# A Recursive Bayesian Estimation Method for Measuring Kinetics of Amyloid Fibrillogenesis

Laura Kellman[1], Dr. James Stroud[2], and Dr. David Eisenberg[2]

[1]Marlborough School, [2]Molecular Biology Institute, UCLA

## *About Me*

I have long been fascinated by math, and more recently by biology. When my high school presented the opportunity to participate in research at a local university two years ago, I looked for a project that could help me see how the mathematics I learned in the classroom could be applied to help us better understand questions in biology.

My advisor found Dr. David Eisenberg's lab at the Molecular Biology Institute at UCLA, a lab studying, among other things, amyloid fibers and Alzheimer's disease. I was introduced to Dr. James Stroud, who had developed a method applying Bayes' theorem to data of amyloid fiber formation. With James' help, I began working on writing code to create realistic simulations to test and refine the method. By testing the method and improving it where possible, we created a method that can be used to analyze real data.

When I began working in the lab, I knew next to nothing about Bayes' theorem or amyloid fibers. Diving into the project meant learning things from an area of mathematics completely foreign to me, and simultaneously attempting to apply it to the real world. With a lot of help from my mentor, I came to understand and even contribute to the project.

Mathematics was necessary at every step, from the method itself, to making realistic simulations for testing, to evaluating our results. This research has made me see how essential math is. Without it, we could not accurately analyze data or understand our results. Participating in this project has made math less abstract – it has applications far beyond getting an A in math class. I believe that math will be a powerful tool in whatever path I pursue.

*The Research*

## Introduction

My research is focused on developing a more accurate method of analyzing data of amyloid fiber formation. Amyloid fibers (Figure 1) form when proteins misfold and bind together, forming fiber like structures (Sipe & Cohen, 2000). These fibers are linked to a number of diseases, including Alzheimer's, Parkinson's, Huntington's, amyloidosis and Type II Diabetes (Ross & Poirier, 2004). Diseases associated with amyloid deposits, in plaques formed outside of cells or intracellular inclusions within, affect millions of people worldwide (Glenner, 1980; Friedrich et al., 2010; Wang, Maji, Sawaya, Eisenberg, & Riek, 2008). It is thought that either the fibers themselves or an intermediate on the pathway of fiber formation is responsible for the cell death seen with these diseases (Ross & Poirier 2004; Bucciantini et al. 2002).

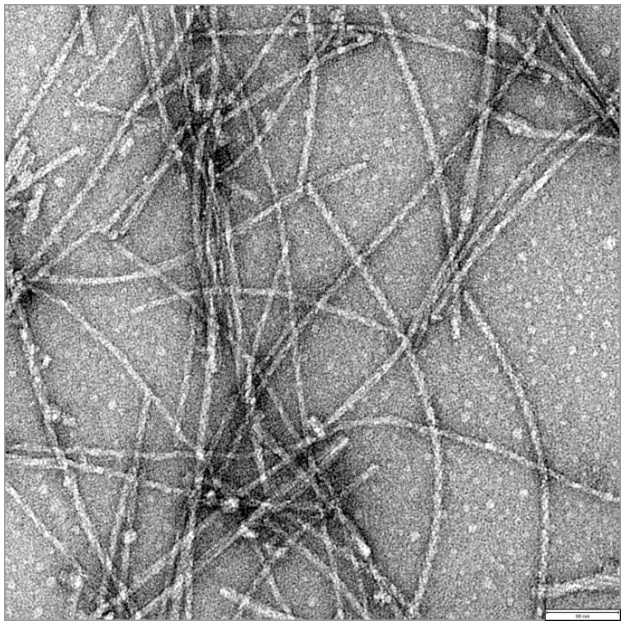

***Figure 1.*** **Electron microscopy of amyloid fibers.**

It is not known exactly how fiber formation proceeds. One prominent hypothesis is that protein monomers misfold and slowly aggregate, eventually reaching a critical size, called the nucleus. Once the nucleus is formed, its structure allows more protein to add very quickly, growing into fibers. Understanding the process by which these fibers forms may help in the discovery of inhibitor molecules to stop the reaction, and ultimately, in the development of medications to treat amyloid related diseases (Jarrett & Lansbury 1992; Jarrett & Lansbury 1993; Ferrone 1999; Morris et al. 2009).

