STOCHASTIC MODELS FOR GENETIC NETWORKS

by

CONSTANTIN CRISTIAN CARANICA

(Under the direction of Jonathan Arnold and Suchendra Bhandarkar)

ABSTRACT

The analysis of gene regulatory networks has emerged as a leading paradigm for understanding how molecules inside cells communicate, process inputs, and coordinate responses to perform processes that sustain cell's life. Such networks are subject to internal noise, which occurs due to small number of molecules taking part in some reactions within a cell. This inherent stochasticity in regulatory networks can have major effects on a cell's fate. It can produce different phenotypes for genetically identical organisms and can lead to different cellular behaviors. Understanding the effects of noise and how cells adapt to it became an important issue in systems biology.

The aim of this thesis is to address two broad challenges posed by stochastic regulatory networks: (1) optimally finding the network characteristics/parameters; (2) quantifying the impact of noise on the observed dynamics. To address the inference of a stochastic network's parameters, two ensemble Markov Chain Monte Carlo (MCMC) methods were developed. They were used to find parameter sets that best describe the observed behavior of an oscillatory

gene network, the clock network of model organism *Neurosporra crassa*. Both methods used the average periodogram as a fitting criterion for simulating the behavior observed in the single cell data. In the first method, Parallel Tempering was successfully used in modeling the clock behavior of single cells observed in the dark (D/D). In the second method, a combination of Particle Swarm Optimization algorithms was used to simulate the light entrainment/synchronization of single cells by light observed in a series of 3 light/dark (L/D) experiments.

For the (D/D) and (L/D) experiments, we were able to test the Stochastic Resonance Hypothesis (SRH) to see whether the intrinsic noise was the main driver of the oscillations. The test showed that stochastic resonance is produced for a single optimal level of noise across all four experiments, a remarkable fact that shows how cells can adapt to the stochastic intracellular noise and use it to their benefit.

INDEX WORDS: Biological Clock, Systems Biology, Genetic Network, Particle Swarm Optimization, Markov Chain Monte Carlo Methods, Stochastic Noise, Stochastic Resonance Hypothesis

STOCHASTIC MODELS FOR GENETIC NETWORKS

by

CONSTANTIN CRISTIAN CARANICA

B.Sc., Bucharest University, Romania, 1997

Ph.D., Louisiana State University, USA, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

STOCHASTIC MODELS FOR GENETIC NETWORKS

by

CONSTANTIN CRISTIAN CARANICA

Major Professors:	Jonathan Arnold
	Suchendra Bhandarkar
Committee:	Heinz-Bernd Schüttler
	Lynne Billard
	Paul Schliekelman
	Liang Liu

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Jonathan Arnold who helped me tremendously during my graduate studies. Without his advice, patience, and encouragement, this thesis would not have been possible. Thank you for believing in me.

I would also like to thank my other committee members: Dr. Heinz-Bernd Schüttler for not getting tired of my numerous questions and for teaching me about rigor in scientific research, Drs. Suchendra Bhandarkar, Lynne Billard, Paul Schliekelman and Liang Liu for their valuable advice during my studies. Also, many thanks to the people in the Georgia Advanced Computing Resource Center (GACRC) for providing the help and equipment (GPGPUs) to carry out the simulations needed to complete this thesis. Dr. Shan-ho Tsai was always ready to go the extra mile to help me run my code on GPUs**.** Thank you very much.

Last but not the least, I want to thank my parents and my friend Liviu Mircea for their continuous support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In living cells, fluctuations of molecular numbers are inevitable under certain conditions. On one hand, such random fluctuations (noise) may impair signal propagation and hamper the coordination of cellular activities. On the other hand, noise in gene expression introduces phenotypic heterogeneity in an isogenic population, which may facilitate cellular differentiation or may be beneficial in heterogeneous environments. Total noise is typically divided into two components: intrinsic and extrinsic. Intrinsic noise, by definition, originates in the randomness associated with discrete, rare biomolecular events (e.g., mRNA synthesis), when few molecules are involved. The remaining noise, which measures fluctuations in the regulation of a gene, is lumped together as extrinsic noise. Cells have learned to adapt to different levels of noise and started to use it to their advantage. For instance, circadian rhythms, which provide internal daily periodicity, are used by a wide range of organisms to anticipate daily changes in the environment. These organisms generate circadian periodicity by similar biochemical networks within a single cell. A model based on the common features of these biochemical networks shows that a circadian network can oscillate reliably in the presence of stochastic biochemical noise even when cellular conditions are altered.

Ability to resist such perturbations might impose several constraints on the oscillatory mechanism. To be able to function reliably in the presence of internal noise means that a very robust oscillatory mechanism has evolved in the cells. Internal noise is very important when there are just a few numbers of molecules in the system, as it usually happens in the cell. To be able to display an oscillatory behavior with the same period in the presence of just few numbers of molecules and in the noisy cellular environment might seem puzzling at first.

In this thesis we try to unravel how circadian rhythms are maintained and affected by the intrinsic cellular noise in the clock network of a model organism, the filamentous fungus *Neurospora crassa*, Predicting and understanding the dynamics of a genetic network describing how the clock functions is a challenge. The problem facing biologists trying to understand such a genetic network is that the model parameters are mostly unknown, and the experimental data are noisy and limited. Most genetic networks, such as that for the biological clock, are part of much larger modules controlling fundamental processes in the cell, such as metabolism, development, or response to environmental signals. We use ensemble methods to develop models that simulate the stochastic behavior of the clock network. Markov Chain Monte Carlo (MCMC) methods and genetic algorithms are used to generate random samples of models. For each model sampled, we use Gillespie algorithm to describe how the clock network evolves in time. We develop novel parallel algorithms on the General Purpose Graphical Processing Unit (or GPGPU) to deploy these ensemble methods.

The organization for the rest of this dissertation is as follows.

In chapter 2 we develop two ensemble methods, one using Metropolis-Hastings algorithm and one using Parallel Tempering algorithm to simulate the behavior of the clock network observed

in the dark. As a goodness-of-fit measure we use the chi-square value given by the distance between average periodogram of the simulated trajectories and average periodogram of the observed cells. We also test the Stochastic Resonance Hypothesis and compare the parameters obtained in our stochastic models with those obtained from a deterministic model.

In chapter 3 we talk about 4 measures of phase and see how the continuized Hilbert phase measure can help us study the synchronization of these biological oscillators. We will also see that this phase measure provides us with a goodness of fit test independent of the test based on average periodogram.

In chapter 4 we study the light entrainment by the clock network observed under dark/dark and 3 light/dark regimes. We propose an ensemble method based on two genetic algorithms to fit the 4 data sets. The proposed method fits the data very well. We also test the Stochastic Resonance Hypothesis once again and discover that there is a single optimal level of noise for all 4 experiments. We will see stochastic resonance in full display.

Chapter 5 summarizes our findings and indicates future work that can be done to improve our findings and algorithms.

CHAPTER 2

ENSEMBLE METHODS FOR STOCHASTIC NETWORKS WITH SPECIAL REFERENCE TO THE BIOLOGICAL CLOCK OF *NEUROSPORA CRASSA*

Caranica, C., A. Al-Omari, Z. Deng, J. Griffith, R. Nilsen, L. Mao, J. Arnold and H-B. Schuttler 2018 *PLoS ONE* 13(5): e0196435.

Reprinted here with permission of publisher

ABSTRACT

A major challenge in systems biology is to infer the parameters of regulatory networks that operate in a noisy environment, such as in a single cell. In a stochastic regime it is hard to distinguish noise from the real signal and to infer the noise contribution to the dynamical behavior. When the genetic network displays oscillatory dynamics, it is even harder to infer the parameters that produce the oscillations. To address this issue, we introduce a new estimation method built on a combination of stochastic simulations, mass action kinetics and ensemble network simulations in which we match the average periodogram and phase of the model to that of the data. The method is relatively fast (compared to Metropolis-Hastings Monte Carlo Methods), easy to parallelize, applicable to large oscillatory networks and large (~2000 cells) single cell expression data sets, and it quantifies the noise impact on the observed dynamics. Standard errors of estimated rate coefficients are typically two orders of magnitude smaller than the mean from single cell experiments with on the order of ~1000 cells. We also provide a method to assess the goodness of fit of the stochastic network using the Hilbert phase of single cells. An analysis of phase departures from the null model with no communication between cells is consistent with a hypothesis of Stochastic Resonance describing single cell oscillators. Stochastic Resonance provides a physical mechanism whereby intracellular noise plays a positive role in establishing oscillatory behavior, but may require model parameters, such as rate coefficients, that differ substantially from those extracted at the macroscopic level from measurements on populations of millions of communicating, synchronized cells.

2.1 INTRODUCTION

Gene regulation is an intrinsically stochastic process [1-3]. The low copy numbers of some molecules, such as genes, involved in gene regulation lead to a noisy time series of numbers of molecular species in a gene regulatory network within a single cell. This randomness can produce different phenotypes for genetically identical organisms [4, 5]and for a single transcription factor [3]. This randomness can also produce coordinated regulation of target genes [6], and for a combination of 2 or more transcription factors, combinatorial regulation by changes in relative pulse timing between transcription factors [7], and have a role in the evolution of genetic networks [8]. To measure this stochasticity and to extract information about the regulatory network from the numbers of molecular species over time has become a major challenge in systems biology[9, 10]. Recent progress in addressing this task has been due mainly to advances in high-throughput single-cell measurement techniques for measuring gene expression, yielding large data sets on gene expression in single cells and the development of computational models used to explain these data [11-15].

Computational models should be able to capture the main features of the experimental data, such as the histories of molecular species in a cell, and provide new insights about the biological process operating in single cells [16, 17]. To build such a model, a critical step is to quantify the many unknown parameters that characterize the behavior of a single cell [18]. For genetic networks describing single cells these parameters include, for example, reaction rate coefficients, initial molecular numbers, mRNA/DNA ratios, and Hill coefficients. These quantities are difficult to measure directly on single cells. Usually only a few of those predicted

by the model are available from experiments, such as the levels of a few proteins or mRNAs, observed through their fluorescence [14, 19].

In the context of gene regulation, we need to simulate the behavior of whole gene networks in single cells to fit these models. One of the earliest methods to simulate stochastic gene networks was developed by Gillespie [20]. It allows exact simulation of stochastic biochemical networks, in principle for any duration of time and network size. By measuring the trajectories of many cells, we can find desired statistical summaries of the period, phase, and amplitude for the time series of molecular numbers in a cell. By comparing these with analogous summaries generated by a stochastic model, we can infer parameters of the underlying stochastic process. The only drawback of Gillespie's method is that it can take a long time to run and generating summary statistics with a high degree of accuracy can be computationally prohibitive. Approximate stochastic simulation methods can be used to speed up the computations but introduce additional errors that are difficult to account for in the model fitting process.

The τ-leaping methods [21] is such an approximation to the exact Gillespie algorithm. Instead of simulating a succession of reactions one at a time, a Poisson distribution is used to approximate the number of times each reaction is occurring in that time interval. This can decrease the simulation time significantly if certain conditions are met. But the Poisson approximation introduces additional error, and supplementary computations are then needed to verify that the approximation is accurate.

Maximum-likelihood methods have also been used for fitting stochastic networks [22]. These methods select those parameter values that maximize the likelihood that the model generates the observed data. The main difference among these approaches is the way the maximum-likelihood estimator is calculated. Some of the methods use Markov Chain Monte Carlo Methods [23, 24], to provide a direct solution of the stochastic model's maximum likelihood estimator or linear approximations thereof [25]. Although these existing methods work well for small networks, they become too cumbersome for larger networks. Many of them produce only point estimates of the network parameters, which cannot capture the behavior of the system due to the noise in these point estimates. Other approximate methods of stochastic network identification have been recently proposed using either moment-closure or volume expansion methods to approximate the chemical master equation describing the stochastic network to simplify the fitting problem [26].

Ensemble methods solve this problem. Based on a bayesian posterior distribution or likelihood function, they produce large samples of parameter values consistent with observed data that can then be model-averaged to capture the system behavior [27]. Consequently, they can produce confidence intervals of the model parameters [28]. More recently, these methods have evolved into Approximate Bayesian Computation (ABC) and have been used successfully in other biological contexts [27, 29]. For large networks, ensemble-based parameter inference methods employ Markov Chain Monte Carlo (MCMC) simulations techniques to draw samples from the high-dimensional model parameter spaces [27, 30, 31]. These MCMC simulations are highly CPU time consumptive and one of the challenges is then to find efficient computational approaches to make these simulations feasible within reasonable computation time limits.

In modeling stochastic molecular time series data, it is important to notice that the individual random trajectories of molecule numbers cannot, in general, be compared directly to individual observed single-cell fluorescence time series. The stochastic variability of the individual trajectories in both model and experiment preclude a meaningful comparison. In the context of potentially oscillatory data, as expected in the biological clock system, it is also not useful to compare the average of model molecular time series to the average over the observed single fluorescent data from all cells. Both model and observed trajectories are typically randomly phase-shifted relative to each other and averaging them tends to cancel out the oscillatory part of the signal. Consequently, it is important to design meaningful summary statistics which preserve the information about the oscillatory signal, including oscillation periods, phases and amplitudes, when averaged over all cells and model trajectories, respectively [32-34].

One aim of this paper is to provide a fast, computationally feasible method for parameter inference in a stochastic oscillatory biochemical network and show its successful application to understanding one of the best studied biological clocks at the molecular level [35]. The method proposed uses different MCMC methods, such as Metropolis-Hastings and parallel tempering, to fit the average periodogram [36], also known as the power spectrum, of the model to the average periodogram of the data, where the average is taken over the periodograms of individual cell fluorescent trajectories. The periodogram is a summary statistic which preserves two of the relevant features of an oscillatory process, its amplitude and period. We use deterministic mass action kinetics to initialize the Markov chains with parameter values that produce small chi-squared values relatively fast on General Purpose Graphics Processor

Units (GPGPUs) [37]. Thus, we can rapidly obtain model parameter sets that capture the important periods and amplitudes in the data. These models can be further used to match the observed phases to test the adequacy of the models.

A second aim of the paper is to explore whether the oscillations in such a network, might actually be caused by or reinforced by the molecular noise in the cell through a Stochastic Resonance-like phenomenon [38-40]. Stochastic Resonance is a theory that arose in physics [38] to explain the behavior of physical oscillators. Under the Stochastic Resonance Hypothesis the stochastic intracellular noise is assumed to have a positive role in generating periodic behavior provided this noise is not too little or too large in magnitude. The key to the Stochastic Resonance Hypothesis is that the presence of oscillations has a nonlinear relation with the level of stochastic intracellular noise.

It is therefore important in explaining oscillations in a stochastic network to infer the impact that noise has on cellular mechanisms and to quantify how these mechanisms respond to different noise levels. Our model and methods to fit a stochastic network (Fig 2.1) were developed with these purposes in mind.

2.2 MODEL

Our models simulate a well-stirred biochemical system with N molecular species $\{S_1, S_2, \dots, S_N\}$ having discrete-valued molecular numbers given by $X = \{X_1, X_2, \dots, X_N\}$. These molecular numbers change in time through the firing of M reactions $\{R_1, R_1, \dots, R_M\}$. The state

of the system at a time t is given by the random vector $X(t) = \{X_1(t), X_2(t), \dots, X_N(t)\}$ with $X_i(t)$ being the number of molecules of species S_i at time t, $i = 1, \dots, N$.

Knowing the state of the system at time t, $X(t) = x$, we assign to each reaction R_j a propensity function $a_j(x)$ whose product with an infinitesimal time increment dt determines the probability that reaction R_j fires in the next infinitesimal time interval $[t, t + dt)$. These propensity functions $a_j(x), j = 1, \dots, M,$, are defined based on mass action kinetics, $a_j(x) = k_j b_j(x)$, where k_j is a kinetic constant specific to reaction $R_j$and $b_j(x)$ counts the number of ways reaction $R_j$can occur given state X. For instance, for mono-molecular, homo-bimolecular and hetero-bimolecular reactions $b(x)$ takes the form x_1, $x_1(x_1 - 1)/2$ and $x_1 x_2$, respectively. The sum of all propensities is denoted by $a_0(x)$.

The parameters we need to infer are initial molecular numbers and the kinetic constants, $\Theta = \{X_1(0), X_2(0), \dots, X_N(0), k_1, \dots, k_M\}$. Gillespie showed that from knowing these parameters we can build an exact sample trajectory of the network to find the molecular numbers at any later time. Thus, a parameter set Θ gives a model of the network's dynamics.

Our study is focused on a well-studied oscillatory network, the clock network of Neurospora crassa [28] . This network is presented in Fig 2.1. The species with superscript r denotes an mRNA; the ones in capital letters are proteins. The rest are genes. As shown in Yu et al. [28], we can consider the protein WC-2 to be constant, so we can ignore the species wc-2^1, wc-$2^{r1}$and the reactions in which they are involved. Also, wc-1^1 is constant. Thus, we reduce the network to 12 molecular species and 22 reactions, which makes our parameter space 34-dimensional. Earlier work [28] has shown that this is a good approximation. This particular

model is in one of two classes of negative feedback models for clocks in different organisms termed a Hill-type transcriptional repression model [41, 42].

Some essential features of the model are captured in the cartoon (Fig 2.1B). The genes *white collar-1* (*wc-1*) and *white collar-2* (*wc-2*) produce a heterodimer WCC = WC-1/WC-2, which activates the oscillator gene *frequency* (*frq*) and a *clock-controlled gene* (*ccg*). The FRQ oscillator protein in turn provides a negative feedback loop to deactivate WCC. The genes *wc-1* and *wc-2* encode the positive elements in the clock mechanism, and *frq* encodes a negative element[28]. The FRQ protein also appears to have a role in stabilizing the *wc-1* mRNA (*wc-1'*)[28].

2.3 MATERIALS AND METHODS

2.3.1 SINGLE CELL DATA OF NEUROSPORA CRASSA

Two single cell data sets were used [13]. One data set has 868 single cells; the second one as a replicate has 1591 single cells. These two data sets were generated through time-dependent oscillatory fluorescent measurement on single *N. crassa* cells encapsulated in aqueous droplets of ~100 um in diameter and physically separated from each other as a result. The measurements were through the use of a fluorescent recorder (mCherry) linked to a promoter on a *clock controlled gene-2* (*ccg-2*) [43]. We obtained one data set including the time series of 868 single cells over ten days, and another data set including the time series of 1,591 single cells over ten days. The second data set is attached as a supplementary excel file. An improved cell-tracking method was used to bring the data set to 1,644 cells from 1,591 cells as originally

described [13]. There are 563 time points per cell taken from time 0 to time 261.5 hours every half hour. Each single cell time series is Rhodamine B normalized, detrended, and bias-corrected [13]. Only 61^{st} to 540^{th} time points were used in the analysis to allow each oscillator the opportunity to reach a stable limit cycle and to maintain cell viability at the end of the series.

2.3.2 RESCALING FROM DETERMINISTIC TO STOCHASTIC UNITS

In previous work [28] our clock network was studied in a deterministic framework. An ensemble of oscillating network models quantitatively consistent with available RNA and protein profiling data was found. If we wish to use concentration results from these deterministic models as inputs in a stochastic framework, then we need to rescale the initial molecular concentrations in the deterministic model to molecular numbers in the stochastic model, that is non-negative integers counting molecules in Fig 2.1, while preserving the deterministic dynamics [44]. This conversion reduces to a change in measurement units for each species such that the total gene concentration of a species in the new molecular number units is 1 in a single cell, and the time-averaged mRNA and protein concentrations in molecular number units are equal to the observed RNA:DNA and protein:DNA ratio, respectively. The RNA:DNA and protein:DNA ratios were determined experimentally, as described below, and summarized in Table 1. The RNA:DNA and protein:DNA ratios used below were 128.7 and 412, respectively, averages from Table 2.1.

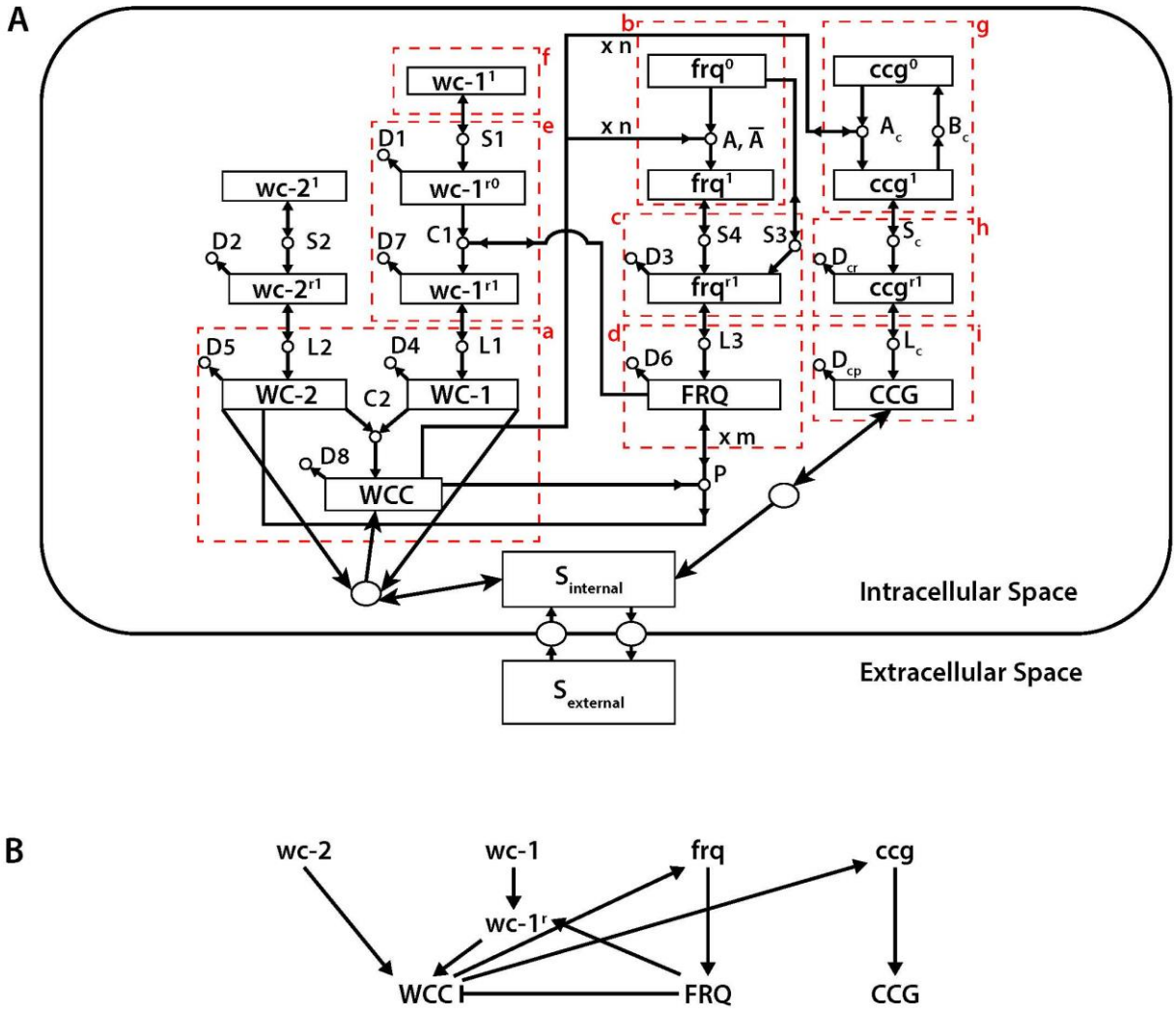

Figure 2.1 (A) The stochastic clock network of Neurospora crassa in a single cell. From[13, 28] The boxes outlined (in red) with dashed lines and labeled a through i in red, define regions of the network between which there is little or no net flow of molecules. (B) Cartoon of the stochastic network highlighting important features in panel A. Arrows are used to indicate a positive effect. A line with a bar (-|) is used to indicate a negative effect

Following an established notation [28], the 12 species concentrations [wc-1^{r0}], [wc-1^{r1}], [WC-1], [WCC], [frq^0], [frq^1], [frq^{r1}], [FRQ], [ccg^0], [ccg^1], [ccg^{r1}], and [CCG] are abbreviated here to $u_{r0}, u_{r1}, u_p, w, f_0, f_1, f_r, f_p, g_0, g_1, g_r, g_p$, respectively, with constant total

gene concentrations $f_G = f_0 + f_1$ and $g_G = g_0 + g_1$. To convert the parameters of the network from deterministic model units to molecular number units of counts of molecules in a cell we do the following: divide the network into separate components or "boxes" in such a way that two different boxes will be connected only through catalytic reactions (Fig 2.1). So, there will be no net flow of molecules between different boxes. Thus, the scales of measurement units between boxes can be varied independently without changing network dynamics. We obtain 9 boxes denoted by letters from *a* to *i*. They are:

a: w, v_p, u_p, b: f_0, f_1, c: f_r, d:f_p, e: u_{r0}, u_{r1}, f: v_r, g: g_0, g_1, h:g_r and i: g_p.

If we denote the model units by mu and real molecular number units by ru, then knowing the value of a concentration parameter expressed in mu, we need to convert it to a value that uses ru. We just need to find the ratio $\frac{mu}{ru}$. Each box will have its own $\frac{mu}{ru}$ ratio. For an arbitrary box *z*, we denote by $\left(\frac{mu}{ru}\right)_z$ its conversion ratio.

For the box containing a gene, like box *b* with species f_0, f_1, we have the values

$$\frac{f_0}{mu}, \frac{f_1}{mu} = (.356365, .0824576)$$

from the deterministic model in Table 1 (column 2), so we know

$$f_{G,mu} = \frac{f_0+f_1}{mu} = 0.4388226.$$

Because there is just one *frq* gene in the cell, we also have

$$\frac{f_0+f_1}{ru} = 1$$

Then

$$\left(\frac{mu}{ru}\right)_b = \frac{1}{f_{G,mu}} = 2.278825.$$

When applying this conversion factor, a gene is converted to the nearest whole gene so there is not a fractional gene.

For boxes with an mRNA we take the average value of an mRNA concentration parameter over a simulated trajectory obtained using the deterministic model and compare it to the RNA:DNA ratio. If the simulated trajectory contains a transient signal, we discard the corresponding part of the trajectory. Since the deterministic models display sustained oscillations, we want our mRNA deterministic values to come from the purely oscillatory part of the solution (not the transient part).

Thus, for box *c* we find

$$f_{r,mu} = \frac{\bar{f_r}}{mu} = \frac{1}{t_1 - t_0}\int_{t_0}^{t_1} \frac{f_r(t)}{mu}\, dt,$$

where $f_r(t)$ is the value of f_r at time t in the deterministic simulation of the network between times t_0 and t_1. In the time interval $[t_0, t_1)$ the deterministic trajectory traced 10 complete cycles.

Also,

$$\frac{\bar{f_r}}{ru} = R_{RNA:gene}$$

is the RNA:DNA ratio for *frq* species, namely 128.7.

This ratio is experimentally determined from Table 2.1. Then,

$$\left(\frac{mu}{ru}\right)_c = \frac{R_{RNA:gene}}{f_{r,mu}} = \quad 128.7/\ 0.02319352=5548.963.$$

Similarly, for boxes with a protein we will use Protein:DNA ratio of 412.1 of the corresponding species. All box ratios can be found in this way. For instance, for box d**,**

$$\left(\frac{mu}{ru}\right)_d = \frac{R_{Prot:DNA}}{f_{p,mu}} = \frac{412.1}{0.46295} = 890.1612$$

The ratios RNA:DNA and Protein:DNA were found experimentally from Table 2.1.

Then, to convert a molecular concentration given by a deterministic model to a molecular number we just multiply the molecular number by the conversion ratio of the box to which the species belongs. To change from $\frac{S}{mu}$ to $\frac{S}{ru}$, we need to multiply the former term by $\frac{mu}{ru}$, but this ratio depends on the box in which the species S resides. The value $\frac{S}{ru}$ is then rounded to the closest integer. If, as above, the deterministic model contains a transient signal, we discarded it. We took $\frac{S}{mu} \equiv s(t_0)$, that is concentration to be converted is the value of species S at the beginning of oscillatory part of deterministic trajectory of S.

To convert the reaction rates from model units to real units we use the law of mass action and the conversion ratios found at step 2.

We will show how the method works using L_3, the translation reaction to FRQ.

We have

$$f_r \xrightarrow{L_3} f_r + f_p.$$

Then

$$\frac{df_p}{dt} = L_3 * f_r.$$

Using model units, we have

$$\frac{df_p/mu_d}{dt/hr} = \frac{hr}{mu_d} * L_3 * \frac{f_r}{mu_c} * mu_c = L_3 * hr * \frac{mu_c}{mu_d} * \frac{f_r}{mu_c} = L_{3,mu} * \frac{f_r}{mu_c},$$

where *hr* stands for hour, our unit of time, and mu_d and mu_c are the model unit of concentration for species in box *d* and *c*, respectively.

$L_{3,mu}$ and $\frac{f_r}{mu_c}$ are the values of L_3 and f_r expressed in model units. They are found from the deterministic model.

When L_3 is expressed in molecular number units, we have

$$L_{3,ru} = \frac{L_3}{1/hr} = L_3 * hr = L_{3,mu} * \frac{mu_d}{mu_c} = L_{3,mu} * \frac{(mu/ru)_d}{(mu/ru)_c} = 3.02387\frac{890.1612}{5548.963} = 0.4851.$$

where $L_{3,ru}$ is the value of L_3 expressed in molecular number units.

Likewise, the other reaction rates can be converted from model units to molecular number units using the deterministic values and the conversion ratios found in step 2.

Note that when converting the reaction rates, we keep the ratios $\frac{dS}{dt}$ the same. Here S is the concentration of a species. We do not change the qualitative behavior of the system by conversion, just express it in different measurement units.

2.3.3 METHOD OF DETERMINATION OF PROTEIN:DNA AND RNA:DNA RATIOS BY SPECIFYING THE SCALE OF STOCHASTIC MODEL

The protein:DNA and RNA:DNA ratios in a cell were experimentally determined to set the scale parameters for the stochastic network. Protein, RNA, and DNA samples were extracted simultaneously from cultures of *Neurospora crassa*, strain FGSC 1858 "bd" (Fungal Genetics Stock Center, 4024 Throckman Plant Sciences Center, Kansas State University, Manhattan, KS 66506). The cultures were grown over 48 hours in the dark such that the total growth time was kept constant as previously described[45] under the "cycle 1" experiment. A kit from "Norgen Biotek Corporation,"(3430 Schmon Parkway, Throld, Ontario, Canada L2V 4Y6) was used to extract RNA, DNA and protein from the same sample. The kit used was Product # 47700, "RNA/DNA/Protein Purification Plus Kit." Their protocol was followed, including the step 1F, for cell lysate preparation for fungi. Samples were done from thirteen different time points spaced at 4-hour intervals over 48 hours. A total of three preps were done for each time point, with a usable sample detected in 2-3 of the preps.

The DNA and Protein amounts from each of these preps, were determined on a "Qubit 2.0 Fluorometer" instrument (ThermoFisher Scientific, 168 Third Avenue, Waltham MA 02451). The RNA concentration was determined using an Agilent BioAnalyzer RNA 6000 Nano chip (Agilent, Palo Alto, California). The amounts were converted to nanomoles [46], averaged, and then ratios RNA:DNA and DNA:Protein were calculated (Table 2.1).

2.3.4 STOCHASTIC SIMULATION ALGORITHM-DIRECT METHOD

For simulating exact trajectories of a network's temporal evolution, we used a variant of Gillespie's simulation algorithm called the direct method [13, 20]. Gillespie showed that knowing the state of the network at a time t, we can infer the exact distribution of the time of next reaction, $t + \tau$, and the probability of each reaction taking place at time $t + \tau$. Thus, we obtain an exact distribution of the state of the network at time$t + \tau$. The Direct method uses these distributions to sequentially sample the time of the next reaction and the reaction that occurs next. It works as follows.

Given a set of parameters $\Theta = \{X_1(0), X_2(0), \dots, X_N(0), k_1, \dots, k_M\}$ and a final time T, we do the following:

(1) Initialize the system, i.e. set $t = 0$ and $X = x = \{x_1(0), x_2(0), \dots, x_N(0)\}$..

(2) Calculate the propensities, $a_j(x), j = 1, \dots M,$, and their sum $a_0 = \sum_{j=1}^{M} a_j(x)$.

(3) Draw the random time step value to the next reaction, τ, as an exponential random variable with mean $1/a_0(x)$ and draw the type of the next reaction to be executed, j_{next}, as a discrete random variable with probabilities $\frac{a_j(x)}{a_0(x)}, j = 1, \dots M$.

(4) Update the state X assuming reaction $R_{j_{next}}$took place. Update the time, $t = t + \tau$.

