Estimating the Rate of Prion Aggregate Amplification in Yeast with a Generation and Structured Population Model

H.T. Banks^{*1}, Kevin Flores^{†1}, Christine R. Langlois²,
Tricia R. Serio³ and Suzanne S. Sindi^{‡4}

¹Department of Mathematics, Center for Quantitative Sciences in Biomedicine,
Center for Research in Scientific Computation, North Carolina State University,
Raleigh, NC 27695, United States

²Max Planck Institute for Biochemistry, 82152 Martinsried, Germany

³Department of Molecular and Cellular Biology, The University of Arizona, Tucson,
AZ 85721, United States

⁴Department of Applied Mathematics, School of Natural Sciences, University of
California, Merced, 5200 N Lake Rd, Merced, CA 95343, United States

October 24, 2016

Abstract

Prions are a special class of proteins capable of adopting multiple (misfolded) conformations, some of which have been associated with fatal diseases in mammals such as bovine spongiform encephalopathy (BSE) or Creutzfeldt-Jakob Disease (CJD). Prion diseases, like protein misfolding diseases in general, are caused by the formation and amplification of ordered aggregates of proteins called amyloids. While such diseases in mammals can take decades to form, yeast have a variety of prion phenotypes that occur over a few hours, making this system an ideal model for protein misfolding disease in general. Most experimental assays of colonies with yeast prions provide steady-state population observations which complicate the inference of biochemical parameters both by the inability to directly measure aggregate amplification and by obscuring heterogeneity between cells. We develop a mathematical and inverse problem formulation to determine the amplification rate with prion aggregates from single-cell measurements observed in propagon amplification experiments. We demonstrate the ability of our formulation to determine heterogeneous amplification rates on simulated and

^{*}htbanks@ncsu.edu

[†]kbflores@ncsu.edu

[‡]ssindi@ucmerced.edu

experimental data. Our results show that aggregate amplification rates for two prion variants are strongly bimodal, suggesting that the generational structure in the yeast population impacts the ability of prion aggregates to amplify.

1. Introduction

Prions are misfolded proteins associated with a variety of fatal, untreatable neurodegenerative disorders in mammals [29, 1, 20]. Many of these diseases are associated with aging and, as such, visible symptoms can take decades to manifest. Because of this, the single celled yeast *S. cerevisiae* has emerged as a model for prion onset because prion phenotypes appear within hours. Additionally, since prions in yeast are associated with harmless heritable phenotypes, rather than disorders, this system allows a unique opportunity to observe prions not possible for mammals. Finally, the experimental tractability of the yeast system facilitates the observations of prion dynamics and manipulations *in vivo* [38, 41].

However, a significant complication introduced by studying prions in yeast is that - unlike the mammalian counterpart - prions in yeast propagate *within* a dividing yeast colony [34, 33, 37]. Most mathematical approaches [35, 25, 22, 36, 24], neglect population heterogeneity and simply model aggregate dynamics in the context of an exponentially growing population of cells [39]. Such models therefore neglect biased protein segregation during cell division as well as asymmetries in the cell cycle - both of which have been shown to be important in the yeast prion system [23, 28]. In addition, most comparisons relied on steady-state population level experiments, which further mask heterogeneity between cells. Although stochastic simulations have proven useful for studying heterogeneity, such models are not amenable to standard analytical tools. Further, stochastic approaches are not readily used for parameter inference, as it is computationally intensive to sample from parameter space. There is a need for robust mathematical approaches to estimate biochemical parameters at the level of individual cells through comparisons with single-cell experimental data.

We provide the first mathematical treatment of yeast prion aggregates in a generational and aggregate structured population model. We demonstrate that we are able to decompose our problem from a system of PDEs into a related system of ODEs plus a single PDE with an explicit analytical solution. We formulate an inverse problem for the estimation of prion aggregate growth rates based on comparisons with propagon amplification assays is formulated. For two prion variants, we identify bimodality in the distribution of amplification rates, suggesting heterogeneity in the ability of individual cells to create aggregates. This work demonstrates that parameter inference on noisy single-cell data is indeed feasible and suggests further lines of study.

2. Model Formulation

2.1. Generation and Structured Population Model

Our model combines intracellular aggregate dynamics with a model structured by the number of divisions a cell has undergone since the start of the experiment. Typically, prion aggregate dynamics have been modeled with the nucleated polymerization equations (NPM) introduced by Masel, Jensen and Nowak [32]. In this model one explicitly considers the evolution of both the number and distribution of aggregate sizes through the processes of conversion of normal protein by existing prion aggregates and fragmentation of prion aggregates into smaller templates. Formally, the NPM consists of a system of infinitely many ODEs when discrete aggregate sizes are modeled, or a coupled ODE/PDE system when continuous aggregate sizes are considered [21, 17]. The NPM model has been adapted to consider prion aggregates existing within a continually dividing colony of yeast cells where prion aggregates are transmitted between cells during division. In yeast the typical approach has been to attempt to quantify prion variants in terms of their steady-state balance between soluble and aggregated protein. However, this single observation does not uniquely specify all parameters [39, 23, 22].

As mentioned in Section 1, our goal is to provide a quantitative model suitable for comparison with single-cell measurements. As will be described in Section 3.1, single-cell experiments are only capable of measuring the number of aggregates in a single cell and not their size. Rather than explicitly model the full aggregate size distribution, or its simplified moment closure (as in [39, 22]), we consider only the number of aggregates present in a cell at a particular time. To capture the behavior of the NPM, we consider the number of aggregates present in a cell at any time as evolving according to the logistic equation where λ represents the maximum rate of aggregate amplification and $K > 0$ the steady-state number of aggregates in a cell as $t \rightarrow \infty$. Let $y(t)$ be the number of aggregates a cell has t minutes after dividing, and following Banks, Flores and Sindi [11], we define this evolution by

$$\frac{dy}{dt} = \frac{\lambda y (K - y)}{K} := v(y) \quad (1)$$

where $y(0) = y_0$. Let $N(t, y)$ be the density of cells with y aggregates at time t . Then this density will evolve in time according to the PDE

$$\frac{\partial}{\partial t} N(t, y) + \frac{\partial}{\partial y} (v(y) N(t, y)) = 0. \quad (2)$$

