment of polymerisation of amyloid β, so that a deterministic approach is natural. Other approaches are empirical, which have the drawback of not being mechanistically relevant. The experimental sigmoidal curves are often fit by a logistic function [Naiki and Gejyo, 1999, Naiki and Nakakuki, 1996, Hasegawa et al., 2002], but it is then not easy to determine the physical meaning of the constants, whether they are related to the initiation of polymerisation or to the autocatalytic part.

Finally, a deterministic approach does not take into account the stochasticity of the initial phase. In [Szabo, 1998] and [Hofrichter, 1986b], the distribution of the time of nucleation is fit to data by assuming that for proteins exhibiting a large variance in the lag phase when polymerising, the tail of the distribution of the lag time is exponential. Then, parameters of growth and secondary processes are fit empirically. More recently, [Szavits-Nossan et al., 2014] proposed a complete stochastic approach for modelling protein self-assembly. However, as far as curve-fitting is concerned, parameters were obtained by a fit on the deterministic prediction of [Knowles et al., 2009].

In this dissertation, we use a minimalistic 2-step model and derive its kinetic parameters in a stochastic framework. In chapter II, we fit the parameters of nucleation and autocatalytic growth on the data presented section 1.2.1.

In the next section, we get into the details of the scaling methods that will be used throughout the two next chapters.

1.4 Scaling methods

In the previous section 1.3.2, we introduced the fluid limit (the definition will be given below) associated with the Markov process $(A_N(t))$ to get back, at the first order, to the law of mass action. Fluid limits are actually a convenient tool to capture the main behaviour of a Markov process $(X_N(t))$ by erasing some stochastic fluctuations.
More generally, scaling in time and space of Markov processes consists in finding appropriate sequences $(\omega_N(t))$ and (ν_N) in order to study the sequence of processes

$$(\overline{X}_N(t)) = \left(\frac{X_N(\omega_N(t))}{\nu_N}\right)$$

when N tends to infinity. Here, our scaling parameter is the molecular volume N previously defined. By appropriate, we mean that the limit of this rescaled process is able to reproduce the main characteristics of the initial Markov process. Renormalisation in time and space is quite classical in statistical physics (see for instance [Comets, 1991]), where it is referred to as 'hydrodynamic limits'. The underlying idea is to study the macroscopic behaviour from the microscopic dynamics.

In this PhD dissertation, since we are dealing with chemical reaction, it is natural to consider that the total number of particules in our system, *i.e* 'the molecular volume' N scales our Markov processes in space. The scaling in time is more complicated. In fact, when studying a network involving many chemical reactions in our case, the time scale $t \mapsto \omega_N(t)$ allows to focus on specific steps of the mechanism. If we consider a reaction requiring the encounter of two molecules, then the transition rates of the associated Markov processes are $O(1)$. Therefore, in order to see the polymerisation happening, *i.e* to observe variations of order N in the quantity of polymerised monomers, the correct time scale is the linear one $t \mapsto Nt$. The scaling in time and space by the initial size of the system is called a *fluid limit*. We give a

more precise definition [Robert, 2003]:

Definition 1.4.1. *A* fluid limit *associated with the continuous-time Markov process* $(X_N(t))$ *such that* $X_N(0) = N$ *is a stochastic process which is one of the limits of the renormalised process*

$$(\overline{X}_N(t)) = \left(\frac{X_N(Nt)}{N}\right)$$

when N *tends to infinity.*

For the general chemical reaction (1.11), we are in the nice case where the fluid limit is unique and deterministic, solution of the ODE describing the law of mass action. Fluid limits provide an asymptotic description when the initial size of the system is large. From then, we can get more information about the fluctuations of the initial Markov process by studying the diffusion around the fluid limit $\overline{X}_N(t)$ [Anderson and Kurtz, 2011]. We will follow this methodology in chapter II. However, it happens, as it will be shown in chapter III, that fluid limits do not give a useful description of the process, so that we have to find another time scale for our study. Indeed, in one of the proposed models of this chapter, we rescale the rates of reaction to slow down the nucleation. As a consequence, the linear time scale is not fast enough to see variations in the quantity of polymers of the order of N.

Coexistence of different time scales. The situation is more complicated when we consider a reaction network with different chemical reactions evolving on different time scales. It is the case when we consider a conformation step, as done in chapter III. Then, the transition rates of the misfolding are of the order of N. As a result, this process is very fast, while the following steps of polymerisation including nucleation and growth are slow. Therefore, locally around a time t, the slow process of polymerisation sees the fast process of misfolding at equilibrium. These two processes are intrinsically linked because the misfolding depends on the quantity of free monomers, *i.e* on the polymerisation, while the polymerisation depends on the number of misfolded monomers. When different time scales coexist in a network in this way, then it is defined as a stochastic averaging problem.
This phenomenon has already been mentioned in biochemistry in 1913 by Michaelis and Menten as 'time-scale separation' [Michaelis and Menten, 1913]. It has also been studied in the deterministic framework by Guckenheimer and Holmes in [Guckenheimer and Holmes, 1990] and in statistical mechanics, in 1961 by Bogolyubov [Bogoliubov, 1961]. In the stochastic context, Khaminskii's studies on averaging for stochastic calculus [Khasminskii, 1968] were further developed by Papanicolaou et *al.* *in* in 1977 [Papanicolaou et al., 1977] and by Freidlin and Wentzell [Freidlin and Wentzell, 1998]. Averaging has also been investigated for loss networks by Hunt and Kurtz in 1994 [Hunt and Kurtz, 1994]. In chapter III, we describe a stochastic averaging principle for a model of polymerisation including a conformation step.

1.5 Presentation of subsequent chapters

This first part of the thesis contains two chapters, each of them corresponding to a paper:

Chapter II. S. Eugène, W. Xue, P. Robert, and M. Doumic. Insights into the variability of nucleated amyloid polymerization by a minimalistic model of stochastic protein assembly. *Journal of Chemical Physics*, vol. 144, iss. 17, p. 175101+, 2016.

Chapter III. M. Doumic, S. Eugène, and P. Robert. Asymptotics of Stochastic Protein Assembly Models. Submitted to *SIAM Appl. Math.*, 2016. http://arxiv.org/abs/1603.06335.

In this section, we introduce these chapters and discuss the main contributions of this thesis.

1.5.1 Chapter II: Introduction of a minimalistic 2-step model

The goal of this chapter is to introduce a simple stochastic model for protein polymerisation in order to confront it to the data presented in section 1.2.1, able to explain the surprisingly high variance (regarding the volumes considered) observed in the initial lag phase.

Our approach departs from the complete model (5.5) and simplifies it to a minimalistic 2-step model, by using the concept of polymerised monomers. We make our usual assumptions (1.10) (including nucleus size $i_0 = 2$), so that the model introduced in this chapter is the following:

$$\begin{aligned} \mathcal{X}_1 + \mathcal{X}_1 &\xrightarrow{\alpha/N^2} 2\mathcal{X}_2 \\ \mathcal{X}_1 + \mathcal{X}_2 &\xrightarrow{\beta/N^2} 2\mathcal{X}_2 \end{aligned} \tag{1.17}$$

with as usual, $\mathcal{X}_1$ being the free monomers, $\mathcal{X}_2$ the polymerised monomers. We write explicitly the dependence on the volume N in the rates of reaction, in order to put emphasis on the encounter of two monomers for the reactions to happen, as it is not always clear in the chemical literature what is hidden behind the word 'rate of reaction'.

We define as usual $X_1^N(t)$ (resp. $X_2^N(t)$) the number of free monomers at time t (resp. the number of polymerised monomers) in a volume N (the dependence in the volume N is from now on written explicitely to avoid confusions in the calculations). We also make the following assumption:

(i) In order to capture the slow initiation (first reaction) as compared to the quick autocatalysis (second reaction), we assume $\alpha \ll \beta$.

(ii) The initial number of introduced monomers at $t = 0$ is denoted by M. We then have a conservation of the total number of monomers for all t: $X_1^N(t) + X_2^N(t) = M_N$.

(iii) The volume N is large and is intended to go to infinity.

(iv) The initial concentration of monomers $m = \lim_{N\to\infty} M_N/N$ remains constant in all our calculations.

Following the approach described in section 1.3.2, we study the corresponding Markov process $(X_1^N(t), X_2^N(t))$,

the transitions of which being given by the law of mass action, *i.e* for $(x_1, x_2) \in \mathbb{N}^2$,

$$\begin{cases} (x_1, x_2) & \longrightarrow (x_1, x_2) + (-2, 2) \quad \text{at rate } \alpha x_1(x_1 - 1)/(2N^2) \\ (x_1, x_2) & \longrightarrow (x_1, x_2) + (-1, 1) \quad \text{at rate } \beta x_1 \times x_2/N^2. \end{cases}$$

The first reaction represents the nucleation step, leading to a lag phase in the experiments shown in picture 1.2. It is often assumed in the literature that the concentration of monomers is more or less constant during the whole process of polymerisation [Morris et al., 2008, Szavits-Nossan et al., 2014], so that the rate of the first reaction can be rewritten as αm^2. Here, we want to cover the whole aggregation process, we do not make any approximation and take into account the depletion of monomers. The order of reaction is then quadratic, and depends on the probability for two monomers to meet, *i.e* the product of the concentrations (*c.f* section 1.3.2).

In the second reaction, $\mathcal{X}_2$ is both a reactant and a product of the reaction, defining an autocatalysis. This captures the steep take-off on the experimental curves (figure 1.2). Hence, the β parameter is a mixture of polymerisation and accelerating secondary processes. The concept of polymerised monomer gives to each monomer of a fibril the same power of polymerisation, so that we do not have to invoke the number of polymers as it is often done in the polymerisation literature [Szavits-Nossan et al., 2014].

Law of mass action for the minimalistic 2-step model. Kinetics of (1.17) are obtained, at the first order, from the law of mass action:

$$\dot{x_1}(t) = -\alpha x_1(t)^2 - \beta x_1(t) \cdot x_2(t). \tag{1.18}$$

with $x_1(t)$ (resp. $x_2(t)$) the deterministic concentration of $\mathcal{X}_1$ (resp. $\mathcal{X}_2$). For clarity, x_1 and x_2 denote respectively the concentrations of free monomers and polymerised monomers in the minimalistic 2-step model, and c_i the concentration of polymers of size i in the complete model (5.5).

From Becker-Döring to our minimalistic model. This minimalistic mechanism is phenomenological and despite its apparent simplicity, it derives directly from Becker-Döring system, which takes into account the size distribution of the fibrils $(X_i)_{i\geq 1}$. Indeed, the idea is that $\mathcal{X}_2$ represents the mass polymerised, *i.e*, from the stochastic point of view

$$\frac{X_2^N(t)}{N} = \sum_{i=2}^{\infty} i \frac{X_i^N(t)}{N}$$

or, in the deterministic framework

$$x_2(t) = \sum_{i=2}^{\infty} i c_i(t).$$

It can easily be seen on the deterministic kinetics. If we go back to a general nucleation-growth model of polymerisation (with assumptions (1.10)) we have:

$$\begin{aligned} \mathcal{X}_1 + \mathcal{X}_1 &\xrightarrow{k_{\text{on}}^n} \mathcal{X}_2 \\ \mathcal{X}_{\text{i}} + \mathcal{X}_1 &\xrightarrow{k_{\text{pol}}^i} \mathcal{X}_{\text{i+1}} \quad \text{for } i \geq 2 \end{aligned}$$

which kinetics are given by equations (1.14), we get:

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=2}^{\infty} i c_i(t) = \sum_{i=2}^{\infty} i \left[J_{i-1} - J_i \right] = 2J_1 + \sum_{i=2}^{\infty} J_i \tag{1.19}$$

where $J_i = k_{\text{pol}}^i c_i \cdot c_1$ for $i \geq 2$, $J_1 = k_{\text{on}}^N c_1^2$. Here, we must remember that the Becker-Döring system models a priori only growth by monomer addition, without any enhancing secondary pathway. To include the autocatalysis, we assume that the rates of reactions are linearly correlated to the length of the polymer (which is coherent with the fact that each polymerised monomer has the same power of polymerisation), *i.e*

$$k_{\text{pol}}^i = \beta i.$$

We also denote $2k_{\text{on}}^N = \alpha$ so that equation (1.19) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=2}^{\infty} i c_i(t) = \alpha c_1^2(t) + \beta c_1(t) \sum_{i=2}^{\infty} i c_i(t)$$

which is precisely what we wanted, *i.e* equation (1.18) by choosing $c_2(t) = \sum_{i=2}^{\infty} i c_i(t)$.

Advantages of the 2-step model. Hence, this simplification captures the main features of the mechanism of polymerisation, with α being an averaged rate of nucleation, β an averaged rate of growth but allows a clear deconvolution of the nucleation and the growth. Finally, this model only has two parameters, which makes it a good candidate for confrontation to data.

Time scale of polymerisation. Before presenting our results, we briefly discuss the time scale of the polymerisation. By using the methods described in section 1.3.2, we can derive the stochastic evolution equations of the process $(X_2^N(t))$:

$$X_2^N(t) = \mathcal{M}_N(t) + \alpha \int_0^t \frac{X_1^N(s)(X_1^N(s)-1)}{N^2} \,\mathrm{d}s + \beta \int_0^t \frac{X_1^N(s) X_2^N(s)}{N^2} \,\mathrm{d}s \tag{1.20}$$

where $\mathcal{M}_N(t)$ is a martingale whose increasing process is given by

$$\langle \mathcal{M}_N \rangle(t) = 2\alpha \int_0^t \frac{X(s)(X(s)-1)}{N^2} \,\mathrm{d}s + \beta \int_0^t \frac{X(s)}{N} \frac{(M - X(s))}{N} \,\mathrm{d}s.$$

On the normal time scale, we can show that, for the convergence in distribution of processes

$$\lim_{N\to\infty}\left(\frac{X_2^N(t)}{N}\right)=0 \qquad \text{and} \qquad \lim_{N\to\infty}\left(\frac{X_1^N(t)}{N}\right)=\lim_{N\to\infty}\left(m-\frac{X_2^N(t)}{N}\right)=m. \tag{1.21}$$

Indeed, let $\epsilon>0$.

$$\mathbb{P}\left(\sup_{0\le t\le T}\mid X_2^N(t)\mid\ge\epsilon N\right)\le\mathbb{P}\left(\sup_{0\le t\le T}\mid\mathcal{M}_N(t)\mid\ge\frac{\epsilon}{2}N\right)+\mathbb{P}\left(\alpha m^2T+\beta mT\ge\frac{\epsilon}{2}N\right).$$

By Doob's inequality, one gets:

$$\mathbb{P}\left(\sup_{0\le t\le T}|\mathcal{M}_N(t)|\ge\frac{\epsilon}{2}N\right)\le\frac{\mathbb{E}\left(\langle\mathcal{M}_N\rangle(T)\right)}{(\epsilon/2)^2N^2}\underset{N\to\infty}{\longrightarrow}0.$$

So that:

$$\mathbb{P}\left(\sup_{0\le t\le T}\mid X_2^N(t)\mid\ge\epsilon N\right)\underset{N\to\infty}{\longrightarrow}0.$$

which proves equation (1.21).

In fact:

Proposition 1.5.1. *On the normal time scale $t\mapsto t$, when N tends to infinity, the process $(X_2^N(t))$ converges in distribution to a Poisson process of parameter αm^2.*

Proof. The proof follows the same steps as theorem 3 of [Feuillet et al., 2014]. Actually, $t\mapsto X_2(t)$ can be seen as a point process with jumps of size 1. Relation (1.20) shows that

$$\left(X_2^N(t)-\alpha\int_0^t\frac{X_1^N(s)(X_1^N(s)-1)}{N^2}\,\mathrm{d}s-\beta\int_0^t\frac{X_1^N(s)X_2^N(s)}{N^2}\,\mathrm{d}s\right)$$

is a martingale with respect to the natural filtration of the associated Poisson processes. The random measure

$$\eta_N(t)=\alpha\int_0^t\frac{X_1^N(s)(X_1^N(s)-1)}{N^2}\,\mathrm{d}s+\beta\int_0^t\frac{X_1^N(s)X_2^N(s)}{N^2}\,\mathrm{d}s$$

is a *compensator* of the point process $t\mapsto X_2^N(t)$ (see [Kasahara and Watanabe, 1986]). Since the sequence of processes $(\eta_N(t))$ is tight by using the fact that $(X_1(t)/N)$ and $(X_2(t)/N)$ are bounded processes, we actually showed that the random measure η_N is converging to the deterministic mesure $\alpha m^2\,\mathrm{d}s$. By theorem 5.1 of [Kasahara and Watanabe, 1986], the result is proved. □

It correlates the fact that polymerisation in a slow process; to see polymers being produced, we have to speed up the time and study the fluid limit of $(X_2^N(t))$, as described in section 1.4.

Main contributions of this chapter. In chapter II we start by deriving the asymptotics of the concentration of monomers on the linear time scale. For this purpose, we introduce the notation

$$\overline{X}_i^N(t) = \frac{X_i^N(Nt)}{N}$$

for $i \in \{1, 2\}$ We prove that, with standard results of stochastic calculus, if $M_N/N \to m$, then

$$\lim_{N\to\infty} \overline{X}_1^N(t) = c_1(t)$$

where $c_1(t)$ is precisely the solution of the law of mass action written in equation (1.18) and then showed the following functional central limit theorem. For the convergence in distribution,

$$\lim_{N\to+\infty} \left(\frac{X_1^N(Nt) - Nc_1(t)}{m\sqrt{N}} \right) = (U(t)),$$

where $U(t)$ is a diffusion.

From then, we derived the asymptotics of the time for δ reaction completion

$$T_N(\delta) = \inf\{t > 0, X_1^N(t) \leq (1-\delta)M_N\}.$$

Remember that our main goal was to derive the variance of this random variable in order to explain the fluctuations of the initial phase of polymerisation. For this end, we derived from the asymptotics of $\left(\overline{X}_1^N\right)$ a central limit theorem for $T_N(\delta)$

$$\lim_{N\to+\infty} \frac{T_N(\delta) - Nt_\delta}{\sqrt{N}} = \frac{U(t_\delta)}{m[\alpha(1-\delta)^2 + \beta\delta(1-\delta)]}$$

where t_δ is the deterministic time for δ reaction completion, *i.e* $t_\delta = c_1^{-1}(\delta)$, allowing us to compute the standard deviation σ_N, when N tends to infinity

$$\lim_{N\to\infty} \sigma_N = \sigma = \sqrt{\frac{3}{2M_N\alpha\beta m^2}}. \tag{1.22}$$

The last result of this chapter is the estimation of the parameters α and β. The main difficulty is that we have a very small set of data (twelve curves per concentration). To find a robust estimation, we showed that actually all the curves superimpose very well, so that the slope of the curve at $T_N(1/2)$ is the same for all the realisations of the process $t \mapsto \overline{X}_2^N(t)$. We based our estimation on this result, and managed to obtained a robust estimation of β, robust in the sense that we obtained almost the same value of β for all the initial concentrations considered. This means that the second reaction of (1.17), despite its drastic simplification, captures the autocatalysis correctly. Unfortunately, the estimation of α is less pretty, since it models the initial take-off, which is highly stochastic.

1.5.2 Chapter III: Asymptotics of stochastic protein assembly models

The minimalistic 2-step model has essentially two limitations.

First, it does not explain the variance observed in the large experimental volumes. By volume, we mean indifferently the physical volume V, $15\mu L$ in our set of experiments, or $N = N_A \cdot V$, or M_N, the initial **number** of monomers introduced (since it is proportional to N). However, we find it more convenient to think in terms of M_N, since it has a real physical meaning (on the contrary to N), and it reveals what we mean by 'large' (on the contrary to V). We reason with the twelve experimental curves obtained for $m = 122\mu M$ showed figure 1.3 to find the typical 'volume' M_N

$$M_N = m \cdot V \cdot 10^{-6} \cdot N_A = 1,1 \cdot 10^{15}.$$

For this order of magnitude, formula (1.22) gives us $\sigma = 1.98 \cdot 10^{-4}\, h$ for the estimated parameters α and β corresponding to the initial concentration $m = 122\mu M$ when the experiments have a standard deviation of $0.9\, h$ (hours). This means that our minimalistic model, with our estimated parameters, is not able to reproduce the variance of the lag time observed in large volumes. It predicts however a variance of order $1h$ for a smaller intial number of monomers, namely $M_N = 10^6$.

The second problem is the regime of validity of our asymptotic expansions. Indeed, in our calculations, we make N tend to infinity and consider that the parameters α and β are constant. However, in practice N is large but not infinite, and the parameters estimated have extreme values to reproduce the sigmoidal shape. By extreme, we mean that α must be very small compared with β in order to capture the slow ignition phase. We estimated for instance $\alpha = 1.33^{-10} h^{-1} \cdot \mu M^{-1}$ for $m = 122\mu M$. As a result, α/β becomes comparable to N and cannot be treated as a constant anymore.

The third chapter of this part tries to tackle these two problems by introducing two variants of the minimalistic 2-step model:

— A first model including a conformation step, which should increase the variance,

— A second model equivalent to the minimalistic 2-step model but with α/N^ν ($\nu > 0$) as a parameter for the nucleation phase instead of α.

Including a conformation step. As said in section 1.2.3, misfolding of proteins is strongly suspected to be the first step towards polymerisation. If we want to refine our study of the initiation of the polymerisation, in order to study its variance, it is then natural to introduce the initial change of conformation of monomers. The rest of the reaction, *i.e* nucleation and growth, is the step as in the simple minimalistic

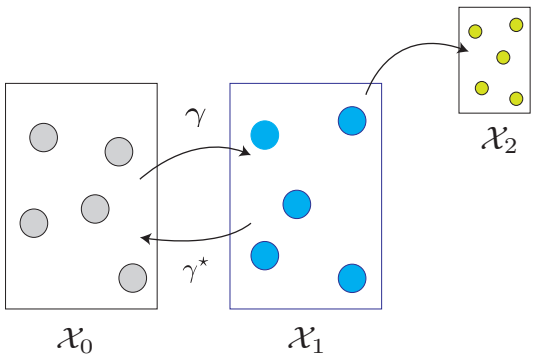


Figure 1.10: Ehrenfest urn for the polymerisation model with misfolding (1.23). The grey balls are the introduced free monomers, the blue ones the misfolded free monomers and the yellow ones the polymerised monomers.

model. The model proposed is then:

$$\begin{aligned} \mathcal{X}_0 &\underset{\gamma^*}{\overset{\gamma}{\rightleftharpoons}} \mathcal{X}_1 \\ \mathcal{X}_1 + \mathcal{X}_1 &\xrightarrow{\alpha/N^2} 2\mathcal{X}_2 \\ \mathcal{X}_1 + \mathcal{X}_2 &\xrightarrow{\beta/N^2} 2\mathcal{X}_2 \end{aligned} \tag{1.23}$$

with $\gamma \ll \gamma^*$. According to the discussion on the rates of reaction in section 1.3.2, the transition rates of the corresponding Markov process $(X^N(t)) = (X_0^N(t), X_1^N(t), X_2^N(t))$ are the following, for an element $x = (x_0, x_1, x_2) \in \mathbb{N}^3$,

$$x \mapsto \begin{cases} x+(1,-1,0) & \text{at rate } \gamma^* x_1 \\ x+(-1,1,0) & \gamma x_0, \end{cases} \qquad x \mapsto \begin{cases} x+(0,-2,2) & \alpha (x_1/N)^2 \\ x+(0,-1,1) & \beta x_1/N \times x_2/N, \end{cases} \tag{1.24}$$

since the change of conformation does not depend on the volume available but on the number of monomers. Hence, the rates of the change of conformation are much larger than the one for the polymerisation.

As a result, this model can be seen as two processes evolving on different time scales: a fast process, the change of conformation $(X_0(t), X_1(t))$, and a slow process, the polymerisation, which is precisely the minimalistic model previously studied. Polymerisation is happening on the linear time scale $t \mapsto Nt$ while misfolding is, locally around a time t, instantaneously at equilibrium. In this chapter, we study the first step of (1.23) as an Ehrenfest urn with rates γ and γ^* the size of which varying in time as the polymerisation occurs, as shown in figure 1.10. These two processes are fully coupled, in the sense that the stochastic evolution of $X_2(Nt)$ depends on the quick equilibrium of the urn, and this equilibrium depends on the size of the urn, $M_N - X_2(Nt)$. We prove a Stochastic Averaging Principle (SAP) (*c.f* section 1.4) and the corresponding central limit theorem.

Rescaled reaction rates. In this section we take into account the extreme values of the parameters involved in the polymerisation mechanism. In fact, in order to capture the long phase observed on the data by only a dimerisation step, the α-parameter must be very small, that is, of order $N^{-\nu}$, with $\nu > 0$. α is therefore changed into α/N^{ν} in the minimalistic 2-step model:

$$\mathcal{X}_1 + \mathcal{X}_1 \xrightarrow{\alpha \cdot N^{-\nu}/N^2} 2\mathcal{X}_2 \tag{1.25}$$

$$\mathcal{X}_1 + \mathcal{X}_2 \xrightarrow{\beta/N^2} 2\mathcal{X}_2 \tag{1.26}$$

with the usual corresponding Markov process with transition rates for $(x_1, x_2) \in \mathbb{N}^2$,

$$\begin{cases} (x_1, x_2) & \longrightarrow (x_1, x_2) + (-2, 2) \quad \text{at rate } \alpha \cdot \boldsymbol{N^{-\nu}} (x_1/N)^2 \\ (x_1, x_2) & \longrightarrow (x_1, x_2) + (-1, 1) \quad \text{at rate } \beta x_1 \times x_2/N^2. \end{cases}$$

In this model, it appears that that the polymerisation will happen on a faster time scale than the linear one, since we have slowed down the nucleation as compared to the minimalistic 2-step model. The time scale of polymerisation depends here actually on ν. In this chapter, we prove that the polymerisation happens on the time scale $t \mapsto N \ln N t$ for $0 < \nu \leq 1$, and on the faster time scale $t \mapsto N^{\nu} t$ for $\nu > 1$.

Intuitively, for $0 < \nu \leq 1$, we understand that on the time scale $t \to N^{1+\delta} t$, the polymerisation has already finished for all $\delta > 0$. Indeed, initially, since the polymerisation happens on a faster time scale than the linear one, $X(Nt)/N = m$. Thus, on $t \mapsto N^{1+\delta} t$, reaction (1.25) leads to a quantity $\alpha m^2 N^{1-\nu+\delta} t$ of polymers. Afterwards, reaction (1.26) consumes all the monomers: it creates a quantity $\beta m N^{1-\nu+\delta} N^{\delta} t$ of polymers, *i.e* X_2 becomes of order $N^{1-\nu+2\delta}$, then a quantity $N^{1-\nu+3\delta}$ etc... X_2 becomes of order $N^{1-\nu+k\delta}$ until it reaches N. Finally, on the time scale $t \to N^{1+\delta} t$, the polymerisation has already finished.

For $\nu > 1$, we are in an extreme regime where only the formation of one polymer is enough to trigger the polymerisation, as studied in [Yvinec, 2012, Szavits-Nossan et al., 2014] so that the lag time is actually an exponential random variable of parameter αm^2.

1.6 Future directions

The minimalistic 2-step model is able to capture the two main phenomena involved in polymerisation, and predict a large variance of the lag time for an initial number of monomers capped to 10^6. However, it does not give a large variance for the experimental initial number of monomers introduced, that is 10^15 monomers. For the other models presented in chapter III, we only provide a theoretical study, but the confrontation to the experimental data and the estimation of parameters are more tedious.

Procedure of estimation of parameters. The model (1.23) with a conformation step suggests a variance

$$\lim_{N\to+\infty} \mathrm{Var}(T_N(\delta)) = \frac{3}{2M\alpha r^2\beta r m^2}.$$

with $r = \gamma/(\gamma+\gamma^*)$, the fraction of misfolded monomers in the Ehrenfest urn composed of free monomers $\mathcal{X}_0$ and misfolded monomers $\mathcal{X}_1$.

Our study of the minimalistic 2-step model proved that the variance of the lag time was highly dependent on the number of monomers able to polymerise. Hence, the parameter r plays an important role in this variance since it quantifies the number of misfolded monomers. It also gives another degree of freedom to the model for the parameter estimation procedure. The idea in the models with rescaled rates is the same. We create a filter on the number of monomers able to polymerise with the parameter ν. As a conclusion, to explain the variance, a possible way would be to find a new estimation of the set of parameters (α, β, r) for the model with the conformation step, or for (α, β, ν) for the rescaled rates, taking, in addition to the sigmoidal shape, the experimental variance as an entry.

Sophistication of the model. The other obvious way for future work in the stochastic modelling of protein polymerisation is to study the complete model

$$\begin{aligned}
&\mathcal{X}_0 \quad \underset{\gamma^*}{\overset{\gamma}{\rightleftharpoons}} \quad \mathcal{X}_1 \\
&\mathcal{X}_1 + \mathcal{X}_1 \quad \underset{k_{\mathrm{off}}^2}{\overset{k_{\mathrm{on}}^1}{\rightleftharpoons}} \quad \mathcal{X}_2 \qquad \Big\} \text{ Nucleation} \\
&\qquad\vdots \\
&\mathcal{X}_1 + \mathcal{X}_{i_0-1} \quad \underset{k_{\mathrm{off}}^{i_0}}{\overset{k_{\mathrm{on}}^{i_0-1}}{\rightleftharpoons}} \quad \mathcal{X}_{i_0} \qquad \Big\} \\
&\mathcal{X}_1 + \mathcal{X}_{i_0} \quad \underset{k_{\mathrm{dep}}^{i_0+1}}{\overset{k_{\mathrm{pol}}^{i_0}}{\rightleftharpoons}} \quad \mathcal{X}_{i_0+1} \qquad \Big\} \\
&\qquad\vdots \\
&\mathcal{X}_1 + \mathcal{X}_i \quad \underset{k_{\mathrm{dep}}^{i+1}}{\overset{k_{\mathrm{pol}}^{i}}{\rightleftharpoons}} \quad \mathcal{X}_{i+1} \qquad \Big\} \text{ Growth} \\
&\qquad\vdots
\end{aligned}$$

in the stochastic framework. It would be more realistic, for a probabilistic modelling, to consider that the nucleation occurs by monomers addition, since the encounter of i_0 monomers, for $i_0 > 2$, is very unlikely, and would have a very low probability.

Chapter 2

Insights into the variability of nucleated amyloid polymerisation by a minimalistic model of stochastic protein assembly

Contents

The format of this chapter obeys to the recommendations of the *Journal of Chemical Physics*, *i.e* results are presented in the main part, while the proofs of the mathematical results are postponed in the supplemental material.

Introduction

The amyloid conformation of proteins is of increasing concern in our society because they are associated with devastating human diseases such as Alzheimer's disease, Parkinson's disease, Huntington's disease, Prion diseases and type-2 diabetes [Knowles et al., 2014a, Chiti and Dobson, 2006]. The fibrillar assemblies of amyloid are also of considerable interest in nano-science and engineering due to their distinct functional and materials properties [Fowler et al., 2007, Schwartz and Boles, 2013, Knowles and Buehler, 2011]. Elucidating the molecular mechanism of how proteins polymerize to form amyloid oligomers, aggregates and fibrils is, therefore, a fundamental challenge for current medical and nanomaterials research.

Amyloid diseases are associated with the aggregation and deposition of mis-folded proteins in the amyloid conformation [Knowles et al., 2014a, Chiti and Dobson, 2006]. Amyloid aggregates form through nucleated polymerization of monomeric protein or peptide precursors (e.g. [Xue and Radford, 2013, Kashchiev and Auer, 2010, Ferrone, 1999, Collins et al., 2004, Knowles et al., 2009]). The slow nucleation process that initiates the conversion of proteins into their amyloid conformation is followed by exponential growth of amyloid particles, resulting in growth of amyloid fibrils that is accelerated by secondary processes such as fibril fragmentation and aggregate surface catalyzed heterogeneous nucleation [Xue et al., 2008, Knowles et al., 2009, Cohen et al., 2013, Xue and Radford, 2013] (Figure 2.2). Current mathematical description of protein assembly into amyloid are based on systems of mass-action ordinary differential equations, and they have been successful in describing the average behaviour of amyloid assembly observed by kinetic experiments (e.g.[Xue et al., 2008, Knowles et al., 2009]). The formation kinetics of amyloid aggregates has been studied extensively by bulk *in vitro* experiments in volumes typically in the range of hundreds of μL or larger [Xue et al., 2008], but has also been observed recently in elegant microfluidic experiments in pL to nL range, more closely mimicking physiological volumes in tissues and cellular compartments [Knowles et al., 2011]. Amyloid growth experiments typically follow the appearance of amyloid aggregates or the depletion of monomers as function of time, yielding information regarding the rate of the exponential growth and the length of the lag phase under different protein concentrations at fixed volumes. A hitherto overlooked piece of information that can be derived from these kinetic experiments is the observed variation between experimental repeats, which may hold the key to understanding the early rare nucleation events of amyloid formation [Xue et al., 2008, Szavits-Nossan et al., 2014, Hofrichter, 1986b, Hofrichter, 1986a]. However, current deterministic models cannot describe variability, thus, unable to address whether the observed variations in lag phase length reflect subtle experimental differences between the replicates, contributions from the stochastic nature of the nucleation mechanism, or a combination of both factors. As shown recently by Szavits-Nossan and co-workers using

a stochastic nucleated growth model, rare nucleation events are expected to dictate the behaviour and variability of amyloid formation in small volumes such as in cellular compartments [Szavits-Nossan et al., 2014]. Understanding these rare initial nucleation events of amyloid formation and the variability resulting from the stochastic nature of nucleation, therefore, is of paramount importance in the fundamental understanding of amyloid diseases and in controlling amyloid formation.

Here, we present a new stochastic protein assembly model with the aim to capture the fundamental features of amyloid self-assembly that includes their stochastic nature, and still allow a fully rigorous mathematical analysis of these processes (Figure 2.1). In this spirit, our model contains minimal possible complexity needed to describe a nucleated protein polymerization process, allowing us to study it theoretically in a mathematically rigorous manner, but still allowing useful comparison to experimental data. From our minimal model, we derive a closed form formula that can describe and predict variability in the lag phase duration of nucleated protein assembly by giving a proof to a central limit theorem for our model. Our results demonstrate how stochasticity at the molecular level may influence the kinetics of the total reaction population at a macroscopic scale depending on the relative rates of nucleation and exponential growth, and on reaction volume. We also show how new information relevant to any specific nucleated amyloid assembly can be gained in a conceptually simple and clear manner by applying our analytical results to the analysis of published $\beta_2 m$ amyloid assembly kinetics data [Xue et al., 2008]. We demonstrate that our model qualitatively captures key features of the data such as parallel progress of the curves and the order of magnitude for the rates of the self-accelerating reactions. We also show that the intrinsic stochastic nature of nucleation alone cannot explain the observed variability in lag phase length for published $\beta_2 m$ amyloid assembly data acquired in large (15 μL) volumes suggesting alternative mechanistic assembly steps and additional experimental sources that contribute to the variability in the observed amyloid growth curves. Our approach represents the basis for the development of extensive and tractable stochastic models, which will allow the variability information from amyloid growth kinetics experiments to be used to inform the fundamental molecular mechanisms of the key rare initial events of amyloid formation that may be involved in producing early on-pathway cytotoxic species associated with amyloid disease.