(5) If $t < T$ go to step 2, else stop.

The Direct method yields a trajectory of the network state $\{x(t_0), x(t_1), \dots x(t_k)\}$ in the time interval $[0, T]$. The trajectory can be thought of as belonging to a single cell. We refer to

this trajectory as a Gillespie trajectory (of a single cell). Here $0 = t_0 < t_1 < \cdots < t_k < T$with t_i,i=1,…,k, being the reaction times of the reactions that fire before T. Such a trajectory completely identifies the network state at any time in the interval $[0, T]$.

2.3.5 THE FITTING METHOD

To analyze the initial behavior of the clock network we collected data on a *CCG* protein from 868 single cells. The fluorescence level of the *CCG* protein in each cell was recorded every half hour for 10 days [13].

As a first pass, we constructed a normalized periodogram for each cell, and then we calculated the average (over 868 cells) of these 868 periodograms. Through normalization we made the sum of normalized periodogram values equal to 1. Normalization enabled us to use periodogram values that are invariant to scaling [13]. We did not use this periodogram normalization in the more sophisticated analysis of the 1591 single cell data set where we applied a bias-correction to the average observed periodogram (See section below on removing the detection noise).

At the level of millions of cells, the level of *CCG* is circadian, i.e. cyclical with a period of ~24 hours. To obtain stationary time series we used the moving average method to remove the 24-hour linear trend from the original time series. The periodograms were calculated on detrended data [13]. We assume the average periodogram describes the dynamical behavior of *CCG* protein because it captures the periods and amplitudes in the system.

We selected this summary statistic in fitting the stochastic network because we looked for models with periodic behavior at the single cell level. This choice was first proposed to describe stochastic oscillatory networks near their Hopf bifurcation (*i.e*., a point in the parameter space where oscillations first appear), and the use of the periodogram as the statistic driving the fitting was successfully used in this context [36]. The periodogram captures two important features of an oscillation, amplitude and period. As we expect the oscillatory trajectory to be a mixture of sinusoids with only few of them being relevant, we think the important features of an oscillatory trajectory are embedded in its periodogram. Also, unlike other methods that try to match individual trajectories produced by a stochastic model [23] [47], we try to fit the average of these periodograms. Our view is that to compare two stochastic models it is better to use summary statistics that relates to an average of the stochastic trajectories rather than comparing individual trajectories for four reasons. One, the individual stochastic trajectories are very noisy. Two, averaging the periodogram of individual trajectories reduces this noise. Three, this fitting approach has already proven successful [36]. Four, fitting using 1000s of individual trajectories by the method of maximum likelihood has not proved computationally tractable.

2.3.6 MARKOV CHAIN MONTE CARLO (MCMC) METHODS

We used MCMC methods [13, 27, 28] to find sets of parameters in the stochastic network (Fig 2.1) that best describe the observed average periodogram of a collection of cells. In each Monte Carlo update we used Gillespie's direct method to simulate 1024 Gillespie trajectories

of the system state using a given set of parameters. Here the parameters are the 12 initial molecular numbers and 22 reaction rates as described earlier in Materials and Methods. We calculated the average periodogram of simulated trajectories and determined how well it matched the cell average periodogram. Then we updated a randomly chosen parameter using the Metropolis –Hastings algorithm. The 1024 simulated trajectories were run in parallel on a GPU.

The ensemble $\mathbb{Q}$ used to fit the average periodogram is

$$\mathbb{Q}(\Theta) = \Omega^{-1} \prod_f \frac{1}{\sqrt{2\pi\sigma_f^2}} exp\left(-\frac{\left(\overline{Q_f^{cell}} - \overline{Q_f^{sim}}\right)^2}{2\sigma_f^2} \right) = exp(-\chi^2/2)\Omega^{-1} \prod_f \frac{1}{\sqrt{2\pi\sigma_f^2}} \qquad (1)$$

where $\overline{Q_f^{cell}}$ and $\overline{Q_f^{sim}}$ are average periodogram values of cell and simulated trajectories, respectively, calculated at frequency f. The parameter σ_f^2 is the variance of cell periodogram values at frequency f and is determined experimentally by bootstrapping the periodograms of single cells. The form of the likelihood entails invoking the Central Limit Theorem. The associated $\chi^2 = -2ln\mathbb{Q} + constant$.

We initialized our Markov chains with working oscillatory networks describing the clock at the macroscopic level of 10^7 cells determined previously by the ensemble method [28]. For each chain, we took a set of parameters from the deterministic ensemble family given in Yu et al.[28] and converted it to a set of stochastic parameters as described above.

To make sure we are covering a broad region of the parameter space, we started 4 Monte Carlo chains, each one with a different set of parameters. After running these Monte

Carlo chains for about 74,000 iterations we ended up with chi-squared values of different levels, see Fig 2.2 A.

We concluded that each of these chains might have been trapped to a different local minimum. To avoid being stuck at a local minimum we introduced an alternate MCMC method, the parallel tempering algorithm[48, 49]. Each set of parameters used to start a Metropolis-Hastings chain was now used to start a parallel tempering algorithm (see Materials and Methods).

Our Metropolis-Hastings algorithms were designed as random walk algorithms. The target density was $\mathbb{Q}$. At iteration k we randomly picked a parameter x_i^k of the parameter set $x^k = \left(x_1^k, x_2^k, \dots, x_{34}^k\right)$ and updated it using a uniform proposal kernel $U\left(x_i^k - \alpha_i, x_i^k + \alpha_i\right)$, i.e. the proposal density was

$$q\left(x_i^{k+1} | x_i^k\right) = \frac{1}{2\alpha_i} I_{\left(x_i^k - \alpha_i, x_i^k + \alpha_i\right)}\left(x_i^{k+1}\right).$$

The proposed parameter set was $y^{k+1} = \left(x_1^k, x_2^k, \dots x_i^{k+1}, \dots, x_{34}^k\right)$. Then, we set

$$x^{k+1} = \begin{Bmatrix} y^{k+1} \; with \; probability \; \rho(x^k, y^{k+1}), \\ x^k \; with \; probability \; 1 - \rho(x^k, y^{k+1}) \end{Bmatrix},$$

where $\rho(x, y) = min\{1, L(y)/L(x)\}$.

It is well known that for random walk Metropolis-Hastings algorithms the step-widths α_i must be fine-tuned to ensure the chain is converging in a manageable time. While we tried to optimize the choice of the $\alpha_i's$, we noticed that it might take too long for some chains to

converge (Fig 2.2A). Parallel tempering algorithm avoids the calibration of these hyperparameters.

2.3.7 PARALLEL TEMPERING AS AN ENSEMBLE METHOD

The idea of a parallel tempering algorithm is to simulate K replicas of the original system, each replica being simulated at a different temperature.

So, each replica is a Markov chain having a tempered target distribution of the form

$$\mathbb{Q}_T(\Theta) = \Omega^{-1} \prod_f \frac{1}{\sqrt{2\pi T \sigma_f^2}} exp\left(-\frac{\left(\overline{Q_f^{cell}} - \overline{Q_f^{sim}}\right)^2}{2T\sigma_f^2} \right) = exp(-\chi^2/2)\Omega^{-1} \prod_f \frac{1}{\sqrt{2\pi\sigma_f^2}}$$

For high "temperature" T, the peaks of $\mathbb{Q}_T$ become flatter and broader, making the distribution easier to sample via MCMC methods. High-temperature replicas can sample large volumes of parameter space, whereas low-temperature chains are usually sampling from a local region of the parameter space which may trap them to a local minimum. Parallel tempering achieves superior results by allowing different replicas to exchange their states. Thus, high-temperature replicas ensure that lower temperature chains can access different regions of the parameter space.

The way that a parallel tempering run is set up is as follows. To a set of K replicas we assign temperatures from a grid $T_1 < T_2 < \cdots < T_k$, with $T_1 = 1$ corresponding to our target replica. Each replica explores its tempered distribution using an MCMC method. After a predetermined number of in-chain iterations, swaps between usually adjacent replicas are

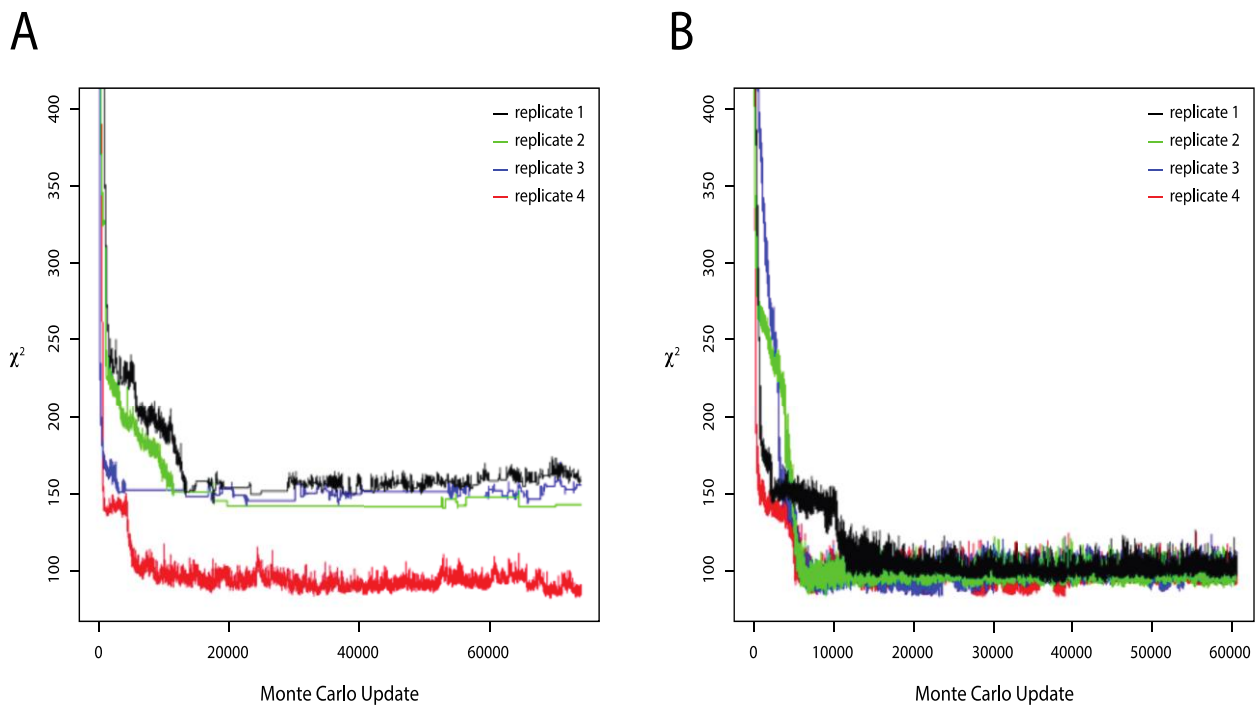

Figure 2.2 Monte Carlo simulations used for fitting average periodogram of the 868 single cell data. (A) 4 chains run using Metropolis-Hastings algorithm. (B) 4 chains starting with the same parameters but run using Parallel Tempering algorithm. No bias-correctio correction was applied (see Materials and Methods)

proposed. A proposed swap between replicas at temperatures T_i and T_j is accepted with probability

$$\rho_{ij} = min\left\{1, \frac{\mathbb{Q}_{T_i}(x_{(j)})\mathbb{Q}_{T_j}(x_{(i)})}{\mathbb{Q}_{T_i}(x_{(i)})\mathbb{Q}_{T_j}(x_{(j)})}\right\},$$

where $x_{(i)}$is the state of i^{th} replica. When a swap is accepted, the replicas exchange their positions in the parameter space; replica i takes configuration $x_{(j)}$ and j assumes the position at $x_{(i)}$. Since hottest replicas can sample big regions of the parameter space, then, if their locations propagate to the coldest replica, they can help it explore different regions of

parameter space. Thus, the goal in choosing an effective grid of temperatures is to make sure the hottest replicas can freely explore the parameter space, i.e. choose T_k big enough, and to choose the intermediate temperatures in such a way that $\rho_{i,j}{}'s$ are big enough to allow each replica to easily move between configurations sampled at different temperatures.

2.3.8 CHOOSING THE GRID IN PARALLEL TEMPERING

Now we show how we chose K, the number of replicas, T_K, the maximum temperature and the temperature grid $T_1 < T_2 < \cdots < T_k$.

Our method is based on a procedure described previously [50].

First, we chose the number of replicas $K = \lceil \sqrt{d} \rceil$, where d is the number of components of θ.

Then we chose maximum temperature T_K. We took $T_k = \frac{\chi^2(\theta^0)}{30}$, i.e. we divided the chi-square value of our initial parameter set $\theta^{(0)}$ by 30. The hottest replica will start with a chi-square value of 30. We wanted the hottest replica to have a high in-chain acceptance rate while having a not too flat distribution. The number of data points used in calculating chi-square values was 85. Then we ran Metropolis-Hastings for a replica with this temperature for 200 iterations.

If acceptance rate was outside the range (0.6, 0.75), then we changed T_K and ran M-H again for 200 iterations. We did this until the acceptance rate fell within the range (0.6, 0.75).

We made a linear grid with K temperatures and set a target swap rate of 0.4 for any two neighboring replicas.

Then we run the parallel tempering in the following way:

1) every replica does an update of its parameter set θ.

2) attempt swaps between replicas 1 and 2, 3 and 4, 5 and 6,…

3) attempt swaps between replicas 2 and 3, 4 and 5, 6 and 7,…

4) repeat steps 1), 2) and 3) 200 times

For every pair of neighboring replicas (*i, i+1*) we calculated

$$\mathbb{Q}_{i,i+1} = \frac{1}{N_{swap}^{i,i+1}} \sum_{l=1}^{N_{swap}^{(i,i+1)}} ln\left(\rho_{i,i+1}^{l}\right),$$

where $N_{swap}^{(i,i+1)}$ is the number of attempted swaps between replicas *i* and *i+1* and $\rho_{i,i+1}^{l}$ is acceptance probability of the l^{th} attempt at swapping *i* and *i+1*.

If $R_{i,i+1} = \left[\sqrt{\frac{\mathbb{Q}_{i,i+1}}{\ln(0.4)}}\right] > 0$, then we add to the grid $R_{i,i+1}$temperatures, evenly spaced between T_i and T_{i+1}.

We run this add-temperature process 3 times to make sure we have enough temperatures in the grid. Then we shifted the temperatures between T_1 and T_k as follows.

1) Run parallel tempering with the new temperature set doing the above steps 1), 2) and 3) 350 times.

2) For each temperature T_i calculate the flow fraction

$$f(T_i) = \frac{n_{up}(T_i)}{n_{up}(T_i)+n_{down}(T_i)},$$

where $n_{up}(T_i)$ and $n_{down}(T_i)$ is the total number of replicas that were drifting upward, respectively downward, when they visited T_i.

3) Linearly interpolate f between temperatures

4) Calculate g the inverse function of f

5) Change the temperature values from T_i to $T_i^{new} = g\left(1 - \frac{i-1}{K-1}\right)$.

This process of shifting the intermediate temperatures was repeated 3 times.

The shifting of temperatures was done to optimize the flow of replicas through the temperature grid. After that, we ran the parallel tempering algorithm for about 60,000 Monte Carlo updates, where by update we mean the steps 1), 2) and 3) described in the add-temperature process.

2.3.9 REMOVING THE DETECTION NOISE FROM THE AVERAGE PERIODOGRAM

A model to calculate the contribution of detector noise to the periodogram variance was derived under mild assumptions, and the detector noise, propagated to the periodogram in the supplement [13]. The assumptions in this calculation were that the total noise could be decomposed additively into stochastic intracellular noise and detector noise, that the stochastic intracellular noise component is independent of the detector noise component, and that the

detector noise in the Rhodamine B level measurements used in normalization of fluorescence measurements can be neglected [13]. The detector noise was independently quantified by replacing cells with fluorescent beads in the microfluidics experiments. In this way a universal system of measurement of gene expression at the single cell level was developed and used here [13].

Suppose we observe the fluorescence values in an experiment with *K* cells and *L* equidistant observation times $t_j = (j-1)\frac{T}{L}$. Here *j*=1,…,*L* and *T* is the duration of the experiment. Denote the frequencies of interest by $f_l = \frac{l}{T}$, *l*=0,…, [*L*/2]. Also assume the cells are treated with rhodamine B to reduce experimental noise.

Then, the detector noise contribution to the periodogram variance at frequency f_l is given by [13]:

$$(\sigma_l^e)^2 = \frac{2\sigma_\epsilon^2}{KL}\left[\langle Q(f_l)\rangle \gamma_Q(l) + Re\left(\langle R(f_l)\rangle \beta_Q(l)^*\right)\right] - \frac{\sigma_\epsilon^4}{KL^2}\left[\left|\gamma_Q(l)\right|^2 + \left|\beta_Q(l)\right|^2\right],$$

where $\langle Q(f_l)\rangle$ and $\langle R(f_l)\rangle$ are the population means of, respectively, average periodogram and average squared Fourier transform of the observed rhodamine B-normalized, detrended fluorescence time series. The quantity σ_ϵ^2 is the variance of the fluorescence signal due to the detector noise averaged over all cells and time points. This variance was determined experimentally by varying the incident light intensity and measuring the resultant variance in fluorescence of fluorescent beads replacing cells in a microfluidics experiment identical to that used for cells [13]. The quantities $\gamma_Q(l)$ and $\beta_Q(l)$ are functions of the weights used in the

moving-average detrending process [51], a standard for the literature. They do not depend of the observed fluorescence signals.

To compare the simulated average periodogram values with observed average periodogram values we need to remove the bias due to detection error. The bias formula is given by

$$Q^{bias}(f_l) = \frac{\sigma_\epsilon^2}{L}\gamma_Q(l).$$

When we try to fit the cell average periodogram, we compare $Q(f_l) - Q^{bias}(f_l)$ to $Q^{model}(f_l)$, with $Q(f_l)$ being the average periodogram over *K* cells calculated at frequency f_l and $Q^{model}(f_l)$ being the average periodogram of the simulated time series calculated at f_l . Unlike the analysis of the 868 single cell data set, the analysis of 1591 single cell data set with the bias correction does not normalize the periodograms. The whole fitting process is done on an absolute scale without periodogram normalization.

The ensemble to fit the average periodogram becomes

$$\mathbb{Q}_{bias-free}(\Theta) = \Omega^{-1}\prod_l \frac{1}{\sqrt{2\pi\sigma_{f_l}^2}} exp\left(-\frac{\left(Q(f_l)-Q^{bias}(f_l)-Q^{model}(f_l)\right)^2}{2\sigma_{f_l}^2}\right) =$$

$$exp\left(-\frac{\chi^2}{2}\right)\Omega^{-1}\prod_l \frac{1}{\sqrt{2\pi\sigma_{f_l}^2}} \qquad (2)$$

Here $\left(\sigma_{f_l}^c\right)^2 = \sigma_{f_l}^2 - \left(\sigma_{f_l}^e\right)^2$. The Central Limit Theorem is being invoked to obtain the approximate distribution of the average periodogram over > 1000 single cells needed to write down the ensemble in (2). See [13] for details.

2.4 RESULTS

2.4.1 PARALLEL TEMPERING AS OPPOSED TO METROPOLIS-HASTINGS IS SUFFICIENT FOR FITTING A STOCHASTIC MODEL WITH MANY PARAMETERS

As seen in Fig 2.2B the parallel tempering algorithms greatly improved the mixing of the chains and helped them escape local minima. These algorithms also converged much faster when compared to M-H algorithms. Their only downside is that they are slightly more computationally intensive and in general take longer to run when compared to the simple Metropolis-Hastings algorithms. However, parallelization can greatly reduce the additional time taken to run a parallel tempering algorithm when compared to Metropolis-Hastings algorithm. Better mixing and faster convergence more than compensate for the longer run time of each iteration.

For the chains in Fig 2.2A the average computation time per iteration were 25.76, 3.62, 0.33 and 2.41 seconds, respectively. For the four chains in Fig 2.2B the computation times per iteration were on average 41.98, 34.07, 30.11 and 16.89 seconds, respectively.

We see that, when using Metropolis-Hastings algorithm, computation time, convergence and mixing varies greatly with the initial parameters in Fig 2.2A.

Even though the computation time when using parallel tempering algorithm was on average longer compared to the time used by the Metropolis-Hastings algorithm, we see that the parallel-tempering chains quickly set to what is likely the region of the parameter space with lowest chi-squared value. Beginning with iteration 20,000 the chi-square value varied

between 82.87 and 125.5 for all chains. The mixing of these chains was excellent, with the swap acceptance rate between replicas at neighboring temperatures being larger than 0.5 for all parallel-tempering chains. The maximum number of replicas used in a parallel-tempering algorithm was 7. The maximum temperature was 10.

The best model given by the parallel-tempering algorithm gave a good fit to the average periodogram of the cells as can be seen in Fig 2.3. The model captures the main frequencies in the data. It does not fit the data very well at high frequencies, but at high frequencies data are very noisy.

2.4.2 THE FIT OF THE STOCHASTIC MODEL CAN BE IMPROVED SUBSTANTIALLY BY INCREASING THE NUMBER OF CELLS AND SUBSTRACTING THE DETECTION NOISE FROM THE PERIODOGRAM

To reduce the noise in the periodogram in Fig 2.3 we increased the number of isolated cells to 1591 in a replicate experiment and removed the detection noise in the periodogram. See Table 2. Without normalization of the periodogram the final χ^2 was 671.332 as opposed to 12,024.9, when the bias correction was not made. With 240 data points and 34 parameters, the chi-squared contribution per data point was 2.80, which is comparable to earlier work on a macroscopic scale [45].

2.4.3 THERE ARE STRONG SIMILARITIES IN THE RATE CONSTANTS BETWEEN THE STOCHASTIC NETWORK AND DETERMINISTIC NETWORK

The ensemble averages of parameters and their standard errors across the ensemble are reported in Table 2.2. Some of the parameters are key to sustained oscillations in the deterministic model [28]. Some of these include the activation (A) and deactivation rate (Abar), the decay rate of the stabilized *wc-1* mRNA (D7)[28], the decay rate (D6) of the FRQ protein [52, 53].

The initial parameter values were computed by MCMC Metropolis Hastings Method [28] for a deterministic model, in which the protein WC-2 was treated as constant to good approximation. The best parameter values in this ensemble were then converted to the molecular number units of the stochastic network in Table 2.2 (column 3). For example, in the stochastic network the initial numbers of molecules in each cell in Table 1 are given as opposed to concentrations used in the deterministic model. This conversion is described in Materials and Methods. Generally there is good agreement between the estimated rate constants estimated from the trajectories of 1,591 cells (column 6) and the initial guess from the deterministic model (column 3), but no such agreement exists for the smaller experiment with only 868 single cell trajectories. All discussion below is for the larger single cell experiment with 1,591 cells. In this discussion below *wc-1* and *wc-2* and their products are positive elements in the clock, while *frq* and its products are negative elements providing negative feedback to *wc-1* and *wc-2* and their products [35] in Fig 2.1.

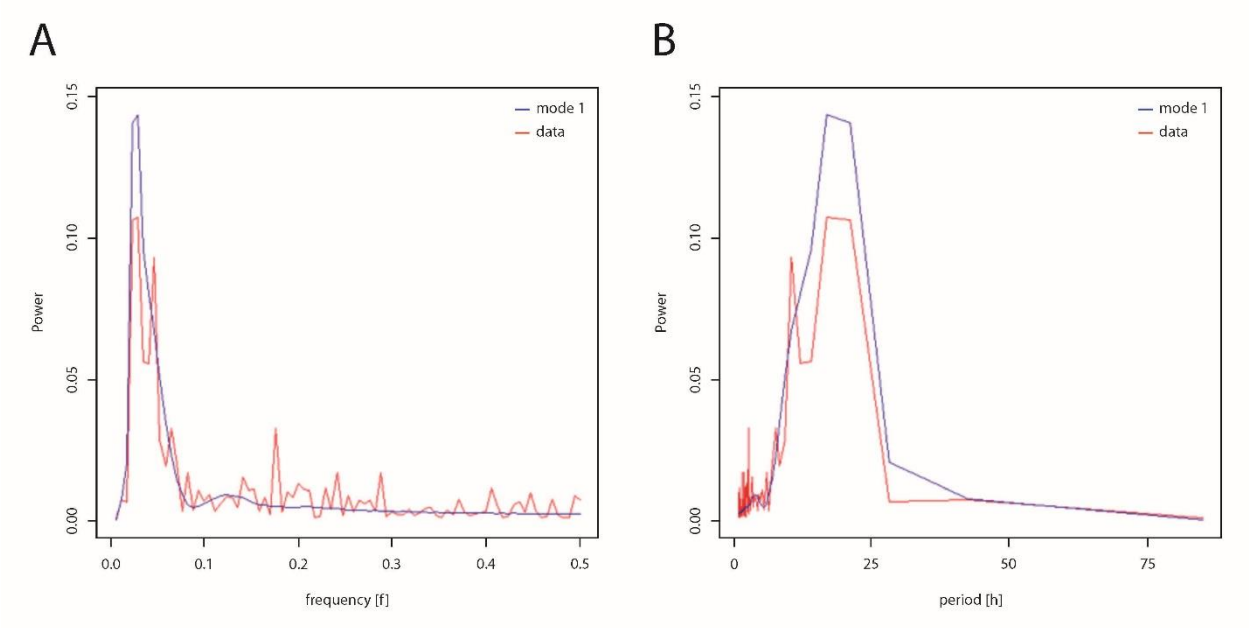

Figure 2.3 Fitting of average periodogram of the cells (red) by average periodogram produced by best parallel-tempering model(blue). (A) in frequency domain (B) using period

The decay rate of FRQ, D6, is thought to determine the period of the clock oscillator [52] and be involved in the phenomenon of temperature compensation in the clock. As the FRQ decay rate D6 decreases, the period is expected to increase. This coupling of period and FRQ may be more complicated[53]. The stochastic network's decay rate (.194 +/- .002) is in quite good agreement with the macroscopic deterministic model (0.152).

The decay rate of the stabilized *wc-1* mRNA, D7, in Fig 2.1 is thought to be critical determinant of clock oscillations [28]. The theory predicted (and experiment confirmed in previous work [28] that there should be small decay rate or a long half-life at the macroscopic level. The decay rate D7 in the stochastic network (2.131 +/- 0.090) appears somewhat higher than measured in the deterministic model (0.138). One possible explanation is that the

constraint on decay rates for isolated cells that experience stochastic intracellular noise in phase may be relaxed relative to that in a deterministic model at the macroscopic level. If the oscillations are actually supported by the noise, by way of some Stochastic Resonance mechanism [38], then this may impose less severe constraints on the decay rate D7.

Another critical parameter for oscillations to occur in the deterministic model is the activation (A) and deactivation rates (Abar) of the oscillator *frq* gene by WCC in Fig 2.1. These activation and deactivation rates in the stochastic model (2.56E-10 +/- 7.31E-12 and 1.590 +/- 0.036) tend to be qualitatively similar (6.06E-13 and .0.547), both being quite small.

Another critical parameter for oscillations in the deterministic model is the rate of deactivation of WCC (the P reaction) by FRQ [28]. The deactivation rate (P) in the stochastic model (2.7E-9 +/- 4.8E-11) is similar to that estimated at the macroscopic scale (3.12E-11), both being small.

The new MCMC method for specifying the parameters from periodogram tends to produce standard errors across the ensemble that are one to two orders of magnitude smaller than the ensemble means. The parameter values are quite tightly specified by the new estimation method and the use of at least 1,500 cells.

It is interesting to see what this model actually looks like. Various views of a single Gillespie trajectory are shown for one of the best fitting models (Fig 2.4). In panel A is shown the molecular counts of the positive element WCC. The counts are correlated with the activation of the FRQ gene in panel B, but the correlation is not perfect. Switching on the *frq* gene is a stochastic event in this model. The result of switching on the *frq* gene is

transcriptional bursts in its mRNA in panel C. There are at least 9 such bursts in Panel C over a 240 h interval. The width of these bursts as shown is wider than the time that a *frq* gene is active in a cell. In turn there are resulting even broader peaks in the FRQ protein production in panel D. The noisiest trajectory is what we actually see, CCG-2 in panel E. These views of a Gillespie trajectory give us a picture of the noise in a single cell under one fitted model in the model ensemble.

2.4.4 THE STOCHASTIC INTRACELLULAR NOISE LEVEL CAN BE EXPERIMENTALLY DETERMINED AS A PARAMETER IN THE MODEL

The mRNA to DNA ratios and protein to DNA ratios in conidia have been previously determined to be ~1:18:50[54]. Our ratios from Table 2.1 tend to be higher as 1:129:412, although the protein/RNA ratio is similar to previous reports. Here we report a higher amplification in the DNA -> RNA step of the Central Dogma. These ratios were then used to set the noise in the stochastic network (Fig 2.5). The network was subdivided into independent blocks that were only linked by catalytic reactions. Then the ratios in Table 2.1 were used to convert concentrations in the deterministic model into molecular numbers within the cell (as described in Materials and Methods for each independent block) as illustrated in Table 2.2.

The ratios of RNA to DNA and protein to DNA set the level of noise in the Gillespie trajectory of CCG-2. As the ratios get smaller the noise increases in the expression of CCG-2 in the Gillespie model trajectories for CCG-2 protein. The level of stochastic intracellular noise is then set by the amplification at each step in the Central Dogma.

Having experimentally determined these ratios, it is natural to ask how these ratios affect the goodness of fit of the model (Fig 2.6). We varied the ratios about the experimental values. A slightly better fit could be obtained by allowing the protein/DNA ratio to be slightly higher. The values of the chi-squared statistics across an ensemble would suggest that the fit of the model ensemble if fairly robust to variation in these ratios.

The robustness of the ensemble across different RNA/DNA and protein/DNA ratios can be understood by the fact that our stochastic models can be well approximated by a chemical Langevin equation, which in turn can be approximated by the ODE equations of the deterministic model. The chemical Langevin equation (CLE)[10, 55] is a stochastic equation that describes the rate of change of the state vector of molecular numbers, *X*, as follows:

$$\frac{dX(t)}{dt} = \sum_{j=1}^{M} v_j a_j\big(X(t)\big) + \sum_{j=1}^{M} v_j \sqrt{a_j\big(X(t)\big)}\Gamma_j(t) \quad (3)$$

The molecular numbers comprised in *X* are treated as continuous random variables. The first term on the right-hand side is just the rate function of the corresponding deterministic model, and the second term represents the noise due to the stochasticity of the reaction events. The v_j is the vector of changes in molecular numbers produced by the firing of reaction R_j and Γ_j's are statistically independent Gaussian white-noise processes. The crucial point here is that the strength of the noise term in the CLE *increases relative to* the deterministic rate term in the CLE, as the molecule numbers for RNA and protein *decrease*.

The change of RNA/DNA and protein/DNA ratios in Fig 2.5 was done by rescaling the RNA and protein numbers in such a way that the corresponding deterministic model remained

unchanged and only the noise term in the CLE was affected. So, if the deterministic term dominates the stochastic term, the noise level will not matter, hence the robustness.

2.4.5 THE HILBERT PHASE VARIATION BETWEEN CELLS PROVIDES AN INDEPENDENT TEST OF THE GOODNESS OF FIT OF THE MODEL

There are three quantities that characterize the periodic behavior of single cells, their period, amplitude, and phase [13]. Two of these quantities, period and amplitude, are captured in the periodogram used for fitting the model ensemble (Fig 2.3). The phase is functionally independent of the periodogram and hence independent of the first two quantities [13]; therefore, the phase can be used as a test of the adequacy of the model. The phase is not used in the fitting (Fig 2.3). The Hilbert phase can be calculated for each single cell trajectory and each Gillespie trajectory and measures the amount of cycles completed in a fixed period of time (and hence is a function of time). So, for example, 4 tires on a car would complete the same number of cycles in a fixed period of time and be in phase. This phase measure does vary with time and is a well-known measure of phase [56].

The histogram of Hilbert phases for single cells and Gillespie trajectories from the model ensemble are compared as a measure of goodness of fit (Fig 2.7). There are two sources of variation in the Gillespie trajectories, the random variation in phase between Gillespie trajectories of one model and the variation in phase between models in the ensemble of fitted models. The histogram of Hilbert phases in Fig 2.7 reflects both sources of variation. For each model, 1024 Gillespie trajectories were simulated, and each Gillespie trajectory has a Hilbert

phase. In addition, the process was then repeated for over 1000 models in the fitted ensemble (Fig 2.3) to generate all the values for the model histogram (Fig 2.7).

We see that the histogram for the Gillespie trajectories for over a thousand models in the fitted ensemble, covers the histogram measured on single cells over an 85-hour window. The difference between the two histograms is significant by a Kolmogorov-Smirnov (KS) nonparametric test. We carried out a KS-test of the difference of the two histograms ($P < 0.0001$) with the maximum difference in cumulative histograms being 0.1747 [57].