Amyloid formation is commonly studied through fluorescence assays (Morris et al. 2009). These experiments use a molecule, such as thioflavin T, that emits light, or fluoresces, when binding to fibers. Measuring the amount of fluorescence provides a way to measure the relative concentration of fiber in a sample. A typical fluorescence experiment begins with protein monomer in solution with thioflavin T. Over time, the monomers aggregate. The thioflavin T fluoresces when fibers form. Graphing fluorescence versus time produces a logistic function, with a long flat section called the *lag phase* before fibers forms, followed by the *signal phase,* which involves a steep climb as fibers form rapidly after nucleation, and a leveling off after the available protein has been almost completely converted to fiber (Figure 2a).

The time it takes for the first fiber to form, the *lag time*, is important in relating models of the reaction to experimental data, and also in investigating inhibitor molecules to see if they slow down fiber formation (Morris et al. 2009).

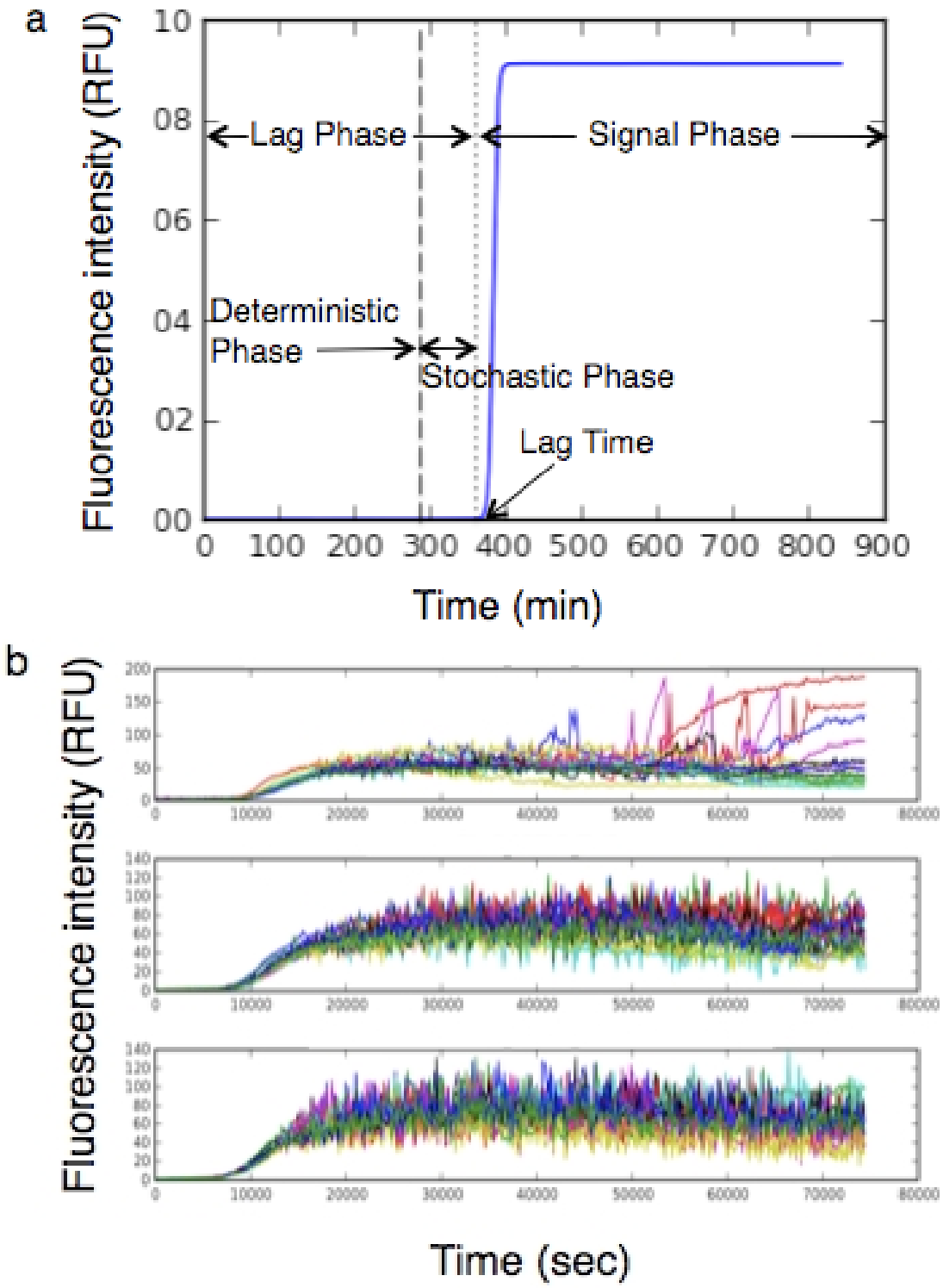


*Figure 2.* **Thioflavin T fluorescence over time of a fibrillogenesis reaction, showing concentration of amyloid fiber. (a) Noiseless reaction, showing the lag phase, or time before fiber is formed, consisting of a deterministic phase, in which monomers aggregate, then a stochastic phase, in which nucleation occurs. (b) Experimental results of fluorescence assays of amyloid beta, a protein related to Alzheimer's disease, showing the range of noise and data anomalies that can occur within the same experiment. The three panels show different runs of the experiment.**

Traditional methods of analyzing fluorescence data take a percentage, usually 10% or 50%, of the maximum fluorescence reached in the reaction, as the lag time. However, limiting conditions like low protein concentration or weak fluorescence mean that fluorescence experiments frequently produce noisy data that challenge analysis by these methods (Figure 2b). To improve the measurement of lag time from noisy fibrillogenesis data, we develop a method that uses recursive Bayesian estimation to more accurately analyze the data.

## The Method

We use a recursive Bayesian estimation scheme to decide the point of transition between the *lag phase* and the *signal phase* (Figure 2a). In this scheme, the series of fluorescence measurements is converted into a series of probabilities that the reaction is still in the lag phase. Central to the method is the use of Bayes' theorem to calculate the posterior probability $p_i(H|E)$ that the $i$th data point is in the lag phase:

$$p_i(H|E) = \frac{p_i(E|H) \cdot p_i(H)}{p_i(E)} \qquad \text{(Eq 1)}$$

where $p_i(H)$ is the prior probability of the hypothesis that the reaction is in the lag phase, $p_i(E)$ is the probability of getting the data points seen, and $p_i(E|H)$ is the probability of getting the data points seen given that the reaction is in the lag phase. Bayes' theorem uses the data and our knowledge of the reaction to give us a more sensitive calculation of the probability of the hypothesis (that we are in the lag phase) given the evidence (the fluorescence data).

### $p_i(H)$

The starting prior, $p_{i=0}(H)$, for the first cycle is empirically set to a low value (*e.g.* $10^{-4}$). Using a low starting prior gives the method sensitivity for cases where the lag phase is very short. The prior for the $i$th update cycle, $p_i(H)$, is the posterior probability from the previous cycle, $p_{i-1}(H|E)$, multiplied

by the mathematical constant, *e* (Euler's number):

$$p_i(H) = p_{i-1}(H|E) \cdot e \qquad \text{(Eq 2)}$$

Although essentially an empirical correction factor, the constant *e* can be thought to arise from a time-dependent *decay* of the probability that time point *i* is in the experimental phase. This decay is equivalent to a time-dependent *growth* of the probability that time point *i* is in the lag phase:

$$\frac{d\{p_i(H)\}}{dt} = \Lambda \cdot p_i(H) \qquad \text{(Eq 3)}$$

Solving this differential equation to get the new $p_i(H)$ gives:

$$\int \frac{d\{p_i(H)\}}{p_i(H)} = \Lambda \cdot \int dt \quad . \qquad \text{(Eq 4)}$$

To satisfy the differential equation, the posterior of the previous cycle, $p_{i-1}(H|E)$, is incorporated into the constant of integration:

$$\ln\{p_i(H)\} = \Lambda \cdot t + \ln\{p_{i-1}(H|E)\} \quad . \qquad \text{(Eq 5)}$$

The growth factor $\Lambda$ is empirically set to 1, directly resulting in the use of *e* in Equation 2. The value of 1 is chosen for simplicity although a range of values for $\Lambda$ produces reasonable performance of the method (data not shown). The unit of time in the analysis is the update *cycle*, making $t$=1 cycle in Equation 5, which then reduces to Equation 2. Empirically, we find that imposing this time-dependent change of the posterior, $p_{i-1}(H|E)$, between update cycles optimizes performance of the method.

