Monte Carlo Studies of Genetic Networks with Special Reference to the Biological Clock of *Neurospora crassa*

by

Yihai Yu

(Under the direction of H. B. Schüttler)

Abstract

With the sequence of genomes of many organisms now available, the major challenge of functional genomics is "re-assembling the pieces". A chemical reaction network is considered to be a very simple and efficient view of a living system. A general purpose kinetic simulator (KINSOLVER) is developed. As a stochastic alternative, an efficient statistical Monte Carlo method is applied to identify an ensemble of deterministic models consistent with RNA and protein profiling data for biological clock of *Neuropora crassa*. Maximally Informative Next Experiment (MINE) is designed and employed to guide new experiments to further improve the quality of the quantitative prediction. A Java Servlet based web site (ENSSOLVER) is developed to visualize and analyze the simulation results.

Index words: Monte Carlo, Ensemble Identification, Genetic Network, *Neuropora crassa*, Biological Clock, Design of Experiment(DOE), Java Servlet

Monte Carlo Studies of Genetic Networks with Special Reference to the Biological Clock of *Neurospora crassa*

by

Yihai Yu

B.S., Fudan University, 2000

M.S., The University of Georgia, 2004

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

2005

Monte Carlo Studies of Genetic Networks with Special Reference to the Biological Clock of *Neurospora crassa*

by

Yihai Yu

Approved:

Major Professor: H. B. Schüttler

Committee: Jonathan Arnold
Michael R. Geller

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2005

Acknowledgments

My chief acknowledgment is to my advisors, Dr. H. B. Schüttler and Dr. Jonathan Arnold, for their guidance, encouragement and insight during this research work. They not only taught me how to conduct research in this multidisciplinary area but also how to enjoy it. I especially would like to express my appreciation for their support and advising for my career development.

I also would like to thank Dr. Michael R. Geller for serving on my advisory committee. I greatly appreciate the Department of Physics and Astronomy not only for granting a teaching assistantship, but also the great opportunity to take courses with so many great professors. I also wish to thank Dr. S. Tsai, Mike Caplinger and Jeff Deroshia for being always so patient and willing to help me with all kinds of problems with computing facilities.

I would like to thank Dr. Thiab Taha, the advisor for my MS of Computer Science, for his stimulating discussion during this research work. I also would like to thank Boanerges Aleman-Meza who implemented the first version of KINSOLVER and gave me a lot of help for my further work. My appreciation also goes to Dr. Wubei Dong, Cara Altimus and Lisa Dudek for their significant input for my research. I also greatly appreciate the inspiring discussion with Dr. Jaxk Reeves, Dr. Shishan Qu and my fellow graduate students in the Department of Physics and Astronomy.

I greatly appreciate Dr. H. B. Schüttler and Dr. Jonathan Arnold for granting a research assistantship. This work is supported by US National Science Foundation under NSF DBI-0243754 and BES-0425762.

Finally, but most importantly, I would like to thank my family. They even don't understand one single word in this dissertation, but without them, this dissertation never would have been written. They give me a life and teach me how to live it.

Table of Contents

Page

Acknowledgments iv

List of Figures vii

List of Tables xiv

Chapter

1 Introduction 1

2 Kinetics Model (KINSOLVER) 4

2.1 Introduction to kinetics model 4

2.2 Numerical methods for kinetics model 6

3 Monte Carlo (MC) Method and Ensemble of Models 10

3.1 Monte Carlo (MC) method 10

3.2 Ensemble of Models 14

4 Monte Carlo (MC) Studies for the Biological Clock of *Neurospora crassa* 18

4.1 Introduction 18

4.2 Genetic Network Model for the Biological Clock 19

4.3 Experimental Methods 20

4.4 *Metropolis Algorithm* for Biological Clock 20

4.5 Results for light-independent biological clock model . . 23

4.6 Light-dependent biological clock model and results . . . 31

4.7 Discussion 33

5 Maximally Informative Next Experiment (MINE) 43
5.1 Introduction 43
5.2 Maximization Criteria 45
5.3 MINE for Biological Clock 47
6 ENSSOLVER 51
Bibliography 55
Appendix
A Stability Analysis for Biological Clock Model 61
A.1 Introduction to Routh-Hurwitz Analysis 61
A.2 Stability Criterion $R > 0$ for Biological Clock Model . . 62
A.3 Stability Criterion $nm > 4$ for Biological Clock Model . 65

List of Figures

2.1 Simple Hydrogen Combustion Model. 8

2.2 A genetic network for the biological clock. The *white-collar-1* (*wc-1*), *white-collar-2* (*wc-2*), *frequency* (*frq*), and *clock controlled gene* (*ccg*) gene symbols can be superscripted 0, 1, r0, r1, indicating, respectively, a transcriptionally inactive (0) or active (1) gene or a translationally inactive (r0) or active (r1) mRNA. Associated protein species are denoted by capitals. Reactions in the network are represented by circles. Arrows entering circles identify reactants; arrows leaving circles identify products; and bi-directional arrows identify catalysts. The labels on each reaction, such as S_4 , also serve to denote the rate coefficients for each reaction. Reactions without products, such as D_8 , are decay reactions. Reactions A and P have cooperative kinetics: (A) $nWCC + frq^0 \rightarrow frq^1$ and (P) $WCC + mFRQ \rightarrow$ *WC-2* + mFRQ. The n and m are Hill coefficients or cooperativities. Only for the A reaction, a backward reaction, $(\bar{A})$ $frq^1 \rightarrow nWCC + frq^0$, is included, with non-zero rate coefficient $\bar{A}$. . . . 9