There are three possible reasons for the discrepancy between phase predicted and phase observed. The systems is experiencing Stochastic Resonance [38-40]. A second possible reason for the discrepancy is cell-to-cell synchronization by quorum sensing. It is unlikely that a quorum sensing mechanism is at work because the cells are physically isolated in droplets in Fig 2.7 [13]. A third reason could be that the Hilbert phase results are dominated by noise fluctuations.

2.4.6 IS THERE AN INTERMEDIATE OPTIMUM IN THE OSCILLATORY SIGNAL AS A FUNCTION OF THE STOCHASTIC INTRACELLULAR NOISE?

One possible explanation for the results on goodness of fit may be synchronization through the Stochastic Resonance mechanism acting on isolated single cells. Under this hypothesis there is a non-monotonic relation between peak height (i.e., signal strength) in the periodogram (Fig 2.3) and the estimated stochastic intracellular noise[38]. Here we examine how the periodogram varies, as the noise is varied (Fig 2.8).

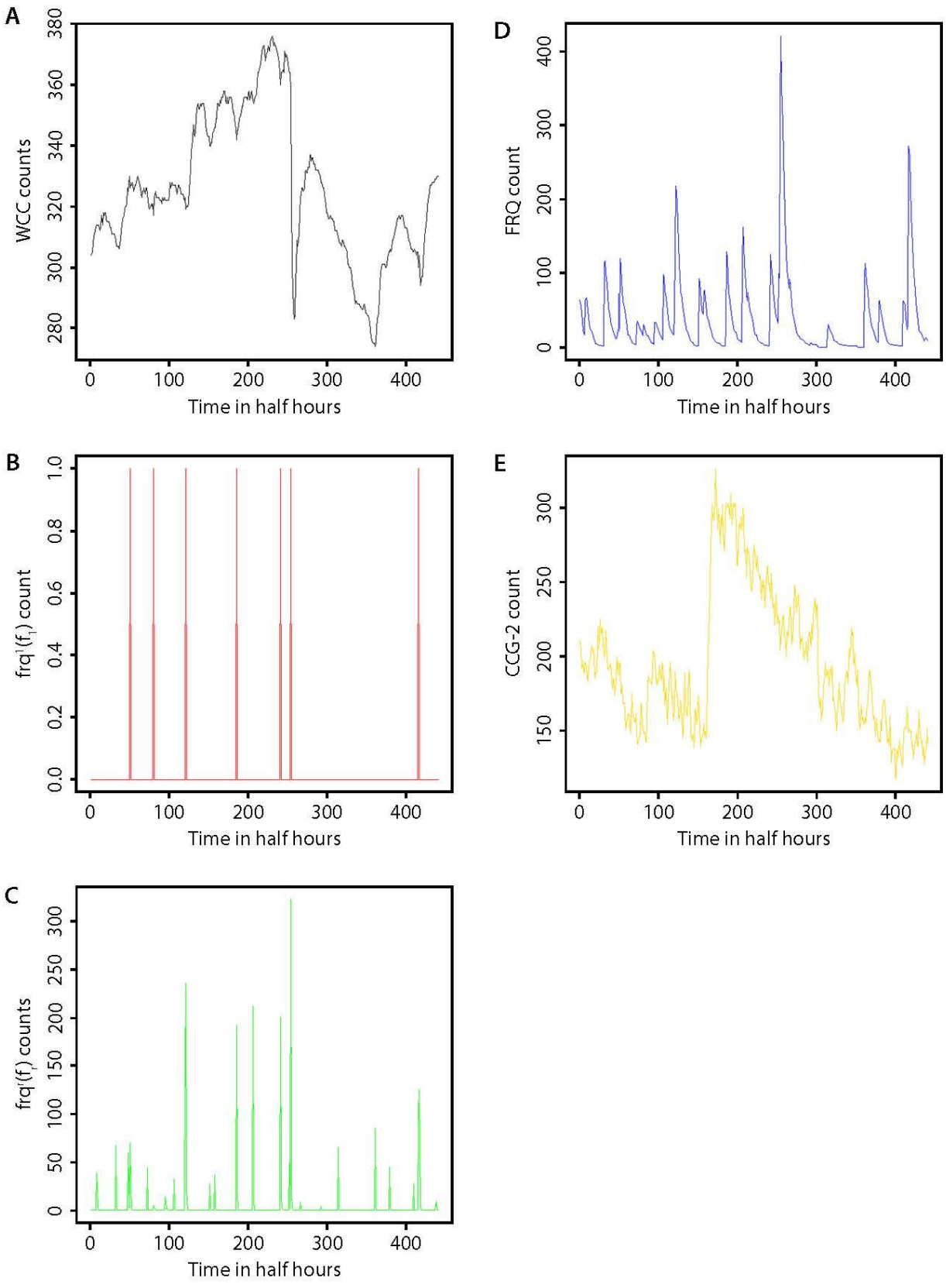
A
WCC counts
280
300
320
340
360
380
0
100
200
300
400
Time in half hours
D
FRQ count
0
100
200
300
400
0
100
200
300
400
Time in half hours
B
frq1(f1) count
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
Time in half hours
E
CCG-2 count
150
200
250
300
0
100
200
300
400
Time in half hours
C
frq1(f1) counts
0
50
100
150
200
250
300
0
100
200
300
400
Time in half hours

Figure 2.4 Multiple views of one stochastic trajectory for one of the best fitting models from the 1591 cell data set. The Gillespie trajectory shown in part is derived from a best fitting model in Table 2 after bias-correction. The chi-squared statistic for this fitted model was 671.332. Time 0 actually corresponds to 20 h, and the last time point, to 240 h. (A) Trajectory of the WCC count. (B) Trajectory of the active *frq* gene count f_1; the gene is either on or off; (C) Trajectory of the *frq* mRNA count f_r. (D) Trajectory of the FRQ protein count. (E) Trajectory of the CCG-2 protein count.

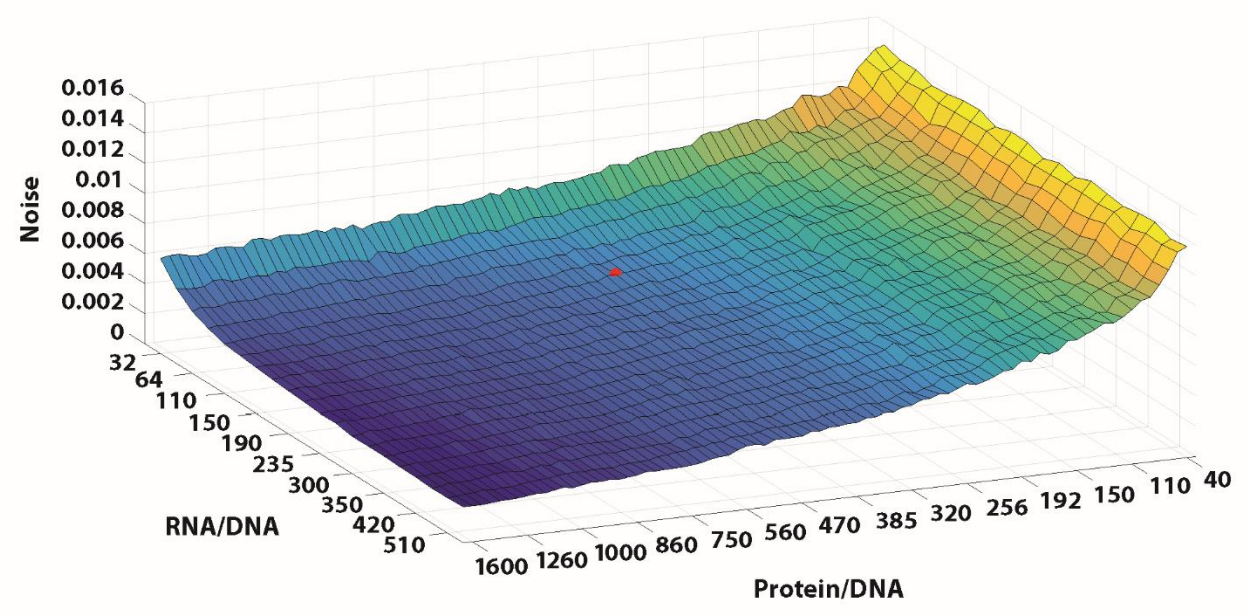

Figure 2.5 Stochastic noise in CCG-2 as a function varies systematically with hypothesized ratios of RNA/DNA and Protein/DNA ratios within a single cell. The total stochastic noise averaged over frequencies (f) in CCG-2 expression is computed from bootstrapping the 1024 Gillespie trajectories. The red dot denotes the experimentally determined ratios (see Table 2.1) and corresponds to a RNA/DNA and protein/DNA ratio of 128.7 and 412, respectively. The model selected was one with minimum chi-squared statistic based on the likelihood in equation (1) for 868 single cells.

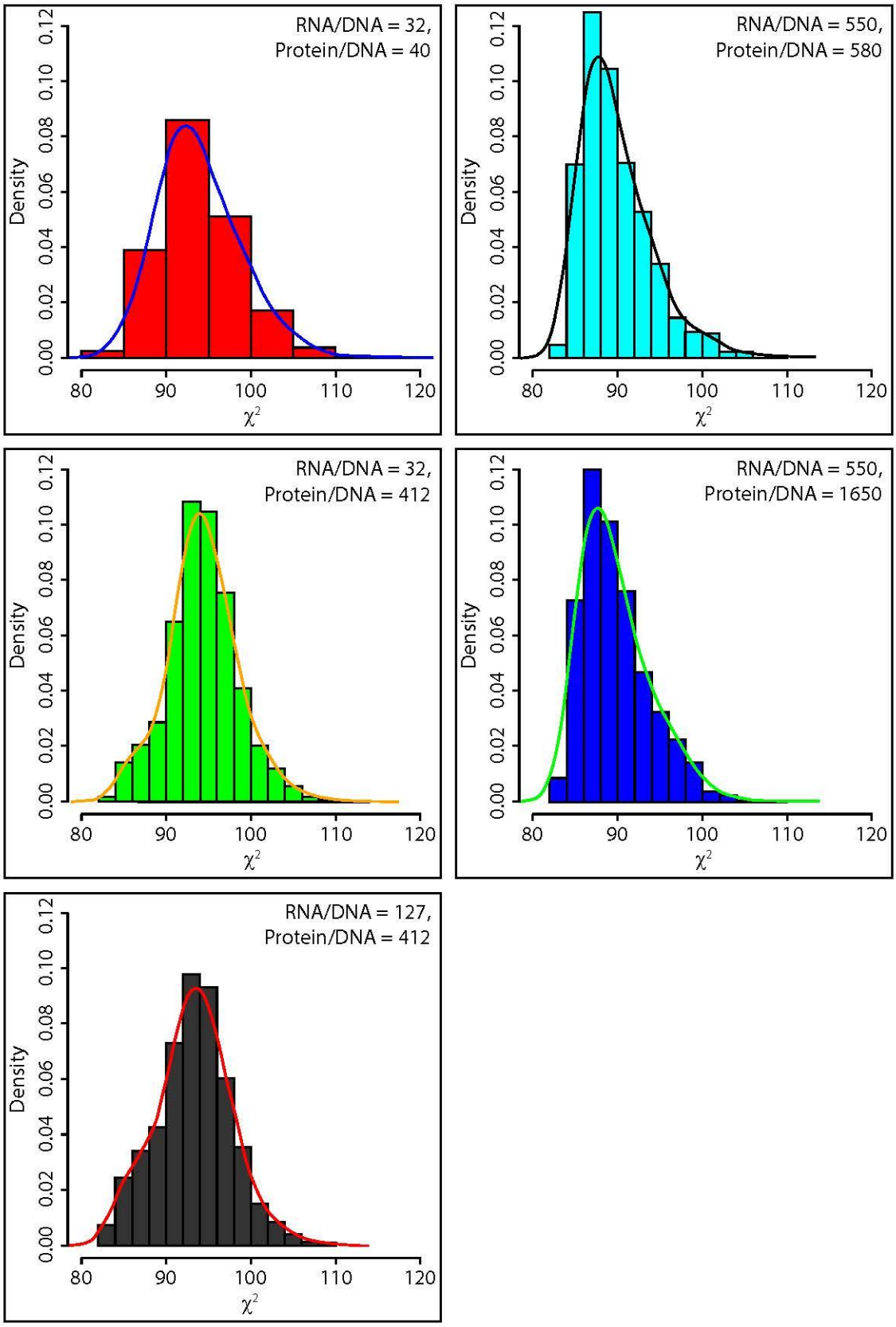
RNA/DNA = 32,
Protein/DNA = 40
RNA/DNA = 550,
Protein/DNA = 580
RNA/DNA = 32,
Protein/DNA = 412
RNA/DNA = 550,
Protein/DNA = 1650
RNA/DNA = 127,
Protein/DNA = 412
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
80
90
100
110
120
χ^2

Figure 2.6 The fit of ensemble of models ($\chi 2$) is robust to variation in the RNA/DNA and protein/DNA ratios. Each ensemble had at least at least 1,400 models derived from an accumulation run. The equilibration runs were done with parallel tempering as described in Materials and Methods. A smooth interpolation is provided for each histogram. The ensembles were derived from Eqn (1) for 868 single cells with no bias correction.

One of the clearest examples of the effects of stochastic resonance is in a simple two-dimensional system, in which the polar coordinates (r,θ) evolve according to the following dynamical system[39]:

$$\dot{r} = r(1 - r^2) + \epsilon_1(t)$$

$$\dot{\theta} = b - r^2 \cos(2\theta) + \epsilon_2(t)$$

where the ϵ-terms are the noise terms. There are two fixed points to this system, and a limit cycle in the deterministic system without the ϵ-terms exists for b > 1 [39]. What is interesting that in the presence of sufficient noise and b < 1, there is directional flow between the fixed points and hence oscillations. If there is too little noise, the dynamical system cannot escape from the stable fixed points and does not oscillate; likewise, too much noise will also wipe out the oscillations. This model illustrates the hypothesis of stochastic resonance.

Consistent with the Stochastic Resonance hypothesis, too little noise may not allow our dynamical system in Fig 2.1 to escape stable fixed points; likewise, too much noise may not allow the dynamical system in Fig 2.1 to settle into a flow between stable fixed points. Yet, if there is the right level of noise, an oscillatory signal may emerge in the periodogram. Here we test this hypothesis in the *N. crassa* clock using the single cell data.

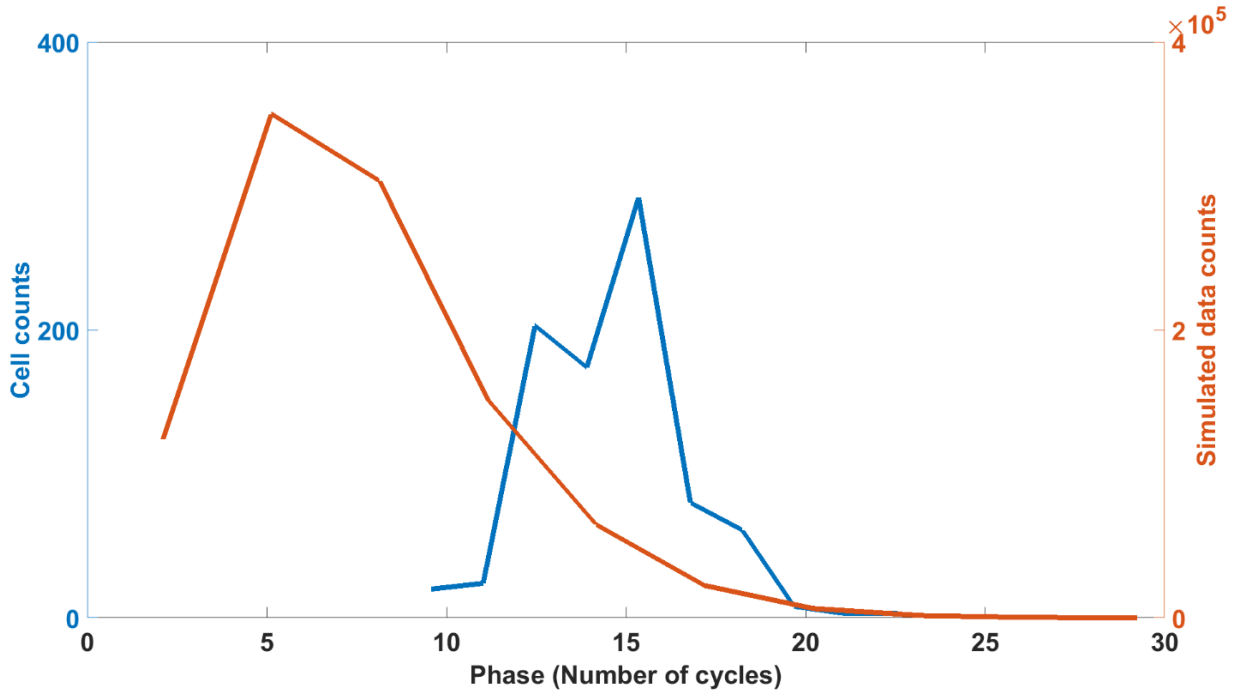

Figure 2.7 Goodness of fit for the model ensemble is tested with the Hilbert Phase for 868 single cells (blue) and Gillespie trajectories (red) under the model with smallest chi-squared statistic in the fitted ensemble (Figure 2.3). The computation of the Hilbert for each trajectory is described previously over a 30 to 115 hour window [13]. The model histogram is that of the Hilbert phases for 1024 Gillespie trajectories on each of > 1000 models in the best fitting model ensemble (Figure 2.3).

What we see in Fig 2.8 is exactly what we would predict under the Stochastic Resonance hypothesis. As the noise is increased above "normal", the peak in the periodogram is diminished, and the oscillatory signal is diminished. If the noise is decreased sufficiently from "normal", the peak in the periodogram is also diminished, and the oscillatory signal is diminished, consistent with the trapping of the real dynamical system in stable fixed points.

Only at an intermediate level of noise do we see oscillations in single cells in Fig 2.8 in the periodogram.

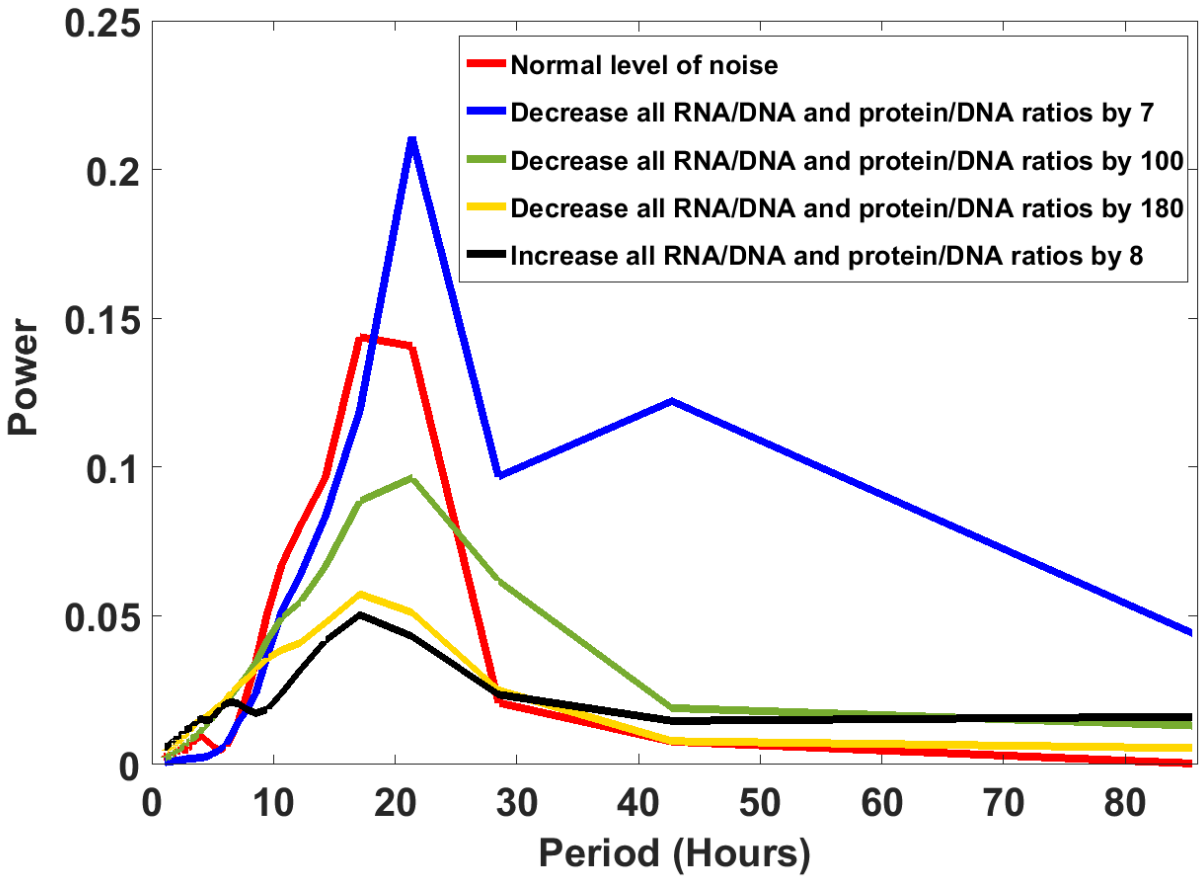

Figure 2.8 There is a non-monotonic relation between the oscillatory signal strength in the normalized periodogram for the CCG protein species and the stochastic intracellular noise. The red curve is for the best fitting model (Figure 2.3B), using the observed RNA/DNA and protein/DNA ratios of 128.7 and 412, respectively. The blue, green and yellow curves have a bigger stochastic intracellular noise than the best fitting model, by shrinking the protein/DNA and RNA/DNA ratios relative to the observed values. The black curve has a smaller stochastic intracellular noise by increasing the protein/DNA and RNA/DNA ratios by a factor of 8.

As a final note, in Fig 2.8 the protein/DNA and RNA/DNA ratios were varied to cause a change in the stochastic intracellular noise (Fig 2.5), while the reaction propensities were left constant. Some might argue from equation (3) that the lead deterministic term in the CLE should be kept constant while varying the stochastic intracellular noise. In this way the limiting deterministic dynamics would be kept constant while the noise is varied. In comparing models with the same deterministic dynamics, it was necessary to vary the propensities so that the lead term in the CLE did not change using the rescaling method in the Materials and Methods. The result of this experiment was the same outcome as in Fig 2.8 with the only change that a much higher RNA/DNA and protein/DNA ratio (~3000) was needed to see the non-monotonic response to stochastic intracellular noise in Fig 2.8.

2.5 DISCUSSION

In order to describe the stochastic behavior of single cell oscillators, a variety of methodological challenges needed to be surmounted. First and foremost, a scalable fitting method that would work with thousands of trajectories on single cells was needed. Existing methods do not operate on this scale. To overcome this challenge a fast, scalable ensemble method for stochastic networks was developed using the periodogram or power spectrum of an average model trajectory to be compared with the average periodogram over single cells. The method is scalable to thousands or tens of thousands of cells on GPUs.

One of the limitations of this approach is that normal Metropolis-Hastings ensemble methods [28] are not adequate for stochastic networks. Parallel tempering methods are shown

to work well on the stochastic networks examined here (Fig 2.2). As more complicated alternatives to the model in Fig 2.1 are considered, surmounting the rate limiting step of generating a Gillespie trajectory may be achieved by other means than through GPUs alone. A promising avenue is generating approximations to the Gillespie trajectory with Quasi Steady-State Approximations (QSSA) to the full stochastic model considered here [58]. The right steady state approximation can help the identifiability of fitting methods for stochastic networks even in simple networks [59].

A second challenge in the fitting process is that there are two sources of error in single cell trajectories, the stochastic intracellular noise and the experimental error [13]. The latter introduces biases into the fitting process. We developed a statistical methodology to remove the experimental error from the fitting process to the periodogram of the model ensemble and thereby achieved a better specification of the model. This new procedure for removing the bias of the experimental error from the periodogram is presented and utilized.

Another challenge of stochastic network identification is characterizing the size of the cell, which sets the level of stochastic intracellular noise, a parameter missing in deterministic models[44]. We developed an empirical approach to identifying the cell size for a stochastic network from the protein/DNA and RNA/DNA ratios for the system under study. We showed that the fitted model was quite robust to variation in these ratios (Fig 2.6). A final challenge is developing protocols to make single cell measurements [60].

In the study of stochastic periodic phenomena we need both ways to fit such models with large amounts of single cell data as well as ways to test the success of these models.

Three statistics provide useful summaries of periodic stochastic networks: period, amplitude, and phase of single cell trajectories. Fortuitously the periodogram is functionally independent of the phase of single cell trajectories [13]. A goodness of fit statistic was then developed using the phase, which included variation in phase across trajectories and the model ensemble.

We then applied these new methodologies to understand the oscillatory behavior of single cells in *N. crassa*. We found that parallel tempering was quite successful in identifying a stochastic network to describe the oscillator in single cells. In many cases the rate constants of macroscopic deterministic models based on millions of cells were similar to those in microscopic stochastic models of single cells (Table 2). Yet, there were also some key differences. For example, in macroscopic models the half-life of the *wc-1* mRNA was measured to be quite long and found to be a critical feature in maintaining oscillations [28]. Yet at the single cell level the half-life was estimated to be much shorter. One possible explanation may be that single cells have other mechanisms to produce oscillations than those that operate at the macroscopic scale. For example, single cells experience stochastic intracellular noise that can move cells from one stationary state to another [39], and this behavior may generate oscillations. This phenomenon depends on not having too much or too little stochastic intracellular noise. We show this phenomenon of Stochastic Resonance [38, 61] may play a role in seeing clock-like behavior in single cells (Fig 2.8). The kinetics of single cells may operate under a set of more relaxed rules for oscillation than those that apply to millions or tens of millions of cells.

Other hypotheses for explaining single cell oscillations need testing [61], such as quorum sensing [62] or cell-cycle gated circadian rhythms [63]. Cell-cycle gating was

controlled by choosing a medium inhibiting cell division [13]. The quorum sensing hypothesis would require communication between cells. Yet, the cells examined here were isolated in droplets. In future work cells will be allowed to share the same droplet, and the associations between cells within droplets provide additional statistics to supplement the periodogram to distinguish between quorum sensing and stochastic resonance. Here we have created both an experimental and methodological arrangement to study the physical theory of Stochastic Resonance in living cells. For the first time, we have a window on a new set of dynamics at the single cell level that play by a different set of rules potentially than ensembles of millions of cells.

Table 2.1 The protein/RNA/DNA ratios used for specifying the scale parameters in a stochastic network were measured and reported below. For each of 13 evenly spaced time points over 48 hours, a sample was taken. A commercial kit was used to simultaneously extract DNA, RNA, and protein from each sample. The amounts of DNA, RNA, and Protein were then measured. To convert these measurements to nanomoles the average molecular weight of a protein and RNA was computed. The average molecular weight of an amino acid is 128.0452. Hence the average molecular weight of a protein in *N. crassa* was taken as 481*128.0452. The average molecular weight of an RNA was taken as 1673*(propA*329.2+propC*305.2+propG*345.2+propU*306.2)+159, where the proportion of A (propA) etc was taken from[46] .

Time	DNA ng/ μl	RNA ng/μl	Protein ng/ μl	DNA (nanomoles)	RNA(nanomoles)	Prot (nanomoles)	RNA:DNA	Prot:DNA	Prot:RNA
0	8.8	1063	422	1.70495E-05	0.001974891	0.006851792	115.8330089	401.8772813	3.469453874
4	6.5	1511	248	1.25934E-05	0.002807206	0.004026645	222.9116944	319.7437035	1.434396273
8	7.5	1036	292	1.45308E-05	0.001924729	0.00474105	132.4586234	326.2761018	2.463230353
12	7.6	1062	550	1.47245E-05	0.001973033	0.00893006	133.9962578	606.4748197	4.526057888
16	15	1092	347	2.90616E-05	0.002028768	0.005634056	69.80927451	193.8661084	2.777082412

20	4.4	1051	261	8.52473E-06	0.001952596	0.00423772	229.0507852	497.1088646	2.170299762
24	7.2	1084	683	1.39496E-05	0.002013905	0.011089511	144.3705234	794.9720944	5.506470958
28	18.2	1216	220	3.52614E-05	0.002259141	0.003572024	64.06840795	101.3012886	1.581142591
32	5.9	1170	387	1.14309E-05	0.00217368	0.006283515	190.1584354	549.6960677	2.890726707
36	13.5	1311	343	2.61554E-05	0.002435636	0.00556911	93.12165128	212.9237118	2.286511341
40	14	1161	369	2.71241E-05	0.002156959	0.005991259	79.5217501	220.8828551	2.777640769
44	13.8	1159	697	2.67367E-05	0.002153244	0.011316822	80.53526552	423.2698834	5.255708547
48	7.4	904	626	1.4337E-05	0.001679493	0.010164032	117.1435702	708.9347917	6.051845528
Ave.							128.7	412	

Table 2.2 Ensemble means and standard errors indicate that the parameters in stochastic network for single cells are tightly specified by the new fitting method.

Parameter	Initial_Parameter value from MCMC Deterministic model ensemble (Yu et al., 2007)	Initial Parameter values from Deterministic model ensemble (column 2) re-scaled to_molecular number_units of stochastic network (column 3)	Mean parameter values from model ensemble computed by Parallel tempering_	Standard error (SE) of parameter value across ensemble computed by parallel tempering	Mean parameter values from model ensemble computed by Parallel tempering_	Standard error (SE) of parameter value across ensemble computed by parallel tempering
Number of cells	-	-	868	868	1591	1591
u_r0	3.99924	113	5015.225951	39.46834951	2156.705728	68.14603254
u_r1	0.442441	18	5427.14616	38.81394987	22.46137677	0.872953544
u_p	4.24E-07	459	5052.398397	39.86454551	2144.149238	68.74768856
f_0	0.356365	1	0.498322148	0.006827531	0.465055176	0.01143672

f_1	0.0824576	0	0.501677852	0.006827531	0.534944824	0.01143672
f_r	4.90E-06	31	5099.803691	39.65320564	59.15869679	2.637452857
f_p	3.0804	345	5661.033184	37.79536413	2534.336311	77.72114325
w	9.24126	101	5070.154735	39.25742344	55.40042039	1.674320488
g_0	0.0066195	1	0.498508576	0.006827539	0.71623752	0.010337149
g_1	2.59E-06	0	0.501491424	0.006827539	0.28376248	0.010337149
g_r	1.17E-06	26	3086.66182	44.02007694	35.67840252	1.030258983
g_p	1.37E-05	102	5272.938106	41.68301079	59.19075145	4.920903774
A	0.000658482	6.06E-13	1.84E-10	3.59E-12	2.56E-10	7.31E-12
Abar	0.546986	0.546986	50.03130674	0.39493562	1.589532708	0.035661845
S1	0.061594783	83.70771546	75.50835675	0.195849227	80.12566921	0.302471515
S3	0.00146575	3.569116497	10.79752752	0.13705406	0.400641074	0.036565894
S4	2.2396	5453.449297	8229.792075	51.85745823	8316.020583	100.2852188
D1	0.723678	0.723678	11.6411168	0.125746972	1.294999006	0.030289616
D3	0.299703	0.299703	13.54885771	0.126263685	4.382612039	0.181101578
C1	0.0428595	4.81E-05	0.002859534	2.05E-05	0.000932789	2.47E-05
L1	31.7758	4.244678204	11.73643559	0.073999739	4.777735371	0.106626479
L3	3.02387	0.485087349	1.339312652	0.010385343	0.665600817	0.011127036
D4	0.00323262	0.00323262	0.077027847	0.001450665	0.08474029	0.004700587
D6	0.15183	0.15183	4.114778877	0.068424489	0.193685712	0.002236097
D7	0.138387	0.138387	3.008652124	0.046190832	2.130911791	0.090030385
D8	0.00248668	0.00248668	0.159664546	0.00266928	0.007744621	0.000182717
C2	0.162687246	0.162687246	6.323785664	0.069435513	1.515554675	0.077548547
P	19.5648	3.12E-11	1.15E-09	1.88E-11	2.72E-09	4.83E-11
Ac	4.06813	7.82E-09	2.68E-07	4.65E-09	1.86E-08	2.55E-09
Bc	2.52197	2.52197	66.30689649	0.666502458	2.581096866	0.040197442
Sc	1.01E-06	73.80414613	1097.22398	6.307615581	61.51499414	1.109629713
Lc	1.15E-08	2.231095711	3.743285409	0.038176388	1.61524392	0.017335914
Dcr	0.219758	0.219758	0.625739758	0.02113632	0.150810052	0.00291715
Dcp	0.696903	0.696903	1.982857699	0.007835124	0.54063952	0.006141903

2.6 REFERENCES

1. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002;297(5584):1183-6.

2. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010;467(7312):167-73.

3. Kepler TB, Elston TC. Stochasticity in Transcriptional Regulation: Origins, Consequences, and Mathematical Representations. Biophysical Journal. 2001;81:3116-36.

4. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. Genetics. 1998;149(4):1633-48.

5. McAdams HH, Shapiro L. Circuit stimulation of genetic networks. Science. 1995;269(5224):650.

6. Cai L, Dalal CK, Elowitz MB. Frequency-modulated nuclear localization bursts coordinate gene regulation. Nature. 2008;455(7212):485-90.

7. Lin Y, Sohn CH, Dalal CK, Cai L, Elowitz MB. Combinatorial gene regulation by modulation of relative pulse timing. Nature. 2015;527(7576):54-8.

8. Levy SF, Ziv N, Siegal ML. Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. PLoS Biology. 2012;10(5):1001325.

9. Voit EO. The best models of metabolism. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2017.

10. David S, Guido S, Ramon G. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. Journal of Physics A: Mathematical and Theoretical. 2017;50(9):093001.

11. Paliwal S, Iglesias PA, Campbell K, Hilioti Z, Groisman A, Levchenko A. MAPK-mediated bimodal gene expression and adaptive gradient sensing in yeast. Nature. 2007;446(7131):46-51.

12. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187-201.

13. Deng Z, Arsenault S, Caranica C, Griffith J, Zhu T, Al-Omari A, et al. Synchronizing stochastic circadian oscillators in single cells of *Neurospora crassa*. Scientific Reports. 2016;6:35828.

14. Bennett MR, Hasty J. Microfluidic devices for measuring gene network dynamics in single cells. Nature Reviews Genetics. 2009;10(9):628-38.

15. Vera M, Biswas J, Senecal A, Singer RH, Park HY. Single-cell and single-molecule analysis of gene expression regulation. Annual review of genetics. 2016;50:267-91.

16. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nature biotechnology. 2011;29(12):1120-7.

17. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nature biotechnology. 2016;34(11):1145-60.

18. Andrews SS, Dinh T, Arkin AP. Stochastic models of biological processes. Encyclopedia of Complexity and Systems Science: Springer; 2009. p. 8730-49.

19. Pothoulakis G, Ceroni F, Reeve B, Ellis T. The Spinach RNA Aptamer as a Characterization Tool for Synthetic Biology. ACS Synthetic Biology. 2014;3(3):182-7. doi: 10.1021/sb400089c.

20. Gillespie DT. Exact Stochastic Simulation of Coupled Chemical-Reactions. J Phys Chem-Us. 1977;81(25):2340-61. PubMed PMID: ISI:A1977EE49800008.

21. Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. The Journal of Chemical Physics. 2001;115(4):1716-33.

22. Wilkinson DJ. Stochastic modelling for systems biology: CRC press; 2011.

23. Daigle BJ, Roh MK, Petzold LR, Niemi J. Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. BMC Bioinformatics. 2012;13(1):68.

24. Tian T, Xu S, Gao J, Burrage K. Simulated maximum likelihood method for estimating kinetic rates in gene expression. Bioinformatics. 2007;23(1):84-91.

25. Reinker S, Altman R, Timmer J. Parameter estimation in stochastic biochemical reactions. Systems Biology, IEE Proceedings. 2006;153. doi: 10.1049/ip-syb:20050105.

26. Fröhlich F, Thomas P, Kazeroonian A, Theis FJ, Grima R, Hasenauer J. Inference for Stochastic Chemical Kinetics Using Moment Equations and System Size Expansion. PLOS Computational Biology. 2016;12(7):e1005030. doi: 10.1371/journal.pcbi.1005030.

27. Battogtokh D, Asch DK, Case ME, Arnold J, Schuttler HB. An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of *Neurospora crassa*. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(26):16904-9. PubMed PMID: 12477937.

28. Yu Y, Dong W, Altimus C, Tang X, Griffith J, Morello M, et al. A genetic network for the clock of *Neurospora crassa*. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(8):2809-14.

29. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface. 2009;6(31):187-202. PubMed PMID: ISI:000262757200006.

30. Liu JS. Monte Carlo Strategies in Scientific Computing. 2001;Springer.

31. Robert CP, Casella G. Monte Carlo Statistical methods. 1999;Springer.

32. Golightly A, Wilkinson DJ. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. Interface Focus. 2011;1(6):807-20.

33. Fearnhead P, Prangle D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012;74(3):419-74.

34. Komorowski M, Miekisz J, Kierzek A: Translational Repression Contributes Greater Noise to Gene Expression than Transcriptional Repression. Biophysical Journal 2009., 96(2): 10.1016/j.bpj.2008.09.052.

35. Dunlap JC. Molecular bases for circadian clocks. Cell. 1999;96(2):271-90. PubMed PMID: 9988221.

36. Thomas P, Straube AV, Timmer J, Fleck C, Grima R. Signatures of nonlinearity in single cell noise-induced oscillations. Journal of Theoretical Biology. 2013;335:222-34. doi: https://doi.org/10.1016/j.jtbi.2013.06.021.

37. Al-Omari A, Griffith J, Judge M, Taha T, Arnold J, Schuttler H. Discovering regulatory network topologies using ensemble methods on GPGPUs with special reference to the biological clock of *Neurospora crassa*. Access, IEEE. 2015;3:27-42.

38. Benzi R, Sutera A, Vulpiani A. The mechanism of stochastic resonance. Journal of Physics A: Mathematical and General. 1981;14(11):L453.

39. Rappel W-J, Strogatz SH. Stochastic resonance in an autonomous system with a nonuniform limit cycle. Physical Review E. 1994;50(4):3249.

40. Gang H, Ditzinger T, Ning C-Z, Haken H. Stochastic resonance without external periodic force. Physical Review Letters. 1993;71:807.

41. Kim Jae K, Kilpatrick Zachary P, Bennett Matthew R, Josić K. Molecular Mechanisms that Regulate the Coupled Period of the Mammalian Circadian Clock. Biophysical Journal. 106(9):2071-81. doi: 10.1016/j.bpj.2014.02.039.

42. Kim Jae K. Protein sequestration versus Hill-type repression in circadian clock models. IET Systems Biology. 2016;10(4):125-35.

43. Castro-Longoria E, Ferry M, Bartnicki-Garcia S, Hasty J, Brody S. Circadian rhythms in *Neurospora crassa*: Dynamics of the clock component *frequency* visualized using a fluorescent reporter. Fungal Genetics and Biology. 2010;47(4):332-41.

44. Wu J, Vidakovic B, Voit EO. Constructing stochastic models from deterministic process equations by propensity adjustment. BMC Systems Biology 2011;5:187.

45. Dong W, Tang X, Yu Y, Nilsen R, Kim R, Griffith J, et al. Systems biology of the clock in *Neurospora crassa*. PloS one. 2008;3(8):e3105. PubMed PMID: 18769678.

46. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature. 2003;422(6934):859-68. PubMed PMID: 12712197.

47. Tian T, Xu S, Gao J, Burrage K. Simulated maximum likelihood method for estimating kinetic rates in gene expression. Bioinformatics. 2007;23. doi: 10.1093/bioinformatics/btl552.

48. Swendsen RH, Wang J-S. Replica Monte Carlo smiulation of spin-glasses. Phys Rev Lett. 1986;57:2607-9.

49. Earl DJ, Deem MW. Parallel tempering: theory, applications, and new perspectives. Physical Chemistry Chemical Physics. 2005;7(23):3910-6.

50. Hamze F, Dickson N, Karimi K. Robust parameter selection for parallel tempering. International Journal of Modern Physics C. 2010;21(05):603-15. doi: 10.1142/s0129183110015361.

51. Izumo M, Sato TR, Straume M, Johnson CH. Quantitative Analyses of Circadian Gene Expression in Mammalian Cell Cultures. PLoS Comput Biol. 2006;2(10):e136. doi: 10.1371/journal.pcbi.0020136.

52. Ruoff P, Loros JJ, Dunlap JC. The relationship between FRQ-protein stability and temperature compensation in the Neurospora circadian clock. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(49):17681-6. doi: 10.1073/pnas.0505137102.

53. Larrondo LF, Olivares-Yañez C, Baker CL, Loros JJ, Dunlap JC. Decoupling circadian clock protein turnover from circadian period determination. Science. 2015;347(6221). doi: 10.1126/science.1257277.

54. Hautala JA, Conner BH, Jacobson JW, Patel GL, Giles N. Isolation and characterization of nuclei from Neurospora crassa. Journal of bacteriology. 1977;130(2):704-13.

55. Gillespie DT. The chemical Langevin equation. The Journal of Chemical Physics. 2000;113(1):297-306. doi: 10.1063/1.481811.

56. Gabor D. Theory of communication. Part 1: The analysis of information. Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of. 1946;93(26):429-41.

57. Kendall M, Stuart A. The Advanced Theory of Statistics, Volume 2, Inference and Relationship. Macmillan, NY. 1979:530.

58. Kim JK, Josić K, Bennett MR. The relationship between stochastic and deterministic quasi-steady state approximations. BMC Systems Biology. 2015;9(1):87. doi: 10.1186/s12918-015-0218-3.

59. Choi B, Rempala GA, Kim JK. Beyond the Michaelis-Menten equation: Accurate and efficient estimation of enzyme kinetic parameters. Scientific Reports. 2017;7(1):17018. doi: 10.1038/s41598-017-17072-z.

60. Lee KK, Labiscsak L, Ahn CH, Hong CI. Spiral-based microfluidic device for long-term time course imaging of *Neurospora crassa* with single nucleus resolution. Fungal Genetics and Biology. 2016;94:11-4.

61. Ullner E, Buceta J, Díez-Noguera A, García-Ojalvo J. Noise-induced coherence in multicellular circadian clocks. Biophysical Journal. 2009;96(9):3573-81.

62. Whiteley M, Diggle SP, Greenberg EP. Progress in and promise of bacterial quorum sensing research. Nature. 2017;551:313-20.

63. Paijmans J, Bosman M, Wolde PRt, Lubensky DK. Discrete gene replication events drive coupling between the cell cycle and circadian clocks. PNAS USA. 2015;113:4063-8.

CHAPTER 3

WHAT IS PHASE IN CELLULAR CLOCKS?

Caranica, C., J.H. Cheong, X. Qiu, E. K. Krach, Z. Deng, L. Mao, H-B. Schüttler and J. Arnold

2019 *Yale Journal of Biology and Medicine* 92(2): 169-178

Reprinted here with permission of publisher

ABSTRACT

Four inter-related measures of phase are described to study the phase synchronization of cellular oscillators, and computation of these measures is described and illustrated on single cell fluorescence data from the model filamentous fungus, *Neurospora crassa*. One of these four measures is the phase shift ϕ in a sinusoid of the form x(t) = $A\cos(\omega t + \phi)$, where t is time. The other measures arise by creating a replica of the periodic process x(t) called the Hilbert transform $\tilde{x}(t)$, which is 90 degrees out of phase with the original process x(t). The second phase measure is the phase angle $F^H(t)$ between the replica $\tilde{x}(t)$ and x(t), taking values between -π and π. At extreme values the Hilbert phase is discontinuous, and a continuous form $F^C(t)$ of the Hilbert Phase is used, measuring time on the nonnegative real axis (t). The continuous Hilbert Phase $F^C(t)$ is used to define the phase $M^C(t_1, t_0)$ for an experiment beginning at time t_0 and ending at time t_1. Because phase differences at time t_0 are often of ancillary interest, the Hilbert Phase $F^C(t_0)$ is subtracted from $F^C(t_1)$. This difference is divided by 2π to obtain the phase $M^C(t_1, t_0)$ in cycles. Both the Hilbert Phase $F^C(t)$ and the phase $M^C(t_1, t_0)$ are functions of time and useful in studying when oscillators phase-synchronize in time in signal processing and circadian rhythms in particular. The phase of cellular clocks is fundamentally different from circadian clocks at the macroscopic scale because there is an hourly cycle superimposed on the circadian cycle.

3.1 INTRODUCTION

Recently single cells have been shown to have circadian oscillators using fluorescent markers in clock-related genes [1]. These kinds of fluorescent and luminescent measurements on single cells have been made in a variety of clock systems recently [2-6]. There are three basic properties of oscillators in single cells: the amplitude, period, and phase of each cell's oscillations. The most elusive of these quantities is the phase. How the phase behaves can provide information on how circadian oscillators synchronize in whole tissues or organisms [7]. Oscillators in single cells each appear to have substantial variation in phase between cells [1], but at the macroscopic level of 10^7 cells the ensemble of circadian oscillators appear synchronized when assayed in liquid culture or in race tubes, for example [8]. How does this phase synchronization arise? In order to address this question, there is a need for a clear notion of phase and how to calculate the phase of a cellular oscillator [9]. Whether or not phase synchronization happens may have profound consequences for our health and successful aging [10].

One particular notion of phase relied on heavily in this work is the Hilbert Phase, which was originally developed for signal processing applications [11]. It has come to the fore recently in the analysis of single cell data in *Mus musculus* [12] and *N. crassa* [1]. There has also been increasing interest in its use for identifying phase response curves describing synchronization to light and other entrainment signals [13,14].

Here we develop four inter-related notions of phase of a cellular oscillator, the phase shift, the Hilbert Phase, the continuous Hilbert Phase, and phase in cycles. We use single cell

oscillators in the model system, *Neurospora crassa*, to illustrate each of these four interrelated measures of phase and their function in describing "phase". We use these phase measures to examine: (1) phase variation in single cells; (2) the effect of the social environment of cellular oscillators on phase; (3) phase as a function of time as when there is synchronization of oscillators. Each of these topics illustrate how all four notions of phase are measured in concert to provide insights into cellular oscillators.

3.2 MEASURES OF PHASE

A fluorescent or luminescent measurement x(t) at each time point t is made on the fluorescence or luminescence of a clock-related gene in single cells. Typically these measurements are taken every half hour over ten, 24 hour days, in the model system, *N. crassa*, used to illustrate the phase calculations(1). The simplest model for these measurements is a sinusoid of the form $x(t) = Acos(\omega t + \phi)$, where A is the amplitude of the signal, ω, the frequency of the oscillation, and ϕ, the phase shift of the process. This is sometimes referred to as the hidden periodicity model [15]. If the phase ϕ were constant over time, then the phase shift would capture all of the phase information about each cell and could be extrapolated safely to later times after the zero-time point to examine phase relations between cells. The challenge is when the phase shift is not constant in time, as when cells synchronize their phases in time.

Let us suppose we could create a replica $\tilde{x}(t)$ of the original process x(t) that is 90 degrees out of phase with the original process. For example, if $x(t)$ $were$ cos (ωt), then the

replica $\tilde{x}(t) = \sin(\omega t)$ would be 90 degrees out of phase (Fig. 3.1A). We can think of each process as a mark on a spinning tire on a car; one tire is marked, and another tire is marked

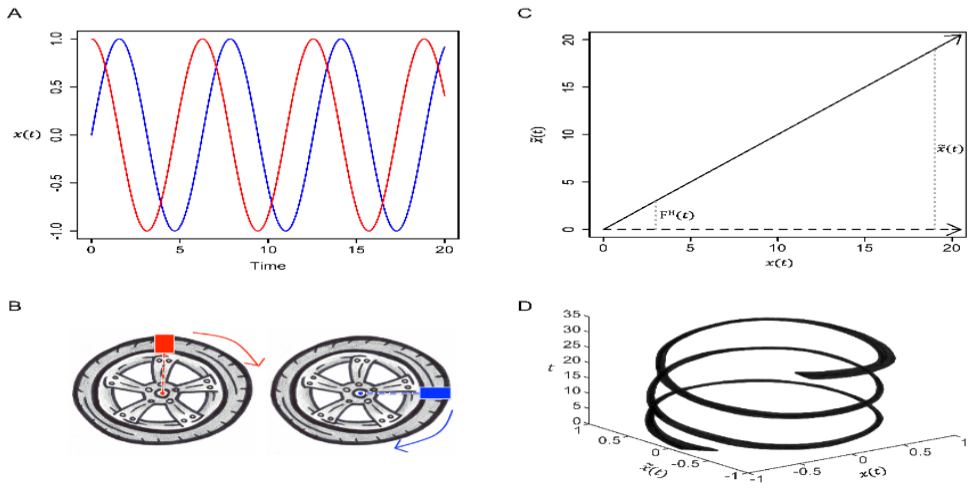

Figure 3.1 Hilbert Phase. (A) The original process x(t) in red is replicated by a Hilbert transform to $\tilde{x}$(t) in blue. (B) Creating the replica is analogous to putting two marks on two different tires, which are at 12:00 and 3:00 o'clock to start. (C) The angle $F^C(t)$ between the two marks is the Hilbert Phase. One mark corresponds to the original process x(t) in red. The second replica mark corresponds to the Hilbert transform $\tilde{x}(t)$ in blue. (D) The pair $(x(t), \tilde{x}(t))$ defines a number in the complex plane, and as the tires rotate, the pair form a spiral over the complex plane known as the analytic signal.

with its mark at 90 degrees from the other mark at time t = 0 (Fig 3.1B). The two marks, x(t) and $\tilde{x}(t)$, are followed over time. Then we could observe how the original process and the replica change position with respect to each other in time. This is usually done by way of an angular measurement between the marks on each tire. In other words, we can watch the two marks on two tires change position (*i.e.*, angle) with respect to each other in time. The surprise is that under very general conditions (described in Materials and Methods) such a replica $\tilde{x}(t)$ can be created and is called the Hilbert Transform of x(t) [11].

The Hilbert Phase $F^H(t)$ is then defined as the phase angle between the original process x(t) and the replica $\tilde{x}(t)$ (Fig 3.1C) [16]:

$$F^H(t) = tan^{-1}\frac{\tilde{x}(t)}{\mathrm{x(t)}}.$$

This phase angle (*i.e.*, between the two marks on two tires) can change over time along with the original fluorescent series x(t) and be computed for the fluorescent series on each cell by a Fast Fourier Transform [17]. The range of values for the Hilbert Phase by convention are usually taken to be from $-\pi$ to π. If the process were $x(t) = \mathrm{Acos}(\omega t + \phi)$, then we could calculate the Hilbert Phase:

$$F^H(t) = tan^{-1}\frac{\tilde{x}(t)}{\mathrm{x(t)}} = tan^{-1}\frac{\mathrm{Asin}(\omega t+\phi)}{\mathrm{Acos}(\omega t+\phi)} = tan^{-1}\tan(\omega t + \phi) = \omega t + \phi$$

In this case of a sinusoidal process the Hilbert Phase has a simple linear relation with time with the y-intercept being the phase shift ϕ and with the slope being the frequency.

To visualize the relation of the process x(t) and its replica $\tilde{x}(t)$ this pair of values is used to form a complex number of the form $(x(t), \tilde{x}(t))$, where x(t) is the real part and $\tilde{x}(t)$ is

the imaginary part in the complex plane at time t (Fig. 3.1C). This pair as a function of time is sometimes referred to as the analytic signal(16). As time t advances, the curve $(x(t), \tilde{x}(t))$ traces out a cycle in the complex plane about its origin (Fig. 3.1D). As the curve approaches $-\pi$ or π, it tends to have discontinuities (Fig. 3.2). To stitch together these discontinuities, the Hilbert phase $F^H(t)$ is continuized as described in Materials and Methods and shown in Fig. 3.2 (in red). The continuous Hilbert phase is denoted by $F^C(t)$. A plot of this average continuous Hilbert phase with an average (over cells) of the original process is shown in the middle panel (Fig. 3.2 in red). With time in the z-direction and the complex plane extending in the x- and y directions, the curve $\left(x(t), F^C(t)\right)$ appears as a spiral (or tornado) in time (see Fig. 3.1D).

3.3 MATERIALS AND METHODS

3.3.1 DATA

Fluorescent measurements were made on cells with a mCherry recorder gene attached to a *clock-controlled gene-2* (*ccg-2*) promoter in strain MFNC9 of *N. crassa*(18). Much of the data were published in an excel spread sheet for ~1,591 isolated cells, each measured every half hour over ten days [7]. Cells are across columns; time is down rows. The fluorescent data were Rhodamine B normalized to control for uncontrolled periodic and aperiodic factors and detrended [1]. Subsequently, improved cell tracking slightly increase the data set to 1,644 cells. All data were loaded into a MATLAB workspace publicly available via GitHub at https://github.com/XiaoQiu2019/Matlab-code-for-phase-paper

3.3.2 CALCULATING PHASE

The calculation of phase is presented [1]. The calculation of phase is detailed again. First a replica $\tilde{x}(t)$ (of the process x(t)) is created using the Hilbert Transform:

$$\tilde{x}(t) = PV\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{x(t)}{t-\tau}d\tau,$$

where the integral is calculated in the principal value sense [19]. This replica $\tilde{x}(t)$ is 90 degrees out of phase with the original process; moreover, it is uniquely specified by the process when it exists. In that the replica $\tilde{x}(t)$ is purely imaginary in Fig. 3.1C and completely out of phase with x(t), it can be derived from the original x(t) using the convolution theorem:

$$\tilde{x}(t) = -iFT^{-1}(FT(x(t)),$$

where FT denotes the Fourier Transform, FT^{-1} is the inverse Fourier Transform and $i = \sqrt{-1}$. This relation is how the replica is computed with the commands fft and ifft in MATLAB [17].

The Hilbert Phase is the phase angle between the original process x(t) and the replica $\tilde{x}(t)$:

$$F^H(t) = tan^{-1}\frac{\tilde{x}(t)}{\mathrm{x(t)}}$$

As an example, if the process x(t) were $\mathrm{A}\cos(\omega t + \phi)$, then the Hilbert transform would be $\mathrm{A}\sin(\omega t + \phi)$**.** It follows that the Hilbert Phase reduces to $F^H(t) = \omega t + \phi$. As the Hilbert Phase passes near $-\pi$ or π, there are usually discontinuities in the Hilbert Phase. To surmount this problem the Hilbert Phase was continuized. The continuous Hilbert Phase is defined recursively by:

$$F^C(t+1) = F^C(t) + m^C(t)2\pi,$$

where time t is an integer value indicating the number of elapsed half hours in each ~ten-day experiment (containing ~480 half hour time points). The multiple $m^C(t)$ was chosen to minimize the following differences with respect to $m^C(t)$:

$$Df_m = |F^H(t+1) - F^H(t) + 2\pi m^C(t)|$$

This was done by the MATLAB code accessible in GitHub.

With the continuous Hilbert phase $F^C(t)$ in hand, then the phase $M^C(t_1,t_0)$ in cycles can be calculated:

$$M^C(t_1, t_0) = \frac{[F^C(t_1) - F^C(t_0)]}{2\pi},$$

where the divisor of 2π insures that the phase $M^C(t_1, t_0)$ counts cycles completed by the continuous phase angle $F^C(t)$ and where the subtracted quantity acts like a generalized phase shift ϕ to remove the phase differences at the beginning of an experiment. The phase $M^C(t_1, t_0)$ in cycles varies with the time interval of the experiment. Allowing the phase to vary with time t_1 permits the examination of phase synchronization between cells as illustrated in the next section.

3.4 RESULTS

3.4.1 PHASE VARIATION IN SINGLE CELLS

It is useful to consider whether or not the continuation of the Hilbert phase has any effect. The original Hilbert phase $F^H(t)$ and continuous Hilbert phase $F^C(t)$ were computed

for a single cell and for each of the ~1,591 single cell trajectories and averaged (Fig. 3.2). A Hilbert phase $F^H(t)$ for a single randomly selected cell is shown on the top panel and is extremely ragged as it approaches $-\pi$ or π, but the continuous curve $F^C(t)$ smooths over the discontinuities. The Hilbert phase $F^H(t)$ is an angle and confined to the interval from $-\pi$ to π (as shown in the lower panel of Fig. 3.2). The average trajectory of the continuous Hilbert phase $F^C(t)$ (middle panel) is smoother over time than the average of the original Hilbert phase $F^H(t)$ too and is a measure of time for a periodic process(20). The average Hilbert phase $F^H(t)$ is also ragged due to the stochastic intracellular variation in each cell(9) as well as the discontinuity in $F^H(t) = tan^{-1}\frac{\tilde{x}(t)}{\mathrm{x(t)}}$ at π and -π. The distribution of Hilbert phase as a consequence appears non-uniform on the unit circle (Fig. 3.2).

Each cell may have its own oscillator [1]. The phase $M^C(t_1, t_0)$ in cycles as defined above can then be calculated for each member of a population of ~1,591 isolated oscillators in *N. crassa* (Fig. 3.3) [1]. The reported phases $M^C(t_1, t_0)$ are in cycles completed in the ten-day interval of the experiment. The cycles completed on average are visible in Fig. 3.2. As can be seen, there was substantial variation in the phase $M^C(t_1, t_0)$ of the oscillators as measured in cycles (Fig. 3.3). At the single cell level each cell is marching to a different drummer. This raises the question of how cells synchronize to generate the coherent behavior at 10^7 cells.

Cellular clocks are fundamentally different from circadian clocks at the macroscopic level. Overlaid on the circadian rhythm is high frequency noise from the cell, much like an hour hand added on to the tolling of bells at the end of a day. The Hilbert phase and phase in cycles $M^C(t_1, t_0)$ count all cycles from all harmonics. As shown earlier an examination of the

periods of cellular oscillators has demonstrated a major circadian harmonic [7] accounts for one cycle per day. The remaining cycles come from the high-frequency intracellular noise [1]. For example, the high frequency noise in a cellular clock allowed it to complete over 7 cycles per 24-hour day (Fig. 3.3). This feature of the experiments also allowed the examination of cellular communication between cells in droplets or neighborhoods. Those in the same droplet are within 74 microns [21] of each other and have the opportunity to communicate and hence perhaps synchronize.

3.4.2 SOCIAL ENVIRONMENT OF CELLS AND ITS EFFECT ON PHASE

In the original experiments cells were placed in the same droplet or in different droplets using a microfluidics device over ten days to observe circadian rhythms, thereby placing cells in different neighborhoods or "social environments" [1]. Those cells in different droplets have no opportunity to communicate and continue marching to their own drummer. It would be interesting to know whether or not 2-cell droplets have a different phase trajectory than 1-cell droplets.

The tornado plots of the analytic signal $(x(t), \tilde{x}(t))$ are shown (Fig. 3.4). The red tornado is an average over 568, 2-cell trajectories in the complex plane; the black tornado is an average over 1,644 single cell trajectories. While the 2-cell average trajectory very closely tracked the average 1-cell trajectory, the average 2-cell trajectory was consistently closer to the origin than the average 1-cell trajectory. This suggests that the 2-cell phase trajectories of the

analytic signal behave differently than 1-cell phase trajectories, but this is analyzed in more detail with synchronization measures derived from the Hilbert phase in the next section.

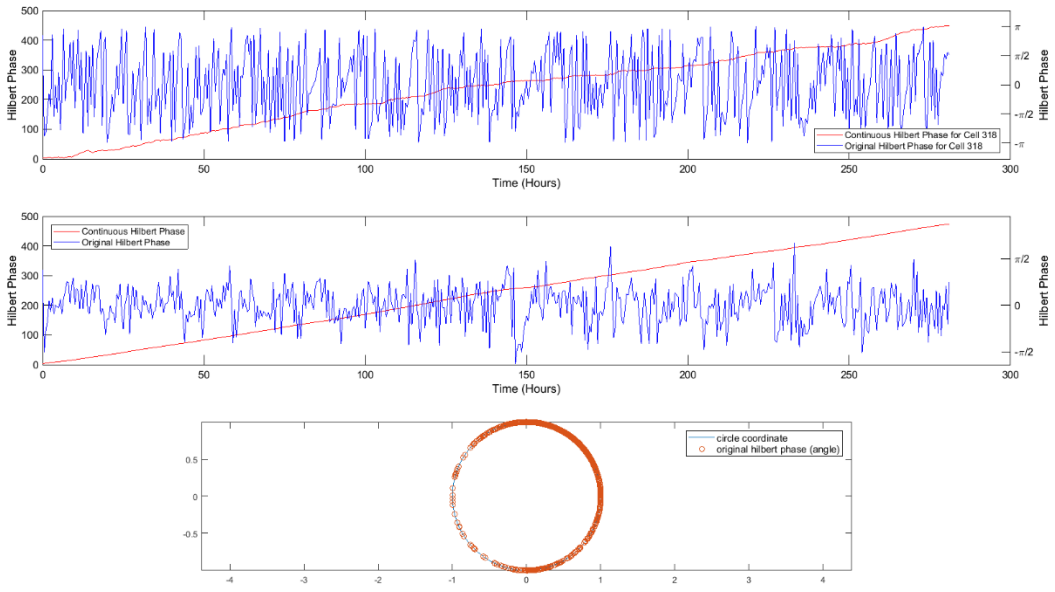

Figure 3.2 The continuous Hilbert phase F^C (t) as a trajectory for a single cell (top panel) or averaged over all 1,591 cells is a smooth function of time (in red) while the original Hilbert phase for cell number 318 (in blue) or averaged over all cells is ragged (in blue) and as an angle is confined to the interval $-\pi$ to π on the upper panel (see bottom panel). For the top and middle panels the values of the continuous Hilbert phase $F^C(t)$ are plotted on the left vertical axis; the values of the original uncontinuized Hilbert phase $F^H(t)$ are plotted on the right vertical axis. On the lower panel is a plot of the original Hilbert phase $F^H(t)$ as a phase angle from $-\pi$ to π on the unit circle**.** Only Hilbert phase values at the final time point are displayed.

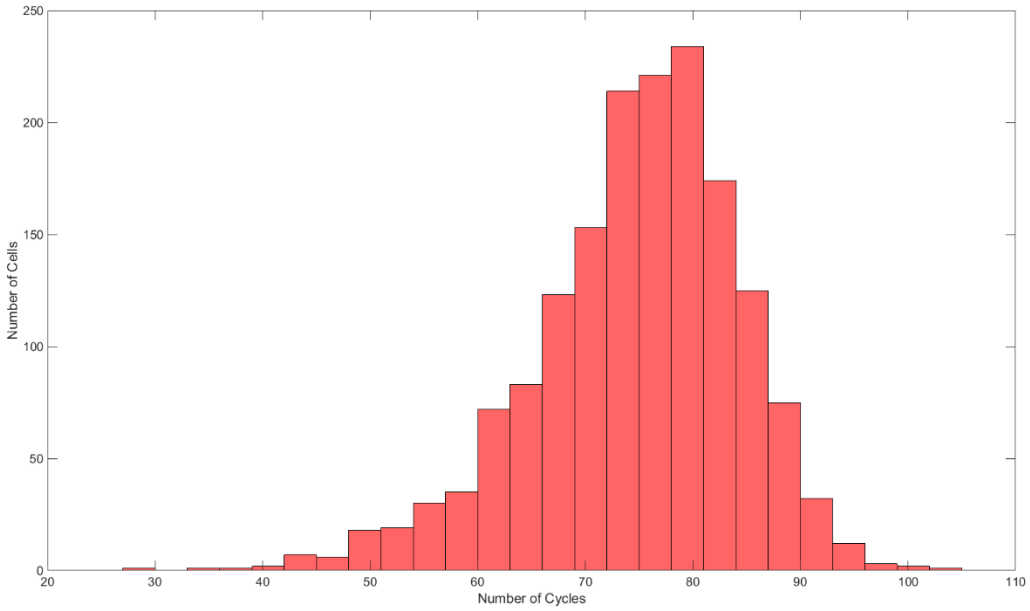

Figure 3.3 There was considerable variation in phase M^C (t_1,t_0) over ~1,591 cells. The phase M^C (t_1, t_0) was computed from t = 0 to t =281 h and is for ~1,591 cells that have no neighbors in a droplet. The mean and standard deviation of this histogram of this histogram of phases for single cells were 74 cycles and 10 cycles, respectively.

3.4.3 PHASE AS A FUNCTION OF TIME

Another feature of the phase plots is that they are a function of time. Both the continuous Hilbert phase $F^C(t)$ and the phase $M^C(t_1, t_0)$ in cycles are plotted as a function of time t_1 with t_0 = 0 averaged over all ~1,591 cells (Fig. 3.2 and Fig. 3.5). The phase $M^C(t_1, t_0)$ in cycles resembles the Hilbert phase, but passes through the origin in Fig. 3.5. As a basis of comparison, the phase in cycles $M^C(t_1, t_0)$ for the data without a sinusoidal assumption and phase in cycles $M^C(t_1, t_0)$ for a sinusoid $x(t) = Acos(\omega t + \phi)$ are plotted as a function of

time in Fig. 3.5. The parameters in the sinusoid were determined by least-squares for all 1,591 cells. As can be seen, the phase curves of the cells are approximately sinusoidal (Fig. 3.5).