***$p_i(E)$***

The marginal prior, $p_i(E)$, in Equation 1 is the probability of observing a value greater than or equal to the value $v_i$ at time point *i* given that $v_i$ comes from a window composed of a hypothetical lag phase combined with an immediately following signal phase of equal length. The limits of this window, $\xi_i$ and $\eta_i$, are described below. The probability $p_i(E)$ is calculated by first calculating an intermediate probability, $p'_i(E)$,

$$p'_i(E) = \int_{V=v_i}^{\infty} P_{norm}(V \mid \langle v \rangle_i, \sigma_i^2)\, dV \qquad \text{(Eq 6)}$$

given that the value $v_i$ comes from the normal distribution $P_{norm}(V \mid \langle v \rangle_i, \sigma_i^2)$ that has a mean

$$\langle v \rangle_i = \frac{\sum_{j=\xi_i}^{\eta_i} v_j}{N_i} \qquad \text{(Eq 7)}$$

and a variance of

$$\sigma_i^2 = \frac{\sum_{j=\xi_i}^{\eta_i} (\langle v \rangle_i - v_j)^2}{N_i} \quad . \qquad \text{(Eq 8)}$$

To prevent outliers and imprecision in estimating $p'_i(E)$ from causing the method to prematurely detect the lag time, we apply a Chauvenet filter to eliminate outlying data points, and also limit the probability $p_i(E)$ by testing it against a very low minimum value $\varepsilon_{min}$ (*e.g.* $10^{-10}$):

$$p_i(E) = \left\{ \begin{array}{l} \varepsilon_{min} \;\; if \;\; p'_i(E) < \varepsilon_{min} \\ p'_i(E) \;\; otherwise \end{array} \right| \quad . \qquad \text{(Eq 9)}$$

The window from which $p_i(E)$ is calculated has limits set so that the hypothetical lag phase is equal in number of data points (before Chauvenet filtering of the lag phase) to the hypothetical signal phase. This ensures that the window from which $p_i(E)$ is calculated is divided into two phases of equal numbers of data points: the hypothetical lag phase and the hypothetical signal phase. Enforcing this equality prevents excessively long lag phases from desensitizing the method or excessively long signal phases from sensitizing it.

***$p_i(E|H)$***

The operand $p_i(E|H)$ in Equation 1 is the probability of observing the value $v_i$ or greater at the time point $i$ given the hypothesis that the time point $i$ is in the hypothetical lag phase:

$$p_i(E|H) = \int_{\chi=v_i}^{\infty} P_{norm}(V \mid \langle v \rangle_{lag,i}, \sigma^2_{lag,i})\, d\chi \quad . \qquad \text{(Eq 12)}$$

The function $P_{norm}(V \mid \langle v \rangle_{lag,i}, \sigma^2_{lag,i})$ is the normal distribution function centered at the mean value of the hypothetical lag phase,

$$\langle v \rangle_{lag,i} = \frac{\sum_{j=\xi_i}^{i} v_j}{1+i-\xi_i} \quad , \qquad \text{(Eq 13)}$$

with a variance equal to that of the lag phase:

$$\sigma^2_{lag,i} = \frac{\sum_{j=\xi_i}^{i} (\langle v \rangle_{lag,i} - v_j)^2}{1+i-\xi_i} \quad . \qquad \text{(Eq 14)}$$

At the point when $p_i(H|E) \leqslant \varphi_h$, where $\varphi_h$ is the *hard cutoff* (*e.g.* $10^{-10}$), the method decides that the experiment is well into the signal phase. To estimate the time when the experiment makes the transition, a linear regression is applied to the probability series between the hard cutoff point, $i=h$, and the point, $i=s$, where $p_i(H|E) \geqslant \varphi_s$. The cutoff value $\varphi_s$ is the *soft cutoff* which is set to a relatively high probability (typically 0.1). The linear fit uses time as the x-value and $p_i(H|E)$ as the y-value, taking the x-intercept of this function as the uncorrected lag time, $T'$.