Supplemental material presents the mathematical background of these results, in particular the rigorous proofs of the convergence results, the precise mathematical characterization of the variability of the assembly process and, finally, some simulations of these stochastic processes.

2.1 A phenomenological stochastic model

To make the model as simple as possible, we consider two distinct types of monomers, we call these species monomers and polymerised monomers, respectively. The polymerised monomers represent all monomers in the amyloid conformation in the aggregates. Its amount may be viewed as representing the total polymerised mass, captured for instance by Thioflavine T (ThT) measurements, as in Figure 2.2. Such a simplification is also justified by the fact that current kinetics measurements of amyloid growth

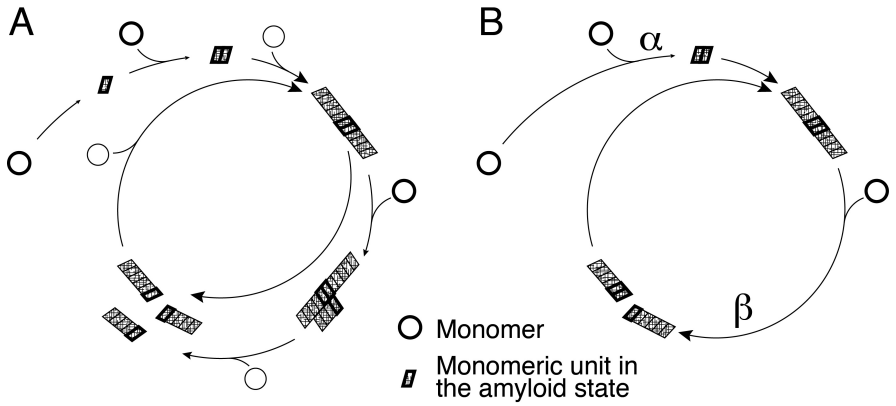


Figure 2.1: (A) Schematic illustration of a full amyloid assembly model, including conformational exchange, nucleation, elongation growth, secondary surface nucleation and fibril fragmentation processes. (B) Schematic illustration of the minimal model represented by reactions (1) and (2). The phenomenological parameters α and β represent the rate constant of the ignition phase, and the rate constant of all possible accelerated growth pathways to the formation of polymers, respectively. The circles represent the un-polymerised monomer $\mathcal{X}$ and the parallelograms represent the monomeric units in the amyloid formation $\mathcal{Y}$ in (1) and (2). Some monomeric units are highlighted with bold outlines to highlight few possible paths a monomeric unit in (1) and (2) can take through the aggregation process.

exhibit variability on the timecourse of the total polymerised mass, without giving any information on the size distribution of fibrils. However, such a simplification do not contain any contributions from spatial dynamics, molecular motion and transport processes, which may add complexity to the stochastic behaviour in small volumes. Previous studies (see for instance [Prigent et al., 2012], Supplemental material (S.M.) 2) have shown that the detail of the reactions of secondary pathways, such as a fragmentation kernel, may have a major impact on the size distribution of polymers, but comparably smaller effects on the timecourse of the polymerised mass. Overall, with this simplification, we can distill the problem down from infinite number of species to two species, which subsequently can describe the ability of the amyloid state to convert normal unpolymerised monomers to the amyloid state without invoking polymer ends or number. Thus, our model is phenomenological and aims to give new insights into the key determinants of stochastic behavior of protein aggregation and suggests simple ways to extract information from experimental data. Our approach departs from the mechanistic modelling approach used in conventional deterministic models of protein aggregation but is complementary to those models (e.g. [Xue et al., 2008, Knowles et al., 2009, Szavits-Nossan et al., 2014]), and the simplifications allows tractable mathematical derivation of closed expressions.

We thus consider two distinct species in our model: monomers, $\mathcal{X}_1$, and polymerised monomers, $\mathcal{X}_2$. We then consider $X_1^N(t)$ and $X_2^N(t)$ to be the respective numbers of particles of each species at time t in a fixed volume V. Initially, it is assumed that there are only M monomers: $X_1^N(0) = M$ and $X_2^N(0) = 0$. We denote $m = M/(V \cdot N_A)$ the initial molar concentration of monomers, where N_A is the Avogadro constant. For convenience in the calculations hereinafter, we introduce the notation $N = V \cdot N_A$. N will be our scaling parameter.

Thus, the chemical reactions associated with this simple model are as follows:

$$\mathcal{X}_1 + \mathcal{X}_1 \xrightarrow{\alpha/N^2} 2\mathcal{X}_2, \tag{2.1}$$

$$\mathcal{X}_1 + \mathcal{X}_2 \xrightarrow{\beta/N^2} 2\mathcal{X}_2, \tag{2.2}$$

where α/N^2 and β/N^2 are rates of the reactions with rate constants of $\alpha > 0$ and $\beta > 0$. These reactions describe the following features of a nucleated polymerisation of proteins that characterises amyloid assembly (see Figure 2.1 for an explanatory scheme of the reactions):

— Reaction (2.1): We call this step 'ignition' since it models the starting point of the polymerisation process. Here, we represents this step as the simplest possible concentration dependent nucleation step that converts two monomers into two monomers that are growth competent (equivalent to two polymerised monomers). The initiation step (1) is equivalent to a nucleation step involving dimer formation. This is a common simplification that has been applied in a number of deterministic model (e.g. [Knowles et al., 2009, Cohen et al., 2013]), and is also motivated by the fact that the first molecular attachment step towards the nucleation barrier may have the biggest energetic penalty according to the classical nucleation theory. In our model, this reaction will occur in a stochastic way. Following the principles of the law of mass action, the encounter of two chemical species occurs at a rate proportional to the product of the *concentrations* of each species. Therefore two given monomers disappear to produce two polymerised monomers at a rate α/N^2.

— Reaction (2.2): We call this second step 'conversion', which we modelled as a self-accelerating autocatalytic process. Here, given a monomer and a polymerised monomer, the monomer converts into a polymerised monomer at a rate β/N^2. This is representative of a range of accelerating secondary pathway reactions such as fragmentation, lateral growth, and aggregate surface catalyzed second nucleation. In this sense, our model may be viewed as a simplification and amalgamation of several mechanistic models. For example, in the case of fragmentation accelerated growth, fibril fragments that interact with monomers $\mathcal{X}_1$ are generated, in a first order approximation, proportional to the number of breakage sites [Xue and Radford, 2013], which in-turn depends on the number of monomeric units in the amyloid fibrils. In the case of secondary fibril surface nucleation, the sites that promote surface nucleation is proportional to available surface [Cohen et al., 2013], which is also dependent on the number of monomeric units in the amyloid fibrils (Fig 1). In particular, we expect our model to behave qualitatively similarly to the mechanistic model described in [Szavits-Nossan et al., 2014], which includes nucleation, polymerization, and fragmentation as a representative self-accelerating secondary process motivated by its experimental analysis [Knowles et al., 2011]. It is however not intended for reaction (2) to be associated to any specific microscopic meaning as described above.

Stochastic Evolution. Any given pair of monomers reacts together by Reaction (2.1) at rate α/N^2, whereas for a given pair of monomer/polymerized monomer reacts by Reaction (2.2) at rate β/N^2. Let M_N be the initial number of monomers and the random variable describing the number of monomers

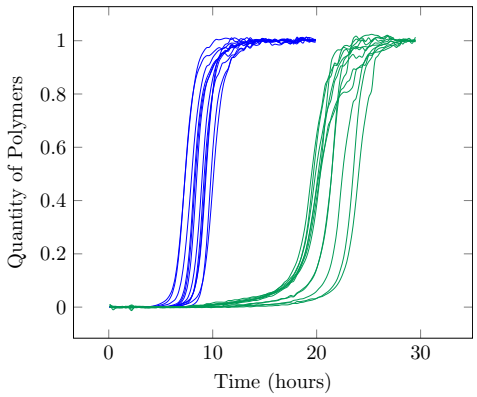


Figure 2.2: Twelve experimental timecourse of polymerised mass for two given initial concentrations of monomers: 122 μM (blue) and 30.5 μM (green) published in [Xue et al., 2008].

remaining at time t is denoted by $X_1^N(t)$. By taking into account the $X_1^N(X_1^N-1)/2$ monomers pairs, and the $X_1^N(M_N-X_1^N)$ monomers/polymerised pairs, the variable $X_1^N(t)$ has jumps of size -2 or -1 which occur at the following rates

$$X_1^N \mapsto \begin{cases} X_1^N-2 & \text{at rate } \dfrac{X_1^N(X_1^N-1)}{2} \times \dfrac{\alpha}{N^2}, \\ X_1^N-1 & \text{“ } \quad X_1^N(M_N-X_1^N) \times \dfrac{\beta}{N^2}. \end{cases} \tag{2.3}$$

The conservation of mass gives the additional relation $X_1^N(t)+X_2^N(t)=M_N$. As noticed previously, in the description of Reactions (2.1) and (2.2) above, this representation is completely coherent with the law of mass action.

2.1.1 Asymptotic evolution of the number of monomers

Assuming that the volume V is large and the initial concentration of monomers remains constant and equal to $m > 0$, i.e. the initial number of monomers M_N is such that $M_N/N \sim m$, we can derive the following:

Polymerisation occurs on the time scale t↦Nt. Let $(\overline{X}_1^N(t))$ be the scaled process defined by

$$\overline{X}_1^N(t) = \frac{X_1^N(Nt)}{N}. \tag{2.4}$$

In Equation (2.4), the time scale of the process $\left(X_1^N(t)\right)$ is accelerated with a factor N. As it will be seen, as N gets large, $t \to Nt$ is the correct time scale to observe the decay of $(X_1^N(t))$ on the space scale proportional to N.

Assuming for the moment that $\left(\overline{X}_1^N(t)\right)$ is converging in distribution, Relations (2.3) then suggest that

its limit $(x_1(t))$ should satisfy the following ODE

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = -\alpha x_1(t)^2 - \beta x_1(t)(m - x_1(t)), \text{ with } x_1(0) = m. \tag{2.5}$$

The following result shows that this is indeed the case.

Proposition 2.1.1 (Law of large numbers)**.** *If the initial number M_N of monomers is such that*

$$\lim_{N\to+\infty} \frac{M_N}{N} = m > 0,$$

then, as N goes to infinity, the process $(\overline{X}_1^N(t))$ converges in distribution to $(x_1(t))$, solution of Equation (2.5), given by the formula

$$x_1(t) = m\frac{\beta}{\alpha}\frac{1}{e^{\beta m t} - 1 + \beta/\alpha}. \tag{2.6}$$

The proof is classical [Ethier and Kurtz, 1986], we recall it in Sections 2.3.1 and 2.3.6 of supplemental material, we comment on the relative influence of the parameters α and β on the deterministic curve, see supplemental figure 2.7.

In order to be able to quantify the variability of experimental replicates, we need to go further, to a second order approximation, i.e. with a central limit result.

Proposition 2.1.2. *If the initial number M_N of monomers is such that*

$$M_N = mN + o\left(\sqrt{N}\right),$$

for $m > 0$, then, for the convergence in distribution,

$$\lim_{N\to+\infty} \left(\frac{X_1^N(Nt) - Nx_1(t)}{m\sqrt{N}}\right) = (U(t)),$$

where $U(t)$ is a diffusion, the unique solution of the following stochastic differential equation (2.12).

The proof is postponed in Section 2.3.2 of supplemental material, together with an explicit formulation and an analysis of the influence of the parameters α and β on the stochasticity of the reactions. We found that the smaller the ratio α/β is, the more important the influence of the stochasticity on the lag-time, but the less important for the following of the reaction. This is quantified in the following study of the stochastic time for δ completion below.

2.1.2 Asymptotics of the time for δ reaction completion

To quantify the effect of α and β on the stochasticity of the reactions, we define the *t*ime for δ reaction completion, where $0 < \delta < 1$ is a percentage, as the following stopping time

$$T_N(\delta) = \inf\{t > 0, X_1^N(t) \leq (1-\delta)M_N\}$$

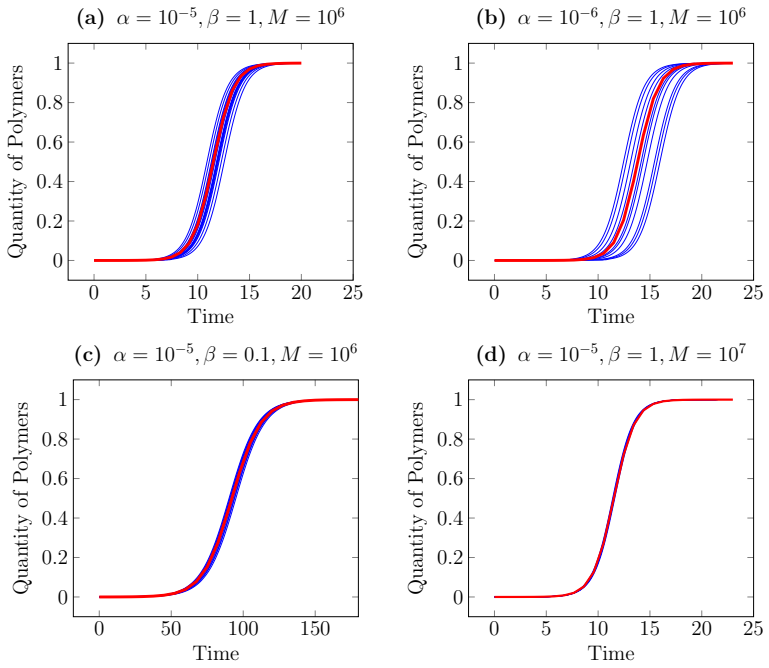


Figure 2.3: The blue curves represent twelve simulations of the model with different parameters. The red curve is the first order obtained in Proposition 2.1.1.

where T_N is the first time when there is a δ fraction of polymers is produced. T_N for δ small - 5 to 10% - represents an alternative definition for the lag-time of the reaction[Prigent et al., 2012].

A law of large numbers and a central limit result for $T_N(\delta)$ as V goes to infinity can be obtained. Note that due to the change in the time scale, we need to rescale T_N by V to get a limit.

Theorem 2.1.1 (Asymptotics of the time for degree of reaction completion δ)**.** *If the initial number M_N of monomers is such that*

$$M_N = mN + o\left(\sqrt{N}\right),$$

for $m > 0$ then, for the convergence in distribution

1. **Law of Large Numbers**.

$$\lim_{N\to+\infty} \frac{T_N(\delta)}{N} = t_\delta \overset{def.}{=} \frac{1}{\beta m} \log\left(1 + \frac{\beta\delta}{\alpha(1-\delta)}\right). \tag{2.7}$$

2. **Central Limit Result**.

$$\lim_{N\to+\infty} \frac{T_N(\delta) - Nt_\delta}{\sqrt{N}} = \frac{U(t_\delta)}{m[\alpha(1-\delta)^2 + \beta\delta(1-\delta)]}$$

where $(U(t))$ is the solution of the SDE (2.12)*.*

The proof of Theorem 2.1.1 is given in Section 2.3.4 of supplemental material. Supplemental Figure 2.6 illustrates and the law of large numbers and the central limit theorem for $T_{1/2}$. Note that the definition of t_δ, which is the limit of the stochastic times $T_N(\delta)/N$ when $N \to \infty$ is coherent with the definition of the *deterministic* time of δ reaction completion as

$$t_\delta = \inf\{t > 0,\, x_1(t) \leq (1-\delta)m\} = x_1^{-1}((1-\delta)m),$$

where $(x_1(t))$ is given by Equation (2.6). Thus, for any given experiment, the distance between a realization of $T_N(\delta)/N$ and t_δ is being given by the explicit formula above. We can therefore derive its stochastic behaviour. The following corollary establishes its variance.

Corollary 2.1.1 (Variance of Time $T_N(\delta)$)**.** *Under the assumptions of the above theorem and with its notations, the variance σ_N^2 of the time for δ completion has a limit σ^2, when $\alpha \ll \beta$*

$$\lim_{N \to +\infty} \sigma_N = \sigma \sim \frac{\sqrt{3}}{\sqrt{2}m\sqrt{M_N \alpha\beta}}. \tag{2.8}$$

The proof of corollary 2.1.1 is given in Section 2.3.5 of supplemental material, together with the exact formula for σ. Interestingly, this result obtained from our minimal model is comparable to the expression on lag-time variations obtained in [Szavits-Nossan et al., 2014] based on a more complex mechanistic model by the mean of a Taylor expansion. This result, therefore, corroborates with the idea that our minimum model with only ignition and conversion contains the key features sufficient in qualitatively describing the stochastic properties of the nucleated protein aggregation processes. Our simplified formula (2.8) and its full general form in the equation (2.18) of the supplemental information, are mathematically fully rigorous, and allows analysis of the intricate interplay between the ignition reaction and the autocatalytic reaction. In fact, it is possible to have a whole range of times when both reactions have an influence over the whole aggregation timecourse, as may be seen on Formula (2.8). It should be noted that this representation of σ is independent of δ. This suggests that the fluctuations do not depend on δ, and therefore, the growth curves predicted by our simple model are all parallel for any given concentration. Figures 2.5 (c) and 2.5 (d) below have been obtained by centering the 12 curves of Figures 2.5 (a) and 2.5 (b) at the half-time corresponding to $\delta = 1/2$. As it can be seen, the times $T_N(\delta)$ for $0.4 \leq \delta \leq 0.7$ are then also superimposed: the curves are identical for this range of values. This is an illustration of the above relation (2.8). The exact mathematical formulation of this phenomenon is shown in supplemental material. Simple as it is, our model captures well this feature experimentally observed. Also, it emphasizes the fact that we can take different values for δ without having an influence on the study. A difficulty however lies in the fact that when the numerical values of the constant α above is in the order of $1/M_N$, then the convergence itself may be a problem, as it can be seen on Figure 2.4.

2.1.3 Estimation of the parameters

In this section, we tested the results obtained with our minimalistic stochastic model on the data published in [Xue et al., 2008]. In these experiments, there are 12 replicate kinetic traces reported for each sample

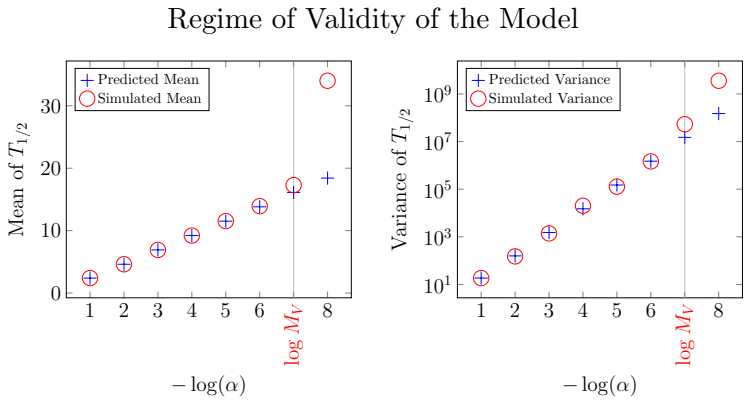


Figure 2.4: Comparison between the simulations and the predictions to see the regime of parameters where the calculations are valid. For these simulations, we fixed $M_N = 10^7$ monomers, $\beta = 1$, and made α varying.

concentration in constant $100\mu L$ reaction volume. The parameters α and β are obtained by fitting the mean half-time $t_{1/2}$ and the mean slope k of the curves at $t_{1/2}$. More precisely, using Formula (2.5) for k and Relation (3.24) for $t_{1/2}$, gives

$$t_{1/2} = \log\left(1 + \beta/\alpha\right)/\beta m \text{ and } k = m\beta\left(1 + \beta/\alpha\right)/4. \tag{2.9}$$

In the experiments in [Xue et al., 2008], there are 12 replicate kinetic traces reported for each sample concentration in constant 15 μL reaction volume. The parameters α and β can be obtained in a straightforward manner by fitting equations (2.9) to the mean half-time $t_{1/2}$ and the mean slope k of the curves at $t_{1/2}$. Table 2.1 shows a summary of our analysis. The constants α and β, and the calculated variance (2.8) are shown for each of the concentrations used. We also carried out a global analysis for α and β, fitting (2.9) simultaneously all of the curves for all concentrations. See Fig. 2.9. The overlay of the experimental curves around the predicted mean is illustrated in Figure 2.5, Figures 2.5 (c) and 2.5 (d) have been obtained by centering the 12 curves of Figures 2.5 (a) and 2.5 (b) at the half-time corresponding to $\delta = 1/2$. As can be seen, the agreement between the calculated and the experimental curves is good for $0.4 \leq \delta \leq 0.7$. This is consistent with the relation (2.8).

Our analysis further demonstrates two important insights. Firstly, we obtained a more well-estimated β parameter. It is remarkable that the numerical value of β, which quantifies the conversion step in our model, does not change much for the 15 concentrations tested in the experiments, considering the simplicity of our model. This is not the case for α,which quantifies the ignition phase, varies between 10^{-7} and 10^{-13}. Here, the parameter α which quantifies the take-off phase (remember that the slope of $(x_1(t))$ at 0 is $-\alpha m^2$) is intrinsically estimated with less precision than β, see Section 2.3.6 of supplemental material. This is a limitation of this simple model, and it also reflect the lack of information content in

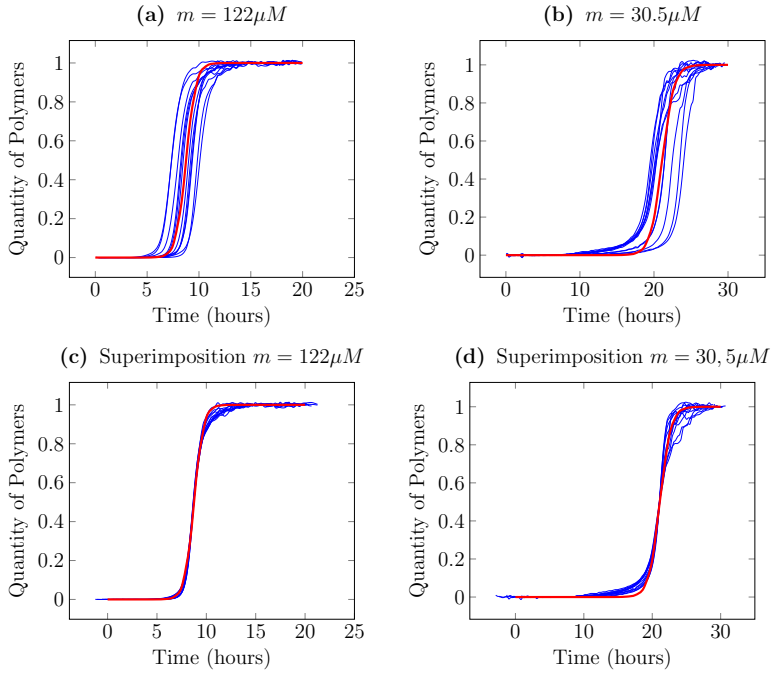


Figure 2.5: (a) and (b): Experimental timecourse of polymerised mass for 12 different experiments. (c) and (d): with a centering at the $T_N(1/2)$ of each curve. Published in [Xue et al., 2008]. The red curve is the predicted mean with the estimated parameters.

$\mathbf{m(10^{-6}M)}$	$\alpha(\mathbf{h^{-1}.M^{-1}})$	$\beta(\mathbf{h^{-1}.M^{-1}})$	**Experimental Std** (h)	**Predicted Std** (h)
12.3	$6.18{\cdot}10^{-7}$	$5.07{\cdot}10^{4}$	7.95	$5.34{\cdot}10^{-2}$
14.6	$2.81{\cdot}10^{-6}$	$4.54{\cdot}10^{4}$	2.98	$2.05{\cdot}10^{-2}$
16.7	$1.59{\cdot}10^{-4}$	$3.75{\cdot}10^{4}$	2.68	$2.45{\cdot}10^{-3}$
17.0	$1.88{\cdot}10^{-3}$	$3.70{\cdot}10^{4}$	1.52	$6.98{\cdot}10^{-4}$
29.5	$1.40{\cdot}10^{-5}$	$3.34{\cdot}10^{4}$	2.13	$3.7{\cdot}10^{-3}$
30.2	$2.89{\cdot}10^{-2}$	$2.96{\cdot}10^{4}$	2.57	$8.40{\cdot}10^{-5}$
30.5	$9.57{\cdot}10^{-8}$	$4.16{\cdot}10^{4}$	1.53	$3.84{\cdot}10^{-2}$
43.7	$7.99{\cdot}10^{-3}$	$2.35{\cdot}10^{4}$	2.10	$1.03{\cdot}10^{-4}$
48.5	$1.68{\cdot}10^{-2}$	$2.01{\cdot}10^{4}$	1.56	$6.55{\cdot}10^{-5}$
61.0	$2.61{\cdot}10^{-2}$	$2.04{\cdot}10^{4}$	1.03	$3.71{\cdot}10^{-5}$
61.0	$2.22{\cdot}10^{-5}$	$2.56{\cdot}10^{4}$	2.55	$1.14{\cdot}10^{-3}$
84.1	$4.53{\cdot}10^{-4}$	$2.24{\cdot}10^{4}$	1.59	$1.66{\cdot}10^{-4}$
102.2	$1.52{\cdot}10^{-3}$	$1.88{\cdot}10^{4}$	0.62	$7.39{\cdot}10^{-5}$
122	$1.33{\cdot}10^{-4}$	$1.75{\cdot}10^{4}$	0.90	$1.98{\cdot}10^{-4}$
123.5	$2.13{\cdot}10^{-4}$	$1.79{\cdot}10^{4}$	0.90	$1.52{\cdot}10^{-4}$
142.1	$2.58{\cdot}10^{-4}$	$1.74{\cdot}10^{4}$	1.11	$1.13{\cdot}10^{-4}$
243.5	$1.75{\cdot}10^{-3}$	$1.09{\cdot}10^{4}$	0.60	$2.46{\cdot}10^{-5}$

Table 2.1: Parameters estimated from experimental data published in [Xue et al., 2008] using our model. The two first columns are the estimated parameters α and β from the model. The third column is the experimental standard deviation of $T_N(1/2)$, while the fourth is the standard deviation predicted by our mathematical results for the model with the estimated parameters. We see that the estimation for β is quite robust, in contrast with that of α.

the kinetics data during the lag phase compared to the growth phase.

Secondly, despite good agreement between our analysis and the data in terms of the shapes of the growth curves, the analysis results in a much smaller order of magnitude for the variability among curves compared with experimental data. Since the relation $\alpha \ll \beta$ holds in the numerical estimations, Equation (2.8) gives the approximation $\sigma^2 \sim 3/(2M_N m^2 \alpha\beta)$ for the variance of the characteristic times of the kinetic traces. A variance of the order of magnitude observed in the experiments [Xue et al., 2008] would be obtained by our model for an initial number of monomers M_N in the order of 10^6. As the number M_N in the experiments of [Xue et al., 2008] performed in $100\mu L$ volumes is closer to 10^{15}, our analysis suggest that the variability observed result from more than a simple stochastic homogeneous nucleation of monomers. This result is consistent with the mechanistic approach used by Szavits-Nossan and co-workers [Szavits-Nossan et al., 2014], where the authors used a stochastic nucleation-polymerization-fragmentation based model. Thus, our model and analysis of the variance suggest alternative initial rare assembly steps that involve additional complexities such as conformational exchange, and/or additional experimental sources that contribute to the variability in the observed amyloid growth curves.

2.2 Conclusion and discussion

Our approach also suggests a straightforward manner in which information regarding the stochastic behaviour of nucleated protein aggregation can be extracted from experimental kinetics data using equations (2.8) and (2.9). We see that the stochasticity influences mainly the ignition step: once the reaction accelerates in the conversion step, all curves become parallel and deterministic, as illustrated both by experiments and the model we presented here. Thus, simple as it is, our model captures well the features experimentally observed for amyloid growth curves. Also, it confirms, as expected, that we can take different characteristic times (such as lag time, or growth mid point) when analysing kinetic growth curves. Our model further informed the need for new mechanistic steps or experimental interpretation of the large observed variations in the lag time lengths. Thus, the variation seen in the kinetic traces must be taken into account in addition to the concentration dependent behaviour of the kinetic traces in evaluating and developing mechanistic understanding of amyloid protein assembly processes.

While our model design was not aimed at describing the reality of any specific amyloid forming system with all of their individual associated complexities, our design by pursuing maximum simplicity are complementary to mechanistic approaches such as in [Xue et al., 2008, Knowles et al., 2009, Szavits-Nossan et al., 2014] in capturing global properties of amyloid assembly. A particularly interesting direction for future work would be to envisage other orders for the reactions, in particular β, which currently is not specific to any particular accelerating growth processes, or an extended model with an initial conformational exchange step, for example. In summary, our current method allows for a rigorous theoretical treatment and understanding, and therefore, provides a basis for future model selection on stochastic 'minimal model', each of these models being the condensation of a family of possible stochastic mechanistic models that are closer to reality but for which analytical formulae are out of reach.

It should be remarked that in many cases the reaction curve shows significant asymmetries about the half-time which cannot be captured without accounting for the correct dependencies of the secondary nucleation rate on the monomer. Such an asymmetry is present e.g. in the data of Figures 2.2 and 2.5, take off is slower than the approach to the plateau, a common characteristic of systems dominated by fragmentation. However, the model predictions, e.g. in Figure 2.3, give curves that are perfectly symmetric about the half time (because of the term $X(M-X)$ in the rate equations). This suggests the possibility of different dependencies of the autocatalytic part on the monomer which we plan to investigate in the future.

2.3 Supplemental material: proofs of the main results and analysis of the parameters

Recall that it is assumed that the parameter M_N is asymptotically proportional to m

$$\lim_{N\to+\infty} \frac{M_N}{N} = m.$$

2.3.1 Proof of the law of large numbers

The proof relies on classical methods of stochastic calculus, see for instance Darling and Norris [Darling and Norris, 2008] or Ethier and Kurtz [Ethier and Kurtz, 1986]. Here, we give a summary of the proof for the completeness.

Stochastic Equations. Let $\mathcal{N}_{\alpha/N^2}(dt)$ [resp. $\mathcal{N}_{\beta/N^2}(dt)$] be a Poisson process with parameter α/N^2 [resp. β/N^2], then Relations (2.3) give that the random variable $X_1^N(t)$ can be represented as a solution of the following stochastic differential equation

$$\mathrm{d}X_1^N(t) = -2 \sum_{i=1}^{X_1^N(X_1^N-1)(t-)/2} \mathcal{N}^i_{\alpha/N^2}(\mathrm{d}t) - \sum_{i=1}^{X_1^N(M-X_1^N)(t-)} \mathcal{N}^i_{\beta/N^2}(\mathrm{d}t), \tag{2.10}$$

with $X_1^N(0) = M$ and $f(s-)$ denotes the limit on the left of f at s. For more discussion and results on related models, see for example Anderson and Kurtz [Anderson and Kurtz, 2011] and Higham [Higham, 2008] and references therein. These SDEs are an equivalent, probabilistic, form of the master equation. To get asymptotics, it is in general easier to work directly with this stochastic differential equation (SDE) rather than with the evolution of the distribution of the process (an infinite system of ODEs) or its generating function.

By using Equation (2.10) and $\overline{X}_1^N$ defined by (2.4), one gets

$$\overline{X}_1^N(t) = \frac{X_1^N(Nt)}{N} = \frac{M_N}{N} + \mathcal{M}_N(t) - \alpha \int_0^t \overline{X}_1^N \left(\overline{X}_1^N - \frac{1}{N}\right)(s)\,\mathrm{d}s - \beta \int_0^t \overline{X}_1^N \left(\frac{M_N}{N} - \overline{X}_1^N\right)(s)\,\mathrm{d}s, \tag{2.11}$$

where $(\mathcal{M}_N(t))$ is the martingale

$$\begin{aligned}\mathcal{M}_N(t) = &-\frac{2}{N}\sum_{i=1}^{\infty}\int_0^{Nt} \mathbb{1}_{\{i\le X_1^N(X_1^N-1)(s-)/2\}}\left(\mathcal{N}^i_{\alpha/N^2}(\mathrm{d}s) - \frac{\alpha}{N^2}\,\mathrm{d}s\right)\\ &-\sum_{i=1}^{\infty}\frac{1}{N}\int_0^{Nt} \mathbb{1}_{\{i\le X_1^N(M-X_1^N)(s-)\}}\left(\mathcal{N}^i_{\beta/N^2}(\mathrm{d}s) - \frac{\beta}{N^2}\,\mathrm{d}s\right).\end{aligned}$$

Its quadratic variation is given by

$$\langle \mathcal{M}_N\rangle(t) = \frac{2\alpha}{N}\int_0^t \overline{X}_1^N\left(\overline{X}_1^N - \frac{1}{N}\right)(s)\,\mathrm{d}s + \frac{\beta}{N}\int_0^t \overline{X}_1^N\left(\frac{M_N}{N} - \overline{X}_1^N\right)(s)\,\mathrm{d}s \le \frac{1}{N}Ct,$$

for some constant C since X_1^N is bounded by M_N. Doob's inequality gives that, with high probability, the martingale $(\mathcal{M}_N(t))$ vanishes uniformly on finite intervals: for $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{0\le s\le t}|\mathcal{M}_N(s)| \ge \varepsilon\right) \le \frac{1}{\varepsilon^2}\mathbb{E}\left(\langle \mathcal{M}_N\rangle(t)\right) \le \frac{1}{N}\frac{Ct}{\varepsilon^2}.$$

We can now show that the sequence $(\overline{X}_1^N)_N$ is tight. Let

$$w_N(\delta) = \sup_{\substack{|u-v|\le\delta\\ u,v\le t}}\left|\overline{X}_1^V(u) - \overline{X}_1^V(v)\right|.$$

Then, Equation (2.11) gives

$$w_N(\delta) \le \sup_{\substack{|u-v|\le\delta\\ u,v\le t}}|\mathcal{M}_N(u) - \mathcal{M}_N(v)| + \delta(\alpha+\beta)\left(\frac{M_N}{N}\right)^2.$$

Therefore, for $\varepsilon > 0$ and $\eta > 0$, there exist δ_0 and N_0 such that if $\delta \le \delta_0$ and $N \ge N_0$ then $\mathbb{P}(w_N(\delta) \ge \varepsilon) \le \eta$. Consequently, the sequence $(\overline{X}_1^N(t))$ is tight, see Ethier and Kurtz [Ethier and Kurtz, 1986] for example. Let $(x_1(t))$ be one of the limiting points of $(\overline{X}_1^N(t))$, it necessarily satisfies the following differential equation

$$\dot{x}_1 = -\alpha x_1^2 - \beta x_1(m - x_1) \text{ with } x_1(0) = m.$$

2.3.2 Proof of central limit result

One proves that, for the convergence in distribution,

$$\lim_{N\to+\infty}\left(\frac{X_1^N(Nt)-Nx_1(t)}{m\sqrt{N}}\right)=(U(t)),$$

where $U(t)$ is the unique solution of the following stochastic differential equation

$$\mathrm{d}U(t)=\frac{\beta\sqrt{\alpha}\sqrt{e^{\beta mt}+1}}{\alpha e^{\beta mt}+\beta-\alpha}\,\mathrm{d}W(t)-\beta m\frac{e^{\beta mt}+1-\beta/\alpha}{e^{\beta mt}-1+\beta/\alpha}U(t)\,\mathrm{d}t, \tag{2.12}$$

with $U(0)=0$ and $(W(t))$ denotes a standard Brownian motion.