To capture the full dynamics of the system, where aggregates exist in dividing cells, we consider $N_i(t, y)$, the density of cells that have undergone i divisions since the start of the experiment. As such, the aggregate dynamics in the population of dividing cells is given

by the following system of weakly-coupled PDEs:

$$\begin{aligned}
\frac{\partial N_0}{\partial t} + \frac{\partial (v(y)N_0(t, y))}{\partial y} &= -(\alpha_0(t) + \beta_0(t)) N_0(t, y) \\
\frac{\partial N_1}{\partial t} + \frac{\partial (v(y)N_1(t, y))}{\partial y} &= -(\alpha_1(t) + \beta_1(t)) N_1(t, y) + R_1(t, y) \\
&\vdots \\
\frac{\partial N_M}{\partial t} + \frac{\partial (v(y)N_M(t, y))}{\partial y} &= -(\alpha_M(t) + \beta_M(t)) N_M(t, y) + R_M(t, y).
\end{aligned} \tag{3}$$

Here the right side of each PDE represents the rates of loss of cells in generation i from division ($\alpha_i(t)$) and death ($\beta_i(t)$) as well as the influx of newly divided cells from generation $i - 1$ to generation i :

$$R_i(t, y) = 2\gamma\alpha_{i-1}(t)N_{i-1}(t, \gamma y).$$

The parameter value $\gamma = 2$ corresponds to unbiased division where aggregates are equally split between cells when a cell of generation $(i - 1)$ creates 2 cells in generation i [13, 40]. We note that this use of generation refers to the number of divisions cells have undergone since the start of the experiment and not - as is common for yeast cells - the total number of divisions since birth. Although the biological system consists of infinitely many generations, in most experimental settings the maximum number of cell divisions can be bounded, and thus our system (3) considers cells only up to a specified maximum generation M . Solution of (3) requires specification of an initial condition for each i ,

$$N_i(0, y) = \phi_i(y). \tag{4}$$

We also need to impose no-flux boundary conditions at $y = 0$ to solve the PDE system

$$v(y)N_i(t, y)|_{y=0} = 0 \tag{5}$$

for all $0 \leq i \leq M$.

2.2. Population Decomposition and Analytic Solution

We first note that our structured model facilitates computation of a variety of quantities related to the yeast cell population. For example, the total number of cells having undergone i divisions is given by

$$\overline{N}_i(t) = \int_0^\infty N_i(t, y)dy$$

and the normalized density of aggregates can be defined by

$$n_i(t, y) = \frac{N_i(t, y)}{\overline{N}_i(t)}, \text{ for } \overline{N}_i(t) > 0$$

and $n_i(t, y) = 0$ otherwise. We further note that these quantities facilitate determining an analytical solution to the system (3). Following previous approaches, [27, 11], we define the initial number of cells:

$$\overline{N}_{0,0} = \int_0^\infty N_{0,0}(y) dy \quad (6)$$

and the initial normalized aggregate density:

$$\overline{n}_{0,0}(y) = \frac{N_{0,0}(y)}{\overline{N}_{0,0}}. \quad (7)$$

The solution to our system is given by the following Theorem (generalized [11] from Hausenauer et al. [27]):

Theorem 2.1 *The solution of the system defined by (3) with initial conditions (6) and (7) is given by:*

$$N_i(t, y) = \overline{N}_i(t) n_i(t, y) \text{ for } 0 \leq i \leq M, \quad (8)$$

in which:

1. $\overline{N}_i(t)$ is the solution of the system of ODEs:

$$\begin{aligned} i = 0 & : \quad \frac{d\overline{N}_0}{dt} = -(\alpha_0(t) + \beta_0(t)) \overline{N}_0 \\ \text{for } 1 \leq i \leq M & : \quad \frac{d\overline{N}_i}{dt} = -(\alpha_i(t) + \beta_i(t)) \overline{N}_i + 2\alpha_{i-1} \overline{N}_{i-1} \end{aligned}$$

where $\overline{N}_0(0) = \overline{N}_{0,0}$ and for $i \geq 1 : \overline{N}_i(0) = 0$.

2. $n_i(t, y)$ is the solution of the PDE

$$\frac{\partial n_i(t, y)}{\partial t} + \frac{\partial (f(y) n_i(t, y))}{\partial y} = 0 \quad (9)$$

with initial conditions for all $i \geq 0 : n_i(0, y) := \gamma^i n_{0,0}(\gamma^i y)$.

In Hausenauer, et al. [27], the authors derive a solution to $n_i(t, y)$ under a specific form of the flux term $f(y) = -ky$. However, this method applies to more general flux terms as described in [11]. In order to prove the theorem, the method of characteristics is first used to determine an explicit solution to equation (9):

$$n_i(t, y) = \gamma^i n_{0,0}(\gamma^i g(t, y)) \frac{e^{-\lambda t}}{K^2} \left(g(t, y) (e^{\lambda t} - 1) + K \right)^2, \quad (10)$$

where $g(t, y) = Ky (Ke^{\lambda t} - ye^{\lambda t} + y)^{-1}$ gives the initial conditions of the characteristics under the assumption of logistic aggregate growth. Direct substitution of $\overline{N}_i(t) n_i(t, y)$ into Equation (3) completes the proof (see [27] and [11] for further details).

3. Inverse Problem Formulation and Analysis

We first describe the experimental propagon amplification assay used to collect the single-cell aggregate measurements. We develop an inverse problem formulation and demonstrate that we can accurately infer kinetic parameters on simulated data sets.