If the cells are in different droplets and unable to communicate, the expectation is that these cellular oscillators may drift out of phase with respect to each other. To test this hypothesis, the 2.5 percentile and 97.5 percentile curves about the mean phase of ~1,591 isolated cells (Fig. 3.5) were computed over time. As expected, the percentiles drifted away from the mean over time. Incidentally the percentiles provide support for the phase curve (dotted) of the sinusoid not being statistically different from the average phase curve (in blue) of isolated cells. The phase curve for the sinusoid is found between the percentiles of the average phase curve for the cells.

The Hilbert phase curves can also be used to examine the synchronization of cells, where phase may change with time. Consider 1,136 cells in droplets with 2-cells per droplet. How do these 568 (=1,136/2) pairs of droplets compare in phase with 1,644 cells that have never known neighbors? The average Hilbert phase for 2-cell droplets steadily diverged from that of the 1-cell droplet in Fig. 3.6 and was smaller than the average of the 1-cell droplets, as would be expected from Figure 3.4. Yet, the percentiles for both the 2-cell and 1-cell droplets still drifted apart with time, indicating an accumulation of phase variation with time even when cells have the opportunity to talk with each other.

Now consider 10 cells that have been placed in isolated droplets within the same data set [1] and never experienced neighbors. Also consider 10 droplets each with 10 cells that have lived together as roommates for 10 days within the same droplet. The average Hilbert phase

curves are shown for the isolated cells and 10 droplets each with 10 roommates (Fig. 3.7). As can be seen, the average continuous Hilbert phase for 10 cells in each of 10 droplets steadily diverged from the average of 10 singletons, who have never known neighbors. Their phases

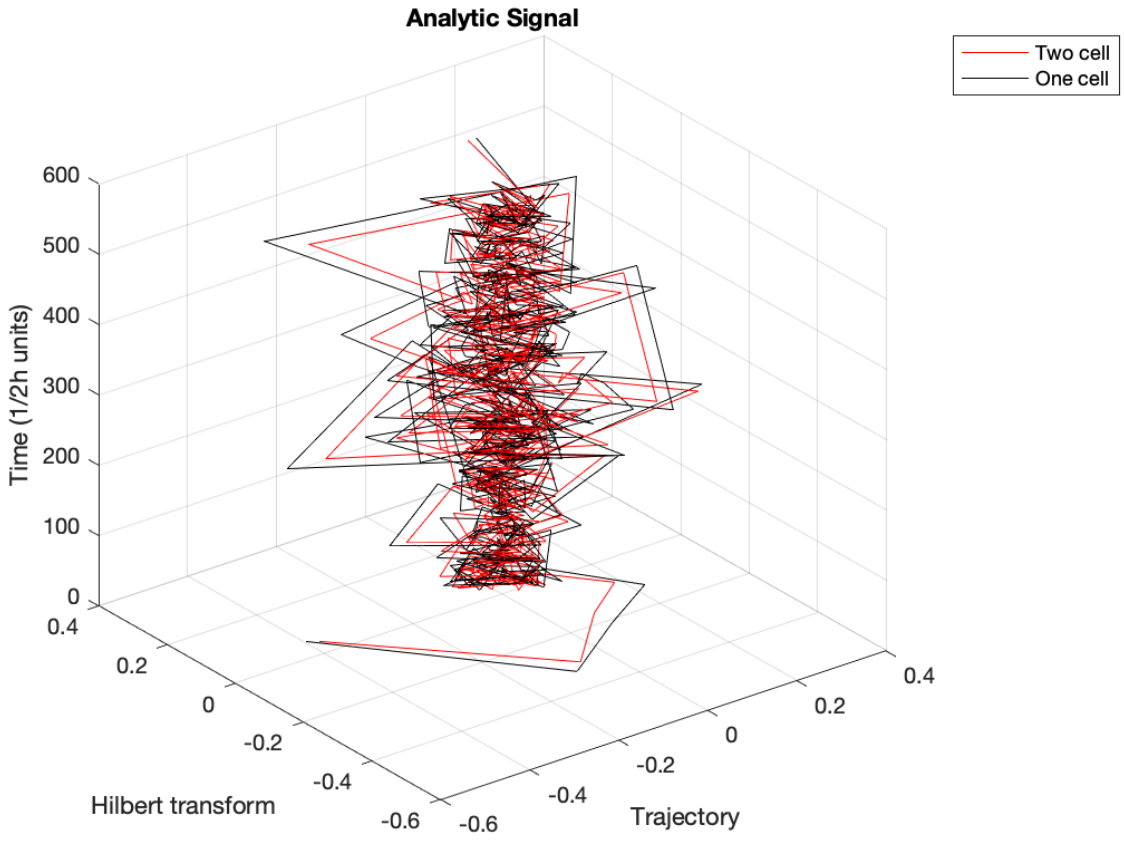

Figure 3.4 The curve (or analytic signal) $(x(t),\tilde{x}(t))$ spiraled over the complex plane for 2-cell (red) and 1-cell (black) trajectories over time with more phase variation in the 1-cell trajectories. The (x,y) plane is the complex plane. The vertical dimension is time. The trajectory of fluorescence x(t) was detrended after Rhodamine B normalization. The red tornado is an average over pairs of cells in the same droplet, and the black tornado is an average over single cells in a droplet [1].

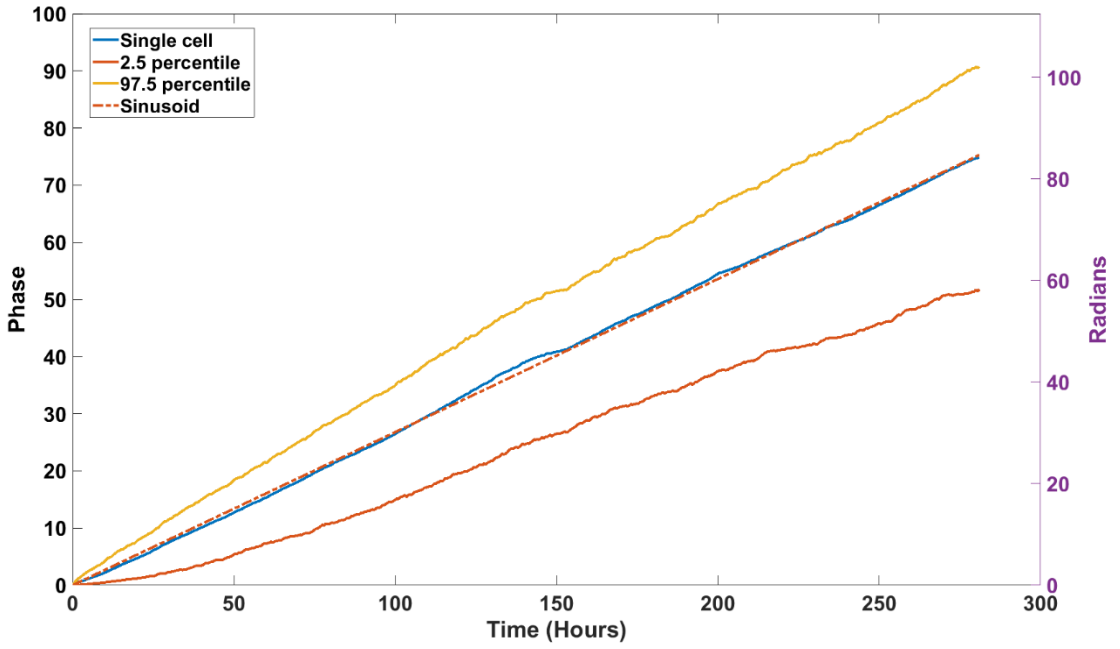

Figure 3.5 The phases of a sinusoid (in red) and that of the data (in blue) were similar, implying the oscillations are approximately sinusoidal. The average frequency ω = 0.3014 was used in x(t)=Acos(ωt+ϕ) to compute the phase (in red). The average frequency ω was computed from fitting $x(t) = Acos(\omega t + \phi)$ individually to ~1,591 cell trajectories by Least Squares using a Python script. The percentiles in yellow and orange are shown about the mean phase.

depend on the cell's social environment. Interestingly the 10-cell mean drifted outside of the 95 percent confidence band about the 1-cell droplet mean. Yet the phase variation also increases with time based on the percentiles of both 10-cell droplets and 10 singletons drifting away from the mean.

Phase also enters directly into the calculation of some synchronization measures [9]. One of the best studied synchronization measures is the Kuramoto order parameter(K) [22]:

$$K = \langle \left| n^{-1} \sum_{j=1}^{n} \exp\left(i\mathrm{F}_j\right) - \langle n^{-1} \sum_{j=1}^{n} \exp\left(i\mathrm{F}_j\right\rangle \right| \rangle.$$

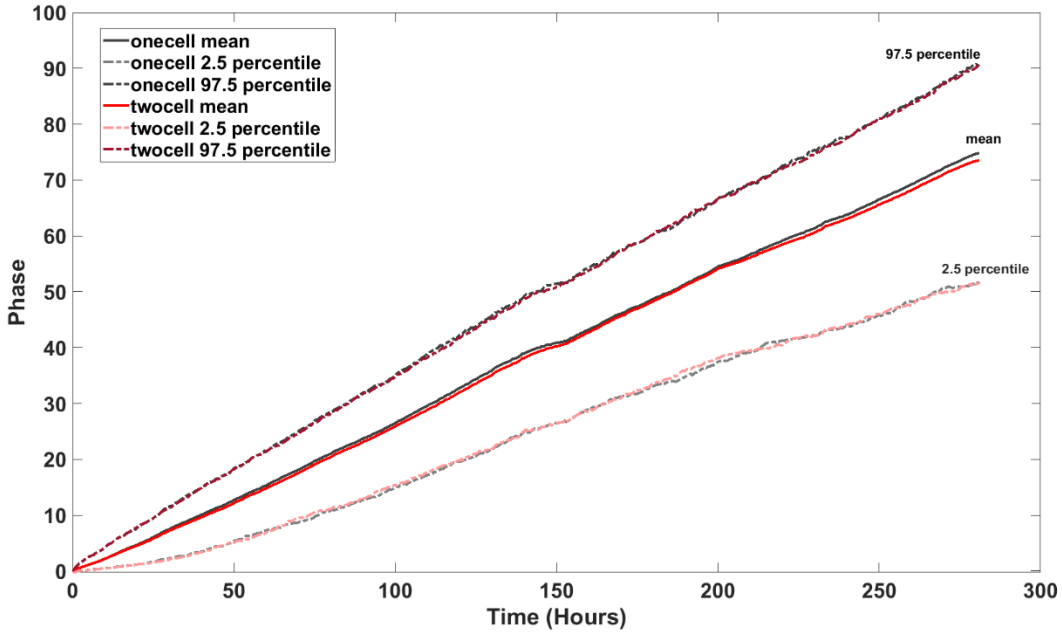

Figure 3.6 The average continuous Hilbert phase F^C (t) for 568 droplets each with 2 cells in one droplet (in red) is plotted against time differs from the average Hilbert phase F^C (t) for 1,644 cells each isolated in a single droplet and never having known neighbors. There was a total of 568 curves being averaged in the first case and 1,644 curves, in the second case. The 2-cell droplets are coded in red; the 1-cell droplets are coded in black. Percentiles are dotted lines with red-dotted lines belonging to the 2-cell droplets and with black-dotted lines belonging to the 1-cell droplets.

The quantity $F_j = F_j^C(t_1)$ is the continuous Hilbert phase for the jth oscillator with t_1 = 480 half hours. There are n oscillators in a particular social environment, and the brackets ⟨ ⟩ denote a time average. As K approaches 1, the collection of n oscillators approaches being fully synchronized. As K approaches 0, the collection of n oscillators approaches no synchronization. A nice feature of biological oscillators is their social environment. Cellular oscillators may live together in a droplet or may be separated into separate droplets.

The Kuramoto K for the 1-cell and 2-cell droplets was computed as follows to examine the effect of sociality on oscillators. For the 1-cell droplets, the n in equation above was set to 1, and the K was computed for each isolated cell. From the resulting 1,644 K values the mean (0.0322) and standard error (0.0007) of K was computed. For the 2-cell droplets, the n in the equation above was set to 2, and the K was computed for each cell pair in a droplet. From the resulting 568 =1,136/2 K values, the mean (0.7428) and standard error (0.0022) of K was computed.

It is useful to consider the effects of sociality on synchronization [9], and in particular to explain the difference in phase between 1-cell and 2-cell droplets in Fig. 3.4 and Fig. 3.5. As a control the Kuramoto K was calculated on all 1,644 1-cell droplets with the resulting K = 0.0322 +/- 0.0007, which is near zero as expected. Then the Kuramoto K was calculated on all 1,136 cells with n =2. The Kuramoto K for each pair of cells was then averaged over all droplets to yield K = 0.7428 +/- 0.0022. The conclusion is that for the 2-cell droplets cells within droplets are highly synchronized relative to the negative control provided by the 1-cell droplets, explaining the phase difference between 1- and 2-cell droplets in Fig. 3.4. A z-test of the difference was highly significant ($z = (.7428 - .0322)/\sqrt{(6.737X10^{-4})^2 + (.0022)^2}$= 314.13, P < .00001).

A Wilcoxon Rank Sum test of the two unpaired samples of Kuramoto K's for 1-cell and 2-cell droplets was also highly significant (P < 0.00001). Alternatively, a more conservative test is to create a sample of 568 pairs drawn randomly with replacement from the 1,644 cells in 1-cell droplets. With n=2 for the artificially generated pairs, in which strangers are neighbors, the average Kuramoto K for each pair of isolated cells was K = 0.6853 +/- 0.0015. A z-test of

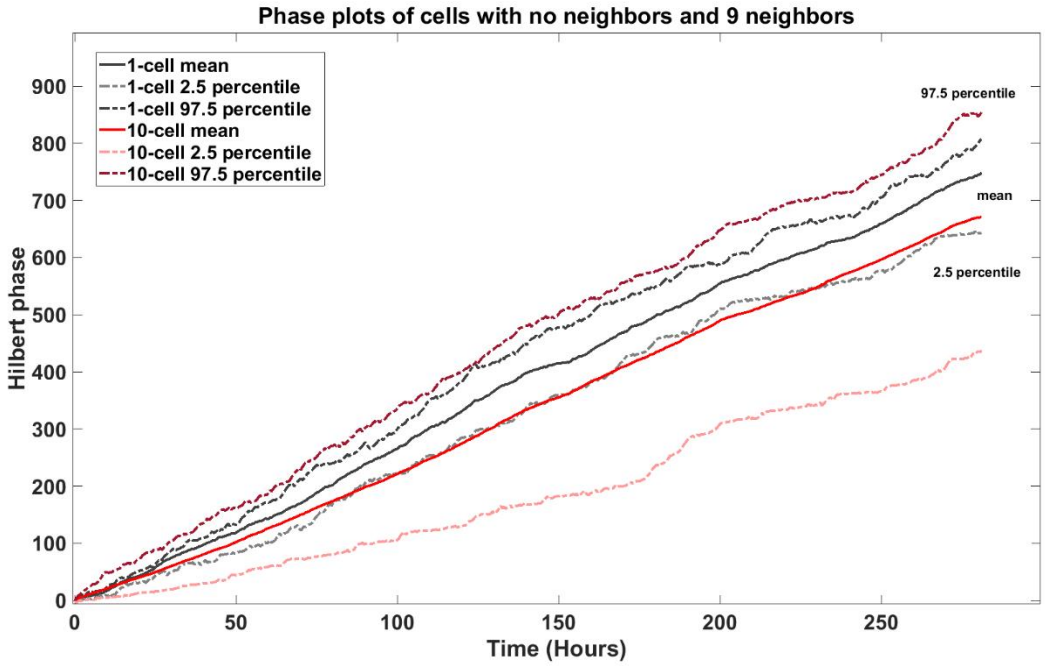

Figure 3.7 The average continuous Hilbert phase F^C (t) for 10 droplets each with ten cells in one droplet (in red) is plotted against time differs from the average Hilbert phase F^C (t) for ten cells each isolated in a single droplet and never having known neighbors. There was a total of ~100 curves being averaged in the first case and 10 curves, in the second case. The 10-cell droplets are coded in red; the 1-cell droplets are coded in black. Percentiles are dotted lines with red-dotted lines belonging to the 10-cell droplets and with black-dotted lines belonging to the 1-cell droplets.

the difference from cells that have truly experienced another cell was highly significant with z = 21.89 (P < .00001). The Wilcoxon Rank Sum test of the two unpaired samples of Kuramoto K's for 2-cell droplets vs artificially created 2-cell droplets was again highly significant (P <.00001). Other synchronization measures that reflect the sociality of the oscillators could be used to test for synchronization as well [9].

3.5 DISCUSSION

The notion of phase used here has been in use since the time of Huygens [23,24] in the description of coupled pendula right down to the present [16]. The notion of Hilbert Phase appeared early in the 20th Century [11]. All 4 phase measures ($\phi, F^H(t), F^C(t), and\ M^C(t_1, t_0)$) are inter-related and graphically summarized in figures 3.2 and 3.5. The first two measures ($\phi, F^H(t)$) are defined on the circle, while the last two measures ($F^C(t), and\ M^C(t_1, t_0)$) are defined on the non-negative part of the real line.

There are several advantages to the notion of phase $M^C(t_1, t_0)$ in cycles described here. The phase $M^C(t_1, t_0)$ can be computed in an integrated way with the period and amplitude of a periodic process using the Fast Fourier Transform [17], in for example, MATLAB (see scripts on GitHub in Materials and Methods). The phase measure does not presume a linear relation with time as does the phase shift ϕ in a sinusoid. Using the phase shift as a measure of phase would be problematic in extrapolating to later times, if the phase were nonlinear in time, as it is likely to be when there is cellular synchronization. The phase measure $M^C(t_1, t_0)$ is functionally independent of the period and amplitude derived from the periodogram or "power spectrum" [1]. This phase measure represents new information about a periodic process not embedded in the amplitude and period [1].

The four phase measures are useful in combination. For example, the use of the phase shift in combination with the phase in cycles allows an assessment of whether or not a periodic process x(t) is sinusoidal (Fig. 3.5). There is a generalization of the phase shift, namely the continuous Hilbert Phase $F^C(t)$ at time t = 0, which coincides with the phase shift, if the

process x(t) were sinusoidal. This generalized phase shift is normally subtracted from the phase because often synchronization experiments are done to minimize this quantity – cells are started in a nearly synchronized state at the beginning of an experiment. Both the continuous phase $F^C(t)$ and phase $M^C(t_1, t_0)$ in cycles are smooth and provide information on when cellular oscillators synchronize (Fig. 3.6). The original Hilbert phase $F^H(t)$ is also in some sense closer to the data and can be plotted on a circle (Fig. 3.2, lower panel). Some scientists also feel more comfortable with a notion of phase on the circle(20). In this case the circular plot reveals a nonuniformity in the phase angle $F^H(t)$ over the circle. Finally, the phase can be used to evaluate various models for how cellular clocks synchronize [7]. When studying synchronization, the phase may be used to construct additional measures of synchronization(9) as illustrated here for the Kuramoto order parameter.

Cellular clocks are fundamentally different from circadian rhythms at the macroscopic scale – they have a high frequency hour hand in addition to the circadian cycle (Fig. 3.3). Much of these high frequency cycles are generated by stochastic intracellular noise in reactions that go on within the cell [1]. There are multiple frequencies on which cellular clocks keep time, much like the two hands on a clock. This can be seen in the upper panel of Fig. 3.2, in which cycles completed is greater than those generated by circadian oscillations present in single cell oscillators [7].

The Hilbert phase has not only been used to study oscillators at the single cell level [1] but has been used to study other periodic biological processes, such as through monitoring the beating heart [25]. The Hilbert phase and its derivative measure of phase here potentially provide useful new information about a variety of biological rhythms.

3.6 REFERENCES

1. Deng Z, Arsenault S, Caranica C, Griffith J, Zhu T, Al-Omari A, et al. Synchronizing stochastic circadian oscillators in single cells of *Neurospora crassa*. Scientific Reports. 2016;6:35828.

2. Abraham U, Granada AE, Westermark PO, Heine M, Kramer A, Herzel H. Coupling governs entrainment range of circadian clocks. Mol Syst Biol. 2010;6(1):438. Epub 2010/12/02. doi: 10.1038/msb.2010.92. PubMed PMID: 21119632; PubMed Central PMCID: PMCPMC3010105.

3. Carr A-JF, Whitmore D. Imaging of single light-responsive clock cells reveals fluctuating free-running periods. Nature Cell Biology. 2005;7:319. doi: 10.1038/ncb1232 https://www.nature.com/articles/ncb1232#supplementary-information.

4. Muranaka T, Oyama T. Heterogeneity of cellular circadian clocks in intact plants and its correction under light-dark cycles. Science Advances. 2016;2(7).

5. Gould PD, Domijan M, Greenwood M, Tokuda IT, Rees H, Kozma-Bognar L, et al. Coordination of robust single cell rhythms in the Arabidopsis circadian clock via spatial waves of gene expression. eLife. 2018;7:e31700. doi: 10.7554/eLife.31700.

6. Gooch VD, Mehra A, Larrondo LF, Fox J, Touroutoutoudis M, Loros JJ, et al. Fully codon-optimized luciferase uncovers novel temperature characteristics of the Neurospora clock. Eukaryotic cell. 2008;7(1):28-37. PubMed PMID: 17766461.

7. Caranica C, Al-Omari A, Deng Z, Griffith J, Nilsen R, Mao L, et al. Ensemble methods for stochastic networks with special reference to the biological clock of *Neurospora crassa*. PloS one. 2018;13(5):e0196435.

8. Dong W, Tang X, Yu Y, Nilsen R, Kim R, Griffith J, et al. Systems biology of the clock in *Neurospora crassa*. PloS one. 2008;3(8):e3105. PubMed PMID: 18769678.

9. Deng Z, Arsenault S, Mao L, Arnold J. Measuring synchronization of stochastic oscillators in biology. J of Physics Conference Series, 29th Annual Workshop, 2016, Recent Developments in Computer Simulational Studies in Condensed Matter Physics, Athens, GA, 22-26 February, 2016. 2016;750(29th Annual Workshop, 2016, Recent Developments in Computer Simulational Studies in Condensed Matter Physics, Athens, GA, 22-26 Feb, 2016):012001. doi: doi:10.1088/1742-6596/750/1/012001

10. Judge M, Griffith J, Arnold J. Aging and the biological clock. Healthy Aging and Longevity. 2017;Circadian Rhythms and Their Impact on Aging:211-34.

11. Gabor D. Theory of communication. Part 1: The analysis of information. Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of. 1946;93(26):429-41.

12. Jeong B, Hong JH, Kim H, Choe HK, Kim K, Lee KJ. Multi-stability of circadian phase wave within early postnatal suprachiasmatic nucleus. Scientific Reports. 2016;6:21463. doi: 10.1038/srep21463

https://www.nature.com/articles/srep21463#supplementary-information.

13. Sisobhan S. Hilbert transform based time series analysis of the the circadian gene regulatory network. IET Systems Biology. 2019;10.1049/iet-syb.2018.5088.

14. Oprisan SA. A consisten definition of phase resetting using Hilbert Transform. Scholarly Research Notices. 2016;2017.

15. Bloomfield P. Fourier analysis of time series : an introduction. New York: Wiley; 1976. xiii, 258 p. p.

16. Kreuz T, Mormann F, Andrzejak RG, Kraskov A, Lehnertz K, Grassberger P. Measuring synchronization in coupled model systems: A comparison of different approaches. Physica D: Nonlinear Phenomena. 2007;225(1):29-42.

17. Marple SL. Computing the discrete-time “Analytic” Signal via FFT. IEEE Transactions. 1999;47(9):2600-3.

18. Castro-Longoria E, Ferry M, Bartnicki-Garcia S, Hasty J, Brody S. Circadian rhythms in *Neurospora crassa*: Dynamics of the clock component *frequency* visualized using a fluorescent reporter. Fungal Genetics and Biology. 2010;47(4):332-41.

19. Hille E. Analytic Function Theory. New York, New York: Chelsea Publishing Company; 1959.

20. Winfree AT. *The Geometry of Biological Time*: Springer Science & Business Media; 2001.

21. Deng Z. Single-cell analysis on the biological clock using microfluidic droplets. University of Georgia PhD Dissertation. 2017.

22. Shinomoto S, Kuramoto Y. Phase Transitions in Active Rotator Systems. Progress of Theoretical Physics. 1986;75(5):1105-10. doi: 10.1143/PTP.75.1105.

23. Huygens C. *Horologium Oscillatorium sive de motu pendulorum,*. In: English translation by Richard J Blackwell (1986) *TPCoGDCtMoPaAtC,* Iowa State University Press, Ames, editor.: F. Muguet, Paris; 1673.

24. Oliveira HM, Melo LV. Huygens synchronization of two clocks. Scientific Reports. 2015;5:11548. doi: 10.1038/srep11548.

25. Mojtaba Jafari T, Eero L, Tero H, Juho K, Jonas E, Mikko P, et al. A real-time approach for heart rate monitoring using a Hilbert transform in seismocardiograms. Physiological Measurement. 2016;37(11):1885.

CHAPTER 4

A STOCHASTIC CLOCK NETWORK WITH LIGHT ENTRAINMENT IS IDENTIFIED FOR SINGLE CELLS OF *NEUROSPORA CRASSA* BY ENSEMBLE METHODS

Caranica, C., A. Al-Omari, H-B. Schuttler and J. Arnold

Submitted to *Nature Scientific Reports*

ABSTRACT

Stochastic networks for the clock were identified by ensemble methods using genetic algorithms that captured the amplitude and period variation in single cell oscillators of *Neurospora crassa*. The genetic algorithms were at least an order of magnitude faster than ensemble methods using parallel tempering and appeared to provide a globally optimum solution from a random start in the initial guess of model parameters (*i.e.*, rate constants and initial counts of molecules in a cell). The resulting goodness of fit χ^2 was roughly halved versus solutions produced by ensemble methods using parallel tempering, and the resulting χ^2 per data point was only χ^2/n = 2708.05/953 =2.84. The fitted model ensemble was robust to variation in proxies for "cell size". The fitted neutral models without cellular communication between single cells isolated by microfluidics provided evidence for only *one* Stochastic Resonance across days from 6 h to 36 h of light/dark (L/D) or in a D/D experiment. When the light-driven phase synchronization was strong as measured by the Kuramoto (K), there was a degradation in the single cell oscillations about the stochastic resonance. The rate constants for the clock stochastic network were consistent with those determined on a macroscopic scale of 10^7 cells.

4.1 INTRODUCTION

One of the main challenges of systems biology is explaining the dynamic behavior of single cells with their stochastic intracellular variation [1, 2]. This stochastic intracellular variation has profound consequences on the regulation and phenotypes of genetically identical individual cells[3, 4]. One example is the effects of stochastic intracellular variation on the dynamics of genes and their products involved in the biological clock [5, 6]. While populations of 10^7 cells/ml display highly synchronized behavior producing regular oscillations at the macroscopic scale, the behavior of individual cells is quite different. There is now evidence that individual cells in *Neurospora crassa* have clocks [5], but there is substantial variation in phase between the clocks in different cells. What mechanisms at the single cell level explain how cells oscillate, and how do these cells come to oscillate in phase on a macroscopic scale?

There are three hypotheses for how cells come to oscillate as they transition from the single cell level to the macroscopic level. One possibility is that there is some form of chemical signal shared between cells that allows cells with different clock phases to reinforce and synchronize each other [6-8]. A second possibility is that the noise itself can play a positive role in generating oscillations[9], and the mechanism for noise producing oscillations can invoke a physical hypothesis for biological oscillators known as Stochastic Resonance[10]. A third possibility is that there is some cell cycle gated mechanism that imposes regular oscillations on single cells [11-14].

These three mechanisms can be examined using flow focusing microfluidics [15] to capture individual cells under particular conditions for observation and to manipulate the environment of the cell to test individually these hypotheses under the effects of a variety of factors, such as light [6]. The conditions of the experiment here are used to isolate and test the Stochastic Resonance Hypothesis. Single cells are isolated in different droplets for observation so that they cannot communicate. Also single cells are maintained in media so that they cannot divide [6]. In this way the effects of cell-to-cell communication and cell cycle-gating on the clock can be eliminated. Only the mechanism of Stochastic Resonance remains to be examined [10].

The Stochastic Resonance Hypothesis can be viewed as a prediction of a reasonable null hypothesis or "Neutral Model" [16] specified by a stochastic clock network (Fig 3.1) that does not invoke any other mechanism to explain clock-like behavior. In previous work the data on ~1591 isolated cells were used to test the adequacy of a clock stochastic network in the dark (D/D) to provide initial evidence for the Stochastic Resonance hypothesis [17]. Here we use additional light entrainment data on single cells to construct a stronger test of this neutral model for explaining clock-like properties and to explore its limitations using light entrainment of single cells under a 6 h, 12 h, and 36 h artificial day with equal amounts of light and dark (L/D) [5]. In addition to providing a stronger test of the neutral model the light entrainment data can also be used as a stronger test of the Stochastic Resonance Hypothesis.

Examining the neutral clock stochastic network without communication and without cell cycle gating is of particular interest under varying light regimes. In some models and experimental systems oscillators are hypothesized to have limits to their ability to entrain to an

external entrainment signal - if the driving signal has a period sufficiently far from the intrinsic period of the cell oscillator, then entrainment fails [18]. These entrainment limits have been examined in the mammalian Suprachiasmatic Nucleus (SCN) [19].

The model filamentous fungus, *N. crassa*, is particularly well suited to test this neutral model for the clock because isolated cells maintain circadian oscillations [6]. The *N. crassa* model system is then complementary to cells in the SCN, which usually cannot sustain oscillations when isolated [20]. The *N. crassa* system is also complementary to another major model system for the clock, the cyanobacterium, *Synechoccocus elongatus*, because *N. crassa* can entrain to light at the single cell level [5]. In contrast *S. elongatus* shuts down transcription in the dark, making it more difficult to study light entrainmen t[21]. Thus, *N. crassa* is particularly well suited to use both dark (D/D) and light entrainment experiments (L/D) to provide a strong test of the neutral hypothesis of the clock stochastic network and Stochastic Resonance.

There are several questions to be addressed about this neutral model: (1) is the stochastic network of the clock consistent with the available single cell data? (2) if not, how does the model fail? (3) Is there a limit to the neutral model's ability to explain light entrainment data? [18] (4) When the amount of single cell data is quadrupled to include light entrainment, how does support for the Stochastic Resonance Hypothesis hold up with light entrainment data and data in the dark? (5) Does stochastic intracellular variation through a Stochastic Resonance mechanism have consequences for circadian oscillations seen in genetically identical cells?

Because this chapter is in some places quite technical, a road map for it is now provided. The end point and message for this work is the last figure, demonstrating a stochastic resonance in the fitted network. Below or above the stochastic resonance the circadian rhythms of single cells are degraded. At the beginning the model is laid out. The novel element to this stochastic network for single cells of *N. crassa* is the inclusion of light as a molecular species. The structure of this network is suggested by earlier deterministic models on the macroscopic scale of 10^7 cells/ml [22]. Then this stochastic network for single cells is fitted to the average periodogram or power spectrum using ensemble methods originally introduced by the authors to systems biology from statistical physics [23]. In the initial implementation of these ensemble methods for the D/D data it was found that the Metropolis-Hastings method of Markov Chain Monte Carlo (MCMC) was insufficient for identification of the model ensemble, but more sophisticated parallel tempering methods were successful in identifying the stochastic network for single cells in the dark (D/D) [17]. So, parallel tempering methods for fitting the clock stochastic network became the natural starting point for fitting model ensembles to L/D data here. These more complicated stochastic networks with a light response proved to be a challenge for parallel tempering methods. It was necessary to develop a novel approach to ensemble methods using genetic algorithms [24]. Genetic algorithms represent a very broad class of optimization methods [25], and here two recently developed genetic algorithms [26, 27] were used to identify model ensembles.

Then an assessment of goodness of fit was made using the Hilbert phase, which is functionally independent of the period and amplitude captured in the average periodogram of single cells [28]. Consideration of the phase over time showed precisely where the model

ensemble succeeded and where the ensemble needed improvement when the effects of light synchronization were weak. Stochastic networks have one additional parameter, the "size of a cell", that determines the level of stochastic intracellular noise in a cell. An empirical approach was developed to identify the level of stochastic intracellular noise in a cell by relating the noise to the cell's RNA/DNA and protein/DNA ratios. The model fitting was shown to be robust to variation in the RNA/DNA and protein/DNA ratios and hence in the level of stochastic intracellular noise. Having identified a promising "neutral model" with no other hypothesized factors affecting cellular phase synchronization, it was demonstrated that there was only one stochastic resonance in the fitted ensemble for a variety of L/D experiments and that as the noise was varied away from this stochastic resonance, circadian oscillations were degraded. These last two observations are the major points of the chapter.