With Equation (2.11), one gets

$$\begin{aligned}U_N(t)=\frac{X_1^N(Nt)-Nx_1(t)}{m\sqrt{N}}=\frac{\sqrt{N}\mathcal{M}_N(t)}{m}-\alpha\int_0^t U_N(s)\left(\overline{X}_1^V(s)+x_1(s)\right)\mathrm{d}s\\ -\beta m\int_0^t U_N(s)\,\mathrm{d}s+\beta\int_0^t U_N(s)(\overline{X}_1^V(s)+x_1(s))\,\mathrm{d}s+\frac{\alpha}{\sqrt{N}}\int_0^t\overline{X}_1^V(s)\,\mathrm{d}s.\end{aligned} \tag{2.13}$$

First note that the process associated to the last term of this expression converges in distribution to zero. Concerning the martingale term of this relation, one has

$$\left\langle\sqrt{N}\frac{\mathcal{M}_N}{m}\right\rangle(t)=\frac{1}{m^2}\left[2\alpha\int_0^t\left(\overline{X}_1^V\right)\left(\overline{X}_1^V-1\right)(u)\,\mathrm{d}u+\beta\int_0^t\overline{X}_1^V\left(m-\overline{X}_1^V\right)(u)\,\mathrm{d}u\right].$$

The law of large numbers which has just been proved gives that this process converges to

$$2\alpha\int_0^t\frac{x_1^2}{m^2}\,\mathrm{d}s+\frac{\beta}{m^2}\int_0^t x_1(s)(m-x_1(s))\,\mathrm{d}s=\frac{\alpha}{m^2}\int_0^t x_1(s)^2\,\mathrm{d}s+\frac{1}{m^2}(m-x_1(t))=\psi(t).$$

Thus, we get from Theorem 1.4 page 339 of Ethier and Kurtz [Ethier and Kurtz, 1986] that, as V goes to infinity, the process $(\sqrt{N}\mathcal{M}_N(t)/m)$ converges in distribution to

$$\int_0^t\sqrt{\dot{\psi}(s)}\,\mathrm{d}W(s),$$

where $(W(t))$ is the standard Brownian motion.

We now prove that the sequence of processes $(U_N(t))$ is tight. Let

$$w_N(\delta)=\sup_{\substack{|u-v|\le\delta\\ u,v\le t}}\left|U_N(u)-U_N(v)\right|,$$

then, by using Equation (2.13), one gets

$$w_N(\delta) \leq \sup_{\substack{|u-v|\leq\delta \\ u,v\leq t}} \left| \frac{\sqrt{N}\mathcal{M}_N(u)}{m} - \frac{\sqrt{N}\mathcal{M}_N(v)}{m} \right| + \sup_{\substack{u,v\leq t \\ |u-v|\leq\delta}} \frac{\alpha}{\sqrt{N}} \left| \int_u^N \overline{X}_1^V(s)\,\mathrm{d}s \right| + C_0 \sup_{\substack{u,v\leq t \\ |u-v|\leq\delta}} \int_u^N |U_N(s)|\,\mathrm{d}s, \tag{2.14}$$

for some fixed constant C_0. Consequently,

$$\sup_{s\leq t}|U_N(s)| \leq \sup_{s\leq t}\left(\left| \frac{\sqrt{N}\mathcal{M}_N(s)}{m} \right| \right) + \frac{\alpha}{\sqrt{N}} mt + C_0 \int_0^t \sup_{\tau\leq s}|U_N(\tau)|\,\mathrm{d}s,$$

by using Gronwall's lemma, one gets

$$\sup_{s\leq t}|U_N(s)| \leq \left[\sup_{s\leq t}\left(\left| \frac{\sqrt{N}\mathcal{M}_N(s)}{m} \right| \right) + \frac{\alpha}{\sqrt{N}} mt \right] e^{C_0 t}.$$

The convergence of the processes $\left(\sqrt{N}\mathcal{M}_N(t)\right)$ shows that the left-hand side of the above expression is bounded with high probability. Relation (2.14) and the tightness of $\left(\sqrt{N}\mathcal{M}_N(t)\right)$ give then directly the tightness of $(U_N(t))$.

Let U be a limiting point of the sequence $(U_N(t))$ when N goes to infinity. Relation (3.16) shows that U must satisfy the following stochastic equation

$$U(t) = \int_0^t b(s)\,\mathrm{d}W(s) + \int_0^t a(s)U(s)\,\mathrm{d}s, \tag{2.15}$$

where

$$b(t) = \sqrt{\dot{\psi}(t)} = \frac{\beta\sqrt{\alpha}\sqrt{e^{\beta m t}+1}}{\alpha e^{\beta m t} + \beta - \alpha}, \qquad a(t) = \beta m \frac{\beta - \alpha - \alpha e^{\beta m t}}{\beta - \alpha + \alpha e^{\beta m t}}. \tag{2.16}$$

This proves that the process $(U_N(t))$ converges in distribution to $(U(t))$.

2.3.3 Explicit solution of the SDE for U

Corollary 2.3.1. *The SDE for U has an explicit solution:*

$$U(t) = \frac{e^{\beta m t}}{\left(\beta/\alpha - 1 + e^{\beta m t}\right)^2} \int_0^t \frac{\beta}{\sqrt{\alpha}} \left[\left(\frac{\beta}{\alpha} - 1 \right) e^{-\beta m s/2} + e^{\beta m s/2} \right] \left[\sqrt{1 + e^{-\beta m s}} \right] \mathrm{d}W_s. \tag{2.17}$$

Straightforward stochastic calculus shows that the right-hand side of Equation (2.17) satisfies the Stochastic Differential Equation associated to Relation (2.15).

2.3.4 Proof of the asymptotics for time for δ reaction completion

It is enough to prove the central limit result. We recall that t_δ is the deterministic time of δ reaction completion, that is

$$t_\delta = x_1^{-1}((1-\delta)m)$$

For $w \geq 0$, since $\{T_N(\delta) \leq w\} = \{X_1^N(w) \leq (1-\delta)M_N\}$,

$$\left\{\frac{T_N(\delta) - Nt_\delta}{\sqrt{N}} \leq w\right\} = \left\{X_1^N\left[N(t_\delta + w/\sqrt{N})\right] \leq (1-\delta)M_N\right\},$$

the probability of the event can therefore be expressed as

$$\mathbb{P}\left(\frac{X_1^N\left[N(t_\delta+w/\sqrt{N})\right] - Nx_1\left(t_\delta+w/\sqrt{N}\right)}{\sqrt{N}} \leq \sqrt{N}\left(x_1(t_\delta) - x_1\left(t_\delta+w/\sqrt{N}\right)\right) + o(1)\right).$$

Hence, by the central limit result, for the convergence in distribution

$$\lim_{N\to+\infty} \frac{X_1^N\left[N(t_\delta + w/\sqrt{N})\right] - Nx_1(t_\delta + w/\sqrt{N})}{m\sqrt{N}} = U(t_\delta),$$

consequently, one gets the convergence

$$\lim_{N\to+\infty} \mathbb{P}\left(\frac{T_N(\delta) - Nt_\delta}{\sqrt{N}} \leq w\right) = \mathbb{P}\left(U(t_\delta) \leq \frac{-\dot{x_1}(t_\delta)}{m} w\right) = \mathbb{P}\left(\frac{U(t_\delta)}{m[\alpha(1-\delta)^2 + \beta\delta(1-\delta)]} \leq w\right)$$

The result is proved.

2.3.5 Proof of asymptotics of variance of the time for δ reaction completion

This is a direct consequence of the above central limit result and of Relation (2.17) and the fact that

$$\begin{aligned}
\sigma^2 &:= \lim_{N\to+\infty} \mathbb{E}\left[\left(\frac{T_N(\delta) - Nt_\delta}{\sqrt{N}}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{U(t_\delta)}{m[\alpha(1-\delta)^2 + \beta\delta(1-\delta)]}\right)^2\right] \\
&= \frac{\alpha}{m^3\beta^2}\left[\left(\frac{\beta}{\alpha} - 1\right)^2 \frac{1}{2\beta}\left(1 - \frac{1}{(1+\beta\delta/(\alpha(1-\delta)))^2}\right) + \left(\frac{\beta}{\alpha} - 1\right)\left(\frac{\beta}{\alpha} + 1\right)\frac{\delta}{\beta\delta + \alpha(1-\delta)}\right. \\
&\qquad \left. + \left(2\frac{\beta}{\alpha} - 1\right)\frac{1}{\beta}\log\left(1 + \frac{\beta\delta}{\alpha(1-\delta)}\right) + \frac{1}{\alpha}\frac{\delta}{1-\delta}\right]. \qquad (2.18)
\end{aligned}$$

One can get a more precise result by using the fact that $(U(t))$ is a Gaussian process, by Equation (2.17)

for example, the following representation holds for the $T_N(\delta)$.

Remark. Provided that $\alpha \ll \beta$ then, as N gets large, the asymptotic behavior of $T_N(\delta)$ is the following:

$$T_N(\delta) \sim Nt_\delta + \sqrt{N}\mathcal{N}\left(0, \frac{3}{2M_N\alpha\beta m^2}\right), \tag{2.19}$$

where $\mathcal{N}(0, x)$ is a center Gaussian random variable with variance x.

We illustrate this remark on Figure 2.6 below. This expansion shows that the stochastic fluctuations, the term associated with $\sqrt{N}$, do not depend on δ. This remarkable property is also true in the experiments: the curves superimpose very well. See Figure 2.5 (c) and (d).
Indeed, the central limit result gives

$$\frac{T_N(\delta) - Nt_\delta}{\sqrt{N}} \sim \frac{U(t_\delta)}{m[\alpha(1-\delta)^2 + \beta\delta(1-\delta)]} \sim \mathcal{N}(0, \sigma^2),$$

by Equation (2.17), where σ is defined above. The expansion follows by using the fact that $\alpha \ll \beta$ in the explicit expression (2.18) of σ.

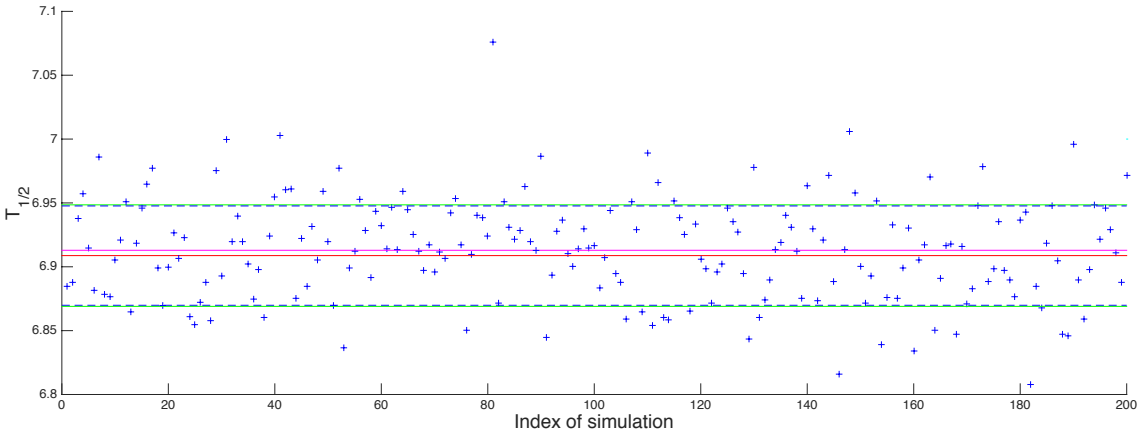


Figure 2.6: Comparison between predicted mean and standard deviation with simulations. We carried out 200 simulations. For each simulation, $T_{1/2}$ is plotted (blue crosses). The predicted mean and the predicted standard deviation of $T_{1/2}$(red line and green lines), and the simulated mean and simulated variance (pink line and dashed line) are also shown with parameters $M_N = 10^6$, $\alpha = 10^{-3}$, $\beta = 1$ and $m = 1$.

2.3.6 Qualitative analysis of the behaviour of x_1 and U

Behaviour of $x_1(t)$

Recall that

$$\frac{dx_1}{dt} = -\alpha x_!^2 - \beta x_1(m - x_1) \text{ with } x_1(0) = m, \text{ i.e. } x_1(t) = m\frac{\beta}{\alpha}\frac{1}{e^{\beta m t} - 1 + \beta/\alpha}.$$

The limit which interests us is when $\alpha \ll \beta$: otherwise, the slope at zero, given by αm^2, is not small

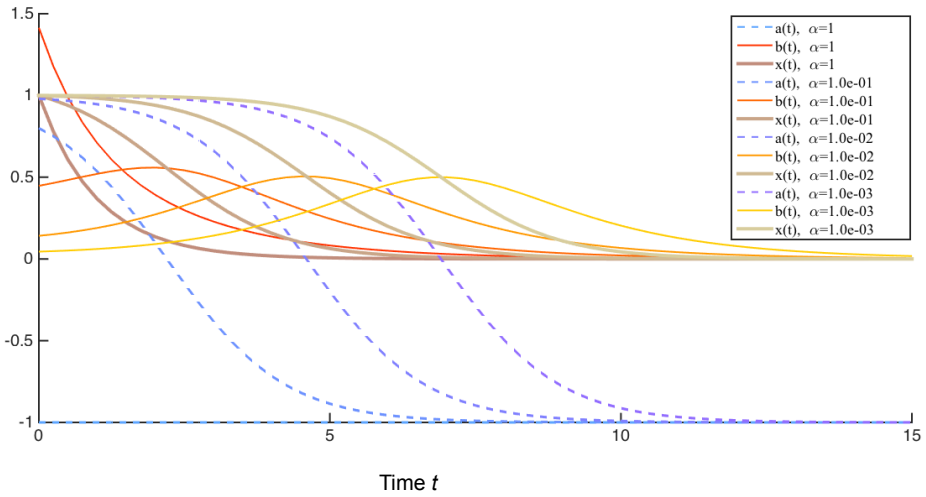


Figure 2.7: Functions $x_1(t)$ $a(t)$ and $b(t)$ for $\alpha = 10^{-k}$, $k = 0, \ldots, 3$. The function $a(t)$ corresponds to the deterministic part of the equation whereas $b(t)$ corresponds to the stochastic part, see Equations (3.18) and (2.16).

compared to the slope at $t_{1/2}$, which is $m^2(\alpha + \beta)/4$, so that there is no lag-time, contrarily to what is observed even for high concentrations. In this limit, the formulae for $t_{1/2}$ and k are

$$t_{1/2} = \frac{1}{\beta m} \log\left(1 + \beta/\alpha\right) \approx \frac{1}{\beta m} \log(\beta/\alpha) \quad \text{and} \quad k = \frac{m\beta}{4}\left(1 + \alpha/\beta\right) \approx \frac{m\beta}{4},$$

so that $\beta = 4k/m$ and $\alpha = \beta \exp(-\beta m t_{1/2}) = 4k \exp(-4k t_{1/2})/m$.

The slope k being measured with little variance between curves of a given concentration, the estimation for β is good, at least for a given concentration. What is remarkable is its goodness through different concentrations: our model thus predicts a linear dependence between k and m. Concerning α, it may change by a typical factor of $\exp(\pm\beta m\sigma)$, so that taking the experimental values of Table 2.1 (first, third and fourth columns) we obtain an uncertainty for α which ranges between 7 and 2.10^5 according to the set of experiments. This high uncertainty in the estimation of α may to a large extent explain the high variability obtained in the estimated α (See Table 2.1, second column). Note also that this uncertainty does not decrease when the initial concentration increases.

Behaviour of U(t)

In Figure 2.7, the functions $(a(t))$ and $(b(t))$ are plotted for fixed $\beta = 1$, $V = 10^5$ and $m = 1$, and various values of α are considered. These functions, defined by Equations (2.16), drive the dynamics of $(U(t))$ by Relation (3.18),

$$\mathrm{d}U(t) = b(t)\,\mathrm{d}W(t) + a(t)U(t)\,\mathrm{d}t.$$

In particular the coefficient $b(t)$ of the Brownian motion $(W(t))$ modulates the stochasticity of $(U(t))$.

We observe that for sufficiently small α :

— $a(t)$ begins at 1, decreases to -1. The curves are translations from one another and the time when $a(t) = 0$ increases when α decreases.

— $b(t)$ is nonnegative, bell-shaped, vanishes at zero and infinity, the curves are translation from one another and its maximum is always the same, around 0.55. The time at which $b(t)$ is maximum increases when α decreases.

— At the crossing time, $a(t) = b(t)$ values a constant, around 0.4 (while $b(t)$ is increasing).

— The smaller the ratio α/β, the higher the average peak value for $|U|$, and the less noisy each path is (see Figure 2.8 for an illustration).

All these facts may be deduced analytically from the approximation values when $\alpha \ll \beta$. Denoting $\varepsilon = \alpha/\beta$

$$b(t) \approx \frac{\sqrt{\varepsilon\beta}\sqrt{e^{\beta m t}+1}}{1+\varepsilon e^{\beta m t}}, \qquad a(t) \approx \beta m \frac{1-\varepsilon e^{\beta m t}}{1+\varepsilon e^{\beta m t}}.$$

For $t = 0$, we have $a(0) \approx \beta m = 1$, $a(t)$ is clearly decreasing and for t large we have $a(t) \to -\beta m$. This implies that $a(t)$, which has the deterministic influence, leads to exponential growth for U around 0 and exponential decrease for U around infinity.

At $t = 0$, $b(t) \approx \sqrt{\varepsilon\beta}$ is very small, b is always positive and at infinity we have $b(t) \approx \sqrt{\beta}\exp(-\beta m t/2)/\sqrt{\varepsilon}$.

Concerning the crossing point, it occurs when

$$\frac{\sqrt{\varepsilon\beta}\sqrt{e^{\beta m t}+1}}{1+\varepsilon e^{\beta m t}} \approx \beta m \frac{1-\varepsilon e^{\beta m t}}{1+\varepsilon e^{\beta m t}}.$$

Denoting $d = \varepsilon\exp(\beta m t)$, assuming $\varepsilon \ll \beta m^2$ and taking the square, we have $d+\varepsilon \approx d \approx \beta m^2(1-d)^2$, and this gives a value for d which is independent of ε and α, and for this value we have $a{=}b \approx \beta m(1-d)/(1+d)$ depending only on β and m.

This also explains the fact that the maximal value for $|U|$ increases in average, whereas the 'noise' in each path decreases. These observations are illustrated in Figure 2.8, where we show for each of the previous values of $\alpha = 10^{-k}$ with $k = 0, \ldots, 4$ five trajectories for U_N in blue and five trajectories for U in red, for $M = 10^5$.

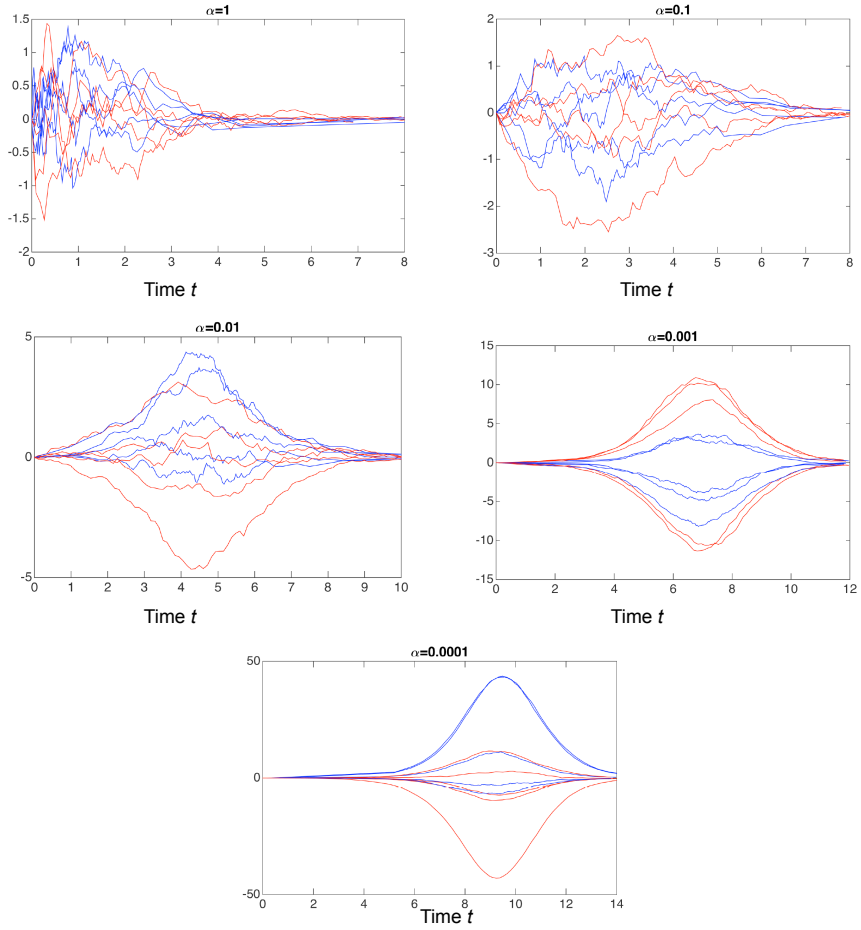


Figure 2.8: Stochasticity of the centered assembly processes $(U_N(t))$ and $(U(t))$. For each of the previous values of $\alpha = 10^{-k}$ with $k = 0, \ldots, 4$, five trajectories for U_N in blue and five trajectories for U in red, for $M = 10^5$. We see that the noise inside each path decreases when α decreases, the stochasticity remaining in the startup of the curves.

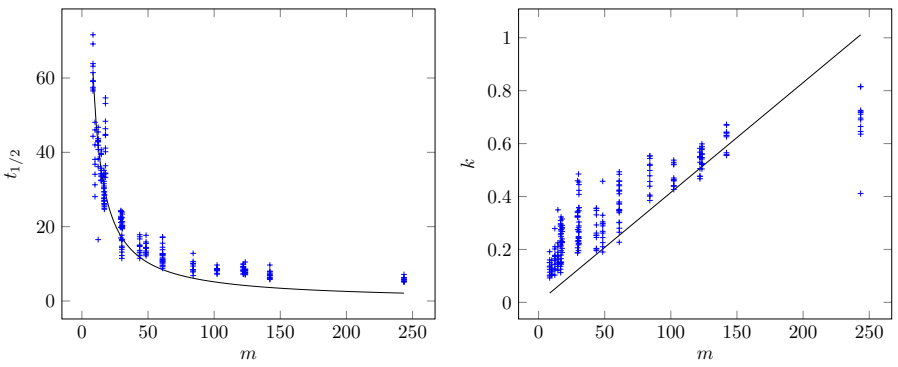


Figure 2.9: Comparison between experimental measurements (blue crosses) and simulated values (black line) for the global best-fit parameters $\alpha = 3.2\, h^{-1}M^{-1}$ and $\beta = 1.7 \cdot 10^4 h^{-1}M^{-1}$. Left: $t_{1/2}$ with respect to the concentration m in μM, in linear scale. Right: the slope k with respect to the concentration m, in linear scale.

Chapter 3

Asymptotics of stochastic protein assembly models

Contents

3.1 Introduction

Protein polymerization is involved is involved in many important biological phenomena: human diseases such as Alzheimer's, Parkinson's, Huntington's diseases and also key biological processes such as actin filamentation, or yet industrial processes, see McManus et al. [McManus et al., 2016] and Ow and Dustan [Ow and Dunstan, 2014]. The initial step of the chain reactions giving rise to polymers consists in the spontaneous formation of a so-called *nucleus*, that is, the simplest possible polymer able to ignite the reaction of polymerization. This early phase is called *nucleation*, and is still far from being understood. As underlined by previous studies, see Szavits-Nossan et al. [Szavits-Nossan et al., 2014], the nucleation step is intrinsically stochastic, leading to an important variability among replicated experiments, not only in small volumes but even in relatively large ones, see Xue et al. [Xue et al., 2008]. The question of building convenient stochastic models, able to render out the heterogeneity observed, and even to predict it, has recently raised much interest in the biological and biophysical community, see Szavits-Nossan et al. [Szavits-Nossan et al., 2014], Yvinec et al. [Yvinec et al., 2016], Pigolotti et al. [Pigolotti et al., 2013] and Eden et al. [Eden et al., 2015].

We start with a simple stochastic model, proposed and studied in Eugène et al. [Eugène et al., 2015] for which we consider extensions to get a deeper understanding on the intricate influence of each reaction considered. In Eugène et al. [Eugène et al., 2015], rigorous asymptotics of the simple model were proved, and it was fitted to the experimental data published in Xue et al [Xue et al., 2008]. It was shown that the predicted variability was much smaller by the model than what was experimentally obtained. One of the conclusions of this work is that other mechanisms have to be taken into account to explain the variability observed in the experiments. We thus propose here two ways to complement the basic model. Let us first recall its definition.

3.1.1 The Basic Model

One of the simplest models to describe the nucleation process considers two populations of chemical components: free monomers and polymerised monomers. Initially there are only free monomers. There are two reactions for the polymerization of a monomer: either two monomers collide to combine into two polymerised monomers or a monomer is polymerised after the encounter of a polymerised monomer. The chemical reactions associated with the basic model can then be described as follows:

$$\begin{cases} \mathcal{X}_1 + \mathcal{X}_1 \xrightarrow{\alpha} 2\mathcal{X}_2, \\ \mathcal{X}_1 + \mathcal{X}_2 \xrightarrow{\beta} 2\mathcal{X}_2. \end{cases} \tag{3.1}$$

These reactions can be represented by the sample paths of a Markov process $(X_1^N(t), X_2^N(t))$, where $X_1^N(t)$ [resp. $X_2^N(t)$] is the number of free [resp. polymerised] monomers at time $t \geq 0$. The scaling variable N should be thought of as the reaction volume. In particular $X_2^N(t)/N$ is the concentration of polymerised monomers at time t. If M_N is the initial number of monomers, it is assumed that the

following regime

$$M_N = mN + o\left(\sqrt{N}\right) \tag{3.2}$$

holds for some $m > 0$. The scaling parameter N can be thought as related to the volume of the system. Since M_N/N is converging to m, m is therefore the initial, asymptotic, concentration of monomers. Throughout the paper m^* denotes an upper bound for the sequence (M_N/N).

The transition rates of $(X_1^N(t), X_2^N(t))$ are given by, for $x = (x_1, x_2) \in \mathbb{N}^2$,

$$x \mapsto \begin{cases} x+(-2,2) & \text{at rate} \quad \alpha \dfrac{x_1(x_1-1)}{2N^2} \\ \\ x+(-1,1) & \text{"} \qquad\quad \beta \dfrac{x_1}{N}\dfrac{x_2}{N}. \end{cases} \tag{3.3}$$

The second coordinate x_2 represents the polymerised mass, i.e., the number of monomers present in any polymer of any size, hence the jump of 2 for x_2 as for x_1 in the first reaction. Note that the conservation of mass implies that the quantity $X_1^N(t)+X_2^N(t)$ is constant and equal to M_N, the total number of initial monomers.

— The first reaction of (3.1) converts two free monomers into two polymerised monomers. In our model, due to thermal noise in particular, these reactions will occur in a stochastic way. Following the principles of the law of mass action, the encounter of two chemical species occurs at a rate proportional to the product of the *concentrations* of each species. Therefore two given monomers disappear to produce two polymerised monomers at a rate $\alpha x_1(x_1-1)/(2N^2)$.

— The second reaction can be seen as an auto-catalytic process. Here, given a monomer at the contact of a polymerised monomer, the monomer is converted into a polymerised monomer at a rate β. Again, by the law of mass action, free monomers disappear at the rate $\beta(x_1/N)(x_2/N)$.

See Eugène et al. [Eugène et al., 2015], Szavits-Nossan et al. [Szavits-Nossan et al., 2014] and Xue et al. [Xue et al., 2008] for a general presentation of these phenomena in a biological context. For more discussion and results on stochastic models associated to chemical reactions, see for example Anderson and Kurtz [Anderson and Kurtz, 2011] and Higham [Higham, 2008] and references therein.

This simple intuitive model of polymerisation has the advantage of having only two parameters to determine. It can be analyzed mathematically by standard tools of probability theory, see Eugène et al. [Eugène et al., 2015]. It has been shown that if $X_2^N(t)$ is the number of polymerised monomers at time t, then the polymerisation process can be described via the following convergence in distribution

$$\lim_{N\to+\infty} \left(\frac{X_2^N(Nt)}{N}\right) = (x_2(t)), \tag{3.4}$$

where $(x_2(t))$ is the non-trivial solution of the following simple ordinary differential equation

$$\dot{x}_2(t) = \alpha\big(m - x_2(t)\big)^2 + \beta\big(m - x_2(t)\big)x_2(t). \tag{3.5}$$

The variable $x_2(t)$ is converging to m has t goes to infinity.

By using these simple mathematical results and the data from experiments with 17 different initial concentrations of monomers (the value of m) and 12 experiments for each concentration, Table I of Eugène et al. [Eugène et al., 2015] shows that, in this setting, the estimation of β is reasonably robust. This is unfortunately not the case for the numerical estimation of α which is varying from $9.57{\cdot}10^{-8}h^{-1}M^{-1}$ to $1.68{\cdot}10^{-2}h^{-1}M^{-1}$. An additional difficulty with this simple model comes from the small values of α obtained. Indeed, for the experiments, the value of the volume N is in the order of 10^{15}, some of the estimated values of α in 10^{-8} are therefore, numerically, of the order of $1/\sqrt{N}$. The asymptotic results are obtained when N gets large and α *fixed*. For this reason, one may suspect a problem of convergence speed in Relation (3.4) when these parameters are used. It turns out that our simulations confirm that the asymptotic regime (3.4) does not seem to represent accurately the system when α is too small.

The purpose of the present paper is to refine this basic model in two different ways.

1. The model can be improved by introducing a key feature of the polymerisation process: misfolding of monomers. Experiments show that, in some cases, monomers can be polymerised only if their 3-D structure has been modified by some events. Such monomers are called misfolded monomers, see Dobson [Dobson, 2003, Dobson, 2006], Knowles et al. [Knowles et al., 2014b]. It turns out that, at a given time, only a small fraction of monomers are misfolded which may also explain that the polymerisation process starts very slowly. In biological cells, this phenomenon of misfolding is reversible, dedicated proteins may correct the misfolded monomers. A misfolded monomer can be turned into a regular monomer and vice-versa. See Bozaykut et al. [Bozaykut et al., 2014] and Lanneau et al. [Lanneau et al., 2010] for example. Section 3.3 is devoted to the mathematical analysis of these models.

2. Another approach is to keep the basic model but with the parameter α being of the order of $1/N^{\nu}$ for some positive ν to take into account that, in practice, the values of this parameter can be very small. Note that this is only a numerical observation, the value of α has no reason to depend on the volume. This model is analyzed in Section 3.3.

The rest of the section is devoted to a brief sketch of the mathematical aspects of these two classes of models. As it will be seen, the models are more challenging from a mathematical point of view, the model with misfolded monomers in particular.

3.1.2 Models with Misfolding Phenomena

Chemical reactions associated with this simple model are as follows:

$$\mathcal{X}_0 \underset{\gamma^*}{\overset{\gamma}{\rightleftarrows}} \mathcal{X}_1, \qquad \begin{cases} \mathcal{X}_1 + \mathcal{X}_1 \xrightarrow{\alpha} 2\mathcal{X}_2, \\ \mathcal{X}_1 + \mathcal{X}_2 \xrightarrow{\beta} 2\mathcal{X}_2. \end{cases}$$

At time $t \geq 0$, $X_0^N(t)$ denotes the number of regular monomers, $X_1^N(t)$ the number of misfolded monomers. As before, the last coordinate $X_2^N(t)$ is the polymerized mass. As a Markov process, $(X_0^N(t), X_1^N(t), X_2^N(t))$ has the following transitions, for an element $x = (x_0, x_1, x_2) \in \mathbb{N}^3$,

$$x \mapsto \begin{cases} x+(1,-1,0) \text{ at rate } \gamma^* x_1 \\ x+(-1,1,0) \quad " \quad \gamma x_0, \end{cases} \qquad x \mapsto \begin{cases} x+(-2,2) \quad \text{at rate } \alpha \dfrac{x_1(x_1-1)}{2N^2} \\ x+(-1,1) \quad " \quad \beta \dfrac{x_1}{N}\dfrac{x_2}{N}. \end{cases} \tag{3.6}$$

It is important to note that the transition between state 0, a regular monomer, and state 1, a misfolded monomer, is spontaneous. Consequently, as it can be seen, the corresponding transition rates *do not* depend on the volume N but simply on the number of components and not on their concentrations. An important consequence of this observation is that the system exhibits a two-time-scales behavior that we will investigate.

An informal description of the asymptotic behavior of $(X_2^N(t))$

The first two coordinates can be seen as an Ehrenfest process with two urns 0 and 1 where each particle in urn 0 (resp. 1) goes to urn 1 (resp. 0) at rate γ (resp. γ^*). See Bingham [Bingham, 1991] and Karlin and McGregor [Karlin and McGregor, 1965] for example. Particles in urn 1 can also go to the urn 2 corresponding to the polymerized mass but this phenomenon occurs at a much slower rate so that, locally, it does not change the orders of magnitude in N of X_2^N.

When $X_2^N \sim x_2 N$, there is a total of $(m-x_2)N$ particles in the urns 0 or 1. The components $(X_0^N(t), X_1^N(t))$ are both of the order of N and are moving on a fast time scale, proportional to N. The transition rates of the process $(X_2^N(t))$ are slower, bounded with respect to N. Because of the fast transition rates of the first two coordinates, the Ehrenfest urn process should reach quickly an equilibrium for which X_0^N has a binomial distribution with parameter $(m-x_2)N$ and r with $r = \gamma/(\gamma+\gamma^*)$, in particular

$$\frac{X_0^N}{N} \sim (1-r)(m-x_2) \text{ and } \frac{X_1^N}{N} \sim r(m-x_2).$$

This suggests that,

a) to see an evolution of X_2^N of the order of N, one has to be on the linear time scale $t \mapsto Nt$: transition rates of the process X_2^N are $O(1)$,

b) if $X_2^N(Nt) \sim x_2(t)N$, in view of transition rates of $(X_2^N(t))$ of Relation (3.6), then $(x_2(t))$ should satisfy the following ordinary differential equation

$$\dot{x}_2(t) = \alpha r^2 (m - x_2(t))^2 + \beta r (m - x_2(t)) x_2(t). \tag{3.7}$$

We recognize the limit equation (3.5) of the simple model, where α, β are respectively replaced by αr^2 and

βr. This result is also true when considering the second order fluctuations of the number of polymers, see Theorem 3.2.2. The proof of the convergence of the process of the concentration of polymerized monomers to the solution of the ODE (3.5) use standard arguments of convergence of a sequence of stochastic processes, see the supplementary material of Eugène et al. [Eugène et al., 2015]. The proof of the corresponding result with misfolding phenomena for the ODE (3.7) is, as we shall see, more delicate to handle.

Stochastic Averaging Phenomenon

To summarize these observations, the coordinates $(X_0^N(t), X_1^N(t))$ form a fast process and $(X_2^N(t))$ is a slow process when the scaling parameter N goes to infinity. This suggests a stochastic averaging principle (SAP) in a fully coupled context.

1. The stochastic evolution of $(X_2^N(Nt))$ is driven by the invariant distribution of an instantaneous associated Ehrenfest process.

2. The parameters of the Ehrenfest process depend on the macroscopic variable $(X_2^N(Nt))$,

see Papanicolaou et al. [Papanicolaou et al., 1977] and Chapter 8 of Freidlin and Wentzell [Freidlin and Wentzell, 1998] for example, see also Kurtz [Kurtz, 1992a].