3.1. Experimentally Quantifying Aggregates

For the Sup35/[PSI^+] yeast prion system, propagon amplification experiments provide single-cell measurements that are a valuable alternative to the standard steady-state population level experiments used to estimate the typical number of transmissible aggregates (propagons) in a single cell [19, 16, 37, 33]. The key feature exploited in these experiments is that when [PSI^+] yeast are treated with guanidine hydrochloride (GdnHCl), the process of aggregate fragmentation is stopped, but other cellular processes such as cell division continue as normal. Since under GdnHCl treatment the total number of aggregates cannot increase (or decrease) and all transmissible aggregates will be diluted through cell division and, neglecting cell death, eventually there will be at most one per cell [15, 37] (see Figure 1). After the removal of GdnHCl, aggregates will again continue to increase in number through the conversion and fragmentation process.

Propagon amplification experiments involve the application of GdnHCl in tandem to monitor the rate of aggregate recovery by measuring the number of propagons in a single cell [30, 39]. First, cells are treated with GdnHCl in order to ensure all cells have at most one aggregate. Second, cells from this GdnHCl treatment are removed and allowed to grow under normal conditions for specified periods of time during which the total number of aggregates will increase and cells may divide. Finally, propagon counting assays (see Figure 1) are performed on cells resulting from the second step. The result is measurements of the number of aggregates in a randomly chosen cell at time $t > 0$ from a population founded by an initial cell having a single aggregate.

In this work, we consider recent propagon amplification experiments performed on two Sup35 variants which differ in the number of oligopeptide repeats in the prion domain. The variant “WT” represents the wild-type (i.e., standard) version of Sup35 with 5.5 oligopeptide repeats and “R15” which has only the first 5 repeats (i.e., the partial oligopeptide repeat is missing). Both Sup35 variants are capable of forming aggregates and conferring the [PSI^+] prion phenotype. Figure 2 shows the result of experiments on both prion variants. The question we pursue in this analysis is: *Are there differences in the rate of aggregate amplification, λ , between these two variants?*

3.2. Growth Rate Distribution Models and Inverse Problem Formulation

In this formulation we make several simplifying assumptions on both the yeast cells and aggregates themselves. While in practice yeast cell division and death rates are time varying, and have been well studied [18, 16, 26], for simplicity we assume that division

Figure 1: **Propagon Amplification Assay**. A two-step process is used to count the number of transmissible prion aggregates (propagons) in a single cell. (Left) A single target cell is isolated, and aggregate fragmentation is stopped through exposure to GdnHCl. Since aggregates (pinwheels) can not increase in number, they are diluted through cell division. (Right) After sufficient dilution, the colony is replated onto solid medium, in the absence of GdnHCl, where each single cell serves as a founder of a distinct yeast colony. Colonies whose founding cell had at least one aggregate will have the $[PSI^+]$ prion phenotype (white) while those founded by a cell with no aggregate will have the $[psi^-]$ non-prion phenotype (red). The number of aggregates in the target cell corresponds to the number of white colonies in the plate.

Figure 2: **Experimental Propagon Amplification Data.** (Left) Propagon counts for the WT variant, (Right) Propagon counts for the R15 variant.

rates $\alpha_i(t)$ are the same for all cells and generations i ($\alpha_i(t) = 0.0077$ 1/min) and that cell death is negligible, $\beta_i(t) = 0$. Further, we assume the initial distribution of aggregates to be given by a truncated normal distribution with $\mu = 1$ and $\sigma = .1$. Finally, while there are two parameters K and λ in the logistic model of prion dynamics (see equation (1)), because prion curing assays have been used to estimate K , the steady-state number of aggregates in a single cell [16, 19, 33], we take this value as known ($K = 200$) and focus effort on only estimating the growth rate λ .

Suppose our experimental observations consists of numbers of aggregates, $\{y_i\}_{i=1}^n$ observed at times $\{t_i\}_{i=1}^n$. We assume that these experimentally observed counts reflect the expected number of aggregates in a cell chosen at that time as given below in equation (11)

$$E[y(t; \lambda)] = \frac{\sum_{i=0}^{\infty} \int_0^{\infty} \xi N_i(t, \xi; \lambda) d\xi}{N(t)}, \quad (11)$$

where $N_i(t, \xi; \lambda)$ is the solution to the aggregate density problem for a given growth rate λ and generation i since the start of the experiment (see equation (9)), and $N(t) = \sum_{i=0}^{\infty} N_i(t)$. Rather than go to infinity, the upper limit of the sums is chosen to be the maximum number of generations M chosen based on the length of the experiments. We are interested in estimating the growth rate λ in the formulation equation (11). Given a set of observations $\{y_i\}_{i=1}^n$ for $E[y(t_i; \lambda)]$, $i = 1, \dots, n$, we are led naturally to an optimization

problem

$$\min_{\lambda} \sum_{i=1}^n |E[y(t_i; \lambda)] - y_i|^2. \quad (12)$$

However, we do not expect a single growth rate to suffice for the aggregate type data in which we are interested. To allow for heterogeneity in estimates of the aggregate growth rates, we follow an approach previously used in [9, 10] relying on the Porohov Metric (which is equivalent to weak star convergence of measures [14, 4, 2]). To this end we assume a family $\mathcal{P}(\Lambda)$ of distributions P of growth rates for our population over a continuum of values Λ . We then use the data in a reformulation of the estimation problem

$$P_0 = \arg \min_{P \in \mathcal{P}(\Lambda)} J_{OLS}^n(P) = \arg \min_{P \in \mathcal{P}(\Lambda)} \sum_{i=1}^n |E[y(t_i; P)] - Y_i|^2, \quad (13)$$

where

$$E[y(t_i; P)] = \int_{\Lambda} E[y(t_i; \lambda)] dP(\lambda). \quad (14)$$

Here our formulation assumes that the data $\{y_i\}$ are the realizations for random variable observations $\{Y_i\}$ in an *i.i.d.* error process given by

$$Y_i = E[y(t_i; P)] + \mathcal{E}_i \quad (15)$$

where $E[\mathcal{E}_i] = 0$, and $\text{cov}[\mathcal{E}_i, \mathcal{E}_i] = \sigma_0^2$.