4.2 MODEL

The neutral model for each genetically identical cell is a stochastic network displayed in Fig 4.1, and the broad outline of its features are given in Fig 4.1A. The network of genes and their products begins with three clock mechanism genes in Fig 4.1A: (1) the gene *frequency* (*frq*) encoding the oscillator protein FRQ; (2) one of two activator genes, *white-collar-1* (*wc-1*) encoding WC-1; and (3) the second of two activator genes, *white-collar-2* (*wc-2*) encoding WC-2. The positive elements WC-1 and WC-2 are transcription factors that form a White-Collar Complex (WCC) [29]. In Fig 4.1A the WCC protein activates the oscillator gene *frq*, which in turn produces ultimately the FRQ protein, which is involved in deactivating the

complex WCC. This negative feedback loop in part explains the origin of oscillations at the macroscopic scale[22]. There is also a positive feedback loop involving FRQ acting on the *wc-1* mRNA (*wc-1*^{r}) in Fig. 4.1A. The "stabilization" of this *wc-1* mRNA by FRQ is crucial to explaining oscillations as well at the macroscopic scale [22].

The details of how the stochastic network functions are given in Fig 4.1B. A single cell is described by the counts of genes and their cognate messenger RNAs (mRNAs) and proteins in a cell. The molecular counts of species change at rates (the labels on circles) associated with the different reactions (circles) in the kinetic network. As examples, all clock mechanism genes (*frq*, *wc-1*, and *wc-2*) are transcribed at a rate Sx (*e.g.*, S4) and translated at a rate Lx (*e.g.*, L3). Messenger RNAs(mRNAs) and proteins decay at a rate Dx. The key reactions for oscillations at the macroscopic level are the rate of activation/deactivation (A and of $\bar{A}$) of the oscillator gene *frq*, the rate of deactivation P of WCC by FRQ, and the decay rate D7 of the stabilized mRNA *wc-1*^{r1} [22]. There are a total of 23 reaction rates and 12 initial conditions for a total of 35 parameters in this model.

Genes under the control of the clock mechanism are called *clock-controlled genes* (*ccg*). One *ccg* of particular interest is the hypothesized gene that produces the autoinducer or quorum sensing signal S_i synchronizing the clocks in different cells (indexed by i). Under the neutral hypothesis the rates of production of this signal are zero (*i.e.*, if $K_{S1} = 0$). There is no communication hypothesized between cells in this paper in that cells are isolated in different droplets. Another is the recorder gene *ccg-2P*:mCherry in the MFNC9 strain for observing the operation of the clock (*i.e.*, the hands on the face of the clock)[30]. In this model the maturation

of the mCherry protein is captured in transcription rate (Sc) and translation rate (Lc). The degradation rate of the mCherry protein is captured by Dcp.

The one novel feature from earlier work [17] is the presence of photons as a light species in Fig 4.1B. This introduces novel light (C2IL) and dark reactions (C2) for the production of WCC as in earlier work [31] (see Fig 4.2 of this earlier work). This slight extension of the model can be shown to be formally equivalent to another network with only one reaction ($C_2 + C_2 f_{IL} s(t)$) producing WCC that varies with time in the following way.

A photon species named *phot* is introduced in Fig 4.1 whose temporal trajectory is not obtained from solving the Master Equation by the Gillespie Algorithm [32] but is given to us. The concentration [*phot*] is an exogenous variable of the form:

$$[phot] = I_L s(t),$$

where I_L is the light intensity and $s(t)$ switches between "Light On" (L), with "On"-intensity I_L, and "Light Off" (D), after every time interval t_{LD}, starting with "L" at time $t_{L,0}$:

$s(t) = 1$ for $t_{L,n-1} \leq t < t_{L,n}$ and $n = 1, 3, 5, 7, \ldots$ (i.e., if n is odd)

$s(t) = 0$ for $t_{L,n-1} \leq t < t_{L,n}$ and $n = 2, 4, 6, 8, \ldots$ (i.e., if n is even).

$t_{L,n} = t_{L,0} + n\, t_{LD}$ for $n=0, 1, 2, 3, 4, \ldots$.

So, for the experiments in [5, 6] as an example for the 12 h day, the specification of the switch s(t) would be:

$t_{LD} = 6h$, $I_L = 5300$ lux, $t_{L,0} = 0$ h.

Here we assume $t_{L,0} = 0$ h is the time when the L/D exposure cycles were started. There is then one rate for a L/D experiment of the form

$$C_2 + C_{2L} I_L s(t) .$$

Because we cannot separate the product $C_{2L} I_L$, we treat it as one parameter called C_{2IL} in Fig 4.1B, which has the same units as C_2. and defined as $C_{2IL} = C_{2L} I_L$. Also, for the L/D exposure to have any substantial effect on the kinetics of the system, it is reasonable to assume that the value of $C_{2L} I_L$ is comparable to C_2, in order of magnitude. For example, if $C_{2L} I_L \ll C_2$ the effect of light exposure on the kinetics is probably negligibly small. On the other hand, if $C_{2L} I_L \gg C_2$ the effect of light exposure would completely dominate the kinetics, *i.e.*, the periodogram of the [CCG]-signal would look very different from the "dark clock" model periodogram, with a strong peak at 12h period for 12h-day experiment, and much less intensity at a 24h period.

So, to get a reasonable estimate for $C_{2L} I_L$, we write

$$C_{2L} I_L = f_{IL} C_2$$

with f_{IL} being a parameter varying in the range (0.01, 100). The parameter C_2 is specified from the D/D experiments, and the parameter f_{IL} affecting illumination was initially set to 2, but then was allowed to float in the fitting. The parameter f_{IL} is kept constant across L/D experiments because these experiments were done on the same apparatus and conditions except for variation in the length of the day [5].

4.3 MATERIALS AND METHODS

4.3.1 SINGLE CELL DATA OF *N. Crassa*

The single cell data from four experiments are used to evaluate the stochastic clock network in Fig 4.1. The cells in these experiments are equipped with an mCherry recorder under the control of a *clock-controlled gene-2* promoter (*ccg-2P*) [30]. This fluorescent mCherry strain is referred to as MFNC9 [30]. Each of the four experiments involved isolating over 1,000 cells in individual droplets, synchronizing cells initially with 26 h of light, and then observing their fluorescence every half hour for at least 10 days [5]. Four experiments were conducted, one in the dark (D/D) and three under 6 h, 12 h, or 36 h L/D regimes with equal amounts of light and dark [5].

The D/D data are available [17], and the L/D entrainment data are available at the IEEE Dataport, https://ieee-dataport.org/documents/single-cells-neurospora-crassa-show-circadian-oscillations-light-entrainment-temperature

4.3.2 RESCALING FROM DETERMINISTIC TO STOCHASTIC MOLECULAR NUMBER UNITS

A method for rescaling initial concentrations and reaction rates of a deterministic network to molecular counts and reaction rates of a stochastic network was described in chapter 2.

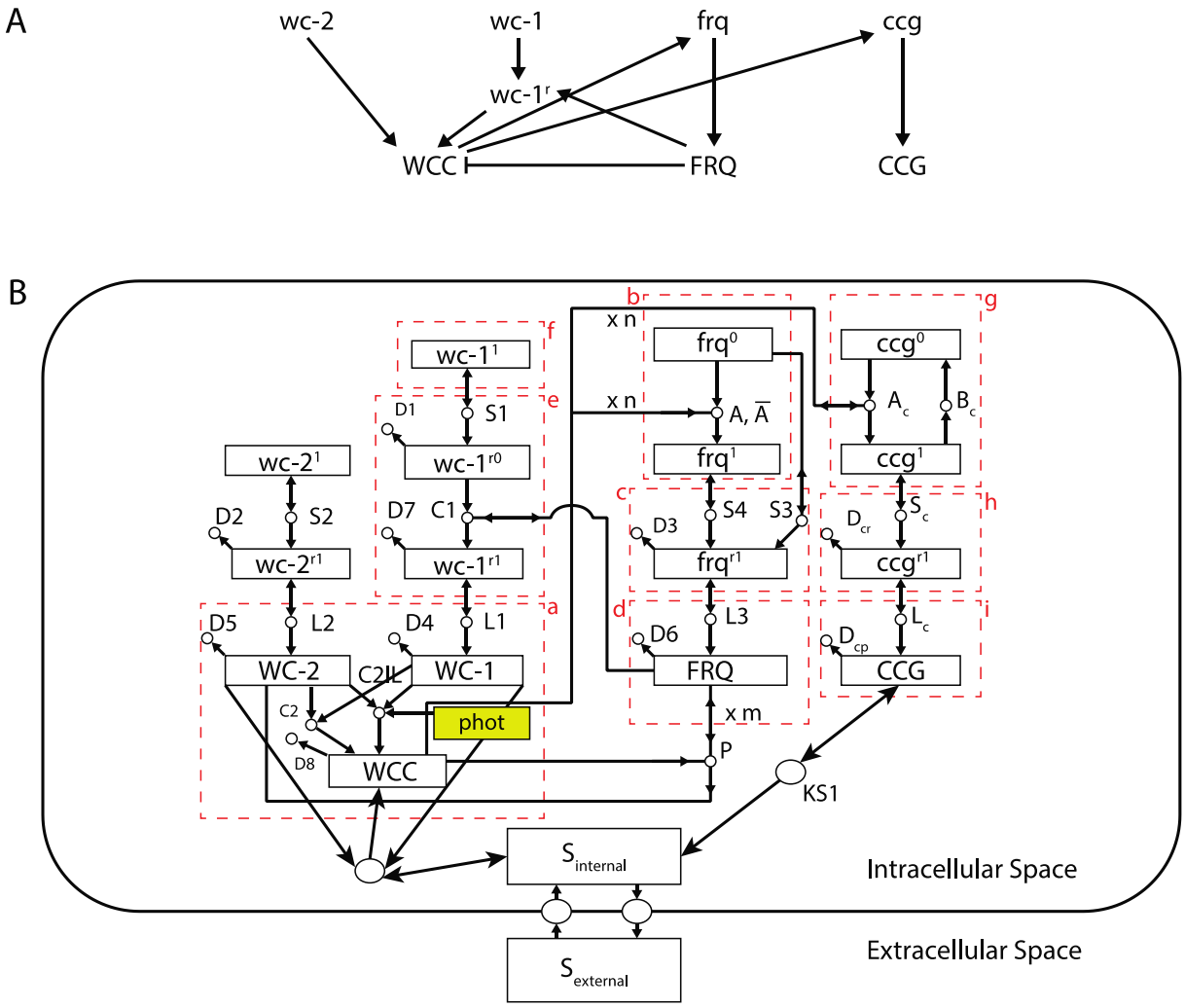

Figure 4.1 (A) The key elements of the clock stochastic network are summarized. There are both a negative feedback loop, in which WCC activates the gene *frq* encoding the oscillator protein and a positive feedback loop in which the FRQ protein stabilizes the *wc-1*^{r} mRNA. The genes *wc-1* and *wc-2* are the positive elements in the clock, while the *frq* gene is the negative element in the clock. (B) The full specification of the model is given by the network in panel B. Circles denote reactions, and boxes represent reactants and products in the network. Double arrows denote catalytic reactions. The labels on reactions do double duty as both label for the reactions and as rate coefficient(s) for a particular reaction. Those reactions with no resultant product constitute decay reactions. All proteins and mRNAs have decay reactions as examples. The *ccg* gene and its cognate products could be *ccg-2* or a hypothetical gene *ccg* encoding a quorum sensing signal as examples. The red dotted boxes denote components of the network across which there is approximately no net flow of molecules. Typically, the dotted boxes are only crossed by catalytic reactions. Modified from [17].

4.3.3 RESCALING WITH THE RNA/DNA AND PROTEIN/DNA RATIOS WITHOUT CHANGING THE NETWORK DYNAMICS

The model above specifies the Master Equation, which describes how the counts of molecular species in Fig. 4.1B change over time [33]. The Master Equation can be approximated by the Chemical Langevin Equation [34], which consists of two components, a deterministic term and a noise term. The deterministic term corresponds to a system of ordinary differential equations. The first term is required to be invariant under rescaling by the RNA/DNA and protein/DNA ratios to leave the network dynamics invariant. Consider one component of the deterministic term, namely the L3 and D6 reactions in dotted box d of Fig 4.1B:

$$frq^{r1} \underset{L3}{\Rightarrow} FRQ + frq^{r1}, FRQ \underset{D6}{\Rightarrow} \varnothing$$

The contribution to the dynamics of FRQ by this reaction is:

$$\frac{d[FRQ]}{dt} = L3[frq^{r1}] - D6[FRQ]$$

The ratios of RNA/DNA ($R_{RNA:DNA}$) and protein/DNA ($R_{Prot:DNA}$) are measured experimentally or changed to vary the stochastic intracellular noise. If the ratios are changed, then the scales of RNA and protein counts change as well so that

$$\frac{dR_{Prot:DNA}[FRQ]}{dt} = L3_{new}R_{RNA:DNA}[frq^{r1}] - D6_{new}R_{Prot:DNA_2}[FRQ],$$

which becomes

$$\frac{d[FRQ]}{dt} = L3_{new}\frac{R_{RNA:DNA}}{R_{Prot:DNA}}[frq^{r1}] - D6_{new}\frac{R_{Prot:DNA}}{R_{Prot:DNA}}[FRQ].$$

One way that the dynamics remain unchanged is if:

$$L3 = L3_{new}\frac{R_{RNA:DNA}}{R_{Prot:DNA}} \text{ or } L3_{new} = L3\frac{R_{Prot:DNA}}{R_{RNA:DNA}} \text{ and } D6_{new} = D6.$$

In this way by stepping through all of the dotted boxes in Fig 4.1, all 23 reaction rates can be rescaled to preserve the original dynamics when the RNA/DNA and protein/DNA ratios are changed to vary the noise in the stochastic network. The rescaling for some other dotted boxes is illustrated in chapter 2.

A bias-corrected periodogram averaged over cells was used as the model-fitting criterion. The bias correction of the periodogram was also described in chapter 2.

4.3.4 USE OF PARALLEL TEMPERING METHOD

One of the methods used to simulate the behavior of the stochastic clock network under the 4 different regimes was parallel tempering. We chose the temperature grid (K) of the parallel tempering method as explained in chapter 2. Parallel tempering algorithm was performed for about 30,000 Monte Carlo updates in Fig 4.2A, where by update we mean the steps 1), 2) and 3) described in the add-temperature process (see chapter 2).

In implementing this temperature grid above, three initial conditions for the parameters were tried, and in one of the MCMC runs the target replica at temperature T_1 stopped swapping with the neighboring temperature late in equilibration. To eliminate this problem the linear temperature grid was allowed to increase again to include 60 temperatures during equilibration in Fig 4.2A.

4.3.5 THE ENSEMBLE FOR STOCHASTIC CLOCK NETWORK IS DETERMINED BY RUNS OF GENETIC ALGORITHMS FOLLOWED BY METROPOLIS-HASTINGS

Since parallel tempering did not provide a satisfactory fitting (see Fig. 4.2A) we employed two genetic algorithms to try to find regions of parameter space with a small χ^2 [26, 27]. The two algorithms used in the simulations are part of the family of genetic algorithms known as Particle Swarm Optimization (PSO) algorithms. PSO algorithms try to optimize a function **f** defined on a domain $\mathcal{D}$ by dividing a population of particles $\mathbf{x_i}$ in $\mathcal{D}$**,** i=1,2,…,sz, into groups called swarms and letting these swarms look for regions in the parameter space that could contain the optimum value of **f**. To make the exploration of the parameter space more effective, the swarms are encouraged to share information among themselves. Usually, 90% of the total number of generations are used to explore the parameter space to find promising region(s) that could contain optimum values of **f**. The exploration phase is followed by exploitation, whereby the algorithm speeds up the convergence of particles to an optimum value of **f**. In the following, we assume that we want to minimize the function **f**, *e.g.*, the χ^2, in equation (2), so $\mathbf{f(x_1)<f(x_2)}$ means $\mathbf{x_1}$ is better than $\mathbf{x_2}$.

The Dynamic Multi-Swarm Particle Swarm Optimizer with Cooperative Learning Strategy (DMS-PSO-CLS) genetic algorithm is now briefly described [27]. It has three features for optimization: (1) swarms of particles, moving in the parameter space with the best particle of a swarm being denoted by ***pbest***; (2) a culling/recombination stage at the end of a generation where each parameter of the two worst particles in each swarm is replaced by the corresponding parameter of one of the ***pbest*** particles; (3) a regrouping between particles (or migration between

swarms) every RR generations Then the process is repeated in each succeeding generation, 600 to 1000 generations.

Swarm movement

Each of sz particles with NN=4 particles per swarm in MM swarms has an inertia of *w*, which decreases linearly with generation, the initial value being w_1 and the final value being w_2. The acceleration constants c_1 and c_2 determine in part the genetic algorithms in Table 4.1. Each particle i has a position component x_i^d and velocity component v_i^d on parameter d in the D-dimensional parameter space, d=1,2,…,D. The dimension D is 35 here. All parameters in the model are rescaled to the unit cube $[0,1]^{35}$. Initial conditions were chosen as part of a Sobol space-filling sequence in the parameter space [44] or randomly from within the unit cube [45].

Equations of motion of the swarm are given by [27]:

$$v_i^d \leftarrow wv_i^d + c_1 \cdot rand1_i^d\left(pbest_i^d - x_i^d\right) + c_2 \cdot rand2_i^d\left(lbest_i^d - x_i^d\right) \qquad (3)$$

$$x_i^d \leftarrow x_i^d + v_i^d$$

for the exploration phase and by

$$v_i^d \leftarrow wv_i^d + c_1 \cdot rand1_i^d\left(pbest_i^d - x_i^d\right) + c_2 \cdot rand2_i^d\left(gbest^d - x_i^d\right) \qquad (4)$$

$$x_i^d \leftarrow x_i^d + v_i^d$$

for the exploitation phase.

The vector $\boldsymbol{pbest_i}$ is a particle's historically best position in the parameter space according to (2). The vector $\boldsymbol{lbest_i}$ is the historically best position in the i[th] particle's swarm.

Table 4.1 Genetic algorithms with characteristics below were used to optimize the likelihood function in (2) and produce an ensemble of models from the 4 experiments described in Materials and Methods. Each run was initialized with θ – parameters either initially positioned on a space-filling Sobol sequence[45] or randomly within the 35-dimensional parameter space including the illumination parameter f_{IL} (see Materials and Methods). All genetic algorithms were run for 600-1,000 generations to equilibrate the search for an optimum to Equation (2).

Method	No. of swarms M	Particles per swarm N	Initialization of particles	Number of generations	Final χ^2
DMS-PSO-CLS[27]	5	4	Sobol	1,000	4607.65*
DMS-PSO-CLS	20	4	Sobol	1,000	2773.07
DMS-PSO-CLS	10	4	Sobol	1,000	2781.44
DMS-PSO-CLS	10	4	random	1,000	3703.9
DMS-PSO-CLS	10	4	random	600	2743.5
PSO-DLS[26]	10	4	random	600	2933.5
PSO-DLS	5	4	Sobol	1,000	2708.05
PSO-DLS	20	4	Sobol	1,000	2797.88
PSO-DLS	10	4	Sobol	1,000	3436.22
PSO-DLS	10	4	random	1,000	2772.33
PSO-DLS	10	4	random	1,000	2880.11
PSO-DLS	10	4	random	1,000	2941.89
PSO-DLS	5	4	random	1,000	6702.86*
PSO-DLS	20	4	random	1,000	2768.15

*These two algorithms had only 20 particles and were eliminated from further consideration.

The vector **gbest** represents the position of the globally best solution, *i.e.*, the best particle in the whole population. Note that when **lbest** and **gbest** are calculated, the historically best solution of all particles in a swarm and historically best of particles in the whole population are

being recorded, respectively. The quantities $rand1_i^d$ and $rand2_i^d$ are uniform random numbers drawn from [0,1] that vary with each update to the velocity of a particle.

Culling and Recombination

The genetic algorithm in row 2 (Table 4.1) with $\chi^2 = 2773.07$ is used to illustrate culling and recombination. For each dimension d, there is a random draw of size 2 from the 80/4 = 20 **pbest** particles moving on the parameter space. For each dimension d of the parameter space, 2 of the **pbest** particles are randomly chosen, and the best of them will donate parameter d to one of the worst particles. The method is repeated for second worst particles. Since this random draw of 2 best particles is redone for each dimension, it is possible that the 2 worst particles will have contributions from more than 2 of the best particles. In other words, there is recombination between all of the best particles in culling the 2 worst particles from each swarm.

Regrouping (or migration)

Particles are randomly regrouped every RR generations. The constants used were: (1) w_1= 0.9; (2) w_2= 0.4; (3) $c_1 = 1.49445$; (4) c_2 = 1.49445; (5) v_{max} = 0.2; (6) v_{min} = -0.2 (7) RR = 5[27]; (8) $w_3 = 0.2.$ [27]; (9) T =1 in Equation (2). Inertia weight *w* was decreased

linearly from w_1 to w_2 during exploration phase and was kept constant at w_3 during the exploitation phase. Pseudocode is available in [27].

An alternative genetic algorithm named Particle Swarm Optimization with Dynamic Learning Strategy (PSO-DLS) [26]was also tried (Table 1). There were only two stages, swarm movement and migration. Sharing of information among swarms is done by communicating particles. With probability 1-*p*, a particle does not communicate with other swarms, and its movement is described by (3). With probability *p* a particle does communicate with other

swarms, and its movement is described by the following:

$$v_i^d \leftarrow wv_i^d + c_1 \cdot rand1_i^d\left(pbest_i^d - x_i^d\right) + c_2 \cdot rand2_i^d\left(\frac{1}{MM}\sum_{m=1}^{MM} lbest_m^d - x_i^d\right) \qquad (5)$$

$$x_i^d \leftarrow x_i^d + v_i^d$$

Equation (4) is used during the exploitation phase.

This is sometimes referred to as the admixture model of migration with *p* as the migration rate[46]. The same acceleration constants, c_1 and c_2, were used and set to 1.49445. The admixture parameter linearly increases with generation t according to *p* = t/iter, where iter is the total number of exploration generations (in our case 540 or 900 in Table 4.1). There is pseudocode available [26].

Metropolis-Hastings accumulation followed equilibration using the 12 best solutions in Table 4.1 and were combined at the end to produce a reconstruction of the likelihood in (2). A total of 14,000 total updates were performed for each of the 12 chains. The first 3,500 updates were used to adjust the parameter step widths in the Metropolis-Hastings algorithm and were discarded [47]. From the remaining 10,500 updates, every 35th model was sampled for a total of 300 samples from each of the 12 chains. The final sample for the accumulation run consisted of 12 x 300 = 3,600 models. Summary statistics on each model parameter for the fitted ensemble are given in Table 4.2.

4.4 RESULTS

4.4.1 OBTAINING THE FITTED STOCHASTIC NETWORK TO THE SINGLE CELL DATA IN BOTH D/D AND L/D EXPERIMENTS

The equilibration process to fitting the ensemble to the D/D data and 3 L/D experiments described in Materials and Methods occurred in three stages by parallel tempering. In the first stage the grid of temperatures was allowed to grow to 17 chains with 15,117 updates.

Beginning with an initial χ^2 = 10,507, the ending achieved was χ^2 = 6,371. In order to promote further communication between replicas at different temperatures, the temperature grid was expanded to 60 chains with 10,567 updates for a final χ^2 = 5,977. In the final stage the illumination parameter f_{IL} was allowed to float in the fitting process from a value of 2. In the final stage the fitting improved to a χ^2 = 5,410 after 8,030 more updates with 60 chains as shown in Fig 4.2A.

4.4.2 STRONG TEST OF THE NEUTRAL MODEL WITH LIGHT ENTRAINMENT ON SINGLE CELLS

In chapter 2 the neutral model in Fig 2.1 was tested against D/D single cell fluorescent data alone on the MFNC9 strain with a CCGp:mCherry recorder using a periodograms (of model and data) with 256 frequencies. The power of the 21 h signal in the model could be varied by changing the amount of amplification occurring during transcription and translation in the network (Fig 2.1B). With little amplification the final molecular counts in the cell of the CCGp:mCherry recorder would be smaller and noisier; with substantial amplification the final molecular counts would be

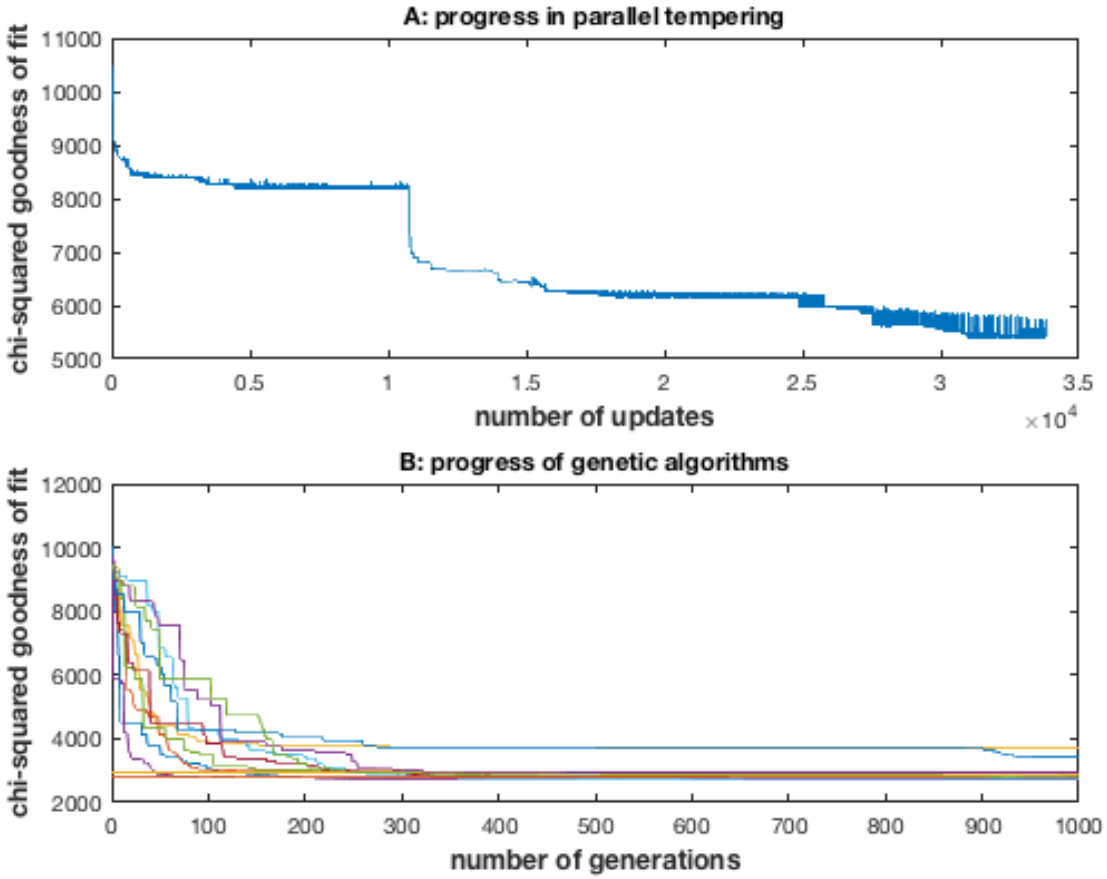

Figure 4.2 The chi-squared goodness of fit statistic improved during a Monte Carlo simulation used for fitting the model ensemble in Fig 4.1 to average periodograms for the D/D experiment and 3 L/D experiments using: (A) parallel tempering or (B) genetic algorithms. In every case the genetic algorithms outperformed parallel tempering.

larger and less noisy. In this way the strength of the circadian signal could be examined versus the stochastic intracellular noise inherent in a cell's molecular counts. The model ensemble fitted to the D/D data alone predicted Stochastic Resonance in which the power in the periodogram spectrum associated with a 21 h peak varied nonlinearly with the level of stochastic intracellular noise.

Here we constructed a much stronger test of the neutral model in Fig 4.1 by introducing light entrainment data for days of varying length: 6 h day, 12 h day, and 36 h day. This model is neutral in the sense that there is no communication between cells because the cells are isolated in droplets [5]. We quadrupled the amount of single cell data used to fit the stochastic clock network in Fig. 2.1A to single cell data on four experiments, in which each experiment provided trajectories on over 1,000 cells every half hour for ten days [5]. Together all four experiments produced 953 frequencies in the power spectrum for fitting the ensemble. The data are publicly available (see Materials and Methods). The data can be thought of as protein levels on a CCG protein every half hour over 10 days in each experiment, although in truth the CCG-2 protein has been replaced with the mCherry recorder [30].

Parallel tempering was used to fit the stochastic network in Fig 4.1B to the experimental periodograms on the four experiments. In the accumulation run the final chi-squared statistic (see Materials and Methods) was χ^2 = 6496.55 with n = 240 + 256 + 201 + 256 = 953 frequencies in the periodograms computed from the best of three independent Monte Carlo runs reaching finishing stage 1 (Fig 4.2A). The number of parameters in Fig 4.1B was 35 with one parameter (f_{IL}) fixed. The contribution of each data point in the periodogram was then χ^2/n =6.82. This fit to the average periodograms of the single cell data was unsatisfactory as shown in Fig 4.3. While the data on the D/D experiment were quite well fitted by the model ensemble under the neutral model in Fig 2.1, the model ensemble did not track well to the data for the 6 h and 12 h day. In both cases there were two peaks in the power spectrum, but the 6 h and 12 peaks in Fig 4.3 were not well predicted by the model. One possibility is that the model ensemble failed to capture fully the entrainment to light in Fig 4.3 because the illumination parameter f_{IL} was fixed at 2. During the third stage of the equilibration process, the illumination parameter was allowed

to float but did not depart from 2 in Fig 4.2A during equilibration. Allowing an extra degree of freedom and more equilibration steps, however, did improve the fit slightly (Fig 4.2A).

There are two further possible explanations for the lack of fit in Fig 4.3. One, parallel tempering is failing to find the best models by optimizing Equation (2) or two, there are two populations of oscillators for the 6 h and 12 h day, indicating a limitation on light entrainment. To test these hypotheses we implemented a novel class of ensemble methods [24] for fitting models to data utilizing genetic algorithms (see Materials and Methods) to maximize Equation (2) during the equilibration phase of Markov Chain Monte Carlo followed up by Metropolis-Hastings Monte Carlo in the accumulation phase.

In a genetic algorithm 20-80 particles (*i.e.*, models) were created in 5, 10, or 20 swarms to live on the 35 dimensional parameter space (with the illumination parameter f_{IL} being fitted as well) [26, 27]. The swarms of particles moved stochastically in the parameter space as specified by Equation (3-4) or (3-5) in Materials and Methods during a generation. The best particle at the end of 600 or 1000 generations (Table 4.1) was used to initiate a Metropolis-Hastings Monte Carlo accumulation run. A distinct genetic algorithm was also tried in Table 4.1 with different dynamics in (3-5) and no recombination. The genetic algorithms were implemented on GPUs.

A total of 14 such independent runs of the genetic algorithms were conducted with 5, 10, or 20 swarms and 4 particles in a swarm, on the same data set used by parallel tempering (Fig. 4.3) to examine the impact of the genetic algorithm and swarm number, for example, on calculation time and finding the optimum to (2). There was no significant difference in the final chi-squared statistics between the two types of genetic algorithms in Table 4.1 by a Wilcoxon Rank Sum Test at the 0.05 [48].