A stochastic averaging principle is indeed proved as well as a corresponding central limit theorem (CLT). In our cases there are some differences with the classical framework of stochastic averaging principles. The state space of the fast process depends on the scaling parameter N, and is not in particular a fixed process (with varying parameters) as it is usually the case. See Hunt and Kurtz [Hunt and Kurtz, 1994] or Sun et al. [Sun et al., 2015] for example. A law of large numbers with respect to N for the invariant distribution of the fast process is driving the evolution of the slow process. The approach used in the paper relies on the use of occupation measures on a continuous state space instead of a discrete space, this leads to some technical complications as it will be seen. Concerning central limit theorems in a SAP context, there are few references available for jump processes. The methods presented in Kang et al. [Kang et al., 2014] or in Sun et al. [Sun et al., 2015] do not seem to be helpful in our case. Instead, an ad-hoc estimation, Proposition 3.2.4, gives the main ingredient to derive a central limit theorem, see Section 3.2.

3.1.3 Models with Scaled Reaction Rates

Again, $X_1^N(t)$ (resp. $X_2^N(t)$) is the number of free (resp. polymerised) monomers at time $t \geq 0$. The transition rates of the Markov process $(X_1^N(t), X_2^N(t))$ associated to these models are the same, except

that the parameter α is replaced by α/N^ν with $\nu > 0$. For $x = (x_1, x_2) \in \mathbb{N}^2$, the rates are given by

$$x \mapsto \begin{cases} x+(-2,2) & \text{at rate} \quad \dfrac{\alpha}{N^\nu}\dfrac{x_1(x_1-1)}{2N^2} \\ \\ x+(-1,1) & \text{"} \qquad\quad \beta\ \dfrac{x_1}{N}\dfrac{x_2}{N}. \end{cases} \tag{3.8}$$

Convergence (3.4) shows that the polymerisation occurs on the linear time scale $t \mapsto Nt$ for the basic model. It will be shown that for (3.8), the phenomenon does not start on this time scale. A slightly more rapid time scale is necessary for this purpose, it is shown that polymerization is happening on the time scale $t \mapsto (N \log N)\cdot t$ for $0 < \nu \leq 1$ and $t \mapsto N^\nu t$ when ν>1. See Section 3.3.

3.2 Stochastic Models with Misfolding Phenomena

3.2.1 Notations and Definitions

The following notations will be used throughout the paper. For $\xi \geq 0$, $\mathcal{N}_\xi(\mathrm{d}t)$ denotes a Poisson process with parameter ξ and $(\mathcal{N}^i_\xi(\mathrm{d}t))$ an i.i.d. sequence of such processes. All the Poisson processes are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

If f is a real valued function on $\mathbb{R}_+$ continuous on the right and with a left limit at any point of $\mathbb{R}_+$, $f(t-)$ denotes its limit on the left of $t > 0$. In the following the jump of f at $t > 0$ is denoted as

$$\mathrm{d}f(t) = f(t)-f(t-).$$

Recall that, at time $t \geq 0$, $X_0^N(t)$ is the number of monomers, $X_1^N(t)$ is the number of misfolded monomers and $X_2^N(t)$ is the polymerized mass. It is not difficult to see that these processes can be seen as the solution of the following stochastic differential equations,

$$\begin{cases} \mathrm{d}X_0^N(t) = \displaystyle\sum_{i=1}^{X_1^N(t-)} \mathcal{N}^i_{\gamma^*}(\mathrm{d}t) - \sum_{i=1}^{X_0^N(t-)} \mathcal{N}^i_{\gamma}(\mathrm{d}t), \\ \mathrm{d}X_2^N(t) = \displaystyle\sum_{i=1}^{X_1^N(t-)(X_1^N-1)(t-)/2} 2\mathcal{N}^i_{\alpha/N^2}(\mathrm{d}t) + \sum_{i=1}^{X_1^N(t-)X_2^N(t-)} \mathcal{N}^i_{\beta/N^2}(\mathrm{d}t), \end{cases} \tag{3.9}$$

with the relation of conservation of mass M_N=$X_0^N(t)$+$X_1^N(t)$+$X_2^N(t)$ and initial conditions $X_0^N(0) = M_N$ and $X_1^N(0) = X_2^N(0) = 0$.

It should be noted that the processes $(X_0^N(t))$, $(X_1^N(t))$ and $(X_2^N(t))$ are taken as cadlag, i.e. for $i = 0$, 1, 2, almost surely the function $t \mapsto X_i^N(t)$ is right continuous and has a left limit on any point of $\mathbb{R}_+$. For $N \geq 1$, the property of martingale mentioned in the following for processes with index N will refer

to the filtration $(\mathcal{F}_t^N)$ defined by, for $t > 0$,

$$\mathcal{F}_t^N = \sigma\left\langle \mathcal{N}_\xi^i([0,s]) : s \le t, i \in \mathbb{N}, \xi \in \{\gamma^*, \gamma, \alpha/N^2, \beta/N^2\}\right\rangle.$$

For some fixed T, we investigate the convergence in distribution of the sequence of processes $(X_2^N(Nt))/N)$ on the finite time interval $[0,T]$ in the space of cadlag processes on $[0,T]$ endowed with the Skorohod topology, see Chapter 3 of Billingsley [Billingsley, 1999]. As it will be seen the convergence proved occurs in fact for the topology of the uniform norm on this space of functions. All statements concerning convergence in distribution of processes in the following will refer to this notion.

3.2.2 Evolution Equations

By integrating Equation (3.9), one gets the relation

$$X_2^N(t) \;=\; X_2^N(0) \;+\; \frac{\alpha}{N^2}\int_0^t X_1^N(s)(X_1^N(s)-1)\,\mathrm{d}s \;+\; \frac{\beta}{N^2}\int_0^t X_1^N(s)X_2^N(s)\,\mathrm{d}s \;+\; \mathcal{M}_2^N(t), \quad (3.10)$$

where $(\mathcal{M}_2^N(t))$ is the martingale

$$\mathcal{M}_2^N(t) \;=\; 2\int_0^t \sum_{i=1}^{X_1^N(s-)(X_1^N-1)(s-)/2} \left[\mathcal{N}^i_{\alpha/N^2}(\mathrm{d}s) - \frac{\alpha}{N^2}\,\mathrm{d}s\right] + \int_0^t \sum_{i=1}^{X_1^N(s-)X_2^N(s-)} \left[\mathcal{N}^i_{\beta/N^2}(\mathrm{d}s) - \frac{\beta}{N^2}\,\mathrm{d}s\right]$$

which can be rewritten as an infinite sum of martingales

$$\mathcal{M}_2^N(t) = \sum_{i=1}^{+\infty}\int_0^t 2\times \mathbb{1}_{\{i\le X_1^N(s-)(X_1^N-1)(s-)/2\}}\left[\mathcal{N}^i_{\alpha/N^2}(\mathrm{d}s) - \frac{\alpha}{N^2}\,\mathrm{d}s\right] + \sum_{i=1}^{+\infty}\int_0^t \mathbb{1}_{\{i\le X_1^N(s-)X_2^N(s-)\}}\left[\mathcal{N}^i_{\beta/N^2}(\mathrm{d}s) - \frac{\beta}{N^2}\,\mathrm{d}s\right].$$

The previsible increasing process $(\langle \mathcal{M}_2^N\rangle(t))$of the martingale $(\mathcal{M}_2^N(t))$, is the unique increasing previsible process $(A(t))$ null at t=0 such that the process

$$\left(\left(\mathcal{M}_2^N(t)\right)^2 - A(t)\right)$$

is a local martingale. See Section VI-34 page 377 of Rogers and Williams [Rogers and Williams, 2000]. It is given by

$$\langle \mathcal{M}_2^N\rangle(t) = 2\frac{\alpha}{N^2}\int_0^t X_1^N(s)(X_1^N(s)-1)\,\mathrm{d}s + \frac{\beta}{N^2}\int_0^t X_1^N(s)X_2^N(s)\,\mathrm{d}s. \quad (3.11)$$

This can be seen by using the independence of the i.i.d. sequences of Poisson processes $(\mathcal{N}^i_{\alpha/N^2})$ and $(\mathcal{N}^i_{\beta/N^2})$ and the fact that, if $(Y(t))$ is a bounded left-continuous adapted process and $\lambda > 0$, then the

previsible increasing process of the martingale

$$\left(\int_0^t Y(s)\left[\mathcal{N}_\lambda(\mathrm{d}s) - \lambda\,\mathrm{d}s\right]\right) \text{ is given by } \left(\lambda\int_0^t Y(s)^2\,\mathrm{d}s\right).$$

See, for example, Theorem (27.6) page 48 of Rogers and Williams [Rogers and Williams, 2000].

For $i = 0,\ 1,\ 2$ and $t \geq 0$, denote

$$\overline{X}_i^N(t) = \frac{X_i^N(Nt)}{N},$$

the main goal of this section is to prove that the process $(\overline{X}_2^N(t))$ is converging in distribution to the solution $(x_2(t))$ of a non-trivial ordinary differential equation. It will show in particular that the polymerization process is occurring on the linear time scale $t \mapsto Nt$.

3.2.3 Random Measures Associated to Occupation Times

Define μ_N the random measure on $[0, m^*]^2 \times [0, T]$ by

$$\langle \mu_N, g\rangle = \int_{\mathbb{R}_+} g\left(\overline{X}_0^N(Nu), \overline{X}_1^N(Nu), u\right)\,\mathrm{d}u.$$

The distribution of the variable μ_N, $N \geq 1$ is an element of the space of the Radon measures on $[0, m^*]^2 \times [0, T]$ whose mass is less than T. The following proposition shows that the sequence of these distributions is tight.

Proposition 3.2.1. *The sequence (μ_N) is tight. Any limiting point μ_∞ of this sequence is such that*

$$\langle \mu_\infty, g\rangle = \int_{\mathbb{R}_+^3} g\,(x, y, u)\ \pi_u(\mathrm{d}x, \mathrm{d}y)\,\mathrm{d}u, \tag{3.12}$$

for any continuous function g on $[0, m^*]^2 \times [0, T]$, where for each $u \geq 0$, π_u is a random Radon measure on $\mathbb{R}_+^2$.

Recall that m^* is an upper bound of the sequence (M_N/N).

Proof. Since $\overline{X}_0^N(t)$ and $\overline{X}_1^N(t)$ are bounded, for any $T > 0$, the measure μ_N has a compact support. Lemma 3.2.8 page 44 of Dawson [Dawson, 1993] gives directly that the sequence (μ_N) of random measures is tight.

Let (μ_{N_k}) be a convergent subsequence with limit μ_∞. By using Skorohod's representation theorem, see Theorem 1.8 of Ethier and Kurtz [Ethier and Kurtz, 1986], one can assume that there exists a negligible measurable set $\mathcal{A}$ of the probability space such that, outside this subset, the convergence of the sequence (μ_{N_k}) of Radon measures towards μ_∞, that is

$$\lim_{k\to+\infty} \langle \mu_{N_k}, g\rangle = \langle \mu_\infty, g\rangle \text{ for all } g \in C([0, m^*]^2 \times [0, T]),$$

holds.

Let $h \in C([0,m^*]^2)$ and $f \in C([0,T])$, denoting $h \otimes f(x,y,u) = h(x,y)f(u)$, for $(x,y) \in [0,m^*]^2$ and $u \in [0,T]$, then, as a limit of the sequence (μ_{N_k}), the Radon measure

$$f \mapsto \langle \mu_\infty, h \otimes f \rangle$$

is absolutely continuous with respect to Lebesgue's measure. Consequently, for any $h \in C([0,m^*]^2)$, there exists some function $(\widetilde{\pi}_u(h), 0 \leq u \leq T)$ such that

$$\langle \mu_\infty, h \otimes f \rangle = \int_0^T \widetilde{\pi}_u(h) f(u) \, \mathrm{d}u.$$

By the differentiation theorem, see Theorem 7.10 in Rudin [Rudin, 1987], the function $(\widetilde{\pi}_u(h))$ can be represented as

$$\widetilde{\pi}_u(h) = \limsup_{\varepsilon \to 0} \frac{1}{\varepsilon} \left\langle \mu_\infty, h \otimes \mathbb{1}_{\{[u-\varepsilon/2, u+\varepsilon/2]\}} \right\rangle, \quad u \in [0,T],$$

consequently, the mapping $(\omega, u) \mapsto \widetilde{\pi}_u(h)(\omega)$ is $\mathcal{F} \otimes \mathcal{B}([0,T])$-measurable.

Let $\mathcal{S}$ be a countable dense subset of $C([0,m^*]^2)$, then there exists a subset E_0 of $[0,T]$ negligible for the Lebesgue measure such that, for all $u \in [0,T] \setminus E_0$ and ϕ_1, $\phi_2 \in \mathcal{S}$,

1. $\widetilde{\pi}_u(p_1\phi_1 + p_2\phi_2) = p_1\widetilde{\pi}_u(\phi_1) + p_2\widetilde{\pi}_u(\phi_2)$, $\forall p_1$, $p_2 \in \mathbb{Q}$,
2. $\widetilde{\pi}_u(\phi_1) \leq \widetilde{\pi}_u(\phi_2)$ if $\phi_1 \leq \phi_2$,
3. $\widetilde{\pi}_u(1) = 1$.

With the same method as in Section II.88 of Rogers and Williams [Rogers and Williams, 1994], for any $u \in [0,T] \setminus E_0$, one gets the existence of a Radon measure π_u on $[0,m^*]^2$ such that $\widetilde{\pi}_u(h) = \pi_u(h)$ for any $h \in \mathcal{S}$. By density of $\mathcal{S}$, the mapping $(\omega, u) \mapsto \pi_u(h)(\omega)$ is also $\mathcal{F} \otimes \mathcal{B}([0,T])$-measurable and the relation

$$\langle \mu_\infty, h \otimes f \rangle = \int_0^T \pi_u(h) f(u) \, \mathrm{d}u.$$

holds for all $h \in C([0,m^*]^2)$ and $f \in C([0,T])$. To identify μ_∞ one concludes with the density of the functions of the form $h \otimes f$ where h [resp. f] belongs to some countable dense subset of $C([0,m^*]^2)$ (resp. $C([0,T])$). The proposition is therefore proved. □

Representation (3.12) is related to Lemma 1.4 of Kurtz [Kurtz, 1992a]. Our proof relies on classical arguments of measure theory, a functional version of Caratheodory's extension theorem in particular which is described in Section II.88 of Rogers and Williams [Rogers and Williams, 1994]. In Kurtz [Kurtz, 1992a], a more sophisticated result, see Morando [Morando, 1969], on the extension of bi-measures is the key ingredient. The notion of bi-measure goes back to Kingman, see Dellacherie and Meyer [Dellacherie and Meyer, 1978] for example. We could also have used it to prove our result, but we preferred to give

here a more self-contained proof. It should be mentioned that Lemma 1.4 of Kurtz [Kurtz, 1992a] gives also additional measurability properties of the family (π_u) which are of no use in our case.

Proposition 3.2.2. *If μ_∞ is a limiting point of (μ_N) with the representation (3.12) then, for any $\mathcal{C}^1$-function f on $\mathbb{R}_+^2$, almost surely*

$$\int_0^t \int_{\mathbb{R}_+^2} (\gamma^* y - \gamma x) \left(\frac{\partial}{\partial x} f(x,y) - \frac{\partial}{\partial y} f(x,y) \right) \pi_u(\mathrm{d}x, \mathrm{d}y)\ \mathrm{d}u = 0, \quad \forall t \geq 0, \tag{3.13}$$

in particular, almost surely,

$$\int_0^t \int_{\mathbb{R}_+^2} (\gamma^* y - \gamma x)^2\, \pi_u(\mathrm{d}x, \mathrm{d}y)\ \mathrm{d}u = 0, \quad \forall t \geq 0. \tag{3.14}$$

Relation (3.14) just says that almost surely and for almost all u, the measure π_u is degenerated on $\mathbb{R}_+^2$ and carried by the subset $\{(x, \gamma x/\gamma^*) : 0 \leq x \leq m^*\}$. Since m^* is only an upper bound of the sequence (M_N/N), it can therefore be taken arbitrarily close to m. The support of the measure π_u is thus contained in the set $\{(x, \gamma x/\gamma^*) : 0 \leq x \leq m\}$.

Proof. for $(i,j) \in \mathbb{Z}^2$, one denotes by Δ_{ij} the discrete differential operator

$$\Delta_{ij}^N(f)(x,y) = f(x + i/N, y + j/N) - f(x,y), \quad (x,y) \in [0, m^*]^2.$$

After some trite calculations, the stochastic differential equations (3.9) give the relation

$$\begin{aligned} f\left(\overline{X}^N(t/N)\right) = f\left(\overline{X}^N(0)\right) &+ \gamma \int_0^t X_0^N(s) \Delta_{-1,1}^N(f)\left(\overline{X}^N(s/N)\right) \mathrm{d}s \\ &+ \gamma^* \int_0^t X_1^N(s) \Delta_{1,-1}^N(f)\left(\overline{X}^N(s/N)\right) \mathrm{d}s \\ &+ \alpha \int_0^t \frac{X_1^N(s)(X_1^N(s)-1)}{2N^2} \Delta_{0,-2}^N(f)\left(\overline{X}^N(s/N)\right) \mathrm{d}s \\ &+ \beta \int_0^t \frac{X_1^N(s)}{N} \frac{X_2^N(s)}{N} \Delta_{0,-1}^N(f)\left(\overline{X}^N(s/N)\right) \mathrm{d}s + \mathcal{M}_f^N(t), \end{aligned} \tag{3.15}$$

where $(\overline{X}^N(t)) = (X_0^N(Nt)/N, X_1^N(Nt)/N)$ and $(\mathcal{M}_f^N(t))$ is the associated martingale. Its previsible

increasing process is given by

$$\begin{aligned}\left\langle \mathcal{M}_f^N \right\rangle (t) = \gamma \int_0^t X_0^N(s)\Delta_{-1,1}^N(f)^2 \left(\overline{X}^N(s/N)\right) \mathrm{d}s \\ + \gamma^* \int_0^t X_1(s)\Delta_{1,-1}^N(f)^2 \left(\overline{X}^N(s/N)\right) \mathrm{d}s \\ + \alpha \int_0^t \frac{X_1^N(s)(X_1^N(s)-1)}{2N^2}\Delta_{0,-2}^N(f)^2 \left(\overline{X}^N(s/N)\right) \mathrm{d}s \\ + \beta \int_0^t \frac{X_1^N(s)}{N}\frac{X_2^N(s)}{N}\Delta_{0,-1}^N(f)^2 \left(\overline{X}^N(s/N)\right) \mathrm{d}s. \quad (3.16)\end{aligned}$$

Note that, for i, $j \in \mathbb{Z}$

$$\Delta_{i,j}^N(f)(x,y) = \frac{1}{N}\left(i\frac{\partial f}{\partial x}(x,y) + j\frac{\partial f}{\partial y}(x,y)\right) + o(1/N),$$

and, since $(\overline{X}_i^N(t))$ is bounded for $i = 0$ and 1, there exists some finite constant $C_0(T)$ such that for $0 \le s \le NT$, one has $\langle \mathcal{M}_f^N \rangle(s) \le C_0(T)$.

By changing the time variable in Nt in Equation (3.15) and by dividing by N one gets the relation

$$\begin{aligned}\frac{1}{N}\left(f\left(\overline{X}^N(t)\right) - f\left(\overline{X}^N(0)\right)\right) \\ = \int_0^t \left[\gamma^*\overline{X}_1^N(s) - \gamma\overline{X}_0^N(s)\right]\left[\frac{\partial f}{\partial x} - \frac{\partial f}{\partial y}\right]\left(\overline{X}^N(s)\right) \mathrm{d}s \\ - \frac{\alpha}{N}\int_0^t \overline{X}_1^N(s)\left(\overline{X}_1^N(s) - 1/N\right)\frac{\partial f}{\partial y}\left(\overline{X}^N(s)\right) \mathrm{d}s \\ - \frac{\beta}{N}\int_0^t \overline{X}_1^N(s)\overline{X}_2^N(s)\frac{\partial f}{\partial y}\left(\overline{X}^N(s)\right) \mathrm{d}s + \frac{\mathcal{M}_f^N(Nt)}{N} + o(1), \quad (3.17)\end{aligned}$$

with $(\overline{X}^N(t)) = (\overline{X}_0^N(t), \overline{X}_1^N(t))$.

Since the previsible increasing process of the martingale $(\mathcal{M}_f^N(Nt)/N)$ is

$$\left(\left\langle \frac{\mathcal{M}_f^N(N\cdot)}{N} \right\rangle (t)\right) = \left(\frac{\left\langle \mathcal{M}_f^N \right\rangle (Nt)}{N^2}\right),$$

Doob's Inequality, see Theorem (70.1) page 177 of Rogers and Williams [Rogers and Williams, 2000], gives that for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{0\le t\le T}\frac{|\mathcal{M}_f^N(Nt)|}{N} \ge \varepsilon\right) \le \frac{1}{\varepsilon^2}\mathbb{E}\left(\left\langle \frac{\mathcal{M}_f^N(N\cdot)}{N} \right\rangle (T)\right) \le \frac{C_0(T)}{N^2},$$

the martingale process $(\mathcal{M}_f^N(Nt)/N)$ is thus converging to 0.

From Equation (3.17), one gets similarly the following convergence in distribution

$$\lim_{N\to+\infty}\left(\int_0^t\left[\gamma^*\overline{X}_1^N(s)-\gamma\overline{X}_0^N(s)\right]\left[\frac{\partial f}{\partial x}-\frac{\partial f}{\partial y}\right]\left(\overline{X}^N(s)\right)\,\mathrm{d}s\right)=0. \tag{3.18}$$

For $t\geq 0$,

$$\int_0^t\left[\gamma^*\overline{X}_1^N(s)-\gamma\overline{X}_0^N(s)\right]\left[\frac{\partial f}{\partial x}-\frac{\partial f}{\partial y}\right]\left(\overline{X}^N(s)\right)\,\mathrm{d}s \\ =\int\left[\gamma^*y-\gamma x\right]\left[\frac{\partial f}{\partial x}-\frac{\partial f}{\partial y}\right](x,y)\,\mathbb{1}_{\{s\leq t\}}\,\mu_N(\mathrm{d}x,\mathrm{d}y,\mathrm{d}s),$$

and, by approximating the indicator function of $[0,t)$ (resp. $[0,t]$) from below (resp. from above) by continuous functions on $[0,T]$, this last term converges in distribution to

$$\int\left[\gamma^*y-\gamma x\right]\left[\frac{\partial f}{\partial x}-\frac{\partial f}{\partial y}\right](x,y)\,\mathbb{1}_{\{u\leq t\}}\,\mu_\infty(\mathrm{d}x,\mathrm{d}y,\mathrm{d}u) \\ =\int_0^t\int_{\mathbb{R}_+^2}(\gamma^*y-\gamma x)\left(\frac{\partial}{\partial x}f(x,y)-\frac{\partial}{\partial y}f(x,y)\right)\pi_s(\mathrm{d}x,\mathrm{d}y)\,\mathrm{d}s.$$

This convergence in distribution also holds for any finite marginals of this process. The convergence of processes (3.18) gives therefore the desired identity (3.13) in distribution. The last assertion of the proposition is proved by taking the function $f(x,y)=\gamma^*x^2-\gamma y^2$. □

3.2.4 A Stochastic Averaging Principle

Relation (3.10) gives the following integral equation for $(\overline{X}_2^N(t))$,

$$\overline{X}_2^N(t) \;=\; \overline{X}_2^N(0)\;+\;\alpha\int_0^t\overline{X}_1^N(s)\left(\overline{X}_1^N(s)-1/N\right)\mathrm{d}s\;+\;\beta\int_0^t\overline{X}_1^N(s)\overline{X}_2^N(s)\,\mathrm{d}s\;+\;\frac{\mathcal{M}_2^N(Nt)}{N}, \tag{3.19}$$

In the same way as before, one can show that the expected value of the previsible increasing process of the martingale $(\mathcal{M}_2^N(Nt)/N)$ is vanishing as N gets large by Equation (3.11). Doob's Inequality gives that the martingale converges in distribution to 0. The criteria of the modulus of continuity, see Theorem 7.2 page 81 of Billingsley [Billingsley, 1999], gives therefore that the sequence of processes $(\overline{X}_2^N(t))$ is tight in the space of cadlag processes endowed with the Skorohod topology, see Chapter 3 of Billingsley [Billingsley, 1999]. Corollary of page 142 of this reference gives in fact that the convergence in distribution of a subsequence of $(\overline{X}_2^N(t))$ occurs for the topology of the uniform norm. It can therefore be assumed, for some subsequence (N_k), that the following convergence holds,

$$\lim_{k\to+\infty}\left(\mu_{N_k},\left(\overline{X}_2^{N_k}(t)\right)\right)=(\mu_\infty,(x_2(t)))$$

for a random measure μ_∞ as in Proposition 3.2.1 and some continuous stochastic process $(x_2(t))$. The rest of the section is devoted to the identification of $(x_2(t))$.

Proposition 3.2.3. *For any continuous function g on $[0, m^*]^2$, the relation*

$$\left(\int_0^t \mathrm{d}u \int g(x,y)\pi_u(\mathrm{d}x,\mathrm{d}y)\right) \overset{dist.}{=} \left(\int_0^t g\left((m-x_2(u))(1-r,r)\right)\,\mathrm{d}u\right)$$

holds, with $r = \gamma/(\gamma+\gamma^*)$.

One concludes that the measure $\pi_u(\mathrm{d}x,\mathrm{d}y)$ of Proposition 3.2.1 is simply the Dirac measure at $[m - x_2(u)](1-r,r)$. This is the rigorous description of the fact described at the beginning of this section that, if the fraction of polymerized mass is $x_2(u)$, then the fraction of regular [resp. misfolded] monomers is $(1-r)(m-x_2(u))$ [resp. $r(m-x_2(u))$].

Proof. The criteria of the modulus of continuity shows that the sequence of processes

$$\left(\int_0^t g\left(\overline{X}^{N_k}(u)\right)\,\mathrm{d}u\right) = \left(\int_0^t g\left(\overline{X}_0^{N_k}(u), \overline{X}_1^{N_k}(u)\right)\,\mathrm{d}u\right)$$

is tight. In particular a possible limit of this sequence is a continuous process. All we have to do is to identify its finite marginals.

For a fixed $t \geq 0$, by using the convergence in distribution of the positive measure (μ_{N_k}), Skorohod's representation theorem and by approximating the indicator function of $[0,t)$ (resp. $[0,t]$) from below (resp. from above) by continuous functions on $[0,T]$, one gets the convergence

$$\lim_{k\to+\infty} \int_0^t g\left(\overline{X}_0^{N_k}(u), \overline{X}_1^{N_k}(u)\right)\,\mathrm{d}u$$
$$= \int_0^t \mathrm{d}u \int g(x,y)\pi_u(\mathrm{d}x,\mathrm{d}y) = \int_0^t \mathrm{d}u \int g\left(x, \frac{\gamma}{\gamma^*}x\right)\pi_u(\mathrm{d}x,\mathrm{d}y) \quad (3.20)$$

by Proposition 3.2.2. The above convergence in distribution also holds for finite marginals for $t_1 \leq t_2 \leq \cdots \leq t_p$, $p \geq 1$. One has to identify the first marginal of (π_u). If f is a continuous function on $[0, m^*]$, by conservation of mass, one has the relation

$$\left(\int_0^t f\left(\overline{X}_0^{N_k}(u) + \overline{X}_1^{N_k}(u)\right)\,\mathrm{d}u\right) = \left(\int_0^t f\left(\frac{M_{N_k}}{N_k} - \overline{X}_2^{N_k}(u)\right)\,\mathrm{d}u\right).$$

Relation (3.20) and the convergence properties of the right hand side of this identity give the following identity of the distribution of processes

$$\left(\int_0^t f\left(x/r\right)\pi_u(\mathrm{d}x,\mathrm{d}y)\,\mathrm{d}u\right) \overset{\text{dist.}}{=} \left(\int_0^t f\left(m-x_2(u)\right)\,\mathrm{d}u\right).$$

The proposition is proved. □

Theorem 3.2.1. *Under the scaling condition (3.2) and if the initial state of the solution of the SDE (3.9)*

is given by $(M_N, 0, 0)$ *then, for the convergence in distribution,*

$$\lim_{N\to+\infty}\left(\frac{X_2^N(Nt)}{N}\right) = (x_2(t)) \overset{\text{def.}}{=} \left(\frac{1-e^{-\beta r m t}}{1+(\beta/\alpha r-1)e^{-\beta r m t}}m\right), \tag{3.21}$$

with $r = \gamma/(\gamma+\gamma^*)$.

Proof. By using Relation (3.19), Proposition 3.2.1 and the above proposition, one gets that any limiting point $(x_2(t))$ of $(X_2^N(Nt)/N)$ satisfies necessarily the following integral equation (integral form of the equation (3.7))

$$x_2(t) = \alpha r^2 \int_0^t (m - x_2(s))^2 \, \mathrm{d}s + \beta r \int_0^t (m - x_2(s)) x_2(s) \, \mathrm{d}s. \tag{3.22}$$

By uniqueness of the solution of this equation, one gets the convergence in distribution of the sequence of processes $(X_2^N(Nt)/N)$. Its explicit expression is easily obtained. □

The following corollary gives the asymptotics of the first instant when a fraction $\delta \in (0,1)$ of monomers has been polymerized. This is a key quantity that can be measured with experiments.

Corollary 3.2.1. *[Asymptotics of Lag Time] Under the conditions of Theorem 3.2.1, if for* $\delta \in (0,1)$,

$$T_N(\delta) = \inf\{t \geq 0 : X_2^N(t)/M_N \geq \delta\}, \tag{3.23}$$

then, for the convergence in distribution

$$\lim_{N\to+\infty}\frac{T_N(\delta)}{N} = t_\delta \overset{\text{def.}}{=} \frac{1}{rm\beta}\log\left(1+\frac{\delta\beta}{\alpha r(1-\delta)}\right). \tag{3.24}$$

3.2.5 Central Limit Theorem

From Proposition 3.2.2, it has been proved that if $f : [0, m^*]^2 \to \mathbb{R}$ is a $\mathcal{C}^1$-function then, for the convergence in distribution

$$\lim_{N\to+\infty}\left(\int_0^t (\gamma^* y - \gamma x)\left(\frac{\partial}{\partial x}f(x,y) - \frac{\partial}{\partial y}f(x,y)\right)\mu_N(\mathrm{d}x, \mathrm{d}y, \mathrm{d}s)\right) = (0),$$

with the above notations. The following proposition is an extension of this result. This is the key ingredient to prove the central limit result of this section.

Proposition 3.2.4. *If* $g : [0, m^*]^2 \times \mathbb{R}_+ \to \mathbb{R}$ *is a* $\mathcal{C}^1$*-function then, for the convergence in distribution,*

$$\lim_{N\to+\infty}\left(\int_0^t (\gamma^* y - \gamma x)\left(\frac{\partial}{\partial x}g(x,y,u) - \frac{\partial}{\partial y}g(x,y,u)\right)\sqrt{N}\mu_N(\mathrm{d}x, \mathrm{d}y, \mathrm{d}u)\right) = (0).$$

Proof. We follow the same lines as in the proof of Proposition 3.2.2. The analogue of Relation (3.17) is,

recall that $\overline{X}^N(t) = (\overline{X}_0^N(t), \overline{X}_1^N(t))$,

$$\begin{aligned}
\frac{1}{\sqrt{N}}\left(g\left(\overline{X}^N(t),t\right) - g\left(\overline{X}^N(0),0\right)\right)
&= \sqrt{N}\int_0^t \left[\gamma^*\overline{X}_1^N(s) - \gamma\overline{X}_0^N(s)\right]\left[\frac{\partial g}{\partial x} - \frac{\partial g}{\partial y}\right]\left(\overline{X}^N(s),s\right)\,\mathrm{d}s \\
&\quad - \frac{\alpha}{\sqrt{N}}\int_0^t \overline{X}_1^N(s)\left(\overline{X}_1^N(s) - 1/N\right)\frac{\partial g}{\partial y}\left(\overline{X}^N(s),s\right)\,\mathrm{d}s \\
&\quad - \frac{\beta}{\sqrt{N}}\int_0^t \overline{X}_1^N(s)\overline{X}_2^N(s)\frac{\partial g}{\partial y}\left(\overline{X}^N(s),s\right)\,\mathrm{d}s \\
&\quad + \frac{1}{\sqrt{N}}\int_0^t \frac{\partial g}{\partial z}\left(\overline{X}^N(s),s\right)\,\mathrm{d}s + \frac{M_g^N(Nt)}{\sqrt{N}} + o(1/\sqrt{N}). \qquad (3.25)
\end{aligned}$$

It is not difficult to check with the analogue of Relation (3.16) for the previsible increasing process of the martingale $(M_g^N(Nt)/\sqrt{N})$ that, for $t \geq 0$,

$$\lim_{N\to+\infty} \mathbb{E}\left(\left\langle \frac{M_g^N(Nt)}{\sqrt{N}} \right\rangle\right) = \lim_{N\to+\infty} \frac{\mathbb{E}\left(\langle M_g^N\rangle(Nt)\right)}{N} = 0.$$

Consequently, by Doob's Inequality, the martingale of Relation (3.25) vanishes when N gets large. By using the fact that, if H is some bounded Borel function on $\mathbb{R}_+^3$, then

$$\int_0^t H\left(\overline{X}^N(s),s\right)\,\mathrm{d}s = \int_{\mathbb{R}_+^2\times[0,t]} H\left(x,y,s\right)\,\mu_N(\mathrm{d}x,\mathrm{d}y,\mathrm{d}s),$$

and that $\overline{X}_2^N(s) = M_N/N - \overline{X}_0^N(s) - \overline{X}_1^N(s)$, the desired convergence of the proposition is then easily derived. □

Theorem 3.2.2 (Central Limit Theorem)**.** *Under Condition (3.2) and if $(x_2(t))$ is the function defined by Relation (3.21) then, for the convergence in distribution,*

$$\lim_{N\to+\infty}\left(\frac{X_2^N(Nt) - Nx_2(t)}{\sqrt{N}}\right) = (U(t)),$$

where $(U(t))$ is the solution of the stochastic differential equation

$$\mathrm{d}U(t) = \sqrt{\sigma(t)}\,\mathrm{d}B(t) + h(t)U(t)\,\mathrm{d}t, \qquad (3.26)$$

and $(B(t))$ is a standard Brownian motion and

$$\begin{cases} \sigma(t) = 2\alpha r^2(m - x_2(t))^2 + \beta r(m - x_2(t))x_2(t) \\ h(t) = r(\beta - 2\alpha r)(m - x_2(t)) - \beta r x_2(t). \end{cases}$$

The corresponding result of Eugène et al. [Eugène et al., 2015] when there is no misfolding phenomenon

shows that the functions σ and h are similar if α and β are respectively replaced by αr^2 and βr.