In our application here, one can use a continuum of growth rates $\Lambda = [0, \lambda_{max}]$. Since Λ with the Euclidean metric is a compact metric space, then the growth family $\mathcal{P}(\Lambda)$ is a compact metric space when taken with the Prohorov metric [2, p. 173]. Hence, since the least squares cost functional J_{OLS}^n is continuous in P , we are guaranteed the existence of a minimizer in (13) for the given $\{Y_i\}$.

The above problem is an infinite dimensional optimization problem which lends itself to a family of approximation problems. For a given finite family $\Lambda^L = \{\lambda_1, \lambda_2, \dots, \lambda_L\}$ of growth rates, we consider approximate families $\mathcal{P}^L(\Lambda^L) = \{P^L = \sum_{k=1}^L p_k \Delta_{\lambda_k} | p_k \geq 0 \text{ and } \sum_{k=1}^L p_k = 1\}$ where Δ_{λ_k} is the Dirac delta distribution with mass at λ_k . We identify the probability distribution P^L with a finite dimensional vector which we denote by the vector $\vec{p}^L = (p_1, \dots, p_L)$. The corresponding approximate problems are given by

$$P_0^L = \arg \min_{P^L \in \mathcal{P}^L(\Lambda^L)} J_{OLS}^n(P^L) = \arg \min_{P^L \in \mathcal{P}^L(\Lambda^L)} \sum_{i=1}^n |E[y(t_i; \Lambda^L, \vec{p}^L)] - Y_i|^2. \quad (16)$$

Again, properties of the Prohorov metric [2, p. 173] can be used to guarantee the existence of a minimizer P_0^L in (16) for the given data process $\{Y_i\}$.

During the last decade, we have developed a full theoretical framework related to convergence and consistency in problems for the Prohorov metric such as those being discussed here. This theoretical framework, called the Prohorov Metric Framework, or PMF for short, is summarized and discussed in the context of fairly general problems in [2, Chapter 5]. We summarize here portions of the theory that are relevant to our current problems. We

first discuss issues related to convergence as $L \rightarrow \infty$ of the approximations P_0^L as defined in equation (16) above. These results were first developed in [5] and subsequently used in other problems related to aggregate data.

We let $\Lambda = [0, \lambda_{max}]$ as discussed above and $\Lambda_D = \{\lambda_k^D\}_{k=1}^\infty$ be an enumeration of a countable dense subset Λ_D of Λ . We define $P_D(\Lambda_D)$ as the collection of all finite convex combinations $\sum_{k=1}^L p_k \Delta_{\lambda_k}$ of Dirac measures on Λ with atoms $\lambda_k^D \in \Lambda_D$ and rational weights $\vec{p}^L$. Then $P_D(\Lambda_D)$ is dense in $P(\Lambda)$. A proof of this result can be found in [5].

For each integer L , we define

$$P_D^L(\Lambda_D) = \left\{ P^L \in P_D(\Lambda_D) \mid P^L = \sum_{k=1}^L p_k \Delta_{\lambda_k^D} \right\}. \quad (17)$$

That is, for each L , $P_D^L(\Lambda_D)$ is the set of all atomic probability measures with nodes placed at the first L elements in the enumeration of the countable dense subset Λ_D . Note that $P_D^L(\Lambda_D) \subset P^L(\Lambda) \subset P(\Lambda)$ for all L . The following theorem provides the desired convergence result.

Theorem 3.1 *There exists a sequence of minimizers $\{P_n^L\}$ of $J_{OLS}^n(P)$ over $\mathcal{P}^L(\Lambda)$. Moreover, this sequence has at least one convergent subsequence. The limit P^{n*} (as $L \rightarrow \infty$) of such a subsequence minimizes J_{OLS}^n over $\mathcal{P}(\Lambda)$.*

This theorem follows immediately from the theoretical framework and convergence theorems of [12] and the results above. The power of this theorem is that it transforms an infinite dimensional minimization problem (over $\mathcal{P}(\Lambda)$) into an L dimensional optimization problem. In practice, L nodes $\{\lambda_k^D\}_{k=1}^L$ from the dense, countable subset $\Lambda_D = \{\lambda_k^D\}_{k=1}^\infty$ are fixed in advance and then the L weights p_k^L are estimated in a minimization problem.

We turn next to the question of consistency [2, Sec. 5.5] which relates to the question of what happens to the minimizer P^{n*} of $J_{OLS}^n(P)$ over $\mathcal{P}(\Lambda)$, as $n \rightarrow \infty$, (i.e., as an increasing amount of data becomes available). These results, in the context of probability measures, assures us that the measures P^{n*} converge in some sense to an underlying “true” probability measure P_0 . Under appropriate conditions (primarily that the data is collected in a manner that the “sampling points fill up the space” Λ and an additional identifiability assumption), one has (see [2, Thm 5.5.2]) that the P^{n*} , as $n \rightarrow \infty$, converge with probability 1 to a “true” measure P_0 . That is, $\text{Prob} \left\{ \omega \mid P^{n*}(\omega) \rightarrow P_0(\omega) \right\} = 1$.

In the above examples, the formulation of the objectives assumes that the observations $\{Y_i\}$ follow an absolute error model. That is, in an **absolute error model** we assume that the data, propagator counts $\{y_i\}$, are realizations of observations of the form

$$Y_i = E[y(t_i; \Lambda^L, \vec{p}^L)] + \mathcal{E}_i \quad (18)$$

where the $\mathcal{E}_i$ represent an *i.i.d.* error process with $E[\mathcal{E}_i] = 0$, and $\text{cov}[\mathcal{E}_i, \mathcal{E}_i] = \sigma_0^2$.