The uncorrected lag time is expected to be later than the time in the experiment when the first fiber appears because noise in the data series masks the signal from very small amounts of fiber. To compensate for this overshoot, a correction, $C$, is subtracted from the uncorrected lag time to give the measured lag time $T = T' - C$:

$$C = \frac{\alpha \cdot m^{\alpha}}{\left(\frac{v_h}{\sigma_h}\right)^{1+\alpha}} + k \quad , \qquad \text{(Eq 15)}$$

where $\bar{v_h}$ is the average value of a small window (*e.g.* seven points) around the data point $i=h$ and $\sigma_{lag,\,h}$

is the square root of the variance of the lag phase for data point $i=h$ where the hard cutoff was exceeded. The values for k, m, and α are empirically determined from simulated data, with $k = 7$, $m = 362$, and $\alpha = 0.9$. Although the correction is estimated from simulated data, it improves accuracy of the method across the range of simulation conditions and noise levels we tested and thus is expected to improve measurements from experimental data.

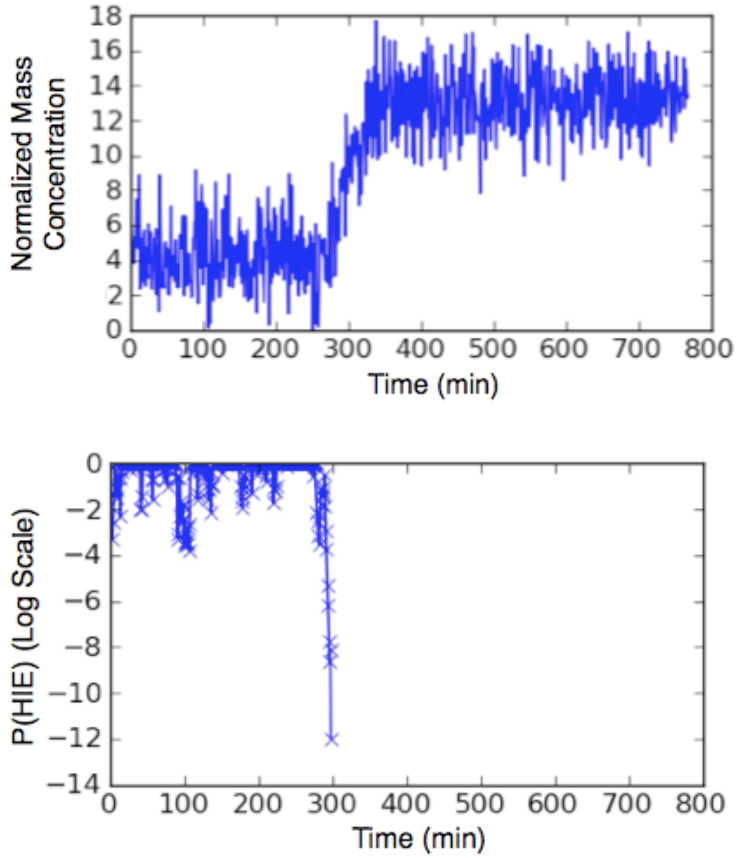


***Figure 3.*** **Simulated reaction and the corresponding probability series used to find point at which the reaction has entered the lag phase.**

**Simulation of Data for Testing**

To develop and evaluate the accuracy of our method and compare it to other methods, we create a simple polymerization reaction using a COPASI biochemical reaction simulator (Hoops et al. 2006). Using this, we created sets of reactions with a wide range of noise levels and a variety of data anomalies, similar to data seen in large scale fluorescence data. We then tested our Bayesian method

and two traditional methods on the data. The traditional methods took 10% and 50% of the maximum fluorescence reached as the lag time.