Proof. Denote

$$U^N(t) = \frac{X_2^N(Nt) - Nx_2(t)}{\sqrt{N}} = \sqrt{N}\left(\overline{X}_2^N(t)) - x_2(t)\right).$$

By combining Equation (3.19),

$$\overline{X}_2^N(t)) = \alpha \int_0^t \overline{X}_1^N(s)^2 \, \mathrm{d}s + \beta \int_0^t \overline{X}_1^N(s)\overline{X}_2^N(s) \, \mathrm{d}s + \frac{\mathcal{M}_2^N(Nt)}{N} + O(1/N)$$

and Relation (3.22),

$$x_2(t) = \alpha r^2 \int_0^t (m - x_2(s))^2 \, \mathrm{d}s + \beta r \int_0^t (m - x_2(s))x_2(s) \, \mathrm{d}s, \tag{3.27}$$

one gets

$$\begin{aligned} U^N(t) = \alpha\sqrt{N} \int_0^t \left(\overline{X}_1^N(s)^2 - r^2(m - x_2(s))^2\right) \mathrm{d}s \\ + \beta\sqrt{N} \int_0^t \left(\overline{X}_1^N(s)\overline{X}_2^N(s) - r(m - x_2(s))x_2(s)\right) \mathrm{d}s + \frac{\mathcal{M}_2^N(Nt)}{\sqrt{N}} + O(1/\sqrt{N}). \end{aligned}$$

Concerning the martingale term, Relation (3.11) gives, for $t \geq 0$,

$$\begin{aligned} \left\langle \frac{\mathcal{M}_2^N}{\sqrt{N}} \right\rangle (Nt) &= 2\alpha \int_0^t \overline{X}_1^N(s)^2 \, \mathrm{d}s + \beta \int_0^t \overline{X}_1^N(s)\overline{X}_2^N(s) \, \mathrm{d}s + O(1/N) \\ &= \int_{\mathbb{R}_+^2 \times [0,t]} \left[2\alpha x^2 + \beta x \left(M_N/N - x - y\right)\right] \mu_N(\mathrm{d}x, \mathrm{d}y, \mathrm{d}s) + O(1/N). \end{aligned}$$

With the same method as in the proof of Theorem 3.2.1, one gets the following convergence in distribution

$$\lim_{N \to +\infty} \left(\left\langle \frac{\mathcal{M}_2^N}{\sqrt{N}} \right\rangle (Nt) \right) = \left(\int_0^t \left[2\alpha r^2 (m - x_2(s))^2 + \beta r x_2(s)(m - x_2(s))\right] \mathrm{d}s \right)$$

by Relation (3.27).

Note also that, for $s \geq 0$,

$$\begin{aligned} \sqrt{N}\left(\overline{X}_1^N(s)\overline{X}_2^N(s) - r(m - x_2(s))x_2(s)\right) \\ = U^N(s)\overline{X}_1^N(s) + \sqrt{N}\left(\overline{X}_1^N(s) - r(m - x_2(s))\right)x_2(s) \end{aligned}$$

and

$$\sqrt{N}\left(\overline{X}_1^N(s) - r(m - x_2(s)\right) = -\frac{\sqrt{N}}{\gamma + \gamma^*}\left(\gamma\overline{X}_0^N(s) - \gamma^*\overline{X}_1^N(s)\right) - rU^N(t). \tag{3.28}$$

The above relation for $(U^N(t))$ can then be rewritten as

$$U^N(t) = \int_0^t U^N(s)\left((\beta-\alpha r)\overline{X}_1^N(s) - \alpha r^2(m-x_2(s)) - \beta r x_2(s)\right)\,\mathrm{d}s \\ -\frac{1}{\gamma+\gamma^*}\int_0^t (\gamma^* y - \gamma x)\left[\alpha(y+r(m-x_2(s)))+\beta x_2(s)\right]\sqrt{N}\mu_N(\mathrm{d}x,\mathrm{d}y,\mathrm{d}s) \\ + \frac{\mathcal{M}_2^N(Nt)}{\sqrt{N}} + O(1/\sqrt{N}). \quad (3.29)$$

The convergence in distribution of the martingale, Proposition 3.2.4 and the criterion of the modulus of continuity give easily the tightness of the sequence $(U^N(t))$. Let $(U(t))$ be a limit of some subsequence $(U^{N_k}(t))$.

A close look at Relation (3.29) shows that the theorem will be proved, with standard arguments, if the following convergence in distribution is proved

$$\lim_{k\to+\infty}\left(\int_0^t U^{N_k}(s)\overline{X}_1^{N_k}(s)\,\mathrm{d}s\right) = \left(r\int_0^t U(s)(m-x_2(s))\,\mathrm{d}s\right).$$

For $k \geq 0$,

$$\int_0^t U^{N_k}(s)\overline{X}_1^{N_k}(s) - rU(s)(m-x_2(s))\,\mathrm{d}s \\ = \int_0^t U^N(s)\left(\overline{X}_1^N(s) - r(m-x_2(s))\right)\mathrm{d}s + \int_0^t r(m-x_2(s))\left(U^N(s)-U(s)\right)\mathrm{d}s,$$

the process associated to the last term of the second part of this identity converges in distribution to 0. By Relation (3.28), the first term can be written as

$$-\int_0^t\left(\overline{X}_2^{N_k}(s)-x_2(s)\right)\left(\frac{\sqrt{N_k}}{\gamma+\gamma^*}\left(\gamma^*\overline{X}_0^{N_k}(s)-\gamma\overline{X}_1^{N_k}(s)\right)+rU^{N_k}(t)\right)\mathrm{d}s \\ = -r\int_0^t\left(\overline{X}_2^{N_k}(s)-x_2(s)\right)U^{N_k}(s)\,\mathrm{d}s \\ -\frac{1}{\gamma+\gamma^*}\int_0^t\left(\frac{M_{N_k}}{N_k}-x-y-x_2(s)\right)(\gamma x-\gamma^* y)\,\sqrt{N_k}\mu_{N_k}(\mathrm{d}x,\mathrm{d}y,\mathrm{d}s).$$

the first term of the right hand side converges in distribution to 0 due to Theorem 3.2.1 and the same property also holds for the second term by Proposition 3.2.4. The theorem is proved. □

As a consequence, one gets the following central limit theorem for the lag time. The notations of Corollary 3.2.1 and Theorems 3.2.1 and 3.2.2 are used.

Corollary 3.2.2. *Under the scaling regime (3.2), for $\delta \in (0,1)$, the convergence in distribution*

$$\lim_{N\to+\infty}\frac{T_N(\delta)-Nt_\delta}{\sqrt{N}} = \frac{-U(t_\delta)}{rm^2(1-\delta)(\beta+\alpha r(1-\delta))} \quad (3.30)$$

holds, where the variables $T_N(\delta)$ *and* t_δ *are defined by* (3.23) *and* (3.24) *and* $(U(t))$ *by* (3.26), *and* $r = \gamma/(\gamma+\gamma^*)$.

Proof. For $z \in \mathbb{R}$ note that, since $(X_2^N(t))$ is a non-decreasing process,

$$\left\{\frac{T^N(\delta) - Nt_\delta}{\sqrt{N}} \geq z\right\} = \left\{X_2^N\left(s_N\right) < \delta M_N\right\} = \left\{\frac{\overline{X}_2^N\left(s_N\right) - Nx_2(s_N/N)}{\sqrt{N}} < \frac{\delta M_N - Nx_2(s_N/N)}{\sqrt{N}}\right\},$$

with $s_N = Nt_\delta + z\sqrt{N}$. From Theorem 3.2.2 one gets the convergence in distribution

$$\lim_{N\to+\infty} \frac{\overline{X}_2^N\left(s_N\right) - Nx_2(s_N/N)}{\sqrt{N}} = U(t_\delta)$$

and the expansion of $(x_2(t))$ at t_δ gives

$$\lim_{N\to+\infty} \frac{\delta M_N - Nx_2(s_N/N)}{\sqrt{N}} = -zrm^2(1-\delta)(\beta + \alpha r(1-\delta)).$$

This completes the proof of the corollary. □

Equation (3.30) shows that the variance of the lag time is inversely proportional to γ/γ^*, a low misfolding rate will thus increase the variability of the polymerisation process.

3.3 Models with Scaled Reaction Rates

For $t \geq 0$, $X_1^N(t)$ is the number of free monomers at time t and $X_2^N(t)$ is the number of polymerized monomers. Recall that, for this model, the transition rates are given by

$$x \mapsto \begin{cases} x+(-2,2) & \text{at rate} \quad \dfrac{\alpha}{N^\nu}\dfrac{x_1(x_1-1)}{2N^2} \\ x+(-1,1) & \text{"} \qquad\quad \beta\,\dfrac{x_1}{N}\dfrac{x_2}{N}. \end{cases}$$

The initial condition is $X_1^N(0) = M_N$ and $X_2^N(0) = 0$. Because of the relation of conservation of mass, one has $M_N = X_1^N(t) + X_2^N(t)$.

It is not difficult to see that the process $(X_2^N(t))$ can be represented as the solution of the following stochastic differential equations,

$$\mathrm{d}X_2^N(t) = \sum_{i=1}^{X_1^N(X_1^N-1)(s-)/2} 2\,\mathcal{N}^i_{\alpha/N^{\nu+2}}(\mathrm{d}t) + \sum_{i=1}^{X_1^N(s-)X_2^N(s-))} \mathcal{N}^i_{\beta/N^2}(\mathrm{d}t). \tag{3.31}$$

By integrating this equation, one gets the relation

$$X_2^N(t) = \frac{\alpha}{N^{2+\nu}} \int_0^t X_1^N(s)(X_1^N(s)-1)\ \mathrm{d}s + \frac{\beta}{N^2} \int_0^t X_1^N(s)X_2^N(s)\ \mathrm{d}s + M^N(t), \tag{3.32}$$

where $(M^N(t))$ is a martingale whose previsible increasing process is given by

$$\left\langle M^N \right\rangle(t) = 2\frac{\alpha}{N^{2+\nu}} \int_0^t X_1^N(s)(X_1^N(s)-1)\ \mathrm{d}s + \frac{\beta}{N^2} \int_0^t X_1^N(s)X_2^N(s)\ \mathrm{d}s. \tag{3.33}$$

The following proposition shows that, on the time scale $t \mapsto Nt$, the polymerised mass is for this model in the order of $N^{1-\nu}$.

Proposition 3.3.1. *Under the scaling condition (3.2), for the convergence in distribution, the relation*

$$\lim_{N\to+\infty} \left(\frac{X_2^N(Nt)}{N^{1-\nu}} \right) = \left(\frac{\alpha m}{\beta} \left(e^{\beta m t} - 1 \right) \right)$$

holds.

Proof. The proof is standard by using the identities (3.32) and (3.33), and the relation $X_1^N(t)+X_2^N(t)=M_N$. See Eugène et al. [Eugène et al., 2015] for example. □

The following lemma introduces a branching process which will be helpful to estimate the order of magnitude in N of the lag time

$$T_N(\delta) = \inf\{t \geq 0 : X_2^N(t)/M_N \geq \delta\},$$

for $0 < \delta < 1$.

Lemma 3.3.1. *For a, $b > 0$, let $(W_{a,b}^N(t))$ be a pure birth process with birth rate*

$$\frac{a}{N^\nu} + \frac{b}{N}x$$

in state $x \in \mathbb{N}$, with $W(0) = 0$ and $0 < \nu \leq 1$. If

$$\tau_{a,b}^N(\delta) \overset{\text{def.}}{=} \inf\left\{t > 0 : W_{a,b}^N(t) \geq \delta N\right\},$$

then the sequence $(\tau_{a,b}^N(\delta)/(N \log N))$ converges in distribution to ν/b.

As it can be seen $(W_{a,b}^N(t))$ is a branching process with immigration. Immigration rate is a/N^ν and the reproduction rate is given by b/N. See Harris [Harris, 2002] for example.

Proof. Let, for $x \in \mathbb{N}$, E_x^N denotes an exponential random variable with parameter $a/N^\nu+xb/N$, assuming

that the random variables E_x^N, $x \geq 0$ are independent, then clearly

$$\tau_{a,b}^N(\delta) \overset{\text{dist}}{=} \sum_{x=0}^{\lfloor \delta N \rfloor} E_x^N.$$

hence after some simple estimations

$$\lim_{N \to +\infty} \frac{\mathbb{E}(\tau_{a,b}^N(\delta))}{N \log N} = \frac{\nu}{b}.$$

In the same way, one checks that the sequence $(\mathrm{Var}(\tau_{a,b}^N(\delta)/N))$ is bounded

$$\mathrm{Var}\left(\frac{\tau_{a,b}^N(\delta)}{N}\right) \leq \sum_{x=0}^{+\infty} \frac{1}{(aN^{1-\nu} + xb)^2}. \tag{3.34}$$

The convergence in distribution follows , by using Chebishev's Inequality. □

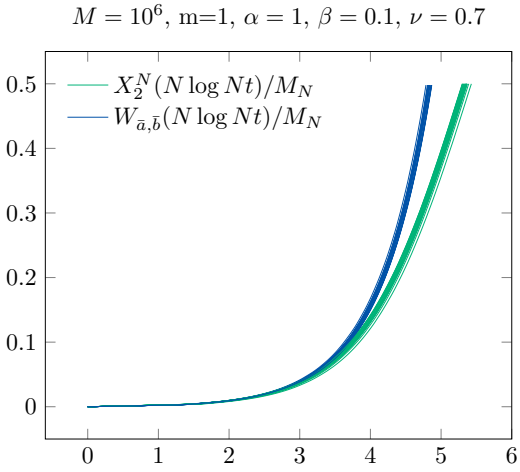


Figure 3.1: In blue, 20 simulations of $(W_{\bar{a},\bar{b}}/M_N)$ and in green, 20 simulations of (X_2^N/M_N) on the time scale $t \mapsto N \log Nt$.

Let $0 < \delta < 1$ and fix some $\underline{\kappa} < 1 < \overline{\kappa}$, one can assume that N is sufficiently large so that $\underline{\kappa} \leq M_N/(mN) \leq \overline{\kappa}$ holds. Recall that $T^N(\delta)$ is the first time that the fraction of the number of polymerised monomers $X_2^N(t)/M_N$ is greater than δ. The transition rates of $(X_2^N(t))$ are given by

$$x \mapsto \begin{cases} x{+}2 \text{ at rate} & \alpha/N^\nu \left[(M_N - x)/N\right]^2 \\ x{+}1 & \beta\, x/N \times (M_N - x)/N. \end{cases} \tag{3.35}$$

By comparing the transition rates, we see that, for $x < \delta M_N$ one has

$$\begin{cases} \alpha/N^\nu \left((M_N - x)/N\right)^2 \geq \alpha/N^\nu \left(\underline{\kappa} m\right)^2 (1-\delta)^2, \\ \beta\, x/N \times (M_N - x)/N \geq \beta \underline{\kappa} m (1-\delta). \end{cases}$$

One can therefore construct a coupling such that, on the event $\{T_N(\delta) > t\}$, the relation $X_2^N(t) \geq W_{\underline{a},\underline{b}}(t)$ holds with $\underline{a}=\alpha\underline{\kappa}m^2(1-\delta)^2$ and $\underline{b}=\beta\underline{\kappa}m(1-\delta)$. One obtains the relation $\tau^N_{\underline{a},\underline{b}}(\delta m) \geq_{st} T_N(\delta)$, where $\geq_{st}$ denotes the stochastic order: if U and V are two real valued random variables

$$U \geq_{st} V \text{ if } \mathbb{P}(V \geq x) \leq \mathbb{P}(U \geq x) \qquad \forall x \in \mathbb{R}.$$

Since $X_2^N(t) \leq 2W_{\bar{a},\bar{b}}(t)$, with $\bar{a}=\alpha(\overline{\kappa}m)^2$ and $\bar{b}=\beta\overline{\kappa}m$, one has $\tau^N_{\bar{a},\bar{b}}(\delta m/2) \leq_{st} T^N(\delta)$. One gets therefore

$$\tau^N_{\bar{a},\bar{b}}(\delta m/2) \leq_{st} T^N(\delta) \leq_{st} \tau^N_{\underline{a},\underline{b}}(\delta m). \tag{3.36}$$

Since the constants $\underline{\kappa}$ and $\overline{\kappa}$ can be chosen arbitrarily close to 1, the following proposition has therefore been proved.

Proposition 3.3.2 (Order of Magnitude of Lag Time)**.** *For $\delta > 0$ and $0<\nu\leq 1$,*

$$\lim_{N\to+\infty} \mathbb{P}\left(\frac{\nu}{\beta m} \leq \frac{T_N(\delta)}{N\log N} \leq \frac{\nu}{\beta m(1-\delta)}\right) = 1.$$

Remark. It is very likely that, to reach the state δN, only the second reaction has a real impact as soon as the variable X_2^N is not 0. If true, simple calculations, as in the proof of the above lemma, would then give that the variable $T^N(\delta)/(N\log N)$ is converging in distribution to $\nu/(\beta m)$ as N get large. Note that the limit in this asymptotic result does not depend on δ which suggests a sharp transition for the polymerisation process.

The birth process $(W_{\bar{a},\bar{b}}(t))$ seems to be close to $(X_2^N(t))$ during the initiation of the polymerisation, as the simulations of Figure 3.1. This suggests that, for δ small, the variables $\tau^N_{\bar{a},\bar{b}}(\delta m)$ and $T^N(\delta)$ are very close. We conclude this part by considering the case $\nu > 1$.

A Very Slow Nucleation Step

Now we assume that $\nu > 1$, in this regime, the first reaction, the nucleation step, is then significantly slowed.

Proposition 3.3.3. *For any $\varepsilon > 0$ and $0 < \delta < 1$, there exist $0 < K_1 < K_2$ such that*

$$\liminf_{N\to+\infty} \mathbb{P}\left(K_1 \leq \frac{T_N(\delta)}{N^\nu} \leq K_2\right) \geq 1-\varepsilon.$$

Proof. By using Relation (3.36), it is enough to derive a corresponding limit theorem for $\tau^N_{a,b}(\delta)/N^\nu$ for some $a>0$ and $b>0$. Let (E_x^1) be a sequence of i.i.d. exponential random variables with parameter 1, then

$$\frac{\tau^N_{a,b}(\delta)}{N^\nu} = \sum_{x=0}^{\lfloor \delta N\rfloor} \frac{E_x^1}{a+xbN^{\nu-1}} = \frac{E_0^1}{a} + \sum_{x=1}^{\lfloor \delta N\rfloor} \frac{E_x^1}{a+xbN^{\nu-1}}. \tag{3.37}$$

The expected value of the last term of the right hand side of the above relation is bounded by $K\log(N)/N^{\nu-1}$

for some constant $K>0$. Consequently, this term becomes negligible in distribution for N large. One gets that the variable $\tau_{a,b}^N(\delta)/N^\nu$ converges in distribution to an exponential random variable. The proposition is proved. □

As we have seen in the proof, the only term that matters in the series in Relation (3.37) is the first one: the time to reach one polymerised monomer. It characterises the order of magnitude of the lag time. This variable has been analysed in Szavits-Nossan et al. [Szavits-Nossan et al., 2014] and Yvinec et al. [Yvinec et al., 2016].

In this article, we proposed and analysed two possible extensions of the model studied in Eugène et al. [Eugène et al., 2015].

The first consists in adding a conformation step, whereas the second investigates the influence of a very slow nucleation reaction - so slow that it may be qualitatively viewed as in the same order of magnitude as a power law of the volume N.

As in [Eugène et al., 2015], we analysed the asymptotic behaviours of these models and proved central limit theorems, whose proofs appeared to be significantly more technical, due to the different time scales involved.

Following our study, there remain many open questions, both theoretical - how can we further enrich the models and prove similar asymptotic results - and linked to experiments - which model variants could finally lead to variances in the same order as experimentally observed in Xue et al. [Xue et al., 2008]. These are directions for future work.

References

[Anderson and Kurtz, 2011] Anderson, D. F. and Kurtz, T. G. (2011). Continuous time Markov chain models for chemical reaction networks. In Koeppl, H., Setti, G., di Bernardo, M., and Densmore, D., editors, *Design and Analysis of Biomolecular Circuits*, pages 3–42. Springer New York.

[Ball et al., 1986] Ball, J., Carr, J., and Penrose, O. (1986). The becker-döring cluster equations: basic properties and asymptotic behaviour of solutions. *Communications in mathematical physics*, 104(4):657–692.

[Ball et al., 2006] Ball, K., Kurtz, T. G., Popovic, L., Rempala, G., et al. (2006). Asymptotic analysis of multiscale approximations to reaction networks. *The Annals of Applied Probability*, 16(4):1925–1961.

[Becker and Döring, 1935] Becker, R. and Döring, W. (1935). Kinetische behandlung der keimbildung in übersättigten dämpfern. *Ann. Phys.*, 24:719–752.

[Bieschke et al., 2005] Bieschke, J., Zhang, Q., Powers, E. T., Lerner, R. A., and Kelly, J. W. (2005). Oxidative metabolites accelerate alzheimer?s amyloidogenesis by a two-step mechanism, eliminating the requirement for nucleation. *Biochemistry*, 44:4977–4983.

[Billingsley, 1999] Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition. A Wiley-Interscience Publication.

[Billingsley, 2009] Billingsley, P. (2009). Convergence of probability measures. *John Wiley & Sons, INC*, pages 1–287.

[Bingham, 1991] Bingham, N. H. (1991). Fluctuation theory for the Ehrenfest urn. *Advances in Applied Probability*, 23(3):598–611.

[Bishop and Ferrone, 1984] Bishop, M. F. and Ferrone, F. A. (1984). Kinetics of nucleation-controlled polymerization. a perturbation treatment for use with a secondary pathway. *Biophysical journal*, 46(5):631.

[Bogoliubov, 1961] Bogoliubov, N. N. (1961). The quasi-averages in problems of statistical mechanics. *Technical report, Joint Institute for Nuclear Research.* Preprint D-781.

[Bozaykut et al., 2014] Bozaykut, P., Ozer, N. K., and Karademir, B. (2014). Regulation of protein turnover by heat shock proteins. *Free Radic Biol Med.*, 77:195–209.

[Caughey and Lansbury, 2003] Caughey, B. and Lansbury, P. (2003). Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders. *Annu. Rev. Neurosci.*, 26:267–298.

[Chiti and Dobson, 2006] Chiti, F. and Dobson, C. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75:333–66.

[Cohen et al., 2013] Cohen, S. I., Linse, S., Luheshi, L. M., Hellstrand, E., White, D. A., Rajah, L., Otzen, D. E., Vendruscolo, M., Dobson, C. M., and Knowles, T. P. (2013). Proliferation of amyloid-beta42 aggregates occurs through a secondary nucleation mechanism. *Proc Natl Acad Sci U S A*, 110(24):9758–9763.

[Collins et al., 2004] Collins, S. R., Douglass, A., Vale, R. D., and Weissman, J. S. (2004). Mechanism of prion propagation: Amyloid growth occurs by monomer addition. *PLoS Biol*, 2(10):e321.

[Comets, 1991] Comets, F. (1991). Limites hydrodynamiques. *Séminaire Bourbaki*, 33:167–192.

[Darling and Norris, 2008] Darling, R. and Norris, J. (2008). Differential equation approximations for Markov chains. *Probability Surveys*, 5:37–79.

[Dawson, 1993] Dawson, D. A. (1993). Measure-valued Markov processes. In *École d'Été de Probabilités de Saint-Flour XXI—1991*, volume 1541 of *Lecture Notes in Math.*, pages 1–260. Springer, Berlin.

[Dellacherie and Meyer, 1978] Dellacherie, C. and Meyer, P.-A. (1978). *Probabilities and potential*, volume 29 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam-New York; North-Holland Publishing Co., Amsterdam-New York.

[Desai and Mitchison, 1997] Desai, A. and Mitchison, T. J. (1997). Microtubule polymerization dynamics. *Annual Review of Cell and Developmental Biology*, 13: 83-117.

[Diaconis et al., 1990] Diaconis, P., Graham, R. L., and Morrison, J. A. (1990). Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random Structures Algorithms*, 1(1):51–72.

[Dobson, 2006] Dobson, C. (2006). The generic nature of protein folding and misfolding. In Uversky, V. and Fink, A., editors, *Protein Misfolding, Aggregation, and Conformational Diseases*, volume 4 of *Protein Reviews*, pages 21–41. Springer US.

[Dobson, 2003] Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, 426:884–890.

[Doumic et al., 2016] Doumic, M., Eugène, S., and Robert, P. (2016). Asymptotics of stochastic protein assembly models. Submitted to SIAM Appl. Math.

[Eden et al., 2015] Eden, K., Morris, R., Gillam, J., MacPhee, C. E., and Allen, R. J. (2015). Competition between primary nucleation and autocatalysis in amyloid fibril self-assembly. *Biophysical Journal*, 108:632 – 643.

[Eigen, 1996] Eigen, M. (1996). Prionics or the kinetic basis of prion diseases. *Biophys. Chem.*, 63:A1–A18.

[Ellis, 1987] Ellis, J. (1987). Proteins as molecular chaperones. *Nature*, 328:378–379.

[Ethier and Kurtz, 1986] Ethier, S. N. and Kurtz, T. G. (1986). *MArkov processes: Characterization and convergence*. John Wiley & Sons Inc., New York.

[Eugène et al., 2015] Eugène, S., Xue, W.-F., Robert, P., and Doumic, M. (2015). Insights into the variability of nucleated amyloid polymerization by a minimalistic model of stochastic protein assembly. *Journal of Chemical Physics*, 144.

[Ferrone, 1999] Ferrone, F. (1999). Analysis of protein aggregation kinetics. *Methods Enzymol.*, 309:256–274.

[Ferrone, 2006] Ferrone, F. (2006). Nucleation: the connections between equilibrium and kinetic behavior. *Methods Enzymol.*, 412:285–299.

[Ferrone et al., 1985] Ferrone, F., Hofrichter, J., and Eaton, W. (1985). Kinetics of sickle hemoglobin polymerization. ii. a double nucleation mechanism. *J Mol Biol.*, 183(4):611–31.

[Ferrone et al., 1980] Ferrone, F., Hofrichter, J., Sunshine, H., and Eaton, W. (1980). Kinetic studies on photolysis-induced gelation of sickle cell hemoglobin suggest a new mechanism. *Biophys. J.*, 32:361–377.

[Feuillet et al., 2014] Feuillet, M., Robert, P., et al. (2014). A scaling analysis of a transient stochastic network. *Advances in Applied Probability*, 46(2):516–535.

[Fowler et al., 2007] Fowler, D. M., Koulov, A. V., Balch, W. E., and Kelly, J. W. (2007). Functional amyloid–from bacteria to humans. *Trends Biochem Sci*, 32(5):217–224.

[Freidlin and Wentzell, 1998] Freidlin, M. I. and Wentzell, A. D. (1998). *Random perturbations of dynamical systems*. Springer-Verlag, New York, second edition. Translated from the 1979 Russian original by Joseph Szücs.

[Frieden and Goddette, 1983] Frieden, C. and Goddette, D. (1983). Polymerization of actin and actin-like systems: evaluation of the time course of polymerization in relation to the mechanism. *Biochemistry.*, 22:5836–5843.

[Glenner and Wong, 1984] Glenner, M. G. G. and Wong, C. W. (1984). Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and Biophysical Research Communications*, 120(3):885–890.

[Goto et al., 2005] Goto, Y., Bellotti, V., and Esposito, R. (2005). Dialysis-related amyloidosis: From molecular mechanism to therapies. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics.*, 1753:1–3s.

[Griffith, 1967] Griffith, J. (1967). Self-replication and scrapie. *Nature*, 215(5105):1043–1044.

[G.Roberts, 2016] G.Roberts, R. (2016). Good amyloid, bad amyloid-what's the difference? *PLOS Biology*, 14(1): e1002362.

[Guckenheimer and Holmes, 1990] Guckenheimer, J. and Holmes, P. (1990). Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. *Applied Mathematical Sciences.*

[Harris, 2002] Harris, T. E. (2002). *The theory of branching processes.* Dover Phoenix Editions. Dover Publications, Inc., Mineola, NY. Corrected reprint of the 1963 original [Springer, Berlin; MR0163361 (29 #664)].

[Hasegawa et al., 2002] Hasegawa, K., Ono, K., Yamada, M., and Naiki, H. (2002). Kinetic modeling and determination of reaction constants of alzheimer's beta-amyloid fibril extension and dissociation using surface plasmon resonance. *Biochemistry*, 41:13489–13498.

[Higham, 2008] Higham, D. J. (2008). Modeling and simulating chemical reactions. *SIAM Review*, 50(2):347–368.

[Hofrichter, 1986a] Hofrichter, J. (1986a). Kinetics of sickle hemoglobin polymerization: I. studies using temperature-jump and laser photolysis techniques. *Journal of Molecular Biology*, 189(3):553 – 571.

[Hofrichter, 1986b] Hofrichter, J. (1986b). Kinetics of sickle hemoglobin polymerization iii. nucleation rates determined from stochastic fluctuations in polymerization progress curves. *J. Mol. Biol*, 189:553–571.

[Hofrichter et al., 1974] Hofrichter, J., Ross, P., and Eaton, W. (1974). Kinetics and mechanism of deoxyhemoglobin s gelation: a new approach to understanding sickle cell disease. *Proc Natl Acad Sci USA*, 71:4864–4868.

[Hunt and Kurtz, 1994] Hunt, P. and Kurtz, T. (1994). Large loss networks. *Stochastic Processes and their Applications*, 53:363–378.

[Jarrett and Lansbury, 1993] Jarrett, J. T. and Lansbury, P. T. (1993). Seeding "one-dimensional crystallization" of amyloid : A pathogenic mechanism in alzheimer's disease and scrapie? *Cell*, 73(6):1055–1058.

[J.M. Berg, 2002] J.M. Berg, J.L. Tymoczko, L. S. (2002). Actin is a polar, self-assembling, dynamic polymer. *Biochemistry, 5th edition, W. H. Freeman, New York*, page 958?960.

[Kang et al., 2014] Kang, H.-W., Kurtz, T. G., and Popovic, L. (2014). Central limit theorems and diffusion approximations for multiscale Markov chain models. *The Annals of Applied Probability*, 24(2):721–759.

[Karlin and McGregor, 1965] Karlin, S. and McGregor, J. (1965). Ehrenfest urn models. *Journal of Applied Probability*, 2:352–376.

[Kasahara and Watanabe, 1986] Kasahara, Y. and Watanabe, S. (1986). Limit theorems for point processes and their functionals. *Journal of the Mathematical Society of Japan*, 38(3):543–574.

[Kashchiev and Auer, 2010] Kashchiev, D. and Auer, S. (2010). Nucleation of amyloid fibrils. *J Chem Phys*, 132(21):215101.

[Kayed and Lasagna-Reeves, 2013] Kayed, R. and Lasagna-Reeves, C. (2013). Molecular mechanisms of amyloid oligomers toxicity. *J Alzheimers Dis.*, 33 Suppl 1:S67–78.

[Khasminskii, 1968] Khasminskii, R. (1968). On averaging principle for itô stochastic di??fferential equations. *Kybernetika Chekhoslovakia*, 4:260–279.

[Knowles and Buehler, 2011] Knowles, T. P. and Buehler, M. J. (2011). Nanomechanics of functional and pathological amyloid materials. *Nat Nanotechnol*, 6(8):469–479.

[Knowles et al., 2014a] Knowles, T. P., Vendruscolo, M., and Dobson, C. M. (2014a). The amyloid state and its association with protein misfolding diseases. *Nat Rev Mol Cell Biol*, 15(6):384–396.

[Knowles et al., 2009] Knowles, T. P., Waudby, C. A., Devlin, G. L., Cohen, S. I., Aguzzi, A., Vendruscolo, M., Terentjev, E. M., Welland, M. E., and Dobson, C. M. (2009). An analytical solution to the kinetics of breakable filament assembly. *Science*, 326(5959):1533–1537.

[Knowles et al., 2011] Knowles, T. P., White, D. A., Abate, A. R., Agresti, J. J., Cohen, S. I., Sperling, R. A., Genst, E. J. D., Dobson, C. M., and Weitz, D. A. (2011). Observation of spatial propagation of amyloid assembly from single nuclei. *Proc Natl Acad Sci U S A*, 108(36):14746–14751.

[Knowles et al., 2014b] Knowles, T. P. J., Vendruscolo, M., and Dobson, C. M. (2014b). The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell Biology*, 15:384–396.

[Kurtz, 1992a] Kurtz, T. (1992a). Averaging for martingale problems and stochastic approximation. In *Applied Stochastic Analysis, US-French Workshop*, volume 177 of *Lecture notes in Control and Information sciences*, pages 186–209. Springer Verlag.

[Kurtz, 1992b] Kurtz, T. G. (1992b). Averaging for martingale problems and stochastic approximation. In *Applied Stochastic Analysis*, pages 186–209. Springer.

[Lanneau et al., 2010] Lanneau, D., Wettstein, G., Bonniaud, P., and Garrido, C. (2010). Heat shock proteins: cell protection through protein triage. *ScientificWorldJournal*, 10:1543–1552.

[Ma and Lindquist, 2002] Ma, J. and Lindquist, S. (2002). Conversion of prp to a self-perpetuating prpsc-like confor- mation in the cytosol. *Science*, 298(5599):1785?8.

[McManus et al., 2016] McManus, J. J., Charbonneau, P., Zaccarelli, E., and Asherie, N. (2016). The physics of protein self-assembly. *Preprint*.

[McQuarrie, 1967] McQuarrie, D. A. (1967). Stochastic approach to chemical kinetics. *Journal of applied probability*, 4(3):413–478.

[Michaelis and Menten, 1913] Michaelis, L. and Menten, M. (1913). Die kinetik der invertinwirkung. *Biochem Z*, 49:333–369.

[Morando, 1969] Morando, P. (1969). Mesures aléatoires. *Séminaire de Probabilités de Strasbourg*, III:190–229.

[Morris et al., 2008] Morris, A., Watzky, M., Agar, J., and Finke, R. (2008). Fitting neurological protein aggregation kinetic data via a 2-step minimal/?ockham's razor? model: the finke-watzky mechanism of nucleation followed by autocatalytic surface growth. *Biochemistry*, 47:2413–2427.

[Morris et al., 2009] Morris, A. M., Watzky, M. A., and Finke, R. G. (2009). Protein aggregation kinetics, mechanism, and curve-fitting: A review of the literature. *Biochimica et Biophysica Acta*, 1794(3):375–397.

[Murphy and LeVine, 2010] Murphy, M. P. and LeVine, H. (2010). Alzheimer's disease and the beta-amyloid peptide. *J Alzheimers Dis.*, 19(1): 311.

[Naiki and Gejyo, 1999] Naiki, H. and Gejyo, F. (1999). Kinetic analysis of amyloid fibril formation. *Methods Enzymol.*, 309:305–318.

[Naiki and Nakakuki, 1996] Naiki, H. and Nakakuki, K. (1996). First-order kinetic model of alzheimer's beta-amyloid fibrilextension in vitro. *Lab. Invest.*, 73:374–383.

[Oosawa and Asakura, 1975] Oosawa, F. and Asakura, S. (1975). *Thermodynamics of the polymerisation of protein*. Waltham. MA : Academic Press.

[Oosawa et al., 1959] Oosawa, F., Asakura, S., Hotta, K., and Nobuhisa, I. (1959). G-f transformation of actin as a fibrous condensation. *J. Polym. Sci.*, 37:323–336.

[Oosawa and Kasai, 1962] Oosawa, F. and Kasai, M. (1962). A theory of linear and helical aggregations of macro- molecules. *J. Mol. Biol.*, 4:10–21.

[Ow and Dunstan, 2014] Ow, S.-Y. and Dunstan, D. E. (2014). A brief overview of amyloids and alzheimer's disease. *Protein Science*, 23(10):1315–1331.