However, because the variance of our data set appeared to increase over time and multiple data points were available at each time (see Figure 2), we considered a **proportional error model** where the error in the experimental observations is proportional to the expected outcome. In this case, the data $\{y_i\}$ are assumed to be realizations of observations of the form

$$Y_i = E[y(t_i; \Lambda^L, \vec{p}^L)] + E[y(t_i; \Lambda^L, \vec{p}^L)]^\delta \mathcal{E}_i, \text{ where } \delta > 0. \quad (19)$$

As discussed in Appendix A, we found our fit was best described by a proportional error model with exponent δ between 0.4 – 0.7. For simplicity we chose $\delta = 0.4$ in our analysis below, but we note the results were not overly sensitive to this value. With the choice of our proportional error model, we adjust our objective from the OLS to the generalized least squares (GLS) formulation:

$$P_0^L = \arg \min_{P^L \in \mathcal{P}^L(\Lambda^L)} J(P^L) = \arg \min \sum_{i=1}^n \left(\frac{E[y(t_i; \Lambda^L, \vec{p}^L)] - Y_i}{E[y(t_i; \Lambda^L, \vec{p}^L)]^\delta} \right)^2. \quad (20)$$

While the original OLS problem can be readily solved with quadratic programming [6, 2], the GLS formulation cannot. In the analysis below we solve our inverse problem by using the Matlab `fmincon` function and constraining our parameters such that $\vec{p}^L$ satisfies $p_k \geq 0$ and $\sum_{k=1}^L p_k = 1$ for different choices for our possible growth rates.

3.3. Inference on Simulated Data

Before moving to experimental data, we demonstrate we are capable of estimating parameters from simulated data sets. We generated simulated data with the same proportional variance model ($\delta = 0.4$) that we chose to use with the experimental data sets and analyzed our ability to correctly predict growth rates under different conditions. In order to maximize consistency between simulations and experiments, we chose the same time points as our experiments and simulated 10 observations at each. We first considered a simulated data set with a unimodal growth rate distribution where we approximated the normal distribution ($\mu = 1.5, \sigma = 0.25$) by equally spaced growth rates in the range $\Lambda = [0.5, 2.5]$. We explored our inverse problem formulation varying the number L of nodes in our estimation from 5 to 100.

In Figure 3 we depict our simulated data set and exact input probability distribution. We see that even though the node set we used for estimation did not contain the same growth rates we used to generate the data set, the distribution predicted (center panel) exhibits the characteristics of our true distribution. In particular, the two cumulative distributions show good agreement; this is consistent with previous results from [7]. That is, as the number of nodes increases from $L = 10$ to $L = 100$ (Figures 4 to 7), we see that the estimated cumulative distributions (right panels) agree increasingly well with the true cumulative distribution, illustrating the convergence outlined above.

Figure 3: **Simulated Data: 5 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution function (blue).

Figure 4: **Simulated Data: 10 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

Figure 5: **Simulated Data: 25 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

Figure 6: **Simulated Data: 50 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

Figure 7: **Simulated Data: 100 Nodes.** (Left) Simulated propagator counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

We next generated simulated data with only two nodes: $\lambda_{\text{low}} \approx 0.90$, (30% of population), and $\lambda_{\text{high}} \approx 2.1$, (70% of population) and solved inverse problems on the same node sets consisting of $L = 5, 10, 25, 50$ and 100 equally spaced nodes in $[0.5, 2.5]$ (see Figures 8 - 12). In all cases the distributions identified in solving the inverse problems were consistent with the true bimodal distribution, although the location of λ_{low} was not always correctly predicted. Based on these results, we conclude that the underlying characteristics of the resulting approximations - such as support and modality - in general, are likely to be in agreement with the true bimodal distribution.

Figure 8: **Simulated Data: 5 Nodes.** (Left) Simulated propagon counts (black) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) vs. estimated cumulative distribution (blue).

Figure 9: **Simulated Data: 10 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

Figure 10: **Simulated Data: 25 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

Figure 11: **Simulated Data: 50 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

Figure 12: **Simulated Data: 100 Nodes.** (Left) Simulated propagon counts (black circles) and best fit model output (blue). (Center) True probability density (black) compared to estimated probability density (blue). (Right) True cumulative distribution (black) compared to estimated cumulative distribution (blue).

4. Parameter Inference on Experimental Data

4.1. Estimating Growth Rates

Motivated by our results on simulated data sets, we next formulated and solved inverse problems for our two prion variants WT and R15 (see Figure 2). Using the proportional error model, we identified the optimal discrete probability distribution $\vec{p}$ for different numbers of equally spaced nodes between 0.5 and 2.5 (see Figures 13-18 solid lines). As the number of nodes increased from 4, we observed bimodality in the probability distribution functions (solid lines) for both WT (black) and R15 (red) variants. Indeed, it seemed that the delta function model was converging for each variant to a bimodal distribution. Based on the results from our simulated data, these estimated node weights appeared consistent with two distinct growth rates rather than a broad unimodal distribution of growth rates.

To determine bimodality more directly, we explored an alternative method of parameter estimation by allowing the location of λ to vary continuously throughout the interval $[0.5, 2.5]$ and identifying the optimal two node distribution of λ (see Figure 19). For reference, we also show the continuously varying values of λ and their associated probabilities in dashed lines in Figures 13-18 (dashed lines). Together, the consistent node placement between the continuous and discrete node models suggest that there are two distinct prion aggregate growth rates for WT and R15 variants. Intriguingly, the bimodal behavior was distinct between variants; assuming cells from the lower and higher λ subpopulations are consistent between variants, this suggests an overall higher rate of aggregate amplification for cells with the WT variant compared to the R15 variant (see Figure 19).

4.2. Model Selection Test to Assess Bimodality

In order to more robustly test the hypothesis of bimodality in the data sets, we performed a model selection test where we compared our model of two continuously varying values for λ with a model where we chose only a single value of λ . In Figure 20 we compare the location of the best choice for a single node (solid line) with the two node model (dashed lines) for both WT and R15. We note that the optimal single node chosen appears to be close to the weighted average for the two node model. To decide between the one or two growth rate model we performed a model selection test on the residual sum of squares (RSS) for our proportional error model (Equation 20) as described in [8, 3, 2]. Our statistic compares the difference between RSS for the one growth rate model (RSS_1 , one parameter λ) and the RSS for the two growth rate model (RSS_2 , three parameters $\lambda_1, \lambda_2, p_1$):

$$\frac{RSS_1 - RSS_2}{RSS_2} \sim \chi^2(2).$$

According to the χ^2 distribution, there was a significantly better fit for the two node model than the one node model for the WT variant ($p = 0.00335$) but not the R15 variant ($p = .1554218$). However, we note that the R15 data still seem to support the bimodality,

Figure 13: **4 Fixed Nodes.** Fit to 4 nodes equally spaced in $\lambda = [0.5, 2.5]$.