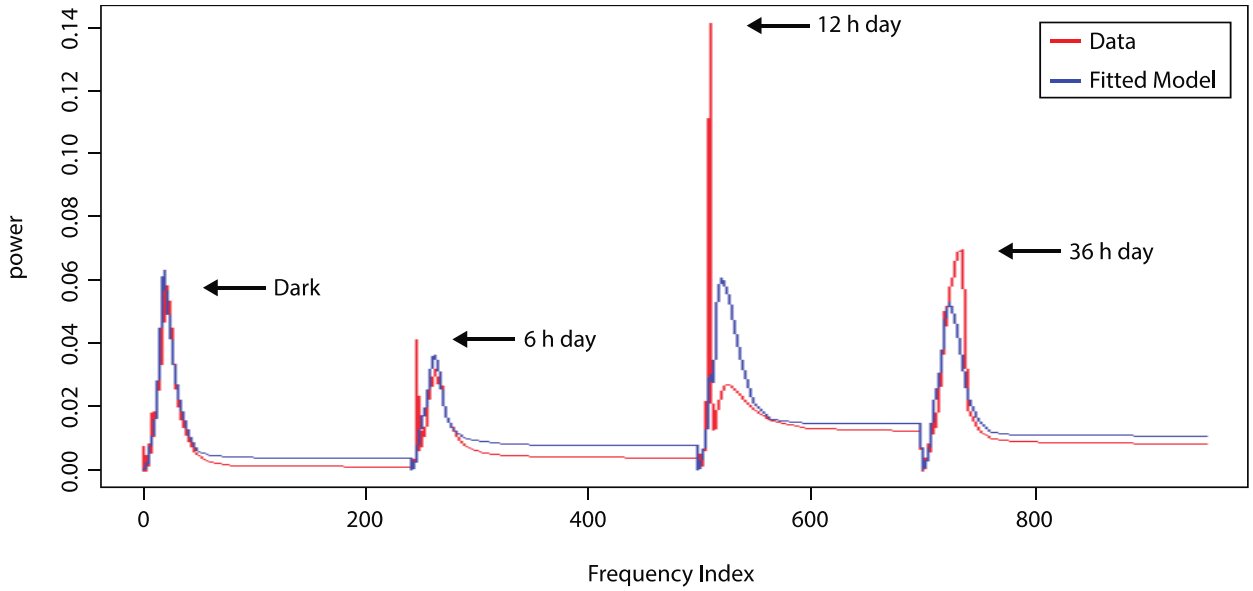

Figure 4.3 The predicted periodograms of the neutral model with no intercell communication were fitted with parallel tempering to the observed periodograms of 4 experiments (described in Materials and Methods), one D/D/ and three L/D with 6 h, 12 h, and 36 h days, respectively, with two major discrepancies for the 6 h day and 12 h days each with its two peaks. Each periodogram represents an average over the individual periodograms of at least 1,000 single cells. The model appeared only to fit one of the peaks of the single cell data in the 6 h day. The fitted periodograms were obtained by an accumulation run with updates from a particular kind of MCMC method called parallel tempering (see Materials and Methods). Observations were taken at half hour intervals over L equidistant observation times. The duration of the experiment is T. The sampled frequencies in the periodogram are denoted by $f_l = \frac{l}{T}$, $l = 1, \ldots \left[\frac{L}{2}\right]$. The first 240 indices $l\ of$ frequencies in the periodogram are for the D/D experiment. The next 256 indices of frequencies are for the L/D experiment with a 6 h day. The next 201 indices are for the L/D with a 12 h day. The last 256 indices are for a 36-h day. For the D/D experiment the x-axis is the index l. The x-axis is l with a shift of 240 for a 6 h day, then with 240+256 for a 12 h day, and finally with 240+256+201 for a 36-h day to separate out the periodograms on the same graph. The periodograms of the experiments and the model were Rhodamine B normalized, detrended, and bias corrected as described in Materials and Methods.

In the first run, the chi-squared statistic was reduced by almost half from the best parallel tempering run with χ^2 = 5410 in Fig 4.2 to χ^2 = 2708.05 by the best genetic algorithm (in red in Table 1). All genetic algorithms, using a random start on the parameter space, outperformed parallel tempering (Table 1) on the same data set. The chi-squared statistic per data point was then χ^2/n = 2708.05/953 =2.84, which is better than other published ensemble fits by

deterministic models on the macroscopic scale [31]. The longest time for an equilibration run with a genetic algorithm for an 80-particle algorithm was 25 h. This is an order of magnitude faster than the equilibration run using parallel tempering in Fig 4.2A. Two 20 particle algorithms were eliminated from the competition for poorer optimization results (Table 1), leaving 12 competing genetic algorithms.

To capture the behavior of the cellular clocks under the model ensemble derived from the best genetic algorithm, the four periodograms were plotted as a function of period (*i.e.*, the inverse of the sample frequency f_l) in Fig 4.4 for ease of interpretation rather than the index of the frequency in Fig 4.3. As can be seen, the fit is extraordinarily good. For example, the model and experimental periodograms are hard to distinguish in Fig. 4.4B. In Fig 4.4B and Fig 4.4C the model tracked quite well to the 6 h and 12 period, respectively. The model succeeded completely in tracking to the period at the driving frequency of the light signal. Over the range of a 6 h day to 36 h day there was no observed limitation to the ability of the model to produce a population of oscillators that tracked to the day experienced, unlike the limit to entrainment for cells in the SCN[18]. In conclusion, the introduction of genetic algorithms appears to support the hypothesis that the limits of entrainment seen in cell tracking in Fig 4.3 to the driving light signal, using parallel tempering, is an artifact of not finding the maximum to equation (2).

One further test was conducted using the remaining 12 independent runs of genetic algorithms to ascertain whether the optimum in Equation (2) was local or global.

As can be seen in Fig 4.2B, all runs converged approximately to the same chi-squared statistic, strongly suggesting a global optimum had been achieved. Each of the 12 independent runs in Table 1 was then used to construct an accumulation run of 10,500 updates with Metropolis-Hastings Monte Carlo[17] and combined to produce a final reconstruction of the likelihood in (2)

together with its summary of the parameter distribution in Table 4.2 (as described in Materials and Methods). The best model in the accumulation run had a χ^2 = 2671.95

One standard control for MCMC experiments is to plot the parameter values in an accumulation run versus sweep (*i.e*., the time taken on average to visit once to each parameter in the model). If the accumulation run were not complete, there would be trends in some parameters with sweep. All of the plots showed no trend, indicating that the accumulation run was successful. The plots also display which parameters are well specified in the ensemble. For example, the *wc-1* stabilized mRNA decay rate (D7) and the protein-protein interaction (C1) are tightly specified, while other parameters, such as the FRQ protein decay rate (D6) is not tightly specified.

4.4.3 KINETIC RULES FOR THE CLOCK AT THE SINGLE CELL LEVEL

The series of 3 light entrainment experiments with the D/D experiment with 953 frequencies in the 4 periodograms in Fig 4.4 provided a strong test of the clock network in a single cell, but they also provided precise estimates of the rate constants and initial gene and cognate mRNA and protein counts in a cell as well (see standard errors in last column of Table 4.2). These parameter estimates (*i.e*., rate constants and initial conditions for molecular counts of species) provided a means to determine if single cells play by the same or different rules than cells in aggregate at the macroscopic level [31].

A comparison was made between a published ensemble only derived from the D/D data using an accumulation run from parallel tempering, in which an adequate fit was obtained[17], with an ensemble (Table 2) derived from D/D data together with the L/D data using an

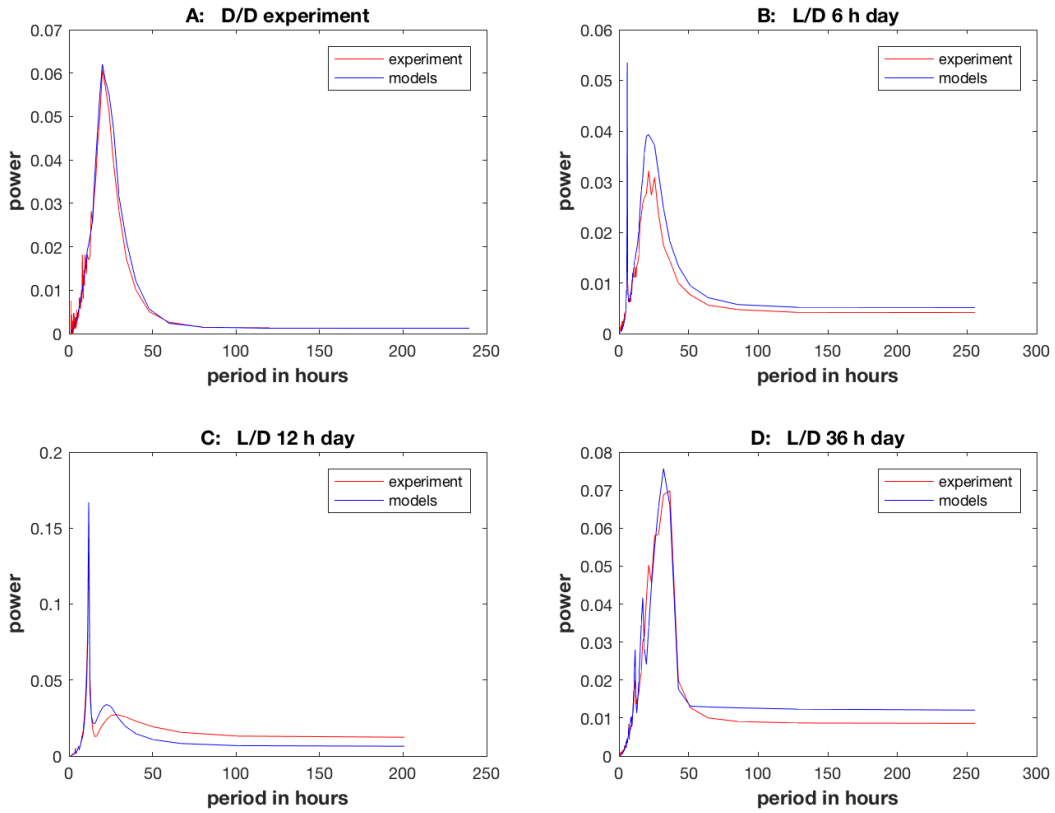

Figure 4.4 The average periodograms for single cells as a function of period for the same four experiments in Fig 4.3 (D/D, L/D with 6 h day, L/D with 12 h day, and L/D with 36 h day) were fitted very well by the model ensemble(χ^2=2708.05) using Genetic Algorithms. (A) D/D experiment; (B) L/D with 6 h day; (C) L/D with 12 h day; (D) L/D with 36 h day. Data are the same as in Fig. 4.3, but power is presented as a function of period in each periodogram. The period is the inverse of the sampled frequency, namely $\frac{1}{f_l}, l = 1, \dots, [L/2]$.

accumulation run from genetic algorithms that also provided a remarkably good fit to the combined data set (Fig. 4.4). In comparing these two model ensembles there was remarkable agreement in the specification of the genetic network, but there are several changes in the rate constants from the estimates based only on the D/D experiment in Table 2. For example, the translation rates (L1, L3, and Lc) were lower on the clock mechanism genes based on the 4 experiments vs. the one D/D experiment with 1,591 cells.

Table 4.2 Ensemble means and standard errors indicate that the parameters in stochastic network for single cells are tightly specified by Markov Chain Monte Carlo using genetic algorithms with the D/D experiment and three L/D experiments described in Materials and Methods.

Parameter	Initial Parameter values from MCMC Deterministic model ensemble (Yu et al., 2007)	Initial Parameter values from published ensemble (column 2) in molecular number units of stochastic network from D/D experiment[17] (column 3)	Mean parameter values from model ensemble computed by Parallel tempering for D/D experiment	Standard error (SE) of parameter value across ensemble computed by parallel tempering for D/D experiment	Mean parameter values from model ensemble computed by genetic algorithms for D/D and L/D experiments in Metropolis-Hastings accumulation run	Standard error (SE) of parameter value across ensemble computed by genetic algorithms for D/D and L/D experiments in Metropolis-Hastings accumulation run
Number of cells	-	1591 (D/D only)	1591 (D/D only)	1591 (D/D only)	4 experiments (D/D + 3 L/D)	4 experiments (D/D + 3 L/D)
u_r0	3.99924	113	2156.705728	68.14603254	249	1.94493283
u_r1	0.442441	18	22.46137677	0.872953544	266.5025	1.24742302
u_p	4.24E-07	459	2144.149238	68.74768856	267.75	1.41999973
f_0	0.356365	1	0.465055176	0.01143672	033333333	0.00785783
f_1	0.0824576	0	0.534944824	0.01143672	0.6666667	0.00785783
f_r	4.90E-06	31	59.15869679	2.637452857	263.66667	1.53551295
f_p	3.0804	345	2534.336311	77.72114325	258.83333	0.96610217
w	9.24126	101	55.40042039	1.674320488	280.11306	1.34419102
g_0	0.0066195	1	0.71623752	0.010337149	0.5	0.00833449
g_1	2.59E-06	0	0.28376248	0.010337149	0.5	0.00833449
g_r	1.17E-06	26	35.67840252	1.030258983	234.41667	1.45223506
g_p	1.37E-05	102	59.19075145	4.920903774	283.93833	1.01849186
A	0.000658482	6.06E-13	2.56E-10	7.31E-12	2.24E-10	1.10E-11
Abar	0.546986	0.546986	1.589532708	0.035661845	0.6046986	0.01326434
S1	0.061594783	83.70771546	80.12566921	0.302471515	82.372323	1.03259487
S3	0.00146575	3.569116497	0.400641074	0.036565894	13.491623	0.42722599

S4	2.2396	5453.449297	8316.020583	100.2852188	77.524423	1.13421469
D1	0.723678	0.723678	1.294999006	0.030289616	71.2760416	2.63473865
D3	0.299703	0.299703	4.382612039	0.181101578	6.16109147	0.17691415
C1	0.0428595	4.81E-05	0.000932789	2.47E-05	9.22E-05	4.25E-06
L1	31.7758	4.244678204	4.777735371	0.106626479	2.60684774	0.04948995
L3	3.02387	0.485087349	0.665600817	0.011127036	0.09315913	0.00269597
D4	0.00323262	0.00323262	0.08474029	0.004700587	0.05674687	0.00194278
D6	0.15183	0.15183	0.193685712	0.002236097	12.0326238	0.28052851
D7	0.138387	0.138387	2.130911791	0.090030385	0.11260178	0.00432734
D8	0.00248668	0.00248668	0.007744621	0.000182717	0.00014153	3.61E-06
C2	0.162687246	0.162687246	1.515554675	0.077548547	95.9668318	3.1203002
P	19.5648	3.12E-11	2.72E-09	4.83E-11	3.46E-08	6.82E-10
Ac	4.06813	7.82E-09	1.86E-08	2.55E-09	3.44E-05	1.89E-06
Bc	2.52197	2.52197	2.581096866	0.040197442	0.88230334	0.0199371
Sc	1.01E-06	73.80414613	61.51499414	1.109629713	11.0853255	0.25892314
Lc	1.15E-08	2.231095711	1.61524392	0.017335914	0.01161864	0.00012954
Dcr	0.219758	0.219758	0.150810052	0.00291715	0.27129774	0.00638981
Dcp	0.696903	0.696903	0.54063952	0.006141903	0.0224811	0.00035852
fIL			-	-	16.4900328	0.38777036

The biggest surprise is in the mRNA stability of *wc-1*. In the fitting of the model to all 4 single cell experiments the derivative mRNA *wc-1*^{r1} was more stable as measured by D7 than in the D/D single cell experiment alone. Having a stable *wc-1*^{r1} mRNA has been argued to be essential for oscillations at the macroscopic scale [22]. In the network fit to all of the single cell data the modified *wc-1* mRNA, *wc-1*^{r1}, is stable.

For example, the decay rate D7 = 0.11 +/- 0.0043 under all 4 experiments with a long lifetime of 1/D7 = 8.88 h as measured macroscopically[22] versus D7 =2.13 +/- 0.09 in the D/D experiment alone.

The single cell data in the D/D experiment alone were not sufficient to confirm this result found macroscopically. For models fitted to the D/D experiment alone, the decay rate (D7) was found to be D7 = 2.13 +/- .09 [17]. Evidence against the parallel tempering method being the cause of the discrepancy in the decay rate (D7) comes from the fact that fitted ensemble achieved by parallel tempering was an adequate fit to the average periodogram of the D/D data. As a caveat, if we had implemented a longer equilibration run, we might have achieved the results of MCMC runs using genetic algorithms reported here. When the lines of different genetic algorithm accumulation runs are close together, as for the translation rate (Lc) for CCG-2p:mCherry, that is indicative that different MCMC runs converged to the same optimizing parameter value. For instance, in the case of the translation rate the ensemble covers the values from .002 to 0.02.

Also a comparison was made between the ensembles computed here using parallel tempering (Fig 4.3) and with those using the genetic algorithms (Table 1) with respect to the illumination parameter (f_{IL}) on a common data set (D/D + 3 L/D experiments). In allowing the illumination parameter to float, the final value of f_{IL} achieved a much larger value of 16.49 +/- 0.39 than that derived under the use of parallel tempering, namely $f_{IL} = 2$.

There are two sources of variation captured in the standard errors in Table 2 on these parameters. There is variation in the standard errors across models, and there is also variation in the parameters estimates due to stochastic intracellular noise. Both sources of error are reflected in the standard errors. Some parameters, such as the decay rate of the stabilized *wc-1* mRNA (D7), are quite tightly specified, while other parameters such as the transcription rate of *frq* (S4), have considerable variation.

Generally in comparing the rate constants obtained from all four experiments (column 6) to those derived from macroscopic experiments (column 2) [22] using Euclidean distance on the parameters in common, the agreement was much better than just based on the D/D experiment alone (column 4). The only rate constant out of line with the macroscopic limit appeared to be the decay rate of FRQ [22]. There is also considerable variation in the estimates of this decay rate. The conclusion is single cells appear to play by similar rules as aggregates of 10^7 cells.

4.4.4 THE STOCHASTIC INTRACELLULAR NOISE LEVEL CAN BE EXPERIMENTALLY DETERMINED AS A PARAMETER IN A MODEL

In chapter 2 evidence was presented that the RNA/DNA and protein/DNA ratios for CCG-2 set the levels of stochastic intracellular noise in a cell, and hence these ratios were measured. They continue to serve a similar role in a system with light entrainment (Fig 4.5). As the RNA/DNA and protein/DNA ratios are increased, leading to larger amplification in RNA counts and protein counts, there was a general decrease in the noise in the system (Fig 4.5). Imagine the red dot as a ball; from most places on the surface the ball rolls to the lowest point in the front left corner of Fig. 4.5. The only caveat is a shallow ridge at low protein/DNA ratios.

The relationship between the stochastic intracellular noise and the RNA/DNA and protein/DNA ratios is not in and of itself surprising[34, 49]; however, exploiting this relationship to determine “size of the cell” appears to be new. This ability to determine empirically the “size of the cell” is why the relation in Fig. 4.5 is presented. In this way these ratios can be used to manipulate the level of stochastic intracellular variation. These ratios were experimentally determined (red dot) previously to set the level of noise in each cell[17].

4.4.5 THE PHASE VARIATION OVER TIME BETWEEN CELLS PROVIDES AN INDEPENDENT TEST OF THE GOODNESS OF FIT

There are three ways to characterize periodic processes, by their period, amplitude, and phase [6]. The period and amplitude are captured in the periodogram or power spectrum (*e.g.*, Fig 4.4), which was used to fit the model ensemble in equation (2). The remaining measure, phase, is functionally independent of the periodogram and was not used to fit the stochastic network to the single cell data and hence is available to test goodness of fit of the stochastic network (see chapter 2).

There are a variety of ways to measure phase, as described in chapter 3. In addition to its independence from the periodogram, the continuized Hilbert phase measure was used to assess whether or not synchronization is taking place between single cells experiencing a common driving light signal [5]. The continuized Hilbert phase measure increases linearly with time for a sinusoidal process, but for a process experiencing synchronization the phase curve is nonlinear [5] – the phases of cellular clocks change towards each other as they synchronize.

To provide an independent test of the stochastic network in Fig 4.1, the average phase and percentiles of the phase distribution at time t were computed over time (t) for cells in all four experiments both for the data and for the model (using 1,024 generated single cell Gillespie trajectories) (Fig 4.6). For the D/D and 6 h day L/D experiments the goodness of fit failed at the 75 h and 125 h mark, as the data (in red) drifted beyond the percentiles of the model phase (blue). In contrast, the percentiles of phase for model and data remained overlapping for the 12 h day and 36 h day L/D experiments.

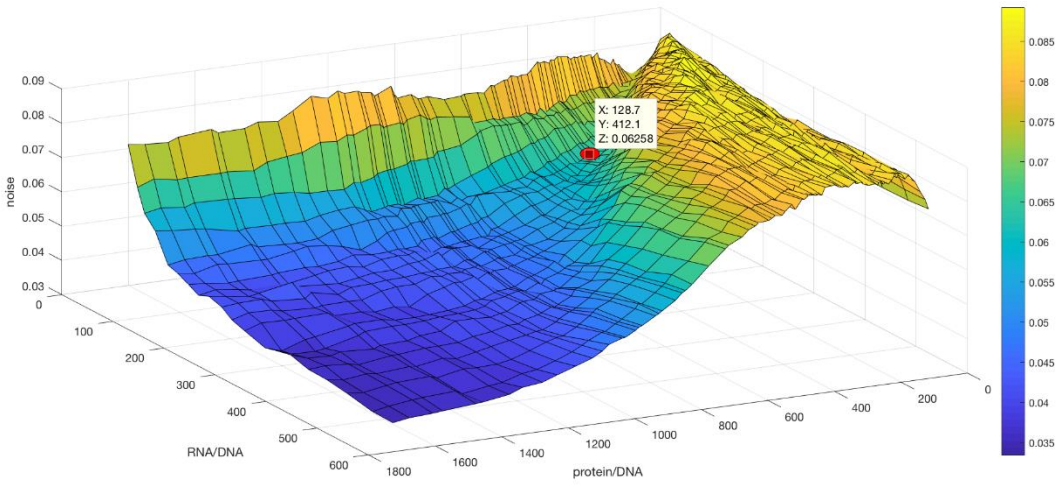

Figure 4.5 Stochastic noise in CCG-2 usually decreases with increases in hypothesized ratios of RNA/DNA and Protein/DNA within a single cell. The total stochastic noise σ_f^2 averaged over frequencies (f) in CCG-2 expression is computed from 1024 Gillespie trajectories from the best model in S Table 1 with a χ^2 = 2671.95. The best model selected was one with minimum chi-squared statistic based on the Likelihood in Equation (2) for the D/D and 3 L/D experiments from an accumulation run based on 12 genetic algorithms in Table 4.1. The red dot denotes the experimentally determined ratios previously[17] and corresponds to RNA/DNA and protein/DNA ratios of 128.7 and 412, respectively. The model with the best chi-squared statistic in the accumulation run was modified to different RNA/DNA and Protein/DNA ratios for each point on the grid above. A total of 1,024 Gillespie trajectories were generated for each model on the grid. The variance in the 1,024 resulting periodogram height was computed for each sample frequency f_l . These variances were summed over all frequencies to produce the noise on the z-axis in Fig. 4.5.

The phase plots also provided information about the cellular clocks in single cells. Phase plots for both the model and data in the 12-h day and 36-h day L/D experiments were bent and hence demonstrated synchronization to the driving light signal. Also, all plots showed increased variation in phase over time, capturing the tug of war between stochastic intracellular noise generating phase variation and light producing changes in phase synchronization and hence the phase mean. The degree of linearity of the D/D and 6-h day L/D experiment (r=.9995 and r=.9998, P < .0001) would also suggest that a sinusoidal approximation would be a good one

[28]. The fact that the D/D and 6 h day L/D experiments did not demonstrate a nonlinear response in time and hence synchronization was consistent with the synchronization measures for the D/D (Kuramoto K = 0.08+/-.0026) and 6 h day L/D (K.=0.30+/-.0066) experiments being smaller than those for the 12 h day L/D (K=0.42+/- 0.0076) and 36 h day L/D (K = 0.33+/-0.0069) experiments[5]. For example, the maximum in light synchronization was measured to take place with a 12-h day, which also show a nonlinear response in the phase curve over time [5].

Stochastic networks have one other dimension to goodness of fit absent in deterministic network models. Having determined what the "size of a cell" is by measuring the RNA/DNA and protein/DNA ratios in Fig 4.5, it is natural to ask how these ratios affected the goodness of fit of the model periodograms to the average of the observed single cell periodograms. These ratios were varied substantially about their measured values to see the effect on goodness of fit (Fig 4.7).

In the Materials and Methods there is a description of how the ratios are varied without altering the dynamics of the system in Fig. 4.1.

The fitting of the D/D data would leave us to hypothesize that the goodness of fit would be robust to variation in the level of stochastic intracellular noise captured by these ratios (see chapter 2). We found the distribution across the fitted model ensemble was quite robust to variation in these ratios (Fig. 4.7). This robustness property can be predicted from the Chemical Langevin Equations that approximate the stochastic network in Fig 4.1.

4.4.6 IS THERE ONE INTERMEDIATE OPTIMUM IN THE OSCILLATORY SIGNAL AS A FUNCTION OF STOCHASTIC INTRACELLULAR NOISE?

The heart of the experiments and calculations in this paper is to examine whether or not the working model in Fig 4.1 displays Stochastic Resonance, *i.e*., a nonlinear relation between the signal/noise ratio captured in the power spectra (Fig 4.4) and the stochastic intracellular noise in the system (Fig 4.5) [10]. The noise is varied by altering the RNA/DNA and protein/DNA ratios in the cell in Fig 4.5. High values of the ratios imply low noise while low ratios imply high noise in Fig 4.5. Reducing the ratios by a constant factor generally decreases the power at the intrinsic frequency (Fig 4.8a) or at the driving frequency (Fig 4.8b-4.8d). In contrast as the ratios are increased, there is a spike in the signal at the resonance, which then fades away as the ratios are increased further. These changes in the ratios were done by altering the initial molecular counts of species without altering the rates constants in the best fitting model to vary the stochastic intracellular noise. For the 36-h day the ratios had to be increased further to see the signal to noise ratio diminish. The results are more easily summarized in Fig 4.8.

The ratios are varied from low (high noise) to high (low noise), and the power in each experiment is presented at the intrinsic frequency of the cellular oscillators (~21 h) or at the driving frequency (~6 h, ~12 h, or ~36 h, depending on the L/D experiment) in Fig. 4.8. There is a clear nonlinear relation for each day that peaks at the same ratio of 15 X the original ratios (128.7 for RNA/DNA and 412 for protein/DNA). The intrinsic frequency of ~21 h is plotted as a control (in red) for the L/D experiments.

The effects of stochastic intracellular noise on the average Gillespie trajectory are shown for a 12-h L/D cycle (Fig. 4.9). This L/D cycle had the highest Kuramoto K order parameter

among the four experiments. As can be seen, away from the Stochastic Resonance there is a degradation in the circadian signal, and at the stochastic resonance there is an amplification of the circadian signal. This is a classic example of stochastic resonance in a biological system [10]. These striking differences in the circadian oscillations arise between cells that are genetically identical!

4.5 DISCUSSION

There are three hypotheses about how oscillations arise at the single cell level. One hypothesis is that stochastic noise contributes to the oscillations, a theory known as Stochastic Resonance [10]. Two teams indicated how Stochastic Resonance could serve as a mechanism to generate such oscillations [50, 51] and possibly to synchronize cellular oscillators. A second hypothesis is that there is a chemical signal through quorum sensing by cellular clocks that isinvolved in synchronization of single cell oscillators [5, 7], thus explaining circadian rhythms on the macroscopic scale of 10^7 cells. A third possibility is cell cycle gating of the single cell oscillators to reinforce the oscillations [11-14]. Under this third possibility there may be no specific genes that induce coordination between different single cell oscillators as for example, in a quorum sensing hypothesis of cell-to-cell communication.

The advantages of this study using the model system, *N. crassa*, is that the single cell environment can be set up to test each of these hypotheses individually using microfluidics [15]. Here a flow focusing, droplet generating microfluidics device was used to isolate *N. crassa* cellular oscillators for testing Stochastic Resonance [5]. The microfluidics device isolated cells in droplets to prevent any form of chemical communication, as under a quorum sensing

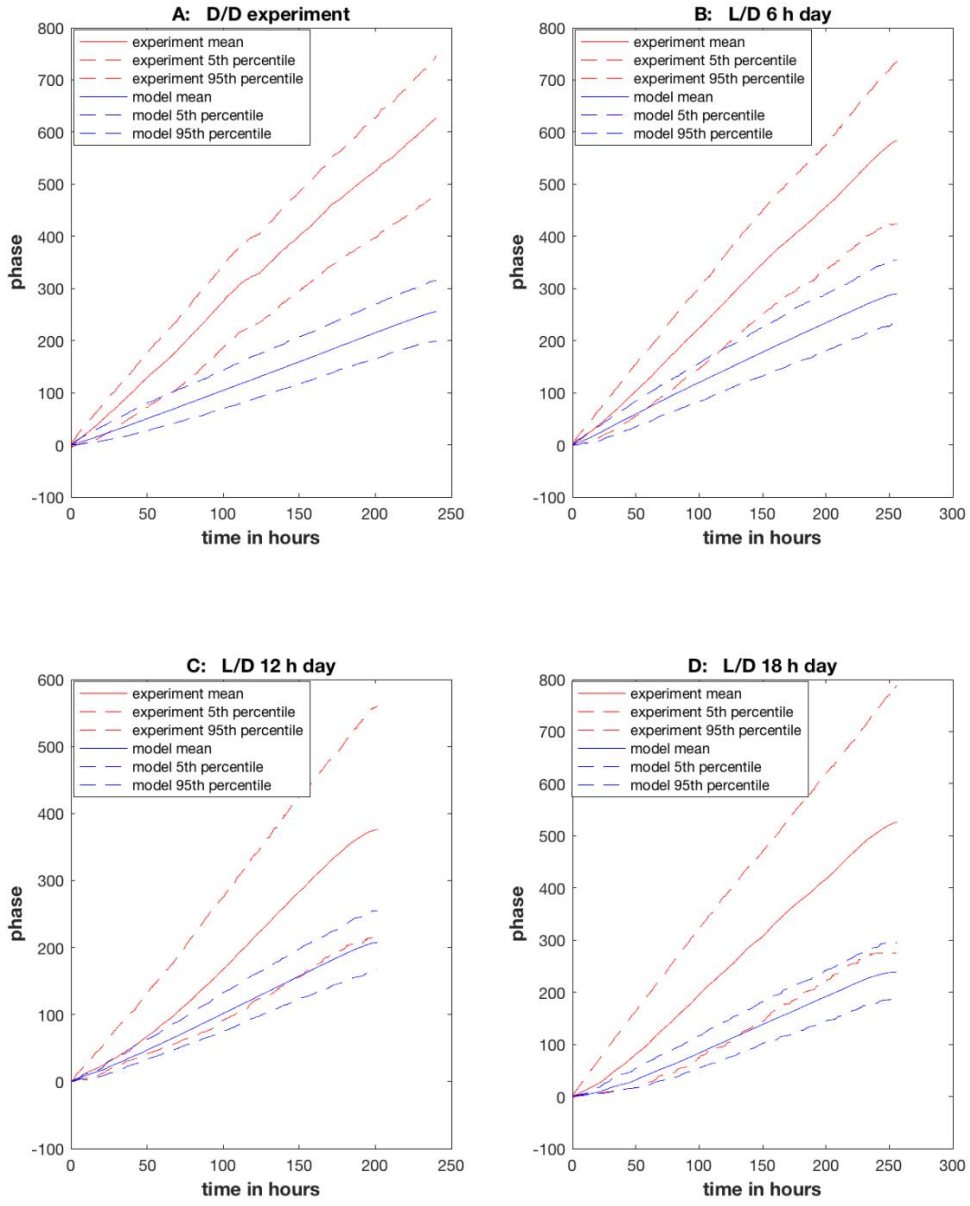

Figure 4.6 The phase plots as a function of time indicated that there are limitations on goodness of fit for the D/D experiment and 6 h day L/D experiment.