[Papanicolaou et al., 1977] Papanicolaou, G. C., Stroock, D., and Varadhan, S. R. S. (1977). Martingale approach to some limit theorems. In *Papers from the Duke Turbulence Conference (Duke Univ., Durham, N.C., 1976), Paper No. 6*, pages ii+120 pp. Duke Univ. Math. Ser., Vol. III. Duke Univ., Durham, N.C.

[Pigolotti et al., 2013] Pigolotti, S., Lizana, L., Otzen, D., and Sneppen, K. (2013). Quality control system response to stochastic growth of amyloid fibrils. {*FEBS*} *Letters*, 587:1405–1410.

[Prigent et al., 2012] Prigent, S., Ballesta, A., Charles, F., Lenuzza, N., Gabriel, P., Tine, L., Rezaei, H., and Doumic, M. (2012). An efficient kinetic model for assemblies of amyloid fibrils and its application to polyglutamine aggregation. *PLoS ONE*, 7(11):e43273.

[Prusiner, 1998] Prusiner, S. B. (1998). Prions. *Proc Natl Acad Sci USA*, 95:13363–13383.

[Ramìrez-Alvarado et al., 2000] Ramìrez-Alvarado, M., Merkel, J. S., and Regan, L. (2000). A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. *Proceedings of the National Academy of Sciences*, 97(16):8979–8984.

[Robert, 2003] Robert, P. (2003). *Stochastic Networks and Queues*. Stochastic Modelling and Applied Probability Series. Springer-Verlag, New York.

[Roberts, 2003] Roberts, C. (2003). Kinetics of irreversible protein aggregation: analysis of extended lumry-eyring models and implications for predicting protein shelf life. *J. Phys. Chem. B*, page 1194?1207.

[Rogers and Williams, 1994] Rogers, L. C. G. and Williams, D. (1994). *Diffusions, Markov processes, and martingales. Vol. 1: Foundations*. John Wiley & Sons Ltd., Chichester, second edition.

[Rogers and Williams, 2000] Rogers, L. C. G. and Williams, D. (2000). *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge. Itô calculus, Reprint of the second (1994) edition.

[Rubinov, 1975] Rubinov, S. (1975). *Introduction to Mathematical Biology*. John Wiley and Sons, New York.

[Rudin, 1987] Rudin, W. (1987). *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition.

[Schwartz and Boles, 2013] Schwartz, K. and Boles, B. R. (2013). Microbial amyloids–functions and interactions within the host. *Curr Opin Microbiol*, 16(1):93–99.

[Serio et al., 2000] Serio, T., Cashikar, A., Kowal, A., Sawicki, G., Moslehi, J., Serpell, L., Arnsdorf, M., and Lindquist, S. (2000). Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science*, 289:1317–1321.

[Shankar et al., 2008] Shankar, G., S. Li, T. M., Garcia-Munoz, A., Shepardson, N., I. Smith, F. B., Farrell, M., M.J. Rowan, C. L., Regan, C., Walsh, D., Sabatini, B., and Selkoe, D. (2008). Amyloid-β protein dimmers isolated directly from alzheimer's brains impair synaptic plasticity and memory. *Nat. Med.*, 14:837?842.

[Sun et al., 2015] Sun, W., Feuillet, M., and Robert, P. (2015). Analysis of large unreliable stochastic networks. *Annals of Applied Probability*. To Appear.

[Sunde et al., 1997] Sunde, M., Serpell, L. C., Bartlam, M., Fraser, P. E., Pepys, M. B., and Blake, C. C. (1997). Common core structure of amyloid fibrils by synchrotron x-ray diffraction. *Journal of molecular biology*, 273(3):729–739.

[Szabo, 1998] Szabo, A. (1998). Fluctuations in the polymerization of sickle hemoglobin a simple analytic model. *J. Mol. Biol*, 199:539–542.

[Szavits-Nossan et al., 2014] Szavits-Nossan, J., Eden, K., Morris, R. J., MacPhee, C. E., Evans, M. R., and Allen, R. J. (2014). Inherent variability in the kinetics of autocatalytic protein self-assembly. *Physical Review Letters*, 113:098101.

[Uversky and Fink, 2004] Uversky, V. and Fink, A. (2004). Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim. Biophys. Acta*, 1698:131–153.

[Uversky et al., 2001] Uversky, V., Li, J., and Fink, A. (2001). Evidence for a partially folded intermediate in alpha-synuclein fibril formation. *J. Biol. Chem*, 276:10737–107744.

[Volmer and Weber, 1926] Volmer, M. and Weber, A. (1926). Keimbildung in übersättigten gebilden. *Z. Physikal. Chemie*, 119:277–301.

[Watzky and Finke, 1997] Watzky, M. A. and Finke, R. G. (1997). Transition metal nanocluster formation kinetic and mechanistic studies. a new mechanism when hydrogen is the reductant: slow, continuous nucleation and fast autocatalytic surface growth. *Journal of the American Chemical Society*, 119(43):10382–10400.

[Wegner and Engel, 1975] Wegner, A. and Engel, J. (1975). Kinetics of the cooperative association of actin to actin filaments. *Biophys. Chem.*, 3:215?225.

[Wegner and Savko, 1982] Wegner, A. and Savko, P. (1982). Fragmentation of actin filaments. *Biochemistry.*, 21(8):1909–13.

[Wright et al., 2015] Wright, M. A., Aprile, F. A., Arosio, P., Vendruscolo, M., Dobson, C. M., and Knowles, T. P. J. (2015). Biophysical approaches for the study of interactions between molecular chaperones and protein aggregates. *Chem. Commun.*, 51:e43273.

[Xue et al., 2008] Xue, W.-F., Homans, S. W., and Radford, S. E. (2008). Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *PNAS*, 105:8926–8931.

[Xue and Radford, 2013] Xue, W.-F. and Radford, S. E. (2013). An imaging and systems modeling approach to fibril breakage enables prediction of amyloid behavior. *Biophys J*, 105(12):2811–2819.

[Yvinec, 2012] Yvinec, R. (2012). *Probabilistic modeling in cellular and molecular biology.* PhD thesis, Université Claude Bernard - Lyon I.

[Yvinec et al., 2016] Yvinec, R., Bernard, S., Hingant, E., and Pujo-Menjouet, L. (2016). First passage times in homogeneous nucleation: Dependence on the total number of particles. *The Journal of Chemical Physics*, 144:034106.

Part II

Stochastic modelling of telomere shortening

Chapter 4

An introduction to telomere shortening

Contents

Most of the time, cells do not divide indefinitely; if they do however, we are in the presence of a potentially pathological phenomenon most commonly known as cancer. The limitation to cell division is generally attributed to the shortening of telomeres, a process that we modelled using stochastic methods. The present chapter offers biological justifications for the particular mathematical model that will be introduced in the next chapter.

4.1 Biology of telomere shortening

We start by giving some useful definitions, before explaining why and how telomere shortening occurs.

4.1.1 Telomeres, replicative senescence: definitions

Telomere. Telomeres are extremities of linear chromosomes in eukaryotic cells. They are essential to genomic stability, as they prevent the ends of chromosomes from being mistaken for DNA breakage. In humans for instance, approximately $10,000$ accidental DNA breaks occur daily [Wellinger and Zakian, 2012a] and must be immediately repaired. For telomeres, it is quite the contrary; they must not be fused with other ends to maintain genome stability. In addition, they play another key role in the cell's fate: at each replication round, they are shortened because the DNA-replication machinery is not able to replicate the extremities (see section 4.1.2). This is why they are also often considered as molecular clocks counting the number of cell divisions.

Replicative senescence. When telomeres become too short because of the so-called telomere 'end-replication problem' [Watson, 1972, Olovnikov, 1973], the cell cannot divide anymore: this state is called replicative senescence. It has been shown in [Abdallah et al., 2009a, Hemann et al., 2001a] that the introduction of a single very short telomere was enough to arrest cell divisions, suggesting that the shortest telomere was triggering senescence. This result allows for a mathematical definition of the time of senescence. In the model presented in the next chapter, we study telomere shortening in *Saccharomyces cerevisiae* which has 16 chromosomes, *i.e* 32 telomeres, and therefore the number of generations before senescence (the time of senescence) can be defined as the random variable T such that

$$T = \inf\left\{n \geq 0, \min(L_n^1, \cdots, L_n^{32}) < S\right\} \tag{4.1}$$

where $(L_n^1, \cdots, L_n^{32})$ is the vector of the lengths of the 32 telomeres of the cell and S is the threshold of senescence.

Telomerase. Most eukaryotic organisms have a specific polymerase called telomerase able to elongate telomeres [Blackburn and Collins, 2011]. Telomerase is a reverse transcriptase, *i.e* an enzyme that creates single-stranded DNA from single-stranded RNA [Autexier and Lue, 2006]. It has been shown in [Teixeira et al., 2004] that telomerase adds new telomere sequences preferentially to shorter telomeres. Hence,

when telomerase is expressed, telomere length is maintained via an equilibrium between the shortening due to the DNA replication machinery and the elongation by telomerase. However, in some multicellular eukaryotes including man, somatic cells inhibit the expression of telomerase so that telomeres are only shortened until replicative senescence [Hayflick, 1965].

4.1.2 The telomere end-replication problem

In this section we explain in more details where the 'telomere end-replication problem' comes from.

DNA replication. DNA replication occurs via a complex DNA replication machinery called the replisome. At each cell division, according to the semiconservative replication mode, the double-stranded DNA opens at several origins generating different replication forks growing simultaneously in two opposite directions (figure 4.1). Each parental strand acts as a template for the synthesis of the new strands.

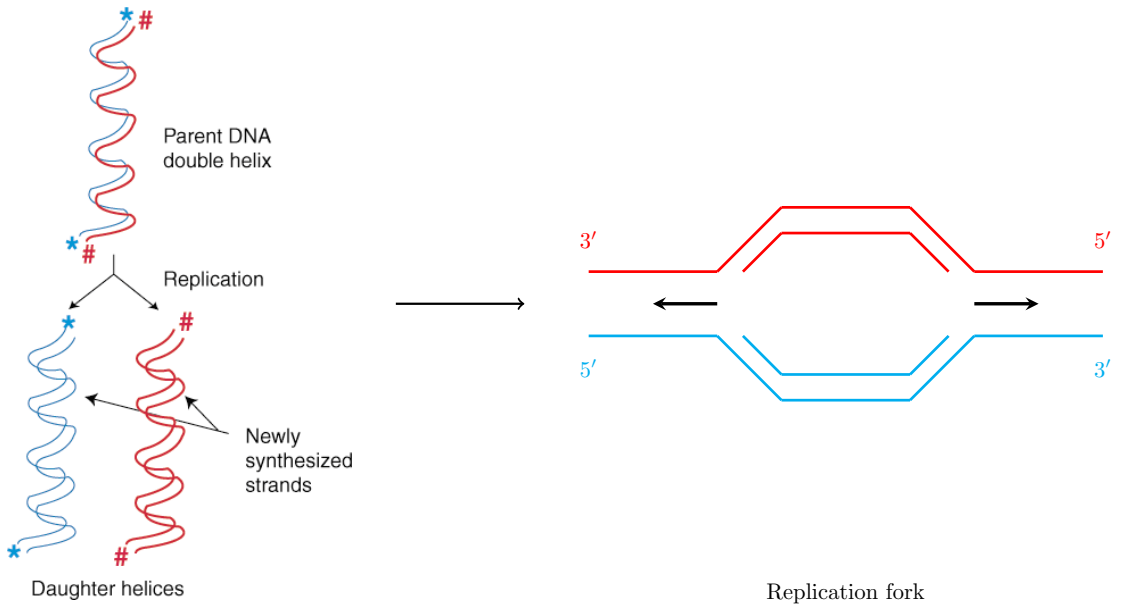


Figure 4.1: Semiconservative mechanism of DNA replication and generation of a replication fork. The left hand side picture is taken from sparknotes.com/biology/molecular/dnareplicationandrepair/section1.rhtml

The 'end replication problem'. Hence, each strand is copied by the replication machinery. However, these strands are oriented from an extremity called $5'$ to the other called $3'$, and are antiparallel.

Since the replisome can only replicate DNA from the extremity 5' to 3', one strand is created continuously (the leading strand) while the other (the lagging strand) is discontinuously replicated into short bits called Okazaki fragments [Okazaki et al., 1968]. Each Okazaki fragment is initiated by a small segment of RNA (blacks fragments in figure 4.2).

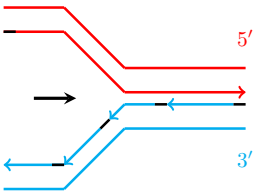


Figure 4.2: Schematic representation of the synthesis of Okazaki fragments. For simplicity, we do not represent the single-stranded overhang in the parental chromosome.

The last step consists in removing these RNA primers and replacing them by DNA except the very last one at the terminus of the chromosome that cannot be replaced by the replication machinery. As a result, the newly synthesised strand via the lagging machinery is shorter than its corresponding parental strand, as shown in figure 4.3.

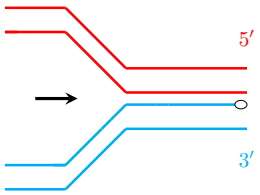


Figure 4.3: Schematic representation of the loss of nucleotides at the chromosome extremity after replication.

As a result, the lagging machinery generates single-stranded overhang while the leading machinery creates a blunt end [Lingner et al., 1995]. In reality, the parental chromosome also has this overhang structure (the previous schemes 4.2 and 4.3 were simplified). Therefore, the length of the template for lagging strand synthesis is actually the same as the parental length. Importantly, on the leading strand, in most organisms (except angiosperm plants [Riha et al., 2000]), additional maturation steps involving resection and fill-in also recreate the overhang structure [Faure et al., 2010, Larrivée et al., 2004].

In conclusion, the mechanism of replication generates two daughter chromosomes having both the same length and single-stranded extremities. This symmetry allows us to focus on a lineage of cells in the next chapter, *i.e* we consider at each generation only one of the daughters; then, exactly one of the extremities has been shortened after replication. We conclude this section by the global picture of telomere shortening shown in picture 4.4, also illustrating our model of telomere shortening in the following chapter, that summarises the previous considerations.

4.1.3 Ageing and Cancer

The problem of telomere shortening has been sparking more and more interest amongst researchers, as it relates to two fundamental problems in biology. Indeed, the shortening mechanism has naturally been associated with ageing, but also with cancer. In fact, the limited number of cell divisions due to the finite length of telomeres is often seen as a barrier against cancer. Hence, cancer cells have exceeded

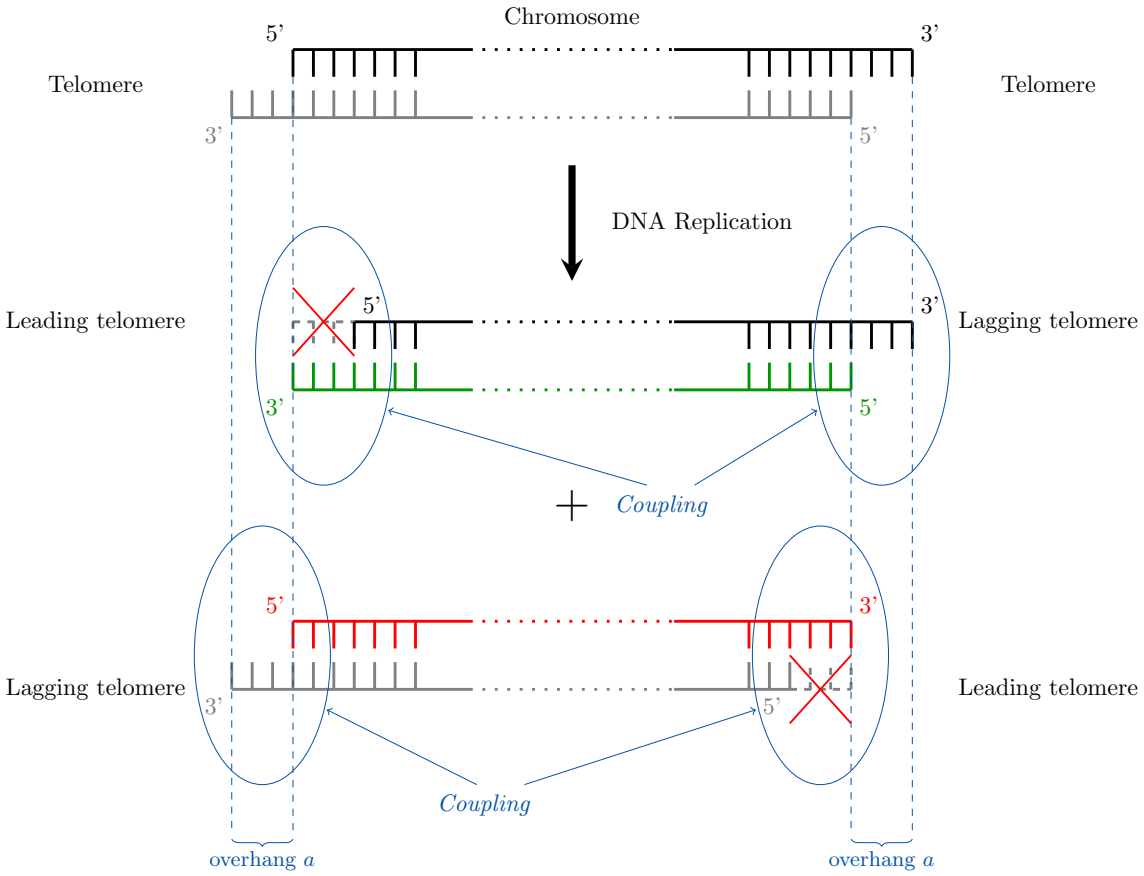


Figure 4.4: Mechanism of DNA replication and telomere shortening. Following one lineage, exactly one extremity of the chromosome is shortened after replication (coupling).

the number of divisions allowed and managed to elongate their telomeres either by telomerase reactivation or by alternative mechanisms not yet entirely identified, but most probably including homologous recombination [de Jesus and Blasco, 2013].

4.2 Some mathematical models of telomere shortening

Telomere-shortening dynamics are an ideal framework for mathematical modelling.. For instance, in 1995, Arino *et al.* [Arino et al., 1995] proposed a mathematical modelling for telomere shortening based on the mechanism proposed in [Levy et al., 1992], shown in figure 4.5.
This model considers that a parental chromosome has a blunt end at one extremity, while the other extremity is single-stranded. It does not take into account the resection recreating the overhang structure after replication observed experimentally in [Faure et al., 2010, Larrivée et al., 2004] for instance. Therefore the two possible asymmetric initial configurations of parental chromosomes must be considered. Moreover, the two daughter cells inherit chromosomes of different lengths. This premise allows

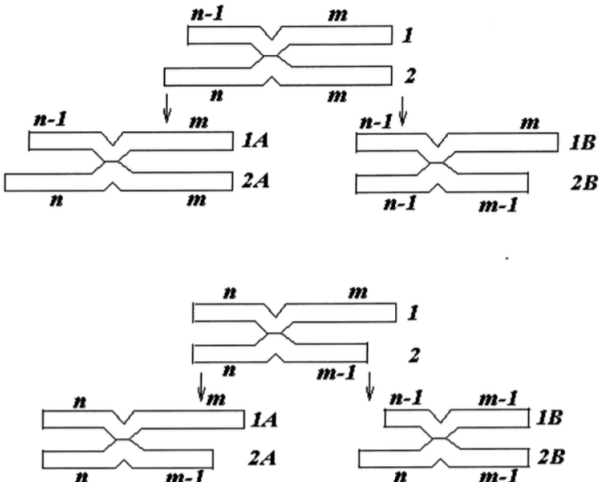


Figure 4.5: Transition rules for telomere shortening in [Levy et al., 1992]. Retrieved from [Arino et al., 1995].

the author to derive quantitative behaviour of the whole cell population by considering at first that cell divisions occur synchronously. Then, in a more realistic version, branching processes are used, in which cell lifetimes are independent identically distributed (i.i.d) random variables. This model was later extended in [Olofsson and Kimmel, 1999] by including cell death in population dynamics. Further, other stochastic growth models of cell populations including telomere loss have been proposed in [Dyson et al., 2007, Portugal et al., 2008]. The latter, for instance, considers that the probability for a cell to divide decreases as telomeres shorten. Another interesting phenomenon occurring in telomere dynamics is the emergence of 'survivors', defined in mutant yeast cells that do not express telomerase but are still able to maintain their telomeres through recombination-based mechanisms [Lundblad and Blackburn, 1993]; these survivors are considered as valuable experimental model for a subset of cancer cells that also rely on recombination, the so-called 'ALT cancer cells'. The emergence of survivors has been mathematically studied for example in [Olofsson and Bertuch, 2010].

4.3 Presentation of chapter V

The subsequent chapter, chapter V, corresponds to the following article:

Chapter V. S. Eugène, T. Bourgeron and Z. Xu. Effects of initial telomere length distribution on senescence onset and heterogeneity. Submitted to *Phys. Rev. Letters*, 2016.

This paper is available on: http://arxiv.org/abs/1606.06842.

In chapter V, we propose a stochastic model for telomere shortening in *Saccharomyces cerevisiae*, based on the mechanism shown in figure 4.4. Here, the two daughter chromosomes have the same length. Therefore, we study at each generation only one of these two daughters, *i.e* we study one lineage and

do not take into account the size of cell population. From then, we propose a mechanism including two phases.

First, telomerase is activated, and telomere length distribution achieves an equilibrium due to the contribution of both the shortening and the elongation. We study the discrete-time Markov chain associated to the length distribution of one telomere $(L_n)_n$, where L_n is the length of a given telomere at generation n.
In a second part, telomeres are initially distributed according to the previous equilibrium and then telomerase is inhibited: telomeres can only shorten until the cell enters replicative senescence. We study the impact of the initial distribution on the number of generations before senescence, called time of senescence and denoted T in this thesis (the stopping time T was already defined at the beginning of this chapter by equation (4.1)). We separately study the effect of the initial mean length of telomeres and of its initial variance.

Chapter 5

Impact of the initial telomere distribution on the onset of senescence

Contents

5.1 Introduction

Telomeres, the ends of eukaryote chromosomes, are poised in a dynamic equilibrium controlled by two processes: limited telomere shortening and elongation by telomerase, a dedicated holoenzyme able to generate *de novo* telomere sequence by reverse-transcribing a template RNA. When telomerase is not expressed, as in human somatic cells, or is mutated, as done experimentally in model organisms such as *Saccharomyces cerevisiae*, [Lundblad and Szostak, 1989], telomeres can only shorten and after many divisions the cell enters replicative senescence, a permanent cell cycle arrest induced by short telomeres that elicits a DNA damage response. Replicative senescence is implicated in organismal ageing and is a potent barrier to cancer emergence, but its remarkable asynchrony and heterogeneity remain a challenge for investigating the exact relationship between initial telomere length distribution and senescence onset. Telomere shortening is the unavoidable consequence of the end replication problem, [Olovnikov, 1973, Watson, 1972, Soudet et al., 2014]. In most examined species, telomeres end with an 5' to 3' singled-stranded DNA overhang (Fig. 5.1), [Hemann and Greider, 1999, Henderson and Blackburn, 1989, Klobutcher et al., 1981, Makarov et al., 1997, McElligott and Wellinger, 1997, Raices et al., 2008, Riha et al., 2000, Wellinger et al., 1993]. When the replication fork reaches the end of the chromosome, the removal of the last Okazaki fragment leaves a gap at the lagging strand, which recreates the single-stranded overhang of the parental telomere (see Fig. 5.1).On the leading strand, after replication, complex maturation steps involving resection and fill-in also regenerate the overhang structure, [Larrivée et al., 2004, Faure et al., 2010, Chai et al., 2006, Wu et al., 2012, Soudet et al., 2014]. Regardless of these maturation steps, the leading strand template for replication is shorter than the lagging strand one, thus generating after replication two new telomeres of different lengths, one unchanged compared to the parental telomere and the other shorter by exactly the length of the overhang, as illustrated in Fig. 5.2. Previous mathematical models of telomere shortening also based on the end-replication problem [Arino et al., 1995, Olofsson and Kimmel, 1999, Arkus, 2005, Levy et al., 1992] did not consider the maturation of the leading strand telomere that generates a 3'-end overhang identical to the one on the lagging strand. This maturation step is widely conserved throughout species with the notable exception of angiosperm plants that display a blunt end at the leading telomere [Riha et al., 2000]. We also note that other mathematical models examined higher level structures such as t-loops [Griffith et al., 1999, Rodriguez-Brenes and Peskin, 2010], or additional telomere states or breaking mechanisms [Kowald, 1997, Rubelj and Vondracek, 1999, Proctor and Kirkwood, 2002, Proctor and Kirkwood, 2003].

On average, each telomere has shortened by a fixed length, which is exactly half of the overhang length. As a corollary, experimental measurements of telomere length, which average over a large subset of telomeres and over a great number of cells (typically $10^7 - 10^8$), should display a constant decrease in telomere length when telomerase is absent and when neglecting other possible shortening mechanisms such as telomere break. This is the case for *S. cerevisiae*, for which telomere shortening in the absence of telomerase was measured to be around 3-4 bp (base pairs) per generation [Marcand et al., 1999], in

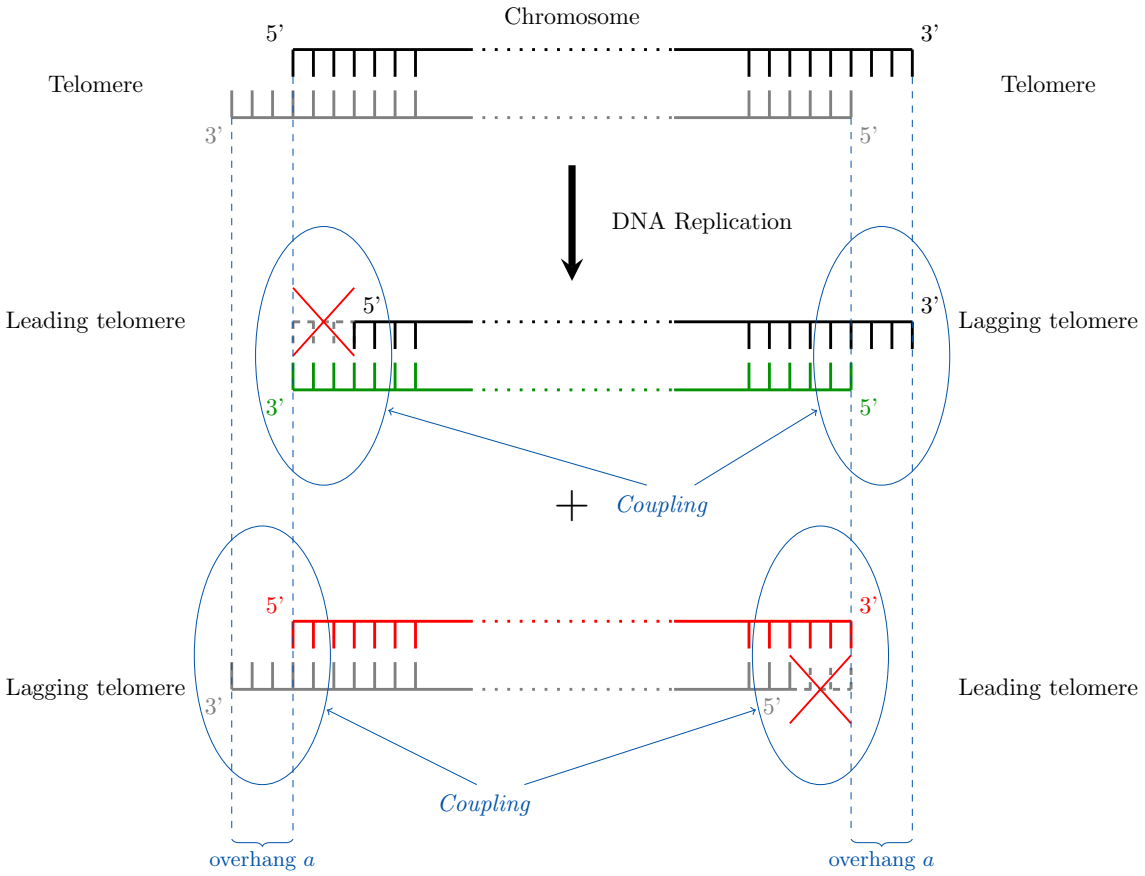


Figure 5.1: Mechanism of DNA replication and Telomere Shortening. Following one lineage, exactly one extremity of the chromosome is shortened after replication (coupling).

agreement with the measured length of the overhang (5-10 nucleotides, [Soudet et al., 2014]). However, while the average dynamics of telomere length is informative of the global regulation and homeostasis of telomere length, it misses important contributions of the asymmetry of telomere replication mechanism to the overall telomere length distribution and to the heterogeneity of the onset of replicative senescence. Taking this asymmetry into account, the shortening of a given telomere in a cell lineage, defined as a random succession of mitotically related cells, [Xu et al., 2015], is not constant and deterministic, but rather probabilistic and follows a Bernoulli process. More precisely, for a given chromosome, after a round of replication, one extremity has been shortened, while the other kept its parental length. Using mathematical modelling and simulations fitted to experimental data, it has previously been showed that the asymmetry of telomere replication *per se* is responsible for variations in the timing of senescence, illustrating the biological relevance of such a mechanism [Bourgeron et al., 2015]. Additionally, if the two ends of a given chromosome are considered together, the 3'-end at one telomere belongs to the same DNA strand as the 5'-end on the other telomere of the same chromosome, implying that the asymmetry of the shortening mechanism during replication at one telomere is inverted compared to the other (see Fig. 5.1). This coupling mechanism between the two ends of the same chromosome will also be included

here.

In this chapter, we study the consequences of the asymmetry and the coupling on the distribution and the dynamics of telomere length in two distinct phases: at steady state in the presence of telomerase and in a strictly shortening regime without telomerase, *i.e.*, senescence. We show that the robustness of telomerase recruitment impacts on the variance of the steady-state distribution of telomere length. In turn, this variance defines different regimes of senescence. In a regime of low initial variance, we find that senescence onset cannot be linearly inferred from the average telomere length or even the length of the shortest telomere. In contrast, a high variance implies a linear correlation between the initial shortest telomere and senescence onset. The ultimate goal is to study the impact of the initial distribution on the onset of senescence and so, in further studies, from measurements of times of senescence for different lineages [Xu et al., 2015], go back to the features of the equilibrium state of telomeres evolving with telomerase.

Hence, our strategy consists in studying separately these two phases. In the first section, we study telomeres being repaired by telomerase and show that the length distribution of telomeres achieves an equilibrium distribution denoted by L_∞ that we explicitly characterise. We also identify the biological mechanisms that affect this distribution. In the second section, we consider a cell of 16 chromosomes, *i.e* 32 telomeres, initially distributed as the previous equilibrium distribution. The telomerase is then inhibited so that telomeres can only be shortened, until the shortest telomere goes below the threshold of senescence and the cell doesn't replicate anymore. Here, we investigate separately the impact of the initial mean and variance on the onset of senescence. For this purpose, we start by considering that all telomeres have initially the same deterministic length $\mathbb{E}(L_\infty)$ and derive an asymptotic expansion of the expected time of senescence. Then, the initial telomere length distribution is chosen to be uniform, centred on $\mathbb{E}(L_\infty)$, in order to study the impact of the initial variance on the time of senescence.

5.2 Telomeres evolving with telomerase

We first describe the most general model, corresponding to a lineage of haploid yeast cells dividing in the presence of active telomerase. Let $(L_n^1, L_n^2) \in \mathbb{N}^2$ denote the lengths of the two extremities of a given chromosome at generation n ('length' stands here for the number of base pairs so that all our processes lie in $\mathbb{N}$). Then, at generation $n+1$:

— exactly one of the extremities is shortened by a length denoted a. The coupling is captured by a Bernoulli random variable B_n with parameter $1/2$: if $B_n = 1$, then L_n^1 is shortened by a nucleotides, whereas L_n^2 is preserved, and conversely if $B_n = 0$.

— telomerase adds new telomere sequences preferentially to shorter telomeres, [Teixeira et al., 2004, Britt-Compton et al., 2009], a behavior that we capture by introducing for each extremity C_n^i, a Bernoulli random variable with parameter $f(L_n^i)$, $(i \in \{1,2\})$, as it is done in [Xu et al., 2013],

where f has the shape shown in Fig. 5.2 **(a)**,

$$\begin{cases} f(l) &= 1 \text{ if } l \leq L_s \\ f(l) &= \dfrac{1}{1+\beta(l-L_s)} \text{ if } l > L_s \end{cases} \tag{5.1}$$

with β and L_s two constants in $\mathbb{R} \times \mathbb{N}$, and L_n^i is the length of the telomere before replication. If $\mathrm{C}_n^i = 1$, telomerase is recruited to telomere L_n^i and elongates it, whereas if $\mathrm{C}_n^i = 0$, the telomere is not elongated and only shortens or stays at the same length. The shape of f is such that below a length threshold L_s, the Bernoulli random variable equals 1 that is telomerase is always active. For a telomere longer than L_s, the probability of $\mathrm{C}_n^i = 1$ decreases to zero, meaning that the longer the telomere, the less likely it is to be elongated by telomerase.

— the number of nucleotides added by telomerase is independent of the length of the telomere, [Teixeira et al., 2004, Xu et al., 2013]. We introduce $\mathcal{G}_n^1$ and $\mathcal{G}_n^2$ two independent geometric random variables of parameter p, independent of all the other quantities (including L_n^1, L_n^2), which correspond to the number of nucleotides added by telomerase to each telomere.

As a result, for any given chromosome the evolution of the $\mathbb{N}^2$-valued process (L_n^1, L_n^2) can be described as follows:

$$\begin{pmatrix} L_{n+1}^1 \\ \\ L_{n+1}^2 \end{pmatrix} = \begin{pmatrix} (L_n^1 - a \cdot B_n)^+ + \mathrm{C}_n^1 \cdot \mathcal{G}_n^1 \\ \\ (L_n^2 - a \cdot (1 - B_n))^+ + \mathrm{C}_n^2 \cdot \mathcal{G}_n^2 \end{pmatrix} \tag{5.2}$$

where $x^+ = \max(0, x)$.