Figure 14: **9 Fixed Nodes.** Fit to 9 nodes equally spaced in $\lambda = [0.5, 2.5]$.

Figure 15: **17 Fixed Nodes.** Fit to 17 nodes equally spaced in $\lambda = [0.5, 2.5]$.

Figure 16: **33 Fixed Nodes.** Fit to 33 nodes equally spaced in $\lambda = [0.5, 2.5]$.

Figure 17: **65 Fixed Nodes.** Fit to 65 nodes equally spaced in $\lambda = [0.5, 2.5]$.

Figure 18: **129 Fixed Nodes.** Fit to 129 nodes equally spaced in $\lambda = [0.5, 2.5]$.

Figure 19: **Continuously Varying Nodes.** Based on the observations for fixed node models, it appeared that the underlying data had a strong preference for two specific nodes.

otherwise the parameter estimated for the two node model would have converged to the signal node distribution.

Figure 20: **1 vs. 2 Nodes.** In order to decide if the underlying data sets were bimodal, we compare the best bimodal nodes (dashed lines) with the best single node model for each case (WT = black, R15 = red).

4.3. Sensitivity Analysis

Having strong evidence in support of the two node model for the WT variant and weaker but still suggestive evidence for R15, we analyzed the sensitivity of our results to our parameter estimates. Following [3] we used the system of differential equations for the local sensitivity of observed propagion counts to changes from optimal parameters. The sensitivities were computed using the complex-step method [31]. In the analysis below we have chosen to represent the bimodal model with the following three parameters: p, λ_1, λ_2 where $\lambda_1 < \lambda_2$ and the probability of growth rate λ_1 is p and λ_2 is $(1 - p)$.

In Figure 21 we plot the normalized sensitivities in time. We first note that, as expected, the model with R15 parameters has lower predicted aggregate counts than the WT parameters (see Figure 21 (Top)). We note differences in the local sensitivity behavior of the variants to p and λ_1 . The behavior of WT is increasingly sensitive to p until around 7 hours where its sensitivity then begins to decrease in contrast to R15 where sensitivity appears to be slowly increasing then leveling out. Further in comparison to WT, R15 has minimal sensitivity to λ_1 . Both WT and R15 exhibit similar behavior to sensitivities in λ_2 where the sensitivity increases and then decreases after the maximum occurs at around 7 hours. We expect strongly decreasing sensitivity to λ values as each of the respective subpopulations approach the same aggregate carrying capacity, which likely explains the λ_2 behavior for both variants.

Figure 21: **Time Varying Local Sensitivities.** (Top) Model output x for optimal parameter choices. Note that despite the difference in the optimal parameters, the model behavior for WT (black) and R15 (red) are quite similar. (Second - Bottom) Time varying sensitivities for WT (black) and R15 (red).

4.4. Confidence Intervals

We next use our sensitivity matrices (of the previous Section) in estimating confidence intervals on our parameters. We follow the asymptotic theory approach given in [2] for estimating confidence intervals for generalized least squares problems. We first describe F , the sensitivity matrix, which is in this case 12×3 matrix for our 12 time points $\{t_1, t_2, \dots, t_{12}\}$ and 3 parameters. Recall, the output of our model is the expected number of aggregates observed at particular time t . For ease in notation, we redefine our model output for the bimodal case (originally defined in Equation (14)) as follows:

$$E[t; \lambda_1, \lambda_2, p] = pE[y(t; \lambda_1)] + (1 - p)E[y(t; \lambda_2)]. \quad (21)$$

The sensitivity matrix is given by

$$F[\lambda_1, \lambda_2, p] = \begin{bmatrix} \frac{\partial E[t_1; \lambda_1, \lambda_2, p]}{\partial p} & \frac{\partial E[t_1; \lambda_1, \lambda_2, p]}{\partial \lambda_1} & \frac{\partial E[t_1; \lambda_1, \lambda_2, p]}{\partial \lambda_2} \\ \dots & \dots & \dots \\ \frac{\partial E[t_{12}; \lambda_1, \lambda_2, p]}{\partial p} & \frac{\partial E[t_{12}; \lambda_1, \lambda_2, p]}{\partial \lambda_1} & \frac{\partial E[t_{12}; \lambda_1, \lambda_2, p]}{\partial \lambda_2} \end{bmatrix} \quad (22)$$

Our matrix of weights is given by a proportional variance model, where we chose $\delta = 0.4$:

$$W = \text{diag}(E[t_1; \lambda_1, \lambda_2, p]^{2\delta}, E[t_2; \lambda_1, \lambda_2, p]^{2\delta} \dots, E[t_{12}; \lambda_1, \lambda_2, p]^{2\delta}). \quad (23)$$

The standard errors of our parameter estimates are given by the square root of the diagonal elements of the covariance matrix Σ , where:

$$\Sigma = \sigma^2 (F^T[\lambda_1, \lambda_2, p] W F[\lambda_1, \lambda_2, p])^{-1}. \quad (24)$$

(If the precise value of σ^2 , the variance of the noise terms in the proportional error model, were known this would be used. Otherwise, as we do in this case, the value σ^2 is estimated from the data.)

Using the formulation above, we computed the standard error and 95% confidence intervals for the WT and R15 parameters (see Tables 1 and 2 respectively). We estimate parameters for the WT variant data with greater confidence than those for R15. This is evident in the smaller standard error (SE) and normalized standard error (SE/estimate). In addition, we note that for WT the 95% confidence intervals for λ_1 and λ_2 are disjoint, further supporting two distinct subpopulations each with their own growth rate. As we describe in greater detail, the most likely explanation for the two subpopulations in the yeast data is that mother and daughter cells have lower and higher rates of aggregate growth, respectively. Since roughly 40% of the population at steady-state consists of mother cells, the estimate of $p = 0.45$ is especially compelling.