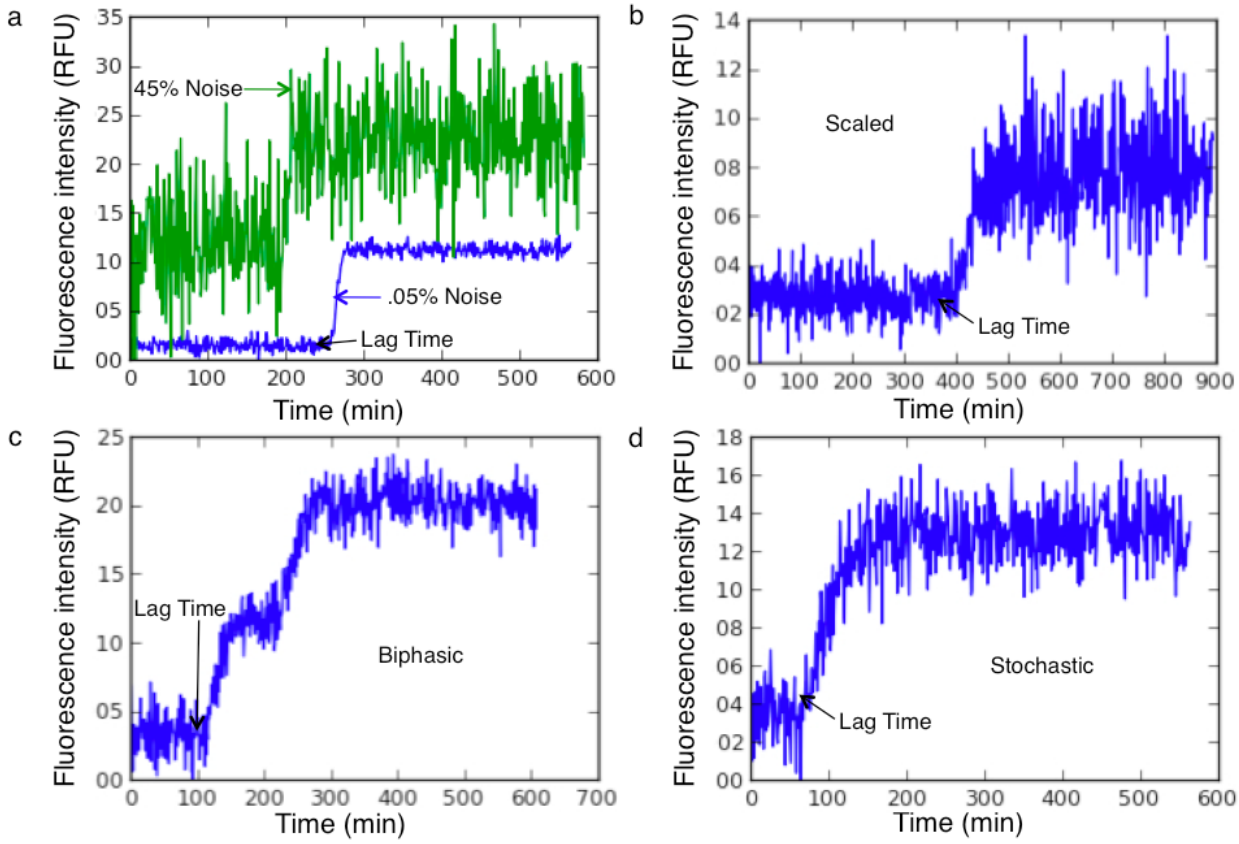


**Figure 4. Problematic data types. (a) Different noise levels, with noise levels, expressed as a fraction of the maximum signal of the simulated data, of .05 (blue), and .45 (green), and scaled to the spread of the data. (b) Example of scaled data, in which noise increases with signal, with a noise level of .25. (c) Example of biphasic data in which the first lag phase and pickup is followed by another pickup. (d) Example of stochastic data, in which there is no deterministic phase in the lag phase.**

**Evaluation of Methods**

The accuracies of the three methods were analyzed using a mean squared difference (MSD) test. For the Bayesian method, the real lag time for the *y*th simulated experiment, $t°_y$, is compared to the lag time, $T_y$, obtained from the method:

$$MSD = \frac{\sum_{y=1}^{Y} (t°_y - T_y)^2}{Y}, \qquad \text{(Eq 17)}$$

where *Y=100*, corresponding to the number of simulated reactions tested for each noise level. For the tenth-time and half-time methods, the value $t°_y$ in Equation 17 is replaced with the time that the noiseless simulated reaction reaches a tenth and half of the maximum value of the series, respectively. For each algorithm, the mean of the MSDs for each noise level is plotted in the left hand panel of Figure 5 for each data type.