The goal of this section is to investigate the properties of the steady state of telomere length distribution. For simplicity, we focus on the behaviour of one telomere, and consider the projection of the first coordinate of a chromosome in order to compute its equilibrium. We will analyse the coupling effect in more depth in the second regime without telomerase. Our model thus becomes:

$$L_{n+1} = (L_n - a \cdot B_n)^+ + \mathrm{C}_n \cdot \mathcal{G}_n \tag{5.3}$$

where L_n is the length of a given telomere at generation n, L_{n+1} the length of one of the two daughter telomeres (at generation $n+1$), $\mathcal{G}_n$ a geometric random variable of parameter $p \in (0,1)$; we also recall its generating function which will be useful later:

$$\mathbb{E}(u^{\mathcal{G}_n}) = \sum_{k=0}^{\infty} u^k \mathbb{P}(\mathcal{G}_n = k) = \sum_{k=0}^{\infty} u^k p(1-p)^k = \frac{p}{1-u(1-p)}.$$

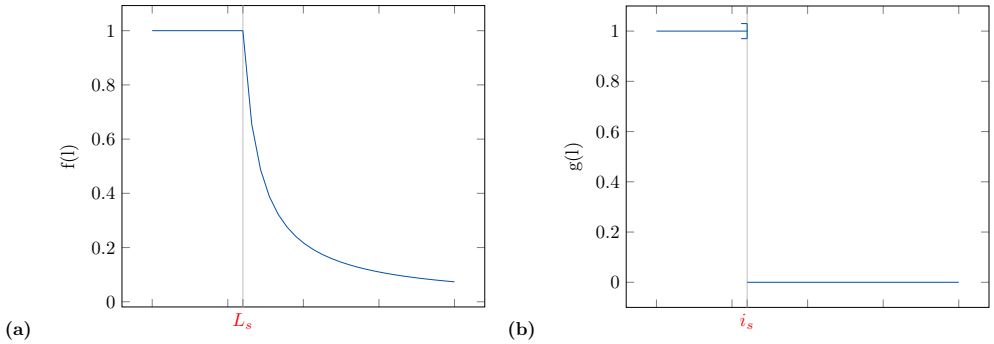
Equation (5.3) corresponds to Lindley's Equation in the queuing networks field (see for instance [Robert,

2003] p.332). An averaged version of this model has been studied in [Xu et al., 2013, Dao Duc and Holcman, 2013], where instead of being stochastic, telomere shortening was chosen to be deterministic with a constant value of $a/2$ (*i.e* $L_{n+1} = L_n - a/2 + \mathrm{C}_n \cdot \mathcal{G}_n$). Finally, to make our computations fully explicit without betraying the principles of the biological mechanism, instead of f, we consider g a sharp threshold at a value i_s (Fig. 5.2 **(b)**)

$$\begin{cases} g(l) &= 1 \text{ if } l \leq i_s \\ g(l) &= 0 \text{ if } l > i_s \end{cases} \tag{5.4}$$

with i_s a constant in $\mathbb{N}$, not necessarily equal to L_s. Our model becomes:

$$L_{n+1} = (L_n - a \cdot B_n)^+ + \mathcal{G}_n \cdot \mathbb{1}_{\{L_n \leq i_s\}}. \tag{5.5}$$



(a) (b)

Figure 5.2: **(a)**: $f(l) = 1$ if $l \leq L_s$, which means that telomeres shorter than L_s are always elongated. For $l > L_s$, $f(l) = 1/(1 + \beta(l - L_s))$. **(b)**: sharp threshold at i_s.

Proposition 5.2.1. *The Markov chain (L_n) defined by (5.5) is aperiodic, irreducible and ergodic. Therefore, it has a unique equilibrium distribution, denoted L_∞, the distribution of which being given by:*

$$(1 + ... + u^{a-1})\mathbb{E}(u^{L_\infty}) = \frac{(1-p)(1+u^a)}{1-u(1-p)} \sum_{k=0}^{i_s} u^k \pi_k + \frac{p}{1-u(1-p)} \sum_{k=0}^{a-1} u^k (1 + u + ... + u^{a-k-1})\pi_k \tag{5.6}$$

where $\pi_k = \mathbb{P}(L_\infty = k)$.

Note that equation (5.6) shows that it is enough to determine the initial vector $(\pi_0, ..., \pi_{i_s})$ to get the whole equilibrium distribution of the size of one telomere in telomerase-positive cells.

Proof. To fully identify our Markov chain, we first write the transition probabilities. Let $p_{ij} = \mathbb{P}(L_{n+1} = j | L_n = i)$ and $p_{ij}^{(n)} = \mathbb{P}(L_n = j | L_0 = i)$. Then:

$$\begin{aligned} &\text{— if } i > i_s,\ p_{ii} = \frac{1}{2} \text{ and } p_{i,i-a} = \frac{1}{2}. \\ &\text{— if } i < a,\ p_{ij} = \frac{1}{2}\mathbb{P}(\mathcal{G}_n = j) + \frac{1}{2}\mathbb{P}(\mathcal{G}_n = j - i) \text{ for } j \geq i, \\ &\qquad \text{and } p_{ij} = \frac{1}{2}\mathbb{P}(\mathcal{G}_n = j) \text{ for } 0 \leq j < i. \end{aligned}$$

— if $a \leq i \leq i_s$, $p_{ij} = 0$ for $j < i - a$,

$$p_{ij} = \frac{1}{2}\mathbb{P}(\mathcal{G}_n - a = j - i) \text{ for } i > j \geq i - a,$$

$$\text{and } p_{ij} = \frac{1}{2}\mathbb{P}(\mathcal{G}_n - a = j - i) + \frac{1}{2}\mathbb{P}(\mathcal{G}_n = j - i) \text{ for } j \geq i.$$

Let us fix $i, j \in \mathbb{N}$. If $i < a$, then $p_{ij} > 0$ for all $j \in \mathbb{N}$. If $i \geq a$, then we consider the event to be shortened at each generation until you go below a, which has a probability greater than $p^{(k)}_{i,i-ka} > 0$ with $k = \lfloor i/a \rfloor$, and then you can reach the state j with the geometric random variable. Irreductibility follows. It is aperiodic because for instance, $p^{(2)}_{ii} > 0$ and $p^{(3)}_{ii} > 0$ for all $i \in \mathbb{N}$.
The ergodicity is a direct consequence of Foster-Lyapunov criteria (see Corollary 8.7 p.214 of [Robert, 2003]). Let $\mathbb{E}_x$ be the expectation of the Markov chain $(L_n)_n$ starting at x. Then:

— $\mathbb{E}_x(L_1 - x) = -a/2$ for $x > i_s$.

— $\mathbb{E}_x(L_1) = x - a/2 + (1-p)/p < \infty$ for $x \leq i_s$.

Let L_∞ be the stationary distribution of $(L_n)_n$, with $\pi_k = \mathbb{P}(L_\infty = k)$. We have:

$$\lim_{n\to\infty} \mathbb{P}(L_n = k) = \pi_k$$

and:

$$L_0 \overset{\mathcal{L}}{=} L_\infty \implies L_1 \overset{\mathcal{L}}{=} L_\infty$$

so that L_∞ satisfies:

$$L_\infty \overset{\mathcal{L}}{=} (L_\infty - a.B_0)^+ + \mathcal{G}_0 \mathbb{1}_{\{L_\infty \leq i_s\}}.$$

and its Laplace transform can be computed by distinguishing three cases for L_∞,$L_\infty < a$, $a \leq L_\infty \leq i_s$ and $L_\infty > i_s$:

$$\begin{aligned}
\mathbb{E}\left(u^{L_\infty}\right) &= \mathbb{E}\left(u^{(L_\infty - a.B)^+ + \mathcal{G}_n \mathbb{1}_{\{L_\infty \leq i_s\}}}\right) \\
&= \mathbb{E}\left(u^{\mathcal{G}_0}.\mathbb{1}_{\{L_\infty < a.B\}}\right) + \mathbb{E}\left(u^{L_\infty - a.B + \mathcal{G}_0}.\mathbb{1}_{\{a.B \leq L_\infty \leq i_s\}}\right) + \mathbb{E}\left(u^{L_\infty - a.B}.\mathbb{1}_{\{L_\infty > i_s\}}\right) \\
&= \frac{1}{2}\mathbb{E}\left(u^{\mathcal{G}_0}\right)\left[\mathbb{P}(L_\infty < a) + \mathbb{P}(L_\infty < 0)\right] + \frac{1}{2}\left[\mathbb{E}\left(u^{L_\infty - a + \mathcal{G}_0}.\mathbb{1}_{\{a \leq L_\infty \leq i_s\}}\right)\right. \\
&\left. + \mathbb{E}\left(u^{L_\infty + \mathcal{G}_0}.\mathbb{1}_{\{0 \leq L_\infty \leq i_s\}}\right)\right] + \frac{1}{2}\left[\mathbb{E}\left(u^{L_\infty - a}.\mathbb{1}_{\{L_\infty > i_s\}}\right) + \mathbb{E}\left(u^{L_\infty}.\mathbb{1}_{\{L_\infty > i_s)\}}\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left(u^{\mathcal{G}_0}\right)\sum_{k=0}^{a-1}\pi_k + \frac{1}{2}\mathbb{E}\left(u^{\mathcal{G}_0}\right)\left[\frac{1}{u^a}\sum_{k=a}^{i_s}u^k\pi_k + \sum_{k=0}^{i_s}u^k\pi_k\right] \\
&\qquad + \frac{1}{2}\left(1 + \frac{1}{u^a}\right)\left[\mathbb{E}\left(u^{L_\infty}\right) - \sum_{k=0}^{i_s}u^k\pi_k\right] \\
&= \frac{1}{2}\left(1 + \frac{1}{u^a}\right)\mathbb{E}\left(u^{L_\infty}\right) + \frac{1}{2}\left(1 + \frac{1}{u^a}\right)\left[\mathbb{E}\left(u^{\mathcal{G}_0}\right) - 1\right]\sum_{k=1}^{i_s}u^k\pi_k + \frac{1}{2}\mathbb{E}\left(u^{\mathcal{G}_0}\right)\sum_{k=0}^{a-1}\pi_k\left(1 - \frac{u^k}{u^a}\right)
\end{aligned}$$

And therefore,

$$(u^a-1)\mathbb{E}\left(u^{L_\infty}\right) = (1-p)(u^a+1)\left(\sum_{k=0}^{i_s} u^k\pi_k\right)\left(\frac{u-1}{1-u(1-p)}\right) + \frac{p}{1-u(1-p)}\sum_{k=0}^{a-1}(u^a-u^k)\pi_k$$

Simplifying by $u-1$, we get expression (5.6). □

5.2.1 Unit shortening, $a = 1$

We start by the simpler case, $a = 1$, where the expressions of the steady state are explicit.

Proposition 5.2.2. *When $a = 1$, the equilibrium distribution $(\pi_k)_{0\le k<\infty}$ of (L_n) is given by:*

$$\begin{aligned} \forall k \in [\![1, i_s]\!],\ \pi_k &= \left(\frac{2(1-p)}{p}\right)^k \pi_0 \\ \forall k > i_s,\ \pi_k &= p(1-p)^k\left(\frac{2}{p}\right)^{i_s+1}\pi_0. \end{aligned} \tag{5.7}$$

Proof. When $a = 1$, equation (5.6) becomes:

$$\begin{aligned} \mathbb{E}(u^{L_\infty}) &= \frac{(1-p)(1+u)}{1-u(1-p)}\sum_{k=0}^{i_s} u^k\pi_k + \frac{p}{1-u(1-p)}\pi_0 \\ &= (1-p)(1+u)\sum_{k=0}^{\infty} u^k(1-p)^k\sum_{k=0}^{i_s} u^k\pi_k + p\pi_0\sum_{k=0}^{\infty} u^k(1-p)^k \end{aligned} \tag{5.8}$$

To find the initial vector $(\pi_0, ..., \pi_{i_s})$, we just identify the coefficients of the power series on both sides of equation (5.8):

$$\sum_{k=0}^{\infty} u^k\pi_k = (1-p)\left[\sum_{k=0}^{\infty}\left\{\sum_{l=0}^{k\wedge i_s}\pi_l(1-p)^{k-l}\right\}u^k + \sum_{k=1}^{\infty}\left\{\sum_{l=0}^{(k-1)\wedge i_s}\pi_l(1-p)^{k-1-l}\right\}u^k\right] + p\pi_0\sum_{k=0}^{\infty}u^k(1-p)^k$$

Fix π_0 (it will be determined by the normalisation condition). The identification of the coefficients of equation (5.8) gives us that for $1 \le k \le i_s$:

$$p\pi_k = (2-p)\sum_{l=0}^{k-1}\pi_l(1-p)^{k-l} + p\pi_0(1-p)^k.$$

By induction, we find:

$$\forall 1 \le k \le i_s, \pi_k = \left(\frac{2(1-p)}{p}\right)^k\pi_0$$

We can now find all the π_k. Let $k \geq i_s + 1$. By identification of the coefficients in (5.8),

$$\pi_k = (2-p)\pi_0 \sum_{l=0}^{i_s} \left(\frac{2(1-p)}{p}\right)^l (1-p)^{k-l} + p\pi_0(1-p)^k = p(1-p)^k \left(\frac{2}{p}\right)^{i_s+1} \pi_0.$$

Finally, since

$$\sum_{k=i_s+1}^{\infty} \pi_k = \pi_0 \left(\frac{2(1-p)}{p}\right)^{i_s+1},$$

we can determine π_0 by the normalisation condition

$$1 = \sum_{k=0}^{i_s} \pi_k + \sum_{k=i_s+1}^{\infty} \pi_k = \pi_0 \sum_{k=0}^{i_s+1} \left(\frac{2(1-p)}{p}\right)^k$$

so that

$$\pi_0 = \frac{p^{i_s+1}(3p-2)}{p^{i_s+2} - (2(1-p))^{i_s+2}}.$$

□

5.2.2 Arbitrary a

We now want to identify the first $(\pi_0, \cdots, \pi_{i_s})$ states of the equilibrium for an arbitrary shortening of length a. In this case, we do not get explicit formulas for the equilibrium distribution as before, but we provide here a method to obtain numerically the distribution of the steady state.

We start by showing that actually only the first a terms $(\pi_0, \cdots, \pi_{a-1})$ are missing.

Lemma 5.2.1. *For all $k \geq a$, there exists a linear function $\phi_k : \mathbb{R}^a \mapsto \mathbb{R}$ such that:*

$$\pi_k = \phi_k(\pi_0, \ldots, \pi_{a-1}).$$

Proof. Equation (5.6) shows that it is enough to prove that $(\pi_a, \cdots, \pi_{i_s})$ depends linearly on $(\pi_0, \cdots, \pi_{a-1})$.

By multiplying equation (5.6) by $(u-1)(1-uq)$, with $q = 1-p$, we get:

$$(u^a - 1)(1-uq)\mathbb{E}\left(u^{L_\infty}\right) = q(u^a+1)(u-1)\left(\sum_{k=0}^{i_s} u^k \pi_k\right) + p\sum_{k=0}^{a-1}(u^a - u^k)\pi_k \tag{5.9}$$

We now identify the coefficients of the power series of each side of (5.9) distinguishing cases for the values of k. Using the identities: $(u^a-1)(1-uq) = -qu^{a+1} + u^a + qu - 1$, $(u-1)(u^a+1) = u^{a+1} - u^a + u - 1$,

the left hand side of (5.9) is decomposed as follows:

$$\begin{aligned}(-qu^{a+1}+u^a+qu-1)\mathbb{E}\left(u^{L_\infty}\right) &= \sum_{k=0}^{\infty}\pi_k\left(-qu^{k+a+1}+u^{k+a}+qu^{k+1}-u^k\right)\\ &= \sum_{k=0}^{\infty}\left(-q\pi_{k-a-1}+\pi_{k-a}+q\pi_{k-1}-\pi_k\right)u^k\end{aligned}$$

and the right hand side:

$$q\Big[\sum_{k=a+1}^{i_s+a+1}\pi_{k-a-1}\,u^k-\sum_{k=a}^{i_s+a}\pi_{k-a}\,u^k+\sum_{k=1}^{i_s+1}\pi_{k-1}\,u^k-\sum_{k=0}^{i_s}\pi_k\,u^k\Big]+p\sum_{k=0}^{a-1}\pi_k\,u^a-p\sum_{k=0}^{a-1}\pi_k\,u^k.$$

The following table gives the recurrence relations obtained after identification of the coefficients.

k	relation
$[\![0,a-1]\!]$	$\emptyset$
a	$\pi_a = 2\dfrac{1-p}{p}\pi_0 - \sum_{k=1}^{a-1}\pi_k$
$[\![a+1,i_s]\!]$	$\pi_k = -2\dfrac{1-p}{p}\pi_{k-a-1}+\dfrac{2-p}{p}\pi_{k-a}$

Hence, for all $a \le k \le i_s$, there exists a linear function $\varphi_k : \mathbb{R}^k \mapsto \mathbb{R}$ such that

$$\pi_k = \varphi_k(\pi_{k-1},\cdots,\pi_0).$$

Finally, by induction, for all $a \le k \le i_s$, there exists a linear function $\phi_k : \mathbb{R}^a \mapsto \mathbb{R}$ such that

$$\pi_k = \phi_k(\pi_0,\ldots,\pi_{a-1})$$

so that the distribution of L_∞ is actually determined by its a first states. □

We are now able to characterise mathematically the equilibrium of the distribution of the length of telomeres evolving with telomerase. By using lemma 5.2.1, we can find $\psi : \mathbb{R}^a \times [0,1] \mapsto \mathbb{R}$ linear in the first a coordinates such that:

$$(1+\cdots+u^{a-1})\mathbb{E}(u^{L_\infty}) = \psi(\pi_0,\ldots,\pi_{a-1},u). \tag{5.10}$$

The polynome $R(u) = 1+...+u^{a-1}$ has exactly $a-1$ roots in the unit disk which are the a^{th} roots of unity except 1. Let $u_k = e^{2i\pi k/a}$ for $1 \le k \le a-1$. The vector $(\pi_0,...,\pi_{a-1})$ is thus solution of the

system:

$$\psi(\pi_0, ..., \pi_{a-1}, u_k) = 0 \text{ for } 1 \leq k \leq a-1 \tag{5.11}$$

where π_0 is, as usual determined, by the normalisation condition. If the system is invertible, then there exists a unique vector $(\pi_0, ..., \pi_{a-1})$ solution of (5.11) to which we added the normalisation condition. Therefore, we can at least numerically obtain the vector $(\pi_0, ..., \pi_{a-1})$. Once we have it, the distribution of the chain follows from (5.10).

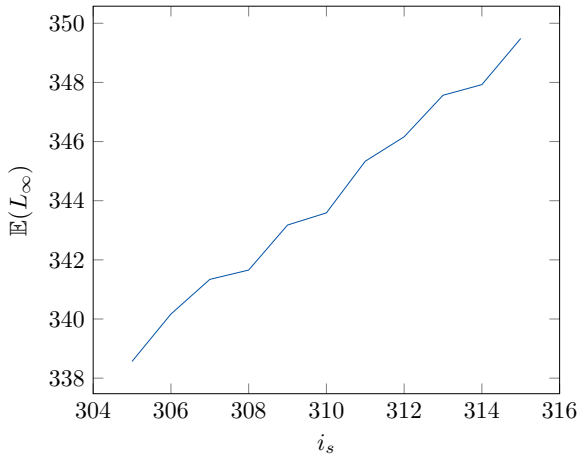
We simplified the general model (5.3) in which telomerase recruitment is governed by f by (5.5) with a sharp threshold for telomerase recruitment in order to have explicit formulas for the equilibrium of telomere length distribution to have an insight of the impact of the parameters (a, p) on the equilibrium distribution of telomere length. However, this simplification also shows the role of the imprecision of the recruitment of the telomerase is the variance of this equilibrium. More precisely, it happens sometimes that telomerase elongates long telomeres. This behavior is captured by the shape of f (5.A) but absent when the threshold is sharp. Hence, a comparison between (5.5) and (5.3) reveals the impact of the imprecision of the telomerase on the equilibrium.

5.2.3 Numerical application

To rigorously compare the variance of the simplified model (5.5) with (5.3), we choose i_s so that the ceiling function of the mean of the equilibrium of (5.5) $\lceil \mathbb{E}(L_\infty) \rceil$ is the same as the one for (5.3), *ie* 342 bp [Xu et al., 2013]. We take the biological parameters obtained in [Teixeira et al., 2004, Soudet et al., 2014] and used in [Xu et al., 2013]

$$a = 7,\ p = 0.026,\ L_s = 90,\ \beta = 0.045.$$

For this purpose, we make i_s vary and run 10^6 numerical simulations of (5.5) per choice of i_s in order to compute the corresponding mean of the equilibrium. This mean is then computed and plotted as a function of i_s:



Finally, we chose the i_s that gives $\lceil \mathbb{E}(L_\infty) \rceil = 342$ bp. This procedure leads to $i_s = 308$ bp.

We ran then 10^6 simulations of (5.3) and (5.5) again with the parameters from [Teixeira et al., 2004, Soudet et al., 2014] and used in [Xu et al., 2013] (Fig. 5.3) and find that the variance obtained using the simplified model (5.5) is smaller than the one with the complete model (5.3) (37 bp as compared to 101 bp) for the biological parameters, demonstrating that the residual recruitment of telomerase to rather long telomeres strongly contributes to the spread of the steady state distribution of telomere length.

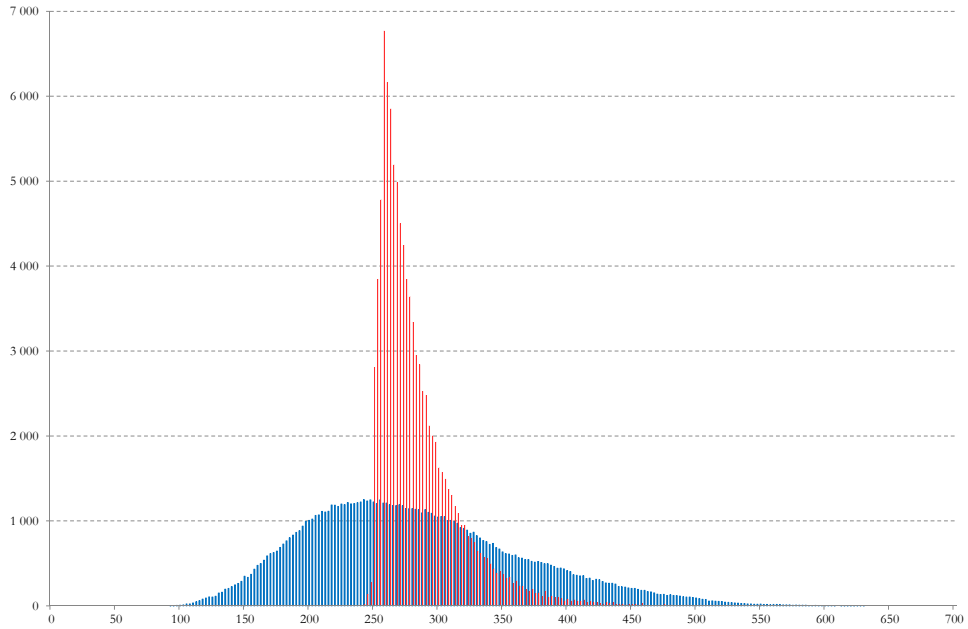


Figure 5.3: In red, the equilibrium distribution of the Markov Chain (5.5). In blue, the equilibrium distribution obtained in [Xu et al., 2013]

As a conclusion, our calculations on model (5.5) exhibit how the equilibrium distribution of telomere length depends on the shortening length a and the elongation by the telomerase via the parameter p. For instance, we showed that for $a = 1$, the geometric elongation gives rise to a geometric behaviour of the equilibrium distribution. In addition, the comparison with model (5.3) shows that, surprisingly the mode of recruitment and activation of telomerase, dependent on the biochemical properties of telomerase and on its interactions with telomeric proteins (*i.e.* Rap1/Rif1/Rif2, Tel1, Cdc13, MRX complex and Ku complex) is also critical for the variance of this equilibrium.

5.3 Impact of the steady state distribution on the onset of senescence

We now analyse the consequences of the steady state distribution on the onset of senescence, meaning the number of generations undergone by a given cell lineage until it enters senescence, which we simply

call time of senescence and we denote T. The main goal of this section is to derive the parameters of the initial distribution from the time of senescence, which is useful for experimentalists because the exact parameters of the initial distribution are usually not accessible while the time of senescence can be measured [Lundblad and Szostak, 1989, Xu et al., 2015]. In senescing cells, telomerase is inactive and when the shortest telomere goes below a threshold S, the cell enters replicative senescence and stops dividing, [Abdallah et al., 2009b, Lundblad and Szostak, 1989, Hemann et al., 2001b, Zou et al., 2004, Armanios et al., 2009]. A haploid yeast cell has 16 chromosomes and thus 32 telomeres. Mathematically, we consider the vector $(L_n^1, L_n^2, \dots, L_n^{32})$ of these 32 telomere lengths at generation n. Because each chromosome behaves independently [Shampay and Blackburn, 1988], we can start by studying one chromosome and the behaviour of the 32 will easily follow. More precisely, the vector $(L_n^1, L_n^2, \dots, L_n^{32})$ can be seen as a family $(X_n^i, Y_n^i)_{1\leq i\leq 16}$ of 16 independent identically distributed couples each representing the two telomeres of a chromosome.

In order to be more general, we consider that chromosomes are initially distributed along the equilibrium of model (5.2) which take into account the coupling of the two extremities of a given chromosome. By using the same arguments as in the proof of proposition 5.2.1, we show that the Markov chain (L_n^1, L_n^2) evolving in $\mathbb{N}^2$ and described by equation (5.2) has a unique equilibrium distribution denoted Π. Hence, for all $1 \leq i \leq 16$, the initial distribution of the couple (X_0^i, Y_0^i) is the following: for $k, l \in \mathbb{N}$,

$$\mathbb{P}(X_0^i = k, Y_0^i = l) = \Pi(X_0 = k, Y_0 = l).$$

The time of senescence is mathematically expressed as:

$$T = \inf\left\{n \geq 0, \min_{1\leq i\leq 16}\left[\min(X_n^i, Y_n^i)\right] < S\right\}.$$

In the following calculations, for simplicity, we will always choose $a = 1$ and $S = 0$. For numerical estimations of the time of senescence, we will divide our results by $a = 7$ to obtain biologically relevant values.

5.3.1 Distribution of the time of senescence

We start by computing the expectation of the time of senescence when the initial distribution of the lengths of telomeres of the same chromosome is Π.

Proposition 5.3.1. *The distribution of the random variable T is given by:*

$$\mathbb{P}(T > n) = \left[\sum_{k+l\geq n} \Pi(x_0 = k, Y_0 = l)\frac{1}{2^n}\sum_{t=n-l}^{k}\binom{n}{t}\right]^{16}. \tag{5.12}$$

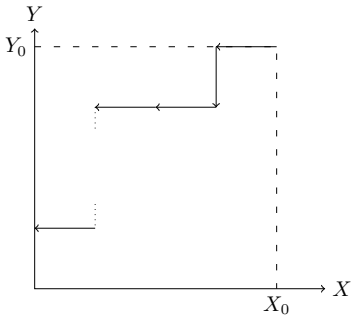
In particular, its expectation can be written as a function of π, the initial distribution as follows:

$$\mathbb{E}(T) = \sum_{n=1}^{\infty} \left[\sum_{k+l \geq n} \Pi(x_0 = k, Y_0 = l) \frac{1}{2^n} \sum_{t=n-l}^{k} \binom{n}{t} \right]^{16}. \tag{5.13}$$

Proof. For the sake of simplicity, we drop the superscript i in this section and start by studying a typical couple (X_n, Y_n). The shortening of these two telomeres can be mathematically translated into the following model:

$$\begin{cases} X_{n+1} &= (X_n - B_n)^+ \\ Y_{n+1} &= (Y_n - (1 - B_n))^+ \end{cases}$$

where B_n is a Bernoulli random variable of parameter $1/2$, and $(x_0, Y_0) \overset{dist.}{\sim} \Pi$. This process is an oriented simple random walk on $\mathbb{Z}^2$ until one of the coordinates reaches zero, and can be written explicitly:



$$\begin{cases} X_n &= X_{n-1} - B_n = x_0 - \sum_{i=1}^{n} B_i = x_0 - \mathcal{B}in(n, 1/2) \\ \\ Y_n &= Y_{n-1} - (1 - B_n) = Y_0 - n + \mathcal{B}in(n, 1/2) \end{cases}$$

where $\mathcal{B}in(n, 1/2)$ is a binomial distribution of parameters n and $1/2$. In this case, let's define the first time one of the coordinates reaches zero, T^1:

$$T^1 = \inf \{n \geq 0, \min(X_n, Y_n) < 0\}$$

. Then,

$$\begin{aligned} \mathbb{P}(T^1 > n) &= \mathbb{P}(\min(X_n, Y_n) \geq 0) \\ &= \mathbb{P}(x_0 - \mathcal{B}in(n, 1/2) \geq 0, Y_0 - n + \mathcal{B}in(n, 1/2) \geq 0) \\ &= \sum_{\substack{k+l \geq n \\ k,l \geq 0}} \Pi(x_0 = k, Y_0 = l) \frac{1}{2^n} \sum_{t=n-l}^{k} \binom{n}{t}. \end{aligned}$$

From here, we easily derive the distribution of the time of senescence by considering all 16 independent

pairs of telomeres:

$$\begin{aligned}\mathbb{P}(T > n) &= \mathbb{P}(\neg\{\text{senescence at the } n^{th} \text{ generation}\}) \\ &= \mathbb{P}(\min_{1\leq k\leq 32} L_k^n \geq 0) = \mathbb{P}(\forall i \in [\![1, \cdots, 16]\!], \min(X_n^i, Y_n^i) \geq 0) \\ &= \mathbb{P}(\min(X_n, Y_n) \geq 0)^{16} = \mathbb{P}(T^1 > n)^{16}.\end{aligned}$$

The expected time of senescence is then:

$$\mathbb{E}(T) = \sum_{n=0}^{\infty} \mathbb{P}(T > n) = \sum_{n=1}^{\infty} \left[\sum_{k+l\geq n} \Pi(x_0 = k, Y_0 = l) \frac{1}{2^n} \sum_{t=n-l}^{k} \binom{n}{t} \right]^{16}.$$

□

We now study separately the influence of the mean and the variance of the initial state on the time of senescence.

5.3.2 Impact of the initial mean on the time of senescence

We start with the effect of the mean of π on the expected time of senescence, without taking into account the variability of the initial distribution, and thus consider that telomeres have initially a deterministic and constant length, denoted x_0 equal to the mean of the initial distribution π, *i.e* for $k \in [\![1, 32]\!]$,

$$L_0^k = \lceil \mathbb{E}(L_\infty) \rceil = x_0. \tag{5.14}$$

We define $T_{x_0}^1$ as the first time one of two coupled telomeres reaches zero both starting from x_0, and T_{x_0} as the time of senescence of the whole cell when the initial state is constant and equals x_0.

Almost surely, $x_0 \leq T_{x_0}^1 \leq 2\,x_0$. This implies that $\mathbb{P}(T_{x_0}^1 > n) = 0$ for $n \geq 2x_0$, and $\mathbb{P}(T_{x_0}^1 > n) = 1$ for $n \leq x_0 - 1$. For $x_0 \leq n \leq 2x_0 - 1$, the law of $T_{x_0}^1$ is given by:

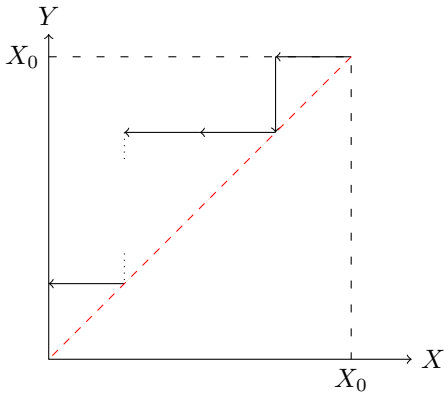
$$\mathbb{P}(T_{x_0}^1 > n) = \frac{1}{2^n} \sum_{k=n-x_0}^{x_0} \binom{n}{k}.$$

The expected time of senescence is then:

$$\mathbb{E}(T_{x_0}) = x_0 + \sum_{n=x_0}^{2x_0-1} \left[\mathbb{P}(T_{x_0}^1 > n)\right]^{16} = x_0 + \sum_{n=x_0}^{2x_0-1} \left[\frac{1}{2^n} \sum_{k=n-x_0}^{x_0} \binom{n}{k} \right]^{16}. \tag{5.15}$$

Numerical simulations with biological parameters from [Teixeira et al., 2004, Soudet et al., 2014]show, *cf.* Fig. 5.3, that the mean of the steady state distribution L_∞ is large (342 bp), so that we can use

an asymptotic expansion of $\mathbb{E}(T_{x_0})$ for large values of x_0. At the first order the mean behaviour prevails, which can be seen as the result of the deterministic process $(\widetilde{X}_n, \widetilde{Y}_n)$ such that (in red on the figure):



$$\begin{cases} \widetilde{X}_n &= \widetilde{X}_{n-1} - 1/2 \\ \\ \widetilde{Y}_n &= \widetilde{Y}_{n-1} - 1/2. \end{cases}$$

This suggests that

$$\mathbb{E}(T^1_{x_0}) \underset{x_0 \to +\infty}{\sim} 2x_0$$

holds.

Proposition 5.3.2. *[Asymptotics of the Time of Senescence for one Chromosome] For the convergence in distribution, when x_0 tends to infinity:*

$$\frac{2x_0 - T^1_{x_0}}{\sqrt{x_0}} \longrightarrow |\mathcal{W}_2|$$

where $(\mathcal{W}_t)$ is a standard Brownian Motion.

Proof. To find the second order of the expansion, we consider a coupling $(\bar{X}_n, \bar{Y}_n)$ such that for $n \leq T^1_{x_0}$,

$$\mathbb{P}\left[(X_n, Y_n) = (x, y)\right] = \mathbb{P}\left[(\bar{X}_n, \bar{Y}_n) = (x, y)\right]$$

where $(\bar{X}_n, \bar{Y}_n)$ is the unbounded random walk on $\mathbb{Z}^2$.
For convenience, the subscript n of the random walks (X_n, Y_n) and $(\bar{X}_n, \bar{Y}_n)$ will from now on be put as an argument, *i.e* we will rather write $(X(n), Y(n))$.

Classical Donsker's theorem [Billingsley, 2009] shows that if $(B_i)_i$ is a sequence of independent Bernoulli random variables of parameter $1/2$, then

$$\left(\frac{\sum_{i=1}^{\lfloor t x_0 \rfloor} (B_i - 1/2)}{\sqrt{x_0}} \right) \underset{x_0 \to +\infty}{\xrightarrow{\mathcal{L}}} \frac{1}{2}(\mathcal{W}_t)$$

where $(\mathcal{W}_t)$ is a standard Brownian Motion.