For the R15 data, the confidence in our model parameters is lower than that for WT. Of particular concern is the 95% confidence interval for λ_1 not only contains the entire 95% confidence interval for λ_2 but allows for negative growth rates. While based on our inverse problem results, there is support for bimodal growth rates with the R15 variant, our data does not allow us to conclude this with statistical certainty.

Parameter	Estimate	SE	Normalized SE	Lower Bound	Upper Bound
p	0.45	0.20	0.45	0.06	0.84
λ_1	1.31	0.23	0.17	0.86	1.75
λ_2	2.15	0.19	0.09	1.78	2.52

Table 1: **WT Variant - Standard Error and 95% Confidence Intervals.**

Parameter	Estimate	SE	Normalized SE	Lower Bound	Upper Bound
p	0.25	0.25	1.00	-0.24	0.74
λ_1	0.82	2.58	3.15	-4.24	5.88
λ_2	1.84	0.10	0.06	1.64	2.04

Table 2: **R15 Variant - Standard Error and 95% Confidence Intervals.**

5. Discussion and Concluding Remarks

Our mathematical model and inverse problem formulation for prion amplification in a yeast colony indicate the possible presence of bimodality in the prion aggregate growth rate distribution. Since previous work has demonstrated a difference in prion aggregate composition between yeast cells in their first generation (daughters) and mature yeast cells (mothers) [23], it is attractive to speculate that these are the same two populations we have identified here. We note that, according to the nucleated polymerization model [22, 32], the amount of soluble (healthy) protein directly impacts the aggregate growth rate, so it is possible that differences in the amount of soluble Sup35 in mother and daughter cells could explain this difference.

We note that while our model provides statistical support for bimodality in the aggregate growth rates for the WT variant but not the R15 variant, it is possible that further experimental data might change this conclusion. In addition, we note that there were several simplifying assumptions we made in the analysis of our model that could be revisited. First, our model of cell division considered only exponential cell growth, a more realistic model of yeast cell growth would employ time varying rates of division. Second, our inverse problem formulation considered only heterogeneity in the aggregate growth rates, but we could consider simultaneously estimating growth rates and the aggregate carrying capacity. In future studies we intend to investigate these, and other, extensions.

6. Acknowledgements

This research was supported in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-15-1-0298, and in part by the National Science Foundation under NSF grant numbers DMS-0946431, DMS-1514929 and INSPIRE BCS-13-44279, and in part by the National Institutes of Health under grant NIGMS R01 GM069802.

References

- [1] Adriano Aguzzi and Magdalini Polymenidou. Mammalian prion biology: one century of evolving concepts. *Cell*, 116(2):313–327, 2004.
- [2] H. T. Banks, Shuhua Hu, and W. Clayton Thompson. *Modeling and Inverse Problems in the Presence of Uncertainty*. CRC Press, 2014.
- [3] H. T. Banks and H.T. Tran. *Mathematical and Experimental Modeling of Physical and Biological Processes*. CRC Press, 2009.
- [4] H.T. Banks. *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*. CRC Press, 2012.
- [5] H.T. Banks and K.L. Bihari. Modeling and estimating uncertainty in parameter estimation. *Inverse Problems*, 17:95–111, 2001.
- [6] H.T. Banks and Jimena L. Davis. A comparison of approximation methods for the estimation of probability distributions on parameters. *American Institute of Aeronautics and Astronautics*, 57:753–777, 2007.
- [7] H.T. Banks and Jimena L Davis. A comparison of approximation methods for the estimation of probability distributions on parameters. *Applied Numerical Mathematics*, 57(5):753–777, 2007.
- [8] H.T. Banks and B.G. Fitzpatrick. Statistical methods for model comparison in parameter estimation problems for distributed systems. *Journal of Mathematical Biology*, 28(5):501–527, 1990.
- [9] H.T. Banks and B.G. Fitzpatrick. Estimation of growth rate distributions in size structured population models. *Quarterly of Applied Mathematics*, 49:215–235, 1991.
- [10] H.T. Banks, B.G. Fitzpatrick, and Yue Potter, Laura K .and Zhang. Estimation of probability distributions for individual parameters using aggregate population data. In *Stochastic Analysis, Control, Optimization and Applications*, pages 353–371. Springer, 1999.
- [11] H.T. Banks, Kevin B. Flores, and Suzanne S. Sindi. On analytical and numerical approaches to division and label structured population models. *Applied Mathematics Letters*, 60:81–88, 2016.
- [12] H.T. Banks and Karl Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Birkhausen, Boston, 1989.