We also evaluated the three methods using a Kolmogorov-Smirnov (KS) test to determine the probability that the distribution of lag times found by each algorithm followed an exponential distribution with the decay constant $\lambda$. The simulated data were generated from an exponential distribution, with each set of generated reactions before adding noise having a p-value of .99 or higher. In this test, a fitted deterministic phase length, $T_d$, is subtracted from each lag time of the set. The parameters $T_d$ and $\lambda$ and are obtained from a non-linear least squares fit of the algorithmic lag times to an exponential distribution.

$$P_{\exp}(t° = t) = \lambda \cdot e^{-\lambda \cdot (t - T_d)} \quad , \qquad \text{(Eq 16)}$$

For each algorithm, the P-value for each noise level is plotted for each data type in the right hand panel of figure 5.

**Results**

The Bayesian algorithm outperformed the other methods for a wide range of noise levels and data types (Figure 5).

**Discussion**

The Bayesian method produced performance superior to current methods of measuring the lag time from fluorescence data of amyloid fiber formation. This was true for the error measure as well as for the measure for recovering the underlying exponential distribution (Figure 5). This result held for normal data, and for the data anomalies tested: scaled, biphasic, and data with a shortened lag phase.

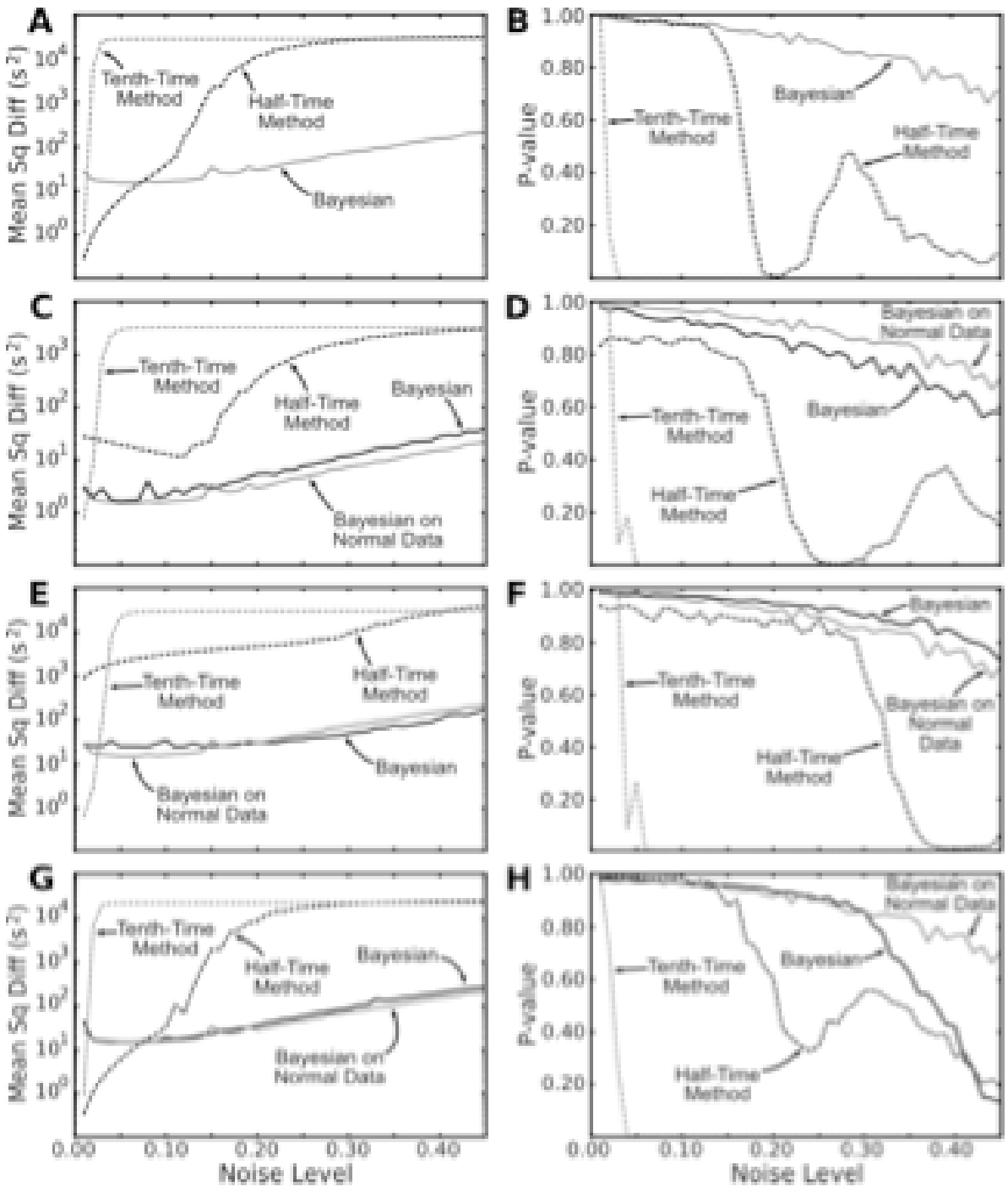


***Figure 5.*** **The Bayesian method outperforms other traditional methods for a wide range of data types and noise levels. (a) Mean squared difference between predicted and actual lag times (MSD) and (b) Kolmogorov-Smirnov (KS) test results for normal data. (c) MSD levels and (d) KS test results for scaled data. (e) MSD levels and (f) KS test results for biphasic data. (g) MSD levels and (h) KS test results for data with a shortened deterministic phase.**

The improved accuracy of our Bayesian method, particularly in its ability to recover the underlying distribution of lag times, will help in fitting fluorescence data to kinetic models of fibrillogenesis. The steady results for the data problems and noise levels suggest that the Bayesian algorithm will be robust to the range of anomalies that occur in large scale fluorescence experiments. Specifically, this method will be useful for finding the starting values for curve fitting to mass concentration in experimental analyses of fibrillogenesis, fitting expressions for the stochastic distribution of lag times to extract rate parameters, and calculating rate parameters from analytical expressions for the lag time.