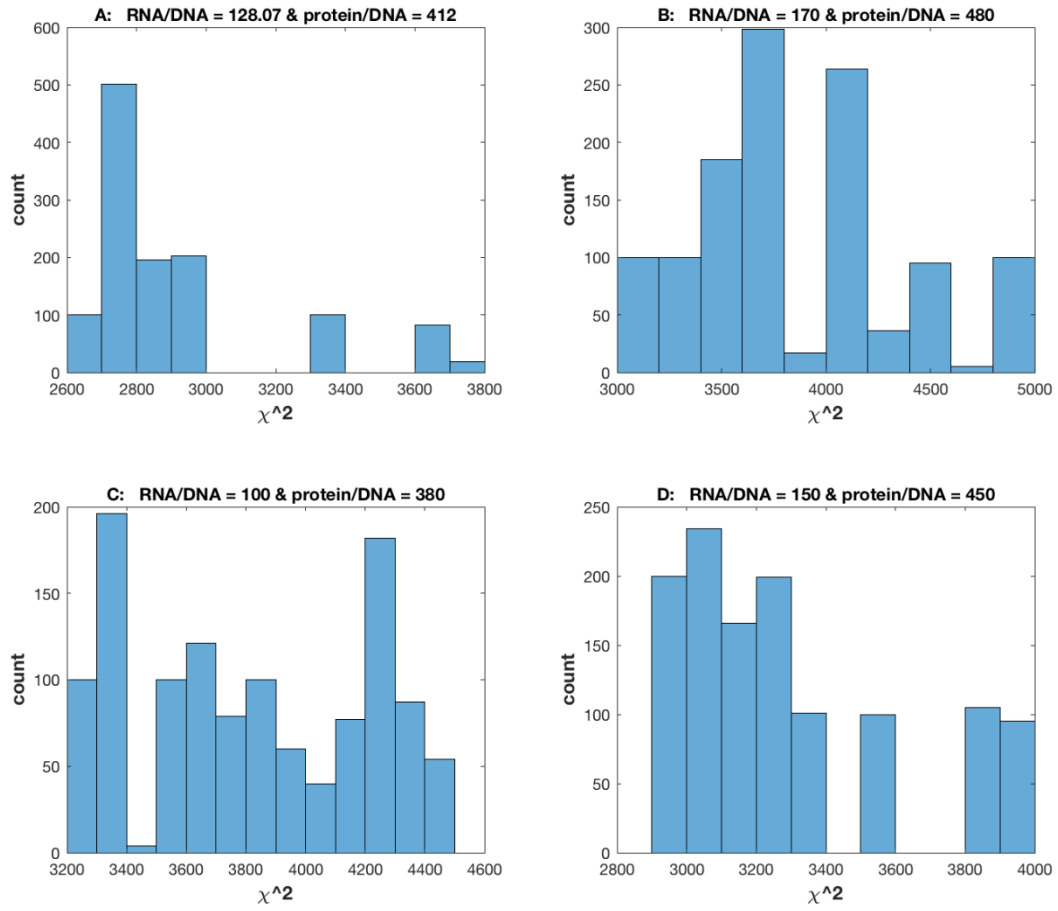

Figure 4.7 The goodness of fit as measured by the chi-squared statistic in (2) was robust to variation in the ratios of RNA/DNA and protein/DNA and hence the stochastic intracellular noise from Fig 4.5. Histograms of the chi-squared statistics of 1,200 models in the accumulation run for determining the chi-squared empirical distribution are shown. The ratios of RNA/DNA and protein/DNA used in each of the 1,200 models was, respectively: (A) 128.7 and 412; (B) 170 and 480; (C) 100 and 380; (D) 150 and 450. A description of how the ratios are varied without altering the rate constants is shown in the Materials and Methods.

hypothesis [5]. The media were selected as well so that there was no cell division to eliminate cell cycle gating as a hypothesis [6]. The model system was exploited in such a way as to be able to take advantage of light entrainment of isolated *N. crassa* single cell oscillators [5], an advantage not present in other model clock systems, such as mammalian [19] or Cyanobacterial

[18] model clock systems. It was possible to demonstrate here strong support for a neutral model without any cellular communication using both light entrainment experiments and D/D experiments to specify a model ensemble in Fig 4.1 describing cellular clocks (Fig. 4.4). In four independent light entrainment experiments the model ensemble was able to capture the period and amplitude behavior of the single cell oscillators from a 6 h L/D cycle to a 36 h L/D cycle at the single cell level (Fig. 4.4). The highly successful fitting was robust to variation in "cell size" present in stochastic networks, as captured in the proxies for cell size, the measured RNA/DNA and protein/DNA ratios (Fig 4.8).

The fitted model ensemble displayed the same stochastic resonance across all four D/D and L/D experiments as the stochastic intracellular noise was varied through the RNA/DNA and protein/DNA ratios (Fig 4.8). Striking differences in the strength of circadian oscillations were seen about the stochastic resonance even when all cells were genetically identical (Fig. 4.9). This finding is reminiscent of the findings of noise effects on damped linear oscillators used to model single cells of the mammalian Suprachiasmatic Nucleus (SCN) [52]. This neutral model with Stochastic Resonance then is a promising framework for testing whether or not Stochastic Resonance can explain by itself the origin of circadian rhythms on a macroscopic scale from the cellular clocks operating on a microscopic scale.

There were some limitations to the neutral model supporting Stochastic Resonance. The periodograms (Fig 4.4) used to fit the stochastic network in Fig 4.1 captured the amplitude and period variation in cellular clocks remarkably well (χ^2 = 2671.95 across 953 frequencies from four periodograms and with 35 model parameters with a chi-squared statistic per data point of $\frac{\chi^2}{n} = 2.80$), but the periodograms are functionally independent of the phase variation [6, 28], when measured by Hilbert phase [53].

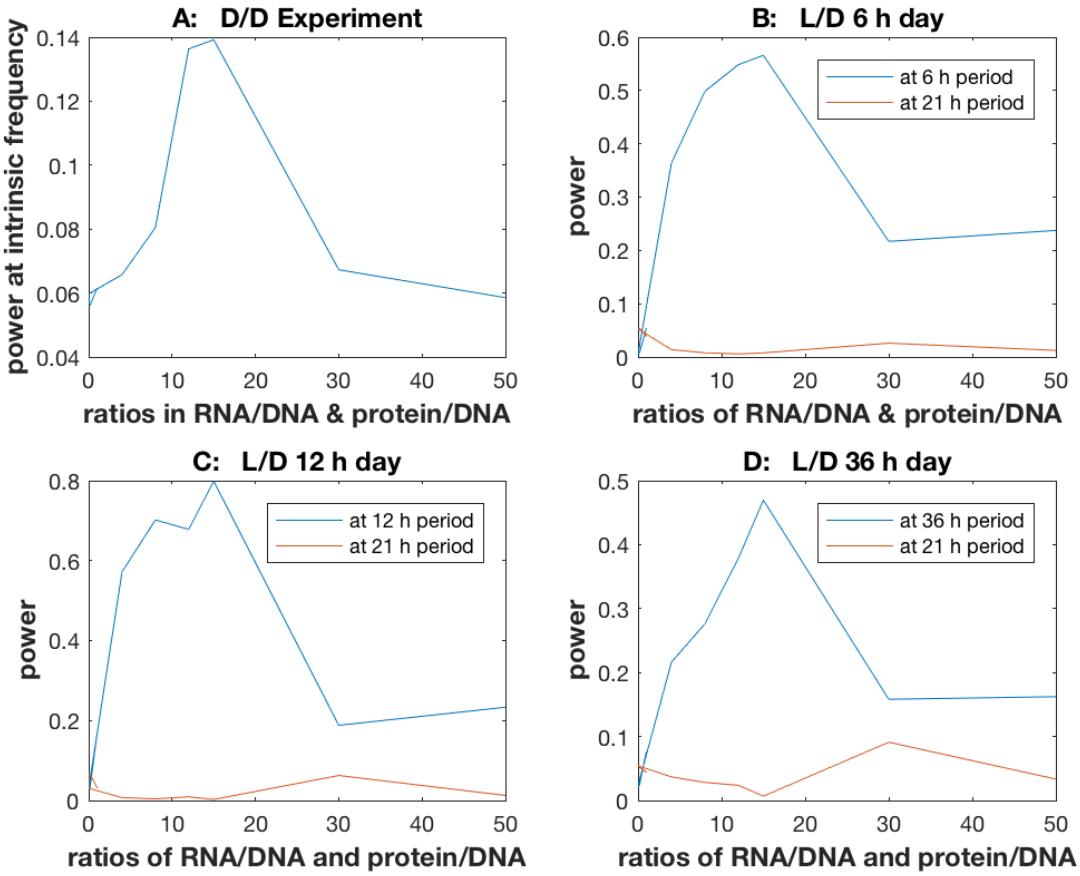

Figure 4.8 The power at the driving frequency or intrinsic frequency for a cellular oscillator is a nonlinear function of the stochastic intracellular noise and takes a maximum at a ratio of 15 for RNA/DNA and protein/DNA for all four independently conducted experiments. The stochastic intracellular noise was varied by changing the initial molecular counts of species (but not the rate constants) by a ratio of: 1/7, 1/10, 1/12, 1/40, 1/100, 1/170, 4, 8, 12, 15, 30, or 50. (A) D/D experiment; (B) L/D 6 h day; (C) L/D 12 h day; (D) L/D 36 h day. The model used to generate the power values above is the best fitting model. The power values are derived from the periodograms. For the L/D experiments the power at the intrinsic frequency of a cellular oscillator is added as a control.

The phase measure used here and derived from the Hilbert phase [53], by virtue of its functional independence of the periodogram [28], was used to test goodness of fit to the single cell experiments (Fig 4.6). The results were that for the L/D 12 h and 36 h day the phase of the

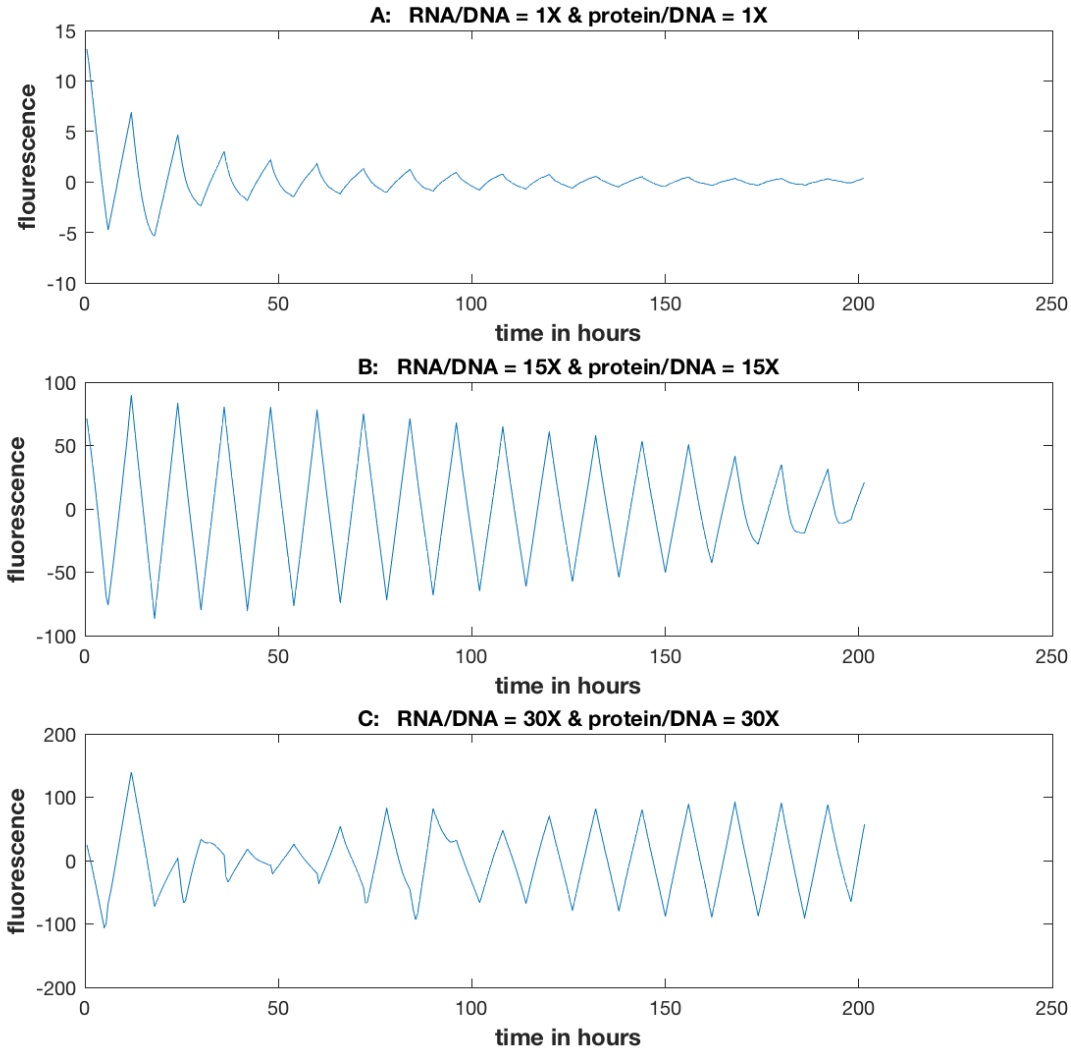

Figure 4.9 The effects of stochastic intracellular variation at the resonance was to amplify the circadian signal, but away from the resonance the signal was degraded. These FRQ trajectories are averages over 1,024 Gillespie trajectories at the best model (S Table 1). The y-axis is the predicted number of the FRQ oscillator protein over time. The RNA/DNA and protein/DNA ratios are at 1X, 15X, and 30X of their measured values of 128.7 and 412, respectively. The stochastic intracellular noise was varied by changing the initial molecular counts as in Fig 4.8.

model ensemble and single cell data over time were consistent with each other (Fig 4.6C and Fig. 4.6D); however, there were some departures from model ensemble predictions of phase over

time after 75 h (Fig 4/6A) for a D/D experiment or 125 h for the single cell data, see Fig. 4.6B for 6 h L/D experiment. As can be seen in the fan shape of the confidence bands for phase over time (Fig 4.6), there is a tug of war between the substantial role of stochastic intracellular noise generating phase variation between cells and the effect of the light signal on the mean phase of the cellular clocks.

One possible explanation for the departure may be that stochastic intracellular variation is winning the war as the system evolves in time, causing the phase of single cells to drift outside the confidence bands of the model ensemble when the synchronization with the light signal is weaker [5] in Fig 4.6A and Fig 4.6B. Some other mechanisms for introducing noise into the kinetics may be needed [54].

The fact that noise plays such a significant role in generating phase variation raised the possibility that the behavior of cellular clocks may be fundamentally different from the rules of clocks at the macroscopic scale of 10^7 cells/ml [22]. We tested this possibility by examining the fitted rate constants derived from single cell data. The result was excellent agreement with the characterized dynamics on the macroscopic scale [22, 31] in Table 2. For example, the prediction that the lifetime of the *wc-1* mRNA being long as measured [22] on a macroscopic scale, held up on a microscopic scale when light entrainment data for single cells were added (Table 2). At this stage there was little evidence for cellular clocks playing by different rules than those at the macroscopic scale of 10^7 cells.

There are a variety of kinds of resonances that could be at play in the circadian system of single cells of *N. crassa* [5]. For example, the resonance could be due to noise acting near a single excitation state in the model [55], as in the phase resetting of cyanobacterial cells [21] or alternatively, due to noise moving the system from one equilibrium point to another as in a

bistable switch [55]. One characteristic of a stochastic resonance, whether it be introduced as in a signal processing tool or naturally occurring, is the presence of at least one stochastic switch[56]. Sriram and Gopinothan [57] were the first to hypothesize such a stochastic resonance in the *N. crassa* circadian system. The basis for such a stochastic switch in Fig. 4.1 lies in the stochastic switching on or off of the oscillator gene *frq* or the *ccg* gene [5]. The one or few copies of genes themselves in Fig. 4.1 provide the basis of the stochastic switching mechanism.

Further experimental and theoretical studies of the model (Fig. 4.1) are required to characterize the resonance. For example, *N. crassa* single cell behavior through microfluidic experiments to examine transcriptional bursting [58, 59] and calculations of the mean amplitude, period, and phase of the model [60] will be needed to arrive at details of the resonance mechanism. Some of this work has already begun on single cell measurements of the mammalian SCN in a phenomenological way by fitting simplified damped or self-sustained oscillators to single cell data on circadian rhythms of the SCN [52].

There are other features of the genetic network (Fig. 4.1) that are hypothesized to mediate the effects of stochastic intracellular noise other than through a resonance [49]. Andreas Wagner [61] has demonstrated by an MCMC analysis of simple two-gene circadian oscillators that have interlocking regulatory connections that are more likely to be robust in period. Liu and colleagues provided experimental evidence on a macroscopic scale that the positive feedback loop by FRQ on *wc-1* mRNA involving (C1) [62] in Fig. 4.1 functions to provide stability and robustness to the clock. It would be interesting to know what effect stochastic intracellular noise has on single cell circadian oscillations when the positive feedback loop in Fig. 4.1 is removed.

Yu et al. [22] reviewed experimental evidence for each reaction in the topology in Fig. 4.1 on a macroscopic scale. As more data are gathered, it may be necessary to alter the topology of the network in Fig. 4.1 as another parameter in the model. Al-Omari et al. [37] developed ensemble methods to identify the topology of the network using the supernet. It may be possible to extend these supernet methods to the single cell level using single cell sequencing [63], allowing a reassessment of the topology in Fig. 4.1 at the single cell level."

What made the results here possible was the development of new fitting methods for stochastic networks [17] in particular and for ensemble methods in general [23]. A longstanding problem (20 years) for ensemble methods applied to oscillatory systems has been the inability to generate successfully a fitted model ensemble without an initially informed guess as to the rate constants and initial species concentrations [23]. We were so limited in the development of finding an ensemble of stochastic networks in Fig 4.1 using existing ensemble methods with parallel tempering in Fig 4.3 [17]. Here by the introduction of genetic algorithms into the equilibration phase of a Markov Chain Monte Carlo reconstruction of the likelihood for a stochastic network in Equation [2], a random initialization of genetic algorithms (Table 1) outperformed existing parallel tempering methods starting with an informed guess as to model parameters [17]. These genetic algorithms also yielded solutions in less time by an order of magnitude (Fig 4.2). As a result the speedup of the genetic algorithms could be used to generate multiple MCMC runs, providing evidence that a global optimum in the fit of the model ensemble was achieved (Fig 4.2B).

The ability to fit stochastic networks to single cell data quickly and efficiently suggests new microfluidics experiments to test the physical hypothesis of Stochastic Resonance in biological systems. The prediction of only one stochastic resonance across light entrainment

experiments in single cells provides a unique opportunity to test the Theory of Stochastic Resonance in a biological system [10].

4.6 REFERENCES

[1] T. B. Kepler and T. C. Elston, "Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations," (in English), *Biophysical Journal,* vol. 81, no. 6, pp. 3116-3136, Dec 2001.

[2] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic Gene Expression in a Single Cell," *Science,* vol. 297, 2002.

[3] L. Cai, C. K. Dalal, and M. B. Elowitz, "Frequency-modulated nuclear localization bursts coordinate gene regulation," *Nature,* vol. 455, no. 7212, pp. 485-490, 2008.

[4] Y. Lin, C. H. Sohn, C. K. Dalal, L. Cai, and M. B. Elowitz, "Combinatorial gene regulation by modulation of relative pulse timing," *Nature,* vol. 527, no. 7576, pp. 54-58, 2015.

[5] Z. Deng *et al.*, "Single cells of *Neurospora crassa* show circadian oscillations as well light entrainment and temperature compensation," *IEEE Access,* vol. 7, pp. 49403-49417, 2019.

[6] Z. Deng *et al.*, "Synchronizing stochastic circadian oscillators in single cells of *Neurospora crassa*," *Scientific Reports,* vol. 6, p. 35828, 2016.

[7] E. Ullner, J. Buceta, A. Díez-Noguera, and J. García-Ojalvo, "Noise-induced coherence in multicellular circadian clocks," *Biophysical Journal,* vol. 96, no. 9, pp. 3573-3581, 2009.

[8] I. T. Tokuda, D. Ono, S. Honma, K.-I. Honma, and H. Herzel, "Coherency of circadian rhythms in the SCN is governed by the interplay of two coupling factors," *PLOS Computational Biology,* vol. 14, no. 12, p. e1006607, 2018.

[9] C. H. Ko *et al.*, "Emergence of noise-induced oscillations in the central circadian pacemaker," *PLoS Biol,* vol. 8, no. 10, p. e1000513, 2010.

[10] R. Benzi, A. Sutera, and A. Vulpiani, "The mechanism of stochastic resonance," *Journal of Physics A: Mathematical and General,* vol. 14, no. 11, p. L453, 1981.

[11] J. Paijmans, M. Bosman, P. R. t. Wolde, and D. K. Lubensky, "Discrete gene replication events drive coupling between the cell cycle and circadian clocks," *PNAS USA,* vol. 113, pp. 4063-4068, 2015.

[12] Q. Yang, B. F. Pando, G. Dong, S. S. Golden, and A. van Oudenaarden, "Circadian gating of the cell cycle revealed in single cyanobacterial cells," *Science,* vol. 327, no. 5972, pp. 1522-1526, 2010.

[13] J. Bieler, R. Cannavo, K. Gustafson, C. Gobet, D. Gatfield, and F. Naef, "Robust synchronization of coupled circadian and cell cycle oscillators in single mammalian cells," *Molecular Systems Biology,* vol. 10, no. 7, p. 739, 2014/07/01 2014.

[14] C. Feillet, G. T. J. van der Horst, F. Levi, D. A. Rand, and F. Delaunay, "Coupling between the Circadian Clock and Cell Cycle Oscillators: Implication for Healthy Cells and Malignant Growth," (in eng), *Frontiers in neurology,* vol. 6, pp. 96-96, 2015.

[15] G. M. Whitesides, "The origins and the future of microfluidics," *Nature,* vol. 442, no. 7101, pp. 368-373, 2006.

[16] M. Kimura, *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press, 1983.

[17] C. Caranica *et al.*, "Ensemble methods for stochastic networks with special reference to the biological clock of *Neurospora crassa*," *PLoS One,* vol. 13, no. 5, p. e0196435, 2018.

[18] U. Abraham, A. E. Granada, P. O. Westermark, M. Heine, A. Kramer, and H. Herzel, "Coupling governs entrainment range of circadian clocks," *Molecular Systems Biology,* 10.1038/msb.2010.92 vol. 6, no. 1, 2010.

[19] M. H. Vitaterna *et al.*, "The mouse *Clock* mutation reduces circadian pacemaker amplitude and enhances efficacy of resetting stimuli and phase–response curve amplitude," *Proceedings of the National Academy of Sciences,* 10.1073/pnas.0603601103 vol. 103, no. 24, p. 9327, 2006.

[20] A. B. Webb, N. Angelo, J. E. Huettner, and E. D. Herzog, "Intrinsic, nondeterministic circadian rhythm generation in identified mammalian neurons," *Proceedings of the National Academy of Sciences USA,* vol. 106, no. 38, pp. 16493-16498, 2009.

[21] S. Gan and E. K. O'Shea, "An Unstable Singularity Underlies Stochastic Phasing of the Circadian Clock in Individual Cyanobacterial Cells," *Molecular Cell,* vol. 67, no. 4, pp. 659-672.e12, 2017/08/17/ 2017.

[22] Y. Yu *et al.*, "A genetic network for the clock of *Neurospora crassa*," *Proc Natl Acad Sci U S A,* vol. 104, no. 8, pp. 2809-2814, 2007.

[23] D. Battogtokh, D. K. Asch, M. E. Case, J. Arnold, and H. B. Schuttler, "An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of *Neurospora crassa*," (in eng), *Proc Natl Acad Sci U S A,* vol. 99, no. 26, pp. 16904-9, Dec 24 2002.

[24] D. P. Landau and K. Binder, "A Guide to Monte Carlo Simulations in Statistical Physics," *Cambridge University Press*, 2009.

[25] J. H. Holland, "Adaptation and Natural and Artificial Systems," *MIT Press, Boston*, 1975.

[26] W. Ye, W. Feng, and S. Fan, "A novel multi-swarm particle swarm optimization with dynamic learning strategy," *Applied Soft Computing*, vol. 61, pp. 832-843, 2017/12/01/ 2017.

[27] X. Xu, Y. Tang, J. Li, C. Hua, and X. Guan, "Dynamic multi-swarm particle swarm optimizer with cooperative learning strategy," *Applied Soft Computing*, vol. 29, pp. 169-183, 2015/04/01/ 2015.

[28] C. Caranica *et al.*, "What is phase in cellular clocks?," *Yale Journal of Biology and Medicine*, vol. i9, no. 2, pp. 169-178, 2019.

[29] S. K. Crosthwaite, J. J. Loros, and J. C. Dunlap, "Light-induced resetting of a circadian clock is mediated by a rapid increase in frequency transcript," (in eng), *Cell*, vol. 81, no. 7, pp. 1003-12, Jun 30 1995.

[30] E. Castro-Longoria, M. Ferry, S. Bartnicki-Garcia, J. Hasty, and S. Brody, "Circadian rhythms in *Neurospora crassa*: Dynamics of the clock component *frequency* visualized using a fluorescent reporter," *Fungal Genetics and Biology*, vol. 47, no. 4, pp. 332-341, 2010.

[31] W. Dong *et al.*, "Systems biology of the clock in *Neurospora crassa*," (in eng), *PLoS One*, vol. 3, no. 8, p. e3105, 2008.

[32] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 81, 1977.

[33] D. T. Gillespie, "Exact Stochastic Simulation of Coupled Chemical-Reactions," (in English), *Journal of Physical Chemistry,* vol. 81, no. 25, pp. 2340-2361, 1977.

[34] D. T. Gillespie, "The chemical Langevin equation," *The Journal of Chemical Physics,* vol. 113, no. 1, pp. 297-306, 2000.

[35] D. T. Gillespie, "Approximate accelerated stochastic simulation of chemically reacting systems," *The Journal of Chemical Physics,* vol. 115, no. 4, pp. 1716-1733, 2001.

[36] A. Al-Omari, J. Griffith, M. Judge, T. Taha, J. Arnold, and H. Schuttler, "Discovering regulatory network topologies using ensemble methods on GPGPUs with special reference to the biological clock of *Neurospora crassa*," *Access, IEEE,* vol. 3, pp. 27-42, 2015.

[37] A. Al-Omari, J. Griffith, C. Caranica, T. Taha, H. Schüttler, and J. Arnold, "Discovering Regulators in Post-Transcriptional Control of the Biological Clock of *Neurospora crassa* Using Variable Topology Ensemble Methods on GPUs," *IEEE Access,* vol. 6, pp. 54582-54594, 2018.

[38] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins, "Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling," *Science,* vol. 301, no. 5629, p. 102, 2003.

[39] M. Izumo, T. R. Sato, M. Straume, and C. H. Johnson, "Quantitative Analyses of Circadian Gene Expression in Mammalian Cell Cultures," *PLoS Comput Biol,* vol. 2, no. 10, p. e136, 2006.

[40] P. Bloomfield, *Fourier analysis of time series : an introduction*. New York: Wiley, 1976, pp. xiii, 258 p.

[41] P. Thomas, A. V. Straube, J. Timmer, C. Fleck, and R. Grima, "Signatures of nonlinearity in single cell noise-induced oscillations," *Journal of Theoretical Biology,* vol. 335, pp. 222-234, 2013/10/21/ 2013.

[42] L. Breiman, *Probability*. Reading, Mass.,: Addison-Wesley Pub. Co., 1968, pp. ix, 421 p.

[43] F. Hamze, N. Dickson, and K. Karimi, "Robust parameter selection for parallel tempering," *International Journal of Modern Physics C,* vol. 21, no. 05, pp. 603-615, 2010.

[44] S. Joe and F. Kuo, "Constructing Sobol Sequences with Better Two-Dimensional Projections," *SIAM Journal on Scientific Computing,* vol. 30, no. 5, pp. 2635-2654, 2008/01/01 2008.

[45] I. M. Sobol, "On the distribution of points in a cube and the approximate evaluation of integrals," *Zh. Vychisl. Mat. Mat. Fiz,,* vol. 7, no. 4, pp. 784-802, 1967.

[46] M. A. Asmussen and J. Arnold, "The effects of admixture and population subdivision on cytonuclear disequilibria," *Theoretical Population Biology,* vol. 39, no. 3, pp. 273-300, 1991/06/01/ 1991.

[47] J. Dai, L. Li, S. Kim, B. Kimball, S. M. Jazwinski, and J. Arnold, "Exact sample size needed to detect dependence in 2 x 2 x 2 tables," (in eng), *Biometrics,* vol. 63, no. 4, pp. 1245-52, Dec 2007.

[48] E. L. Lehmann, "Nonparametrics: Statistical Methods Based on Ranks," Holden-Day: San Francisco, p. p. 300, 1975.

[49] D. Gonze, J. Halloy, and A. Goldbeter, "Robustness of circadian rhythms with respect to molecular noise," *Proceedings of the National Academy of Sciences,* vol. 99, no. 2, p. 673, 2002.

[50] W.-J. Rappel and S. H. Strogatz, "Stochastic resonance in an autonomous system with a nonuniform limit cycle," *Physical Review E*, vol. 50, no. 4, p. 3249, 1994.

[51] H. Gang, T. Ditzinger, C.-Z. Ning, and H. Haken, "Stochastic resonance without external periodic force," *Physical Review Letters*, vol. 71, p. 807, 1993.

[52] P. O. Westermark, D. K. Welsh, H. Okamura, and H. Herzel, "Quantification of circadian rhythms in single cells," *PLoS Comput Biol*, vol. 5, no. 11, p. e1000580, 2009.

[53] D. Gabor, "Theory of communication. Part 1: The analysis of information," *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, vol. 93, no. 26, pp. 429-441, 1946.

[54] Y. Ding, J. Shen, J. Lu, and J. Kurths, "Stochastic resonance in genetic regulatory networks under Lévy noise," *EPL (Europhysics Letters)*, vol. 127, no. 5, p. 50003, 2019/10/10 2019.

[55] L. Gammaitoni, P. Hänggi, P. Jung, and F. Marchesoni, "Stochastic resonance," *Reviews of Modern Physics*, vol. 70, no. 1, pp. 223-287, 01/01/ 1998.

[56] S. Jiao, S. Lei, W. Jiang, Q. Zhang, and W. Huang, "A Novel Type of Stochastic Resonance Potential Well Model and Its Application," *IEEE Access*, vol. 7, pp. 160191-160202, 2019.

[57] K. Sriram and M. S. Gopinathan, "Stochastic resonance in circadian systems," *Theor Chem Acc*, vol. 114, pp. 46-51, 2005.

[58] D. Nicolas, B. Zoller, D. M. Suter, and F. Naef, "Modulation of transcriptional burst frequency by histone acetylation," *Proceedings of the National Academy of Sciences,* vol. 115, no. 27, p. 7153, 2018.

[59] D. J. Jenkins, B. Finkenstädt, and D. A. Rand, "A temporal switch model for estimating transcriptional activity in gene expression," *Bioinformatics,* vol. 29, no. 9, pp. 1158-1165, 2013.

[60] R. L. Stratonovitch, "Theory of Random Noise, Volume II," *Gordon and Breach, London,* 1967.

[61] A. Wagner, "Circuit topology and the evolution of robustness in two-gene circadian oscillators," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 33, p. 11775, 2005.

[62] P. Cheng, Y. Yang, and Y. Liu, "Interlocked feedback loops contribute to the robustness of the Neurospora circadian clock," *Proceedings of the National Academy of Sciences,* vol. 98, no. 13, p. 7408, 2001.

[63] Allon M. Klein *et al.*, "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells," *Cell,* vol. 161, no. 5, pp. 1187-1201, 2015/05/21/ 2015.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In this thesis we used stochastic models to study oscillatory biochemical networks. We focused our study on the clock network of the fungal system *Neurospora crassa*.

In chapter 2, we derived a method for inferring the parameters of the clock network assuming its temporal evolution is stochastic. For this, we used Markov Chain Monte Carlo (MCMC) methods, namely Metropolis-Hastings and Parallel Tempering algorithms to fit the average periodogram of 868 single-cell trajectories of the clock-controlled gene's protein (*CCG*) observed over a 10-day period. We obtained a very good fit using Parallel Tempering algorithm and we were able to test Stochastic Resonance Hypothesis to see whether the intracellular noise was the one driving the oscillations. In the process we developed a method for separating the detection noise from intracellular noise and were able to do a bias-correction of the average periodogram.

In chapter 3 we discussed several measures of phase and we saw how they can be used to measure the synchronization of cellular oscillators. They are important because they can provide a goodness of fit measure between observed and simulated data independent of average periodogram fitting.

In chapter 4 we extended our models from chapter 2 to try to fit the single cell data observed under different light-dark regimes. We saw that parallel tempering method was no longer good enough to fit the extended data. We used as an alternative two genetic algorithms to get a better sampling of the parameter space. These genetic algorithms provided a much better fitting than parallel tempering. We were able to fit very well the main frequencies in the light-dark data and we tested again the Stochastic Resonance Hypothesis. We noticed that there was a single level of noise optimal for all 4 dark/dark and light/dark regimes, an amazing fact that may require further testing/investigation.

In the future, we wish to developed ensemble methods that describe the synchronization of these cellular oscillators. We are going to use quorum sensing as a means of communications between cells and as measure of synchronization we will use intra-class correlation. While we have experimental data for cell isolated in droplets, with droplets containing up to 10 cells, we want to scale the models to 1000s of cells using the Mean Field approximation theory.

Another line of research is to try to fit stochastic models to oscillatory data by developing methods that incorporate phase variation. Just using phase as a primary test for fitting the data will gives as new stochastic models that can be compared to the models we already obtained.

APPENDIX

BIOGRAPHY

Constantin Cristian Caranica finished his undergraduate studies at Bucharest University in June 1997, with a major in mathematics. He earned a Master of Science degree in mathematics from Bucharest University in February 2000. In August 2002 he went to Louisiana State University to pursue graduate studies in mathematics where he earned a Doctor of Philosophy in mathematics in August 2009. He is currently a candidate for the degree of Doctor of Philosophy in statistics, which will be awarded in May 2020.