Therefore, for the convergence in distribution,

$$\lim_{x_0\to+\infty}\left(\frac{x_0-\bar{X}(\lfloor tx_0\rfloor)-\lfloor tx_0\rfloor/2}{\sqrt{x_0}},\frac{x_0-\bar{Y}(\lfloor tx_0\rfloor)-\lfloor tx_0\rfloor/2}{\sqrt{x_0}}\right)=\left(\frac{1}{2}\mathcal{W}_t,-\frac{1}{2}\mathcal{W}_t\right)$$

We can now compute the asymptotic distribution of $T^1_{x_0}$. By definition, since (X_n) and (Y_n) are decreasing, and by the previous coupling,

$$\begin{aligned}\mathbb{P}\left(\frac{2x_0-T^1_{x_0}}{\sqrt{x_0}}<w\right)&=\mathbb{P}\left(T^1_{x_0}>2x_0-w\sqrt{x_0}\right)=\mathbb{P}\left(T^1_{x_0}>\lfloor 2x_0-w\sqrt{x_0}\rfloor\right)\\&=\mathbb{P}\left(X\left(\left\lfloor x_0\left(2-\frac{w}{\sqrt{x_0}}\right)\right\rfloor\right)\geq 0,Y\left(\left\lfloor x_0\left(2-\frac{w}{\sqrt{x_0}}\right)\right\rfloor\right)\geq 0\right)\\&=\mathbb{P}\left(\bar{X}\left(\left\lfloor x_0\left(2-\frac{w}{\sqrt{x_0}}\right)\right\rfloor\right)\geq 0,\bar{Y}\left(\left\lfloor x_0\left(2-\frac{w}{\sqrt{x_0}}\right)\right\rfloor\right)\geq 0\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{x_0\to\infty}\mathbb{P}\left(\frac{2x_0-T^1_{x_0}}{\sqrt{x_0}}<w\right)&=\lim_{x_0\to\infty}\mathbb{P}\left(x_0-\frac{\lfloor(2-w/\sqrt{x_0})x_0\rfloor}{2}-\sqrt{x_0}\frac{1}{2}\mathcal{W}_2\geq 0,\right.\\&\qquad\left.x_0-\frac{\lfloor(2-w/\sqrt{x_0})x_0\rfloor}{2}+\sqrt{x_0}\frac{1}{2}\mathcal{W}_2\geq 0\right)\\&=\mathbb{P}\left(w+\mathcal{W}_2\geq 0,w-\mathcal{W}_2\geq 0\right)\\&=\mathbb{P}\left(|\mathcal{W}_2|\leq w\right)\end{aligned}$$

so that for the convergence in distribution, when x_0 tends to infinity:

$$\frac{2x_0-T^1_{x_0}}{\sqrt{x_0}}\longrightarrow|\mathcal{W}_2|. \tag{5.16}$$

□

Approximation of the expected time of senescence: Since$|\mathcal{W}_2|$ is a random variable whose density is given by $1/\sqrt{\pi}e^{-x^2/4}\mathbb{1}_{\{x\geq 0\}}$, we can get an approximation of the time of senescence for the whole cell from equation (5.15) by replacing $T^1_{x_0}$ by its asymptotic (5.16), when x_0 tends to infinity. This suggests the approximation

$$\begin{aligned}\mathbb{E}(T_{x_0})&\approx x_0+\sum_{n=x_0}^{2x_0-1}\left[\mathbb{P}(2x_0-\sqrt{x_0}\,|\mathcal{W}_2|>n)\right]^{16}\\&\approx x_0+\sum_{k=0}^{x_0-1}\left[\mathbb{P}(|\mathcal{W}_2|<\frac{k}{\sqrt{x_0}})\right]^{16}\approx x_0+\sum_{k=0}^{x_0-1}\left[\operatorname{erf}\left(\frac{k}{2\sqrt{x_0}}\right)\right]^{16}\end{aligned} \tag{5.17}$$

where erf is the error function.

To evaluate whether this expansion is indeed accurate, we ran $10,000$ simulations starting with a constant telomere length distribution, with all telomeres having a length x_0 (Fig. 5.4). As stated earlier, we take $a = 7$ in the simulations, although a was chosen equal to 1 in the calculations. We find that the expansion (5.16) is hardly distinguishable from the theoretical process (compare the dashed line and blue line in Fig. 5.4) and can thus be directly used to estimate the mean of the initial state in experimental studies.

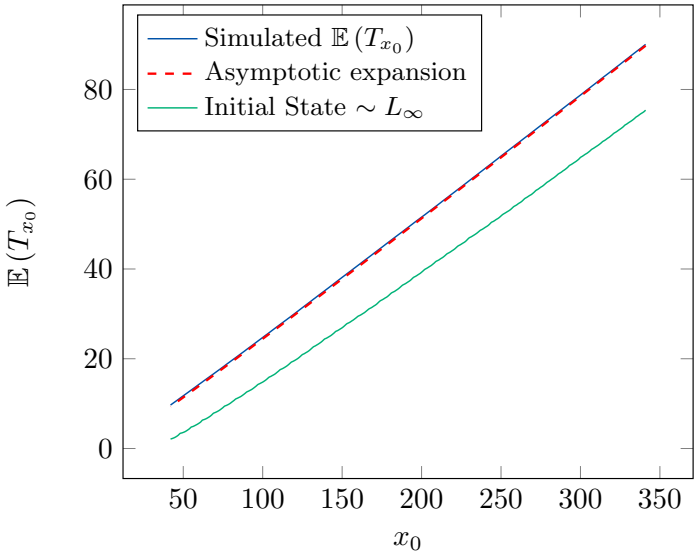


Figure 5.4: The asymptotic expansion (dashed line) corresponds to equation (5.17). The simulation of the process starting from the initial conditions (5.14) is drawn in blue. The green line is the mean time of senescence obtained from an initial distribution L_∞, translated to make its mean vary.

When all telomeres have initially the same length, they have the same probability to be the first to reach the threshold of senescence S. The effect of the coupling between the two extremities of the same chromosome is very strong, so that we have to study the 32 telomeres at once. We see on figure 5.4 that the difference with the process (in green) starting from an initial distribution L_∞ is negligible (about ten generations). In fact, it is the case because of the small variance of L_∞ (37 bp). On the contrary, if the initial variance is large, the intuition is that the senescence will be essentially due to the initial shortest telomere. It has previously been shown in [Bourgeron et al., 2015] that, starting with an experimental distribution of telomere length, in approximately 60 percent of the lineages, the initial shortest telomere among the 32 remains the shortest at the onset of senescence, *i.e.* the signaling telomere. This phenomenon can of course not be explained when considering the initial state as deterministic and constant.

5.3.3 Influence of the initial variance on the time of senescence

Having in mind the previous considerations, we now consider a random initial distribution. To study only the influence of the initial variance, we consider that each initial telomere is uniformly distributed

in the interval $[x_0 - \sigma, x_0 + \sigma]$ and simulate the expected time of senescence as a function of σ (Fig. 5.5).

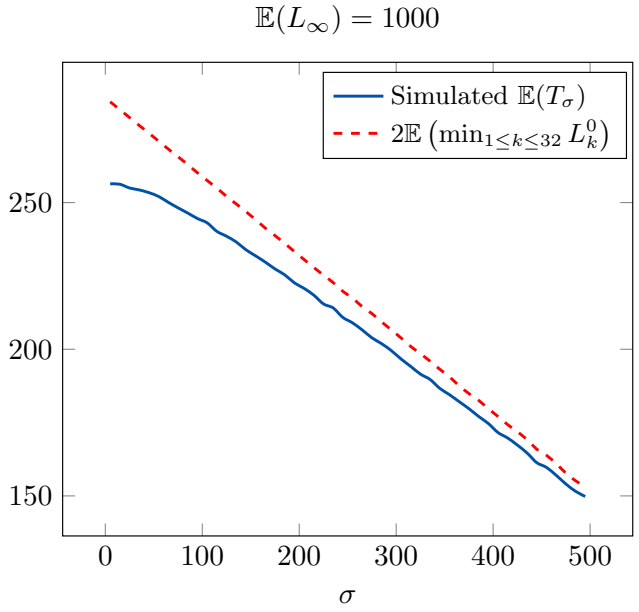


Figure 5.5: Starting from a uniform distribution of variance σ, the time of senescence is computed using equation (5.18), which takes only the mean behaviour of the initial shortest telomere into account, and compared to numerical simulations.

The time of senescence of the cell is denoted T_σ in this subsection.

Approximation of the distribution of the time of senescence. For large σ, *i.e* σ of the order of x_0, the time of senescence depends only on the mean time of senescence of the initial shortest telomere:

$$\lim_{\sigma \to x_0} \mathbb{E}(T_\sigma) = 2\,\mathbb{E}\left[\min_{1\leq k\leq 32} L_k^0\right]. \tag{5.18}$$

In fact, the larger σ is, the most likely it is for the initial shortest to be the signaling telomere. The effect of the coupling is not relevant any more in this case and we can just consider that the initial shortest telomere is performing a simple random walk on $\mathbb{Z}^2$ until it reaches the threshold on senescence. Let $M = \mathbb{E}\left(\min_{1\leq k\leq 32} L_k^0\right)$. If we accept the previous hypothesis, T_σ is now the time for a simple random walk starting from M to reach zero. By using again Donsker's theorem, this leads to the following expansion for T_σ:

$$\frac{2M - T_\sigma}{\sqrt{M}} \underset{M\to\infty}{\overset{\mathcal{L}}{\sim}} \mathcal{W}_2$$

suggesting (5.18). See figure 5.5 for numerical simulations.

5.3.4 Impact of the initial distribution

As a conclusion, depending on the initial variance σ, we found two approximations of the expected time of senescence:

(i) if σ is small, considering that all telomeres have a constant length equal to the mean of the initial distribution allows a good approximation of the expected time of senescence (see figure 5.4). All telomeres have the same probability to be the signalling telomere at the senescence so that we have to consider the coupling between the two extremities of a given chromosome. This coupling induces a second order that is non negligible with the biological parameters from [Teixeira et al., 2004].

(ii) if σ is large, the time of senescence is mainly determined by the shortening of the initial shortest telomere.

With the biological parameters form [Teixeira et al., 2004] and the more general model (5.3) with a smooth threshold at L_s, the variance of the initial state is large enough to allow the use of equation (5.18) for the prediction for the mean time of senescence. Figure 5.6 summarises these phenomena.

5.4 Conclusion and discussion

In summary, we modelled several molecular mechanisms that contribute at various levels to telomere length distribution and dynamics in *S. cerevisiae*, where they are the most exhaustively and quantitatively described. Among these mechanisms, we found that the asymmetry of telomere replication and the coupling between the two telomeres belonging to the same chromosome significantly contributes to senescence heterogeneity and we formally established their links. We also showed that the mode and robustness of telomerase recruitment control the variance of the steady-state telomere length distribution, defining two senescence regimes. With a low initial variance, the 32 telomeres of the cell can be considered as having the same length equal to the initial mean telomere length. In contrast, a high initial variance leads to a major role of the initial shortest telomere in controlling senescence. Because natural telomere length distributions can vary a lot, even within a species, we suggest that depending on the initial variance, the two regimes we describe may operate at the same time during senescence. This work uncovers a new layer of complexity in the relationship between senescence onset and telomere shortening explained by the asymmetry and coupling mechanisms, and proposes methods for assessing the time of senescence or conversely inferring parameters of the initial telomere length distribution.

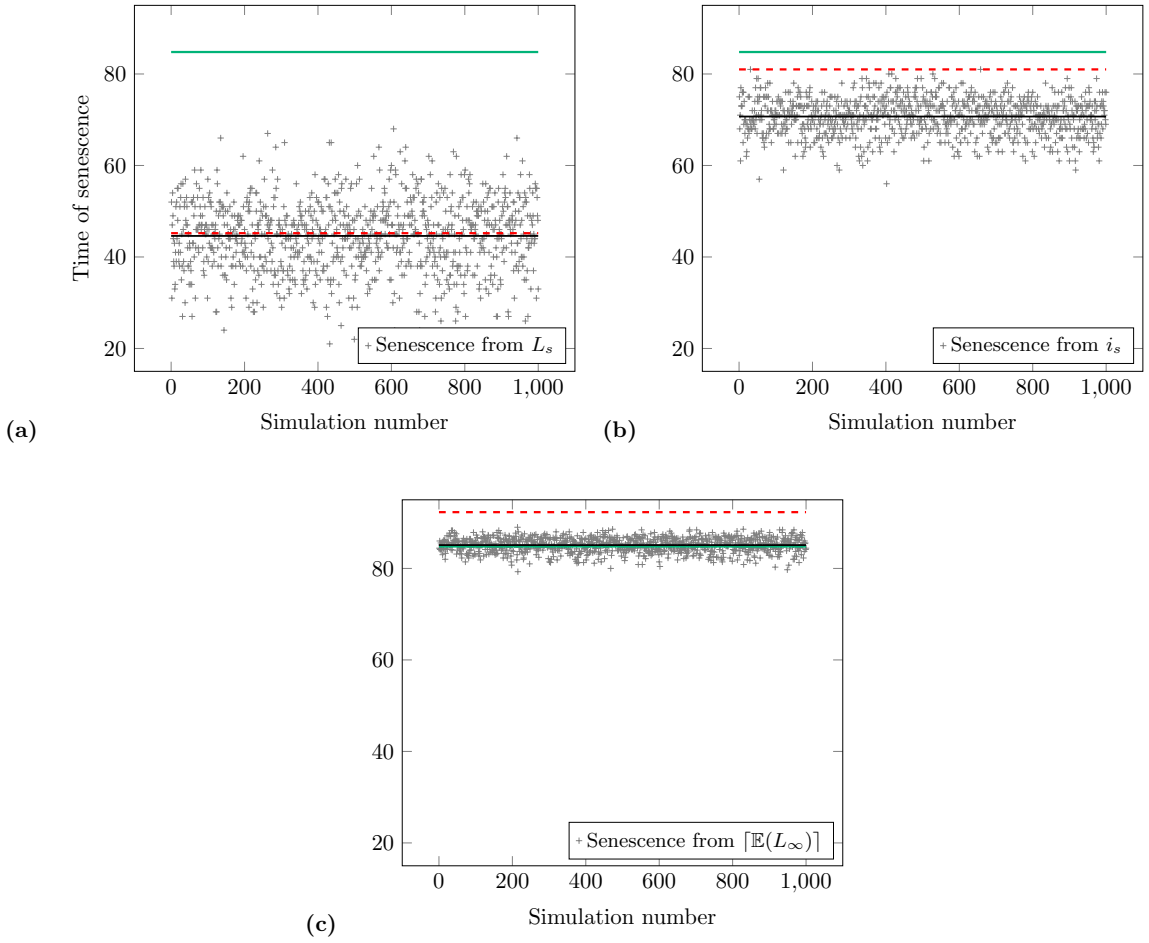


Figure 5.6: Three initial distributions for the telomere length are considered: the equilibrium of (5.3) (which existence is proved appendice 5.A), with a smooth threshold at L_s, L_∞, with a sharp threshold at i_s, and a constant initial state at $\lceil \mathbb{E}(L_\infty) \rceil$. For each case, 1000 simulations of the time are plotted. The threshold of senescence is fixed at $S = 19$ bp. Then, the green line corresponds to equation (5.17), the red dashed line to equation (5.18), and the black line is the simulated expectation. **(a)**: the variance of (5.3) is large enough so that expansion (5.18) captures the mean time of senescence. **(b)**: the variance of (5.5) is not large enough to use expansion (5.18), and not small enough for (5.17). **(c)**: equation (5.17) reproduces the mean time of senescence.

Appendix

5.A Ergodicity of the complete model

The more general model for telomere length distribution when the telomerase is active is given by equation (5.3):

$$L_{n+1} = (L_n - a \cdot B_n)^+ + \mathrm{C}_n \cdot \mathcal{G}_n.$$

Since we use its equilibrium distribution for the numerical simulations presented figure 5.6, we verify here that the Markov chain (5.3) is ergodic.

Lemma 5.A.1. *The Markov chain* (L_n) *defined by* (5.3) *is ergodic.*

Proof. Note that, as before, it is easy to check that the chain is irreductible and aperiodic. The ergodicity is again obtained from Foster-Lyapunov criteria. The shape of the function f governing the mode of recruitment of telomerase

$$\begin{cases} f(l) & = 1 \text{ if } l \leq L_s \\ f(l) & = \dfrac{1}{1+\beta(l-L_s)} \text{ if } l > L_s \end{cases}$$

allows us to find $\gamma < p/(1-p) \cdot a/2$ and K such that if $x \geq K, f(x) \leq \gamma$. Then:

— $\mathbb{E}_x(L_1 - x) = -a/2 + f(x)(1-p)/p \leq -(a/2 - \gamma(1-p)/p)$ for $x > K$, and $a/2 - \gamma(1-p)/p > 0$.

— $\mathbb{E}_x(L_1) \leq x - a/2 + (1-p)/p < \infty$ for $x \leq K$.

The lemma is proved. □

5.B If telomeres were always elongated

The threshold at i_s is a key feature of the dynamics of the telomeres evolving with telomerase. However, it is interesting to see what happens if the telomerase always repairs the shortened telomeres, no matter

how long or short they are. This is equivalent to choose $i_s = \infty$. The model becomes:

$$L_{n+1} = (L_n - a.B_n)^+ + \mathcal{G}_n \tag{5.19}$$

where B_n is still a Bernoulli random variable of parameter $1/2$, independent of all the other quantities.

Proposition 5.B.1. *If $a > 2(1-p)/p$, then the Markov chain (L_n) defined by (5.19) is ergodic and has a unique equilibrium distribution L_∞, the distribution of which being given by:*

$$\left[p(1 + u + \cdots + u^{a-1}) - 2(1-p)u^a\right] \mathbb{E}(u^{L_\infty}) = pu^a \sum_{l=1}^{a} \pi_{a-l} \left(\frac{1}{u} + \cdots + \frac{1}{u^l}\right) \tag{5.20}$$

Proof. We apply again Foster's criterion:

$$\mathbb{E}_x(L_1 - x) = -(a/2 - (1-p)/p) < 0$$

for $a > 2(1-p)/p$. The chain is then ergodic; aperiodicity and irreductibility is, as before, immediate. Let L_∞ be the equilibrium again, $(\pi_k)_k$ its distribution; it satisfies:

$$L_\infty = (L_\infty - a.B_0)^+ + \mathcal{G}_0$$

We can now compute the Laplace transform of L_∞:

$$\begin{aligned}
\mathbb{E}(u^{L_\infty}) &= \mathbb{E}(u^{(L_\infty - a.B_0)^+ + \mathcal{G}_0}) \\
&= \mathbb{E}(u^{\mathcal{G}_0}) \cdot \mathbb{E}(u^{(L_\infty - a.B_0)^+}) \\
&= \frac{1}{2}\mathbb{E}(u^{\mathcal{G}_0}) \left[\mathbb{P}(L_\infty < 0) + \mathbb{P}(L_\infty < a) + \mathbb{E}\left(u^{L_\infty - a} \cdot \mathbb{1}_{\{L_\infty \geq a\}}\right) + \mathbb{E}\left(u^{L_\infty} \cdot \mathbb{1}_{\{L_\infty \geq 0\}}\right)\right] \\
&= \frac{p}{2(1-u(1-p))} \left[\mathbb{P}(L_\infty < a) + \left(\frac{1}{u^a} + 1\right) \mathbb{E}(u_\infty^L) - \frac{1}{u^a} \sum_{k=0}^{a-1} u^k \pi_k\right]
\end{aligned}$$

Then:

$$\begin{aligned}
\left[2u^a(1-u(1-p)) - p(1+u^a)\right] \mathbb{E}(u_\infty^L) &= p \sum_{k=0}^{a-1} (u^a - u^k)\pi_k \\
&= pu^a(u-1) \sum_{l=1}^{a} \pi_{a-l} \left(\frac{1}{u} + \ldots + \frac{1}{u^l}\right)
\end{aligned}$$

One notes that:

$$\begin{aligned}2u^a(1-u(1-p))-p(1+u^a) &= (2-p)u^a - 2u^{a+1}(1-p) - p \\ &= p(u^{a+1}-1) - (u-1)u^a(2-p) \\ &= (u-1)\left[p(1+u+\cdots+u^{a-1}) - 2(1-p)u^a\right]\end{aligned}$$

So that:

$$\left[p(1+u+\cdots+u^{a-1}) - 2(1-p)u^a\right]\mathbb{E}(u^{L_\infty}) = pu^a\sum_{l=1}^{a}\pi_{a-l}\left(\frac{1}{u}+\cdots+\frac{1}{u^l}\right)$$

The distribution of L_∞ is completely determined by the a first terms $(\pi_0,\cdots,\pi_{a-1})$. Since we already have the normalisation constraint, we have to find $a-1$ other equations satisfied by $(\pi_0,\cdots,\pi_{a-1})$. We are going to show that the polynomial $\left[p(1+u+\cdots+u^{a-1}) - 2(1-p)u^a\right]$ has exactly $a-1$ roots in the disk of convergence of the series $\mathbb{E}(u^{L_\infty})$. Let:

$$\begin{aligned}P(u) &= \left[p(1+u+\cdots+u^{a-1}) - 2(1-p)u^a\right] \\ Q(u) &= (u-1)P(u) = (2-p)u^a - (2u^{a+1}(1-p)+p)\end{aligned}$$

Then: P has $a-1$ roots in the unit disk if and only if Q has a roots in the unit disk.
For this, we use Rouché's theorem: suppose there exists $r_0 > 1$ such that:

$$|2u^{a+1}(1-p)+p| < (2-p)|u^a| \ \ \forall u, |u| = r_0$$

then Q has as many roots as u^a inside $D(0,r_0)$, i.e a roots. If $u = re^{i\theta}$:

$$|2u^{a+1}(1-p)+p| < (2-p)|u^a| \iff p + 2r^{a+1}(1-p) < (2-p)r^a$$

Let's introduce $\psi(r) = p + 2r^{a+1}(1-p) - (2-p)r^a$. We want to show that ψ remains negative for some $r_0 > 1$. Quick calculations give us:

$$\begin{aligned}\psi'(r) &= 2(a+1)r^a(1-p) - a(2-p)r^{a-1} \\ \psi'(r) = 0 &\iff r := x_0 = a(2-p)/(2(1-p)(a+1)).\end{aligned}$$

So, ψ decreases on $[0,x_0]$, increases on $[x_0,+\infty)$, and $\psi(1) = 0$. It is clear that it is possible to find such an r_0 if only if $x_0 > 1$, i.e:

$$a > 2\frac{1-p}{p} \tag{5.21}$$

If this condition is satisfied, then all $1 < r_0 < x_0$ work. Since r_0 is arbitrary, by using Rouché's theorem, we prove that Q has at most $a-1$ roots in the unit disk, that we call $(u_1,\cdots,u_{a-1})$. Hence, the vector

$(\pi_0, \cdots, \pi_{a-1})$ is solution of the following system:

$$pu_i^a \sum_{l=1}^{a} \pi_{a-l} \left(\frac{1}{u_i} + \cdots + \frac{1}{u_i^l} \right) = 0$$

for $i = 1 \dots a-1$.
If this system is invertible, then we can fully determine $\mathbb{E}(u^{L_\infty})$. If we look closer at equation(5.21), we see that it is equivalent to say:

$$\mathbb{E}(aB) > \mathbb{E}(\mathcal{G})$$

which is the ergodicity condition. As a conclusion, for $a > 2(1-p)/p$, the Markov chain $(L_n)_n$ is ergodic and its equilibrium distribution L_∞ satisfies (5.20). □

References

[Abdallah et al., 2009a] Abdallah, P., Luciano, P., Runge, K. W., Lisby, M., Géli, V., Gilson, E., and Teixeira, M. T. (2009a). A two-step model for senescence triggered by a single critically short telomere. *Nature Cell Biology*, 11(8):988–993.

[Abdallah et al., 2009b] Abdallah, P., Luciano, P., Runge, K. W., Lisby, M., Géli, V., Gilson, E., and Teixeira, M. T. (2009b). A two-step model for senescence triggered by a single critically short telomere. *Nature Cell Biology*, 11(8):988–993.

[Antal and Krapivsky, 2011] Antal, T. and Krapivsky, P. (2011). Exact solution of a two-type branching process: models of tumor progression. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(08):P08018.

[Arino et al., 1995] Arino, O., Kimmel, M., and Webb, G. F. (1995). Mathematical modeling of the loss of telomere sequences. *Journal of theoretical biology*, 177(1):45–57.

[Arkus, 2005] Arkus, N. (2005). A mathematical model of cellular apoptosis and senescence through the dynamics of telomere loss. *Journal of theoretical biology*, 235(1):13–32.

[Armanios et al., 2009] Armanios, M., Alder, J. K., Parry, E. M., Karim, B., Strong, M. A., and Greider, C. W. (2009). Short telomeres are sufficient to cause the degenerative defects associated with aging. *The American Journal of Human Genetics*, 85(6):823–832.

[Autexier and Lue, 2006] Autexier, C. and Lue, N. F. (2006). The structure and function of telomerase reverse transcriptase. *Annu. Rev. Biochem.*, 75:493–517.

[Billingsley, 2009] Billingsley, P. (2009). Convergence of probability measures. *John Wiley & Sons, INC*, pages 1–287.

[Blackburn and Collins, 2011] Blackburn, E. H. and Collins, K. (2011). Telomerase: an rnp enzyme synthesizes dna. *Cold Spring Harbor perspectives in biology*, 3(5):a003558.

[Bourgeron et al., 2015] Bourgeron, T., Xu, Z., Doumic, M., and Teixeira, M. T. (2015). The asymmetry of telomere replication contributes to replicative senescence heterogeneity. *Scientific Reports*, 5.

[Britt-Compton et al., 2009] Britt-Compton, B., Capper, R., Rowson, J., and Baird, D. M. (2009). Short telomeres are preferentially elongated by telomerase in human cells. *FEBS Letter*, 583(18):3076?3080.

[Canela et al., 2007] Canela, A., Vera, E., Klatt, P., and Blasco, M. A. (2007). High-throughput telomere length quantification by fish and its application to human population studies. *Proceedings of the National Academy of Sciences*, 104(13):5300–5.

[Chai et al., 2006] Chai, W., Sfeir, A. J., Hoshiyama, H., Shay, J. W., and Wright, W. E. (2006). The involvement of the mre11/rad50/nbs1 complex in the generation of g-overhangs at human telomeres. *EMBO reports*, 7(2):225–230.

[Dao Duc and Holcman, 2013] Dao Duc, K. and Holcman, D. (2013). Computing the length of the shortest telomere in the nucleus. *Physical Review Letters*, 111(22):228104.

[de Jesus and Blasco, 2013] de Jesus, B. B. and Blasco, M. A. (2013). Telomerase at the intersection of cancer and aging. *Trends in genetics*, 29(9):513–520.

[Dyson et al., 2007] Dyson, J., Sánchez, E., Villella-Bressan, R., and Webb, G. F. (2007). Stabilization of telomeres in nonlinear models of proliferating cell lines. *Journal of theoretical biology*, 244(3):400–408.

[Eugène et al., 2016] Eugène, S., Bourgeron, T., and Xu, Z. (2016). Effects of initial telomere length distribution on senescence onset and heterogeneity. Submitted to Phys. Rev. Letters.

[Faure et al., 2010] Faure, V., Coulon, S., Hardy, J., and Géli, V. (2010). Cdc13 and telomerase bind through different mechanisms at the lagging-and leading-strand telomeres. *Molecular cell*, 38(6):842–852.

[Griffith et al., 1999] Griffith, J. D., Comeau, L., Rosenfield, S., Stansel, R. M., Bianchi, A., Moss, H., and De Lange, T. (1999). Mammalian telomeres end in a large duplex loop. *Cell*, 97(4):503–514.

[Hayflick, 1965] Hayflick, L. (1965). The limited in vitro lifetime of human diploid cell strains. *Experimental cell research*, 37(3):614–636.

[Hemann and Greider, 1999] Hemann, M. T. and Greider, C. W. (1999). G-strand overhangs on telomeres in telomerase-deficient mouse cells. *Nucleic acids research*, 27(20):3964–3969.

[Hemann et al., 2001a] Hemann, M. T., Strong, M. A., Hao, L.-Y., and Greider, C. W. (2001a). The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell*, 107(1):67–77.

[Hemann et al., 2001b] Hemann, M. T., Strong, M. A., Hao, L.-Y., and Greider, C. W. (2001b). The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell*, 107(1):67–77.

[Henderson and Blackburn, 1989] Henderson, E. R. and Blackburn, E. (1989). An overhanging 3'terminus is a conserved feature of telomeres. *Molecular and Cellular Biology*, 9(1):345–348.

[Klobutcher et al., 1981] Klobutcher, L. A., Swanton, M. T., Donini, P., and Prescott, D. M. (1981). All gene-sized dna molecules in four species of hypotrichs have the same terminal sequence and an unusual 3'terminus. *Proceedings of the National Academy of Sciences*, 78(5):3015–3019.

[Kowald, 1997] Kowald, A. (1997). Possible mechanisms for the regulation of telomere length. *Journal of molecular biology*, 273(4):814–825.

[Larrivée et al., 2004] Larrivée, M., LeBel, C., and Wellinger, R. J. (2004). The generation of proper constitutive g-tails on yeast telomeres is dependent on the mrx complex. *Genes & Development*, 18(12):1391–1396.

[Levy et al., 1992] Levy, M. Z., Allsopp, R. C., Futcher, A. B., Greider, C. W., and Harley, C. B. (1992). Telomere end-replication problem and cell aging. *Journal of molecular biology*, 225(4):951–960.

[Lingner et al., 1995] Lingner, J., Cooper, J. P., and Cech, T. R. (1995). Telomerase and dna end replication: no longer a lagging strand problem? *Science*, 269(5230):1533.

[Lundblad, 2012] Lundblad, V. (2012). Telomere end processing: unexpected complexity at the end game. *Genes & development*, 26(11):1123–1127.

[Lundblad and Blackburn, 1993] Lundblad, V. and Blackburn, E. H. (1993). An alternative pathway for yeast telomere maintenance rescues est1- senescence. *Cell*, 73(2):347–360.

[Lundblad and Szostak, 1989] Lundblad, V. and Szostak, J. W. (1989). A mutant with a defect in telomere elongation leads to senescence in yeast. *Cell*, 57(4):633–643.

[Makarov et al., 1997] Makarov, V. L., Hirose, Y., and Langmore, J. P. (1997). Long g tails at both ends of human chromosomes suggest a c strand degradation mechanism for telomere shortening. *Cell*, 88(5):657–666.

[Marcand et al., 1999] Marcand, S., Brevet, V., and Gilson, E. (1999). Progressive cis-inhibition of telomerase upon telomere elongation. *The EMBO Journal*, 18(12):3509–3519.

[Martens et al., 2000] Martens, U. M., Chavez, E. A., Poon, S. S., Schmoor, C., and Lansdorp, P. M. (2000). Accumulation of short telomeres in human fibroblasts prior to replicative senescence. *Experimental cell research*, 256(1):291–299.

[McElligott and Wellinger, 1997] McElligott, R. and Wellinger, R. J. (1997). The terminal dna structure of mammalian chromosomes. *The EMBO journal*, 16(12):3705–3714.

[Okazaki et al., 1968] Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K., and Sugino, A. (1968). Mechanism of dna chain growth. i. possible discontinuity and unusual secondary structure of newly synthesized chains. *Proceedings of the National Academy of Sciences*, 59(2):598–605.

[Olofsson and Bertuch, 2010] Olofsson, P. and Bertuch, A. A. (2010). Modeling growth and telomere dynamics in saccharomyces cerevisiae. *Journal of theoretical biology*, 263(3):353–359.

[Olofsson and Kimmel, 1999] Olofsson, P. and Kimmel, M. (1999). Stochastic models of telomere shortening. *Mathematical biosciences*, 158(1):75–92.

[Olovnikov, 1973] Olovnikov, A. M. (1973). A theory of marginotomy: the incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. *Journal of theoretical biology*, 41(1):181–190.

[op den Buijs et al., 2004] op den Buijs, J., van den Bosch, P. P., Musters, M. W., and van Riel, N. A. (2004). Mathematical modeling confirms the length-dependency of telomere shortening. *Mechanisms of ageing and development*, 125(6):437–444.

[Ow and Dunstan, 2014] Ow, S.-Y. and Dunstan, D. E. (2014). A brief overview of amyloids and alzheimer's disease. *Protein Science*, 23(10):1315–1331.

[Portugal et al., 2008] Portugal, R., Land, M., and Svaiter, B. F. (2008). A computational model for telomere-dependent cell-replicative aging. *BioSystems*, 91(1):262–267.

[Proctor and Kirkwood, 2002] Proctor, C. J. and Kirkwood, T. B. (2002). Modelling telomere shortening and the role of oxidative stress. *Mechanisms of ageing and development*, 123(4):351–363.

[Proctor and Kirkwood, 2003] Proctor, C. J. and Kirkwood, T. B. (2003). Modelling cellular senescence as a result of telomere state. *Aging cell*, 2(3):151–157.

[Raices et al., 2008] Raices, M., Verdun, R. E., Compton, S. A., Haggblom, C. I., Griffith, J. D., Dillin, A., and Karlseder, J. (2008). C. elegans telomeres contain g-strand and c-strand overhangs that are bound by distinct proteins. *Cell*, 132(5):745–757.

[Riha et al., 2000] Riha, K., McKnight, T. D., Fajkus, J., Vyskot, B., and Shippen, D. E. (2000). Analysis of the g-overhang structures on plant telomeres: evidence for two distinct telomere architectures. *The Plant Journal*, 23(5):633–641.

[Robert, 2003] Robert, P. (2003). *Stochastic Networks and Queues*. Stochastic Modelling and Applied Probability Series. Springer-Verlag, New York.

[Rodriguez-Brenes and Peskin, 2010] Rodriguez-Brenes, I. A. and Peskin, C. S. (2010). Quantitative theory of telomere length regulation and cellular senescence. *Proceedings of the National Academy of Sciences*, 107(12):5387–5392.

[Rubelj and Vondracek, 1999] Rubelj, I. and Vondracek, Z. (1999). Stochastic mechanism of cellular aging–abrupt telomere shortening as a model for stochastic nature of cellular aging. *Journal of theoretical biology*, 197(4):425–438.

[Shampay and Blackburn, 1988] Shampay, J. and Blackburn, E. H. (1988). Generation of telomere-length heterogeneity in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 85(2):534–538.

[Sidorov et al., 2004] Sidorov, I. A., Gee, D., and Dimitrov, D. S. (2004). A kinetic model of telomere shortening in infants and adults. *Journal of Theoretical Biology*, 226(2):169–175.

[Soudet et al., 2014] Soudet, J., Jolivet, P., and Teixeira, M. T. (2014). Elucidation of the dna end-replication problem in saccharomyces cerevisiae. *Molecular cell*, 53(6):954–964.

[Tan, 1999] Tan, Z. (1999). Intramitotic and intraclonal variation in proliferative potential of human diploid cells: explained by telomere shortening. *Journal of theoretical biology*, 198(2):259–268.

[Teixeira et al., 2004] Teixeira, M. T., Arneric, M., Sperisen, P., and Lingner, J. (2004). Telomere length homeostasis is achieved via a switch between telomerase -extendible and -nonextendible states. *Cell*, 117(3):323–35.

[Watson, 1972] Watson (1972). Origin of concatemeric t7 dna. *Nat. New Biol.*, 239(94):197–201.

[Wellinger et al., 1993] Wellinger, R. J., Wolf, A. J., and Zakian, V. A. (1993). Saccharomyces telomeres acquire single-strand tg 1–3 tails late in s phase. *Cell*, 72(1):51–60.

[Wellinger and Zakian, 2012a] Wellinger, R. J. and Zakian, V. A. (2012a). Everything you ever wanted to know about saccharomyces cerevisiae telomeres: beginning to end. *Genetics*, 191(4):1073–1105.

[Wellinger and Zakian, 2012b] Wellinger, R. J. and Zakian, V. A. (2012b). Everything you ever wanted to know about saccharomyces cerevisiae telomeres: beginning to end. *Genetics*, 191(4):1073?1105.

[Wu et al., 2012] Wu, P., Takai, H., and de Lange, T. (2012). Telomeric 3' overhangs derive from resection by exo1 and apollo and fill-in by pot1b-associated cst. *Cell*, 150(1):39–52.

[Xu et al., 2013] Xu, Z., Dao Duc, K., Holcman, D., and Teixeira, M. T. (2013). The length of the shortest telomere as the major determinant of the onset of replicative senescence. *Genetics*, 194(4):847–857.

[Xu et al., 2015] Xu, Z., Fallet, E., Paoletti, C., Fehrmann, S., Charvin, G., and Teixeira, M. T. (2015). Two routes to senescence revealed by real-time analysis of telomerase-negative single lineages. *Nature communications*, 6.

[Zou et al., 2004] Zou, Y., Sfeir, A., Gryaznov, S. M., Shay, J. W., and Wright, W. E. (2004). Does a sentinel or a subset of short telomeres determine replicative senescence? *Molecular biology of the cell*, 15(8):3709–3718.