The sensitivity of our Bayesian method will make it useful in studying fibrillogenesis experimentally. Additionally, since our method is fully autonomous, it will be useful in the processing and analysis of high-throughput data in automated settings. Current methods require sampling of both initial and final concentration levels, and backtracking to some point (e.g. 10% of maximum), requiring the whole data series. Our use of a Bayes'-optimal procedure makes the method ideal for large data sets where computational efficiency is crucial.

***References***


Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C., & Stefani, M. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416(6880), 507-511.

Ferrone, F. (1999). Analysis of protein aggregation kinetics. *Methods in Enzymology*, *309*, 256-274.

Ferrone, F. A., Hofrichter, J., & Eaton, W. A. (1985). Kinetics of sickle hemoglobin polymerization : II. A double nucleation mechanism. *Journal of Molecular Biology*, *183*(4), 611-631. doi:10.1016/0022-2836(85)90175-5

Friedrich, R.P., Tepper, K., Ronicke, R., Soom, M., Westermann, M., Reymann, K., Kaether, C., Fandrich, M. (2010). Mechanism of amyloid plaque formation suggests an intracellular basis of Aβ pathogenicity. *Proceedings of the National Academy of Sciences*, 107(5), 1942 -1947.

Glenner, G.G. (1980). Amyloid deposits and amyloidosis. The beta-fibrilloses (first of two parts). *The New England Journal of Medicine*, 302(23), 1283-1292.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). COPASI — a COmplex PAthway SImulator. *Bioinformatics* 22**,** 3067-74.

Jarrett, J. T., & Lansbury, P. T. (1992). Amyloid fibril formation requires a chemically discriminating nucleation event: studies of an amyloidogenic sequence from the bacterial protein OsmB. *Biochemistry*, 31(49), 12345-12352.

Jarrett, J. T., & Lansbury, P. T. (1993). Seeding "one-dimensional crystallization" of amyloid: A pathogenic mechanism in Alzheimer's disease and scrapie? *Cell*, *73*(6), 1055-1058. doi:10.1016/0092-8674(93)90635-4

Morris, A. M., Watzky, M. A., & Finke, R. G. (2009). Protein aggregation kinetics, mechanism, and curve-fitting: A review of the literature. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, *1794*(3), 375-397. doi:10.1016/j.bbapap.2008.10.016

Ross, C.A. & Poirier, M.A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine*, 10 Suppl, S10-17.

Sipe, J.D. & Cohen, A.S. (2000). Review: history of the amyloid fibril. *Journal of Structural Biology*, 130(2-3), 88-98.

Wang, L., Maji, S. K., Sawaya, M. R., Eisenberg, D., & Riek, R. (2008). Bacterial Inclusion Bodies Contain Amyloid-Like Structure. *PLoS Biol*, 6(8), e195.