- [13] H.T. Banks, Karyn L. Sutton, W. Clayton Thompson, Gennady Bocharov, Dirk Roose, Tim Schenkel, and Andreas Meyerhans. Estimation of cell proliferation dynamics using cfse data. *Bulletin of Mathematical Biology*, 73(1):116–150, 2011.
- [14] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.
- [15] Brian S. Byrne, Lee J. and Cox, Diana J. Cole, Martin S. Ridout, Byron J.T. Morgan, and Mick F. Tuite. Cell division is essential for elimination of the yeast $[psi^+]$ prion by guanidine hydrochloride. *Proceedings of the National Academy of Sciences*, 104(28):11688–11693, 2007.
- [16] Lee J. Byrne, Diana J. Cole, Brian S. Cox, Martin S. Ridout, B.J. Morgan, and Mick F. Tuite. The number and transmission of $[PSI^+]$ prion seeds (propagons) in the yeast *saccharomyces cerevisiae*. *PLoS One*, 4(3):e4670–e4670, 2009.
- [17] Vincent Calvez, Natacha Lenuzza, Marie Doumic, Jean-Philippe Deslys, Franck Mouthon, and Benoit Perthame. Prion dynamics with size dependency–strain phenomena. *Journal of Biological Dynamics*, 4(1):28–42, 2010.
- [18] Chiara Cipollina, Marina Vai, Danilo Porro, and Christos Hatzis. Towards understanding of the complex structure of growing yeast populations. *Journal of Biotechnology*, 128(2):393–402, 2007.
- [19] Diana J. Cole, Byron J.T. Morgan, Martin S. Ridout, Lee J. Byrne, and Mick F. Tuite. Estimating the number of prions in yeast cells. *Mathematical Medicine and Biology*, 21(4):369–395, 2004.
- [20] John Collinge. Prion diseases of humans and animals: their causes and molecular basis. *Annual Review of Neuroscience*, 24(1):519–550, 2001.
- [21] Jason K. Davis and Suzanne S. Sindi. A study in nucleated polymerization models of protein aggregation. *Applied Mathematics Letters*, 40:97–101, 2015.
- [22] Jason K. Davis and Suzanne S. Sindi. A mathematical model of the dynamics of prion aggregates with chaperone-mediated fragmentation. *Journal of Mathematical Biology*, 72(6):1555–1578, 2016.
- [23] Aaron Derdowski, Suzanne S. Sindi, Courtney L. Klaips, Susanne DiSalvo, and Tricia R. Serio. A size threshold limits prion transmission and establishes phenotypic diversity. *Science*, 330(6004):680–683, 2010.
- [24] Hans Engler, Jan Prüss, and Glenn F. Webb. Analysis of a model for the dynamics of prions ii. *Journal of Mathematical Analysis and Applications*, 324(1):98–117, 2006.

- [25] Meredith L. Greer, Laurent Pujo-Menjouet, and Glenn F. Webb. A mathematical analysis of the dynamics of prion proliferation. *Journal of Theoretical Biology*, 242(3):598–606, 2006.
- [26] Mats Gyllenberg. The size and scar distributions of the yeast *saccharomyces cerevisiae*. *Journal of Mathematical Biology*, 24(1):81–101, 1986.
- [27] J. Hasenauer, D. Schittler, and F. Allgöwer. Analysis and simulation of division-and label-structured population models. *Bulletin of Mathematical Biology*, 74(11):2692–2732, 2012.
- [28] Courtney L. Klaips, Megan L. Hochstrasser, Christine R. Langlois, and Tricia R. Serio. Spatial quality control bypasses cell-based limitations on proteostasis to promote prion curing. *eLife*, 3:e04288, 2014.
- [29] Tuomas P.J. Knowles, Michele Vendruscolo, and Christopher M. Dobson. The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell Biology*, 15(6):384–396, 2014.
- [30] Christine R. Langlois, Fen Pei, Suzanne S. Sindi, and Tricia R. Serio. Distinct prion domain sequences ensure efficient amyloid propagation by promoting chaperone binding or processing *In vivo*. *In Review*.
- [31] Joaquim R. R. A. Martins, Ilan M. Kroo, and Juan J. Alonso. An automated method for sensitivity analysis using complex variables. In *AIAA paper 2000-0689, 38th Aerospace Sciences Meeting*, pages 2000–0689, 2000.
- [32] Joanna Masel, Vincent A.A. Jansen, and Martin A. Nowak. Quantifying the kinetic parameters of prion replication. *Biophysical Chemistry*, 77(2):139–152, 1999.
- [33] Peter Olofsson and Suzanne S. Sindi. A crump-mode-jagers branching process model of prion loss in yeast. *Journal of Applied Probability*, 51(2):453–465, 2014.
- [34] K.J. Palmer, M.S. Ridout, and B.J.T. Morgan. Kinetic models of guanidine hydrochloride-induced curing of the yeast $[\text{psi}^+]$ prion. *Journal of Theoretical Biology*, 274(1):1–11, 2011.
- [35] Thorsten Pöschel, Nikolai V. Brilliantov, and Cornelius Frömmel. Kinetics of prion growth. *Biophysical Journal*, 85(6):3460–3474, 2003.
- [36] Jan Prüss, Laurent Pujo-Menjouet, G. Webb, and Rico Zacher. Analysis of a model for the dynamics of prions. *Discrete Contin. Dyn. Syst. Ser. B*, 6(1):225–235, 2006.
- [37] Martin Ridout, Vasileios Giagos, Byron Morgan, Wesley Naeimi, Tobias von der Haar, and Mick Tuite. Modelling prion dynamics in yeast. *Int. Statistical Inst.: Proc. 58th World Statistical Congress, Dublin (Session IPS020)*, 2011.

- [38] Suzanne S. Sindi and Tricia R. Serio. Prion dynamics and the quest for the genetic determinant in protein-only inheritance. *Current Opinion in Microbiology*, 12(6):623–630, 2009.
- [39] Motomasa Tanaka, Sean R. Collins, Brandon H. Toyama, and Jonathan S. Weissman. The physical basis of how prion conformations determine strain phenotypes. *Nature*, 442(7102):585–589, 2006.
- [40] William Clayton Thompson. *Partial differential equation modeling of flow cytometry data from CFSE-based proliferation assays*. North Carolina State University, 2011.
- [41] Mick F. Tuite and Tricia R. Serio. The prion hypothesis: from biological anomaly to basic regulatory mechanism. *Nature Reviews Molecular Cell Biology*, 11(12):823–833, 2010.

A. Fitting Proportional Error Model

Since our experimental data had multiple observations at each time point, we considered fitting a proportional error model. In Figures 22 and 23 we determined **modified residuals** by subtracting the mean of the observed values and dividing by the mean to the appropriate power. That is, if we compute $\bar{y}_i$ is the average of all observed counts at the same time, then the modified residuals are:

$$M(y_i) = \frac{y_i - \bar{y}_i}{\bar{y}_i^\delta}. \quad (25)$$

In a proportional error model, we would expect the residuals to be uniformly dispersed at $\delta = 0$. However, we see that the data are more consistent with $\delta > 0$ as dispersal increases with nonzero δ . In the analysis of our data we chose the value $\delta = 0.4$ but the results were consistent for δ values between 0.4 and 0.6.

Figure 22: Modified Residuals vs Time for WT data.

Figure 23: Modified Residuals vs Time for R15 data.