4.1 Monte Carlo random walk equilibration in the parameter ($\mathbf{\Theta}$) space of the models. Progress towards equilibrium is monitored by $\chi^2 = -2\ln Q$+const which is a measure of the departure between the data and the model prediction of the genetic network in Figure 2.2, for model cooperativities (*i.e.*, Hill Coefficients) $n = m = 4$. As the fit is refined, χ^2 , on average, decreases with progressive Monte Carlo sweeps in the parameter space. One "sweep" comprises one visit, on average, to all M unknown parameter values, $\Theta_1, ...\Theta_M$ (where $M = 47$ in this model), with a consequent random decision whether or not to change the visited parameter, based on the Metropolis updating rule. This random search method for finding a parameter region near a minimum of χ^2 occasionally takes a risk and permits the χ^2 to increase (worsen the fit), thereby allowing the walk to escape from local minima in the χ^2-surface, as seen in the figure. 34

4.2 Comparison to experiments. A model ensemble for the genetic network in Figure 2.2, with cooperativities $n = m = 4$, predicts various profiling data on the biological clock of *Neurospora crassa*. In each panel, predictions of the model ensemble for the lg of concentration (in model concentration units "*cu*") are shown with $+/-$ 2 ensemble standard deviations (shaded areas) about the ensemble mean (continuous lines). The points are the experimental lg data. (A) levels of conidiation for a *bd* mutant, hypothesized to be a measure of a *CCG* protein, over a 7-day interval; (B) levels of *wc-1*^{r1} and *frq*^{r} RNA over time t; (C) levels of *FRQ* and *WC-1* protein over time *t*. 35

4.3 Comparison of model fits. Some model ensembles, using modified or simplified versions of the model in Figure 2.2, outperform the original model in terms of χ^2-values. A histogram of values of $\chi^2 = -2 \ln Q + \text{const}$ is shown for varying model ensembles. (A) The number of molecules of *WCC* (n) or *FRQ* (m) working together cooperatively (*i.e.*, the *Hill coefficients*) in reactions A and P are varied with $n = m$. Also the histogram of χ^2-values for a model ensemble without post-transcriptional regulation of *wc-1* by *FRQ* is reported. Some models with less cooperativity in the action of *WCC* or *FRQ* (*e.g.* $n = m = 3$, dark blue curve) or without post-transcriptional regulation (light blue curve) have smaller χ^2-values on average than the $n = m = 4$ model. We have explored in detail models with $n \neq m$ (for $n, m = 2, \cdots, 4$) with additional Monte Carlo runs as in Figure 4.3; it was unnecessary to consider $n = 1$ and $m = 1, \cdots, 4$ or $m = 1$ and $n = 1, \cdots, 4$ because these cases do not satisfy the condition $nm > 4$ necessary for oscillations (see Appendix A). The best model ensemble (histogram most shifted to the left) to date remained $n = m = 3$. (B) The deactivation reaction (P) is varied, allowing by 4 different deactivation mechanisms (including the one in Figure 2.2, in black) to be compared for their goodness of fit. Each deactivation reaction is defined in the Figure legend. 36

4.4 3D projection of the model ensemble. The 3D parameter sub-space is spanned by re-scaled, dimensionless rate coefficients P', A' and $\bar{A}'$. Points in the ensemble allowing only sustained, undamped oscillations are shown in red; points in the ensemble that allow for damped oscillations (from the Stability Analysis in Appendix A) are shown in blue, for the model of Figure 2.2 with $n = m = 4$. To construct P', A' and $\bar{A}'$, define the "maximal" concentrations of *WCC* and *FRQ*, $w_x := L_1 V_1/(D_7 D_8)$ and $f_{px} := f_G S_4 L_3/(D_3 D_6)$, where $V_1 := S_1 u_1 := S_1[wc\text{-}1^1]$ is the rate of $[wc\text{-}1^{r0}]$-production and $f_G := f_0 + f_1 := [frq^0] + [frq^1]$ is the total, constant *frq*-gene concentration. Using the *FRQ* protein life-time $1/D_6$ as the "unit of time", then define dimensionless, re-scaled rate coefficients, such as P', A', $\bar{A}'$, L_3', S_4' and D_8' , respectively, for the P, A, $\bar{A}$, L_3, S_4 and D_8-reactions as follows: $P' := P(f_{px})^m/D_6$, $A' := A(w_x)^n/D_6$, $\bar{A}' := \bar{A}/D_6$, $L_3' := L_3/D_6$, $S_4' := S_4/D_6$ and $D_8' := D_8/D_6$. The re-scaling permits us to collapse higher-dimensional parameter sub-spaces, of dynamically equivalent models, into lower-dimensional ones. 37

4.5 3D projection of the model ensemble. (A) z-axis is replaced by $S_4' := S_4/D_6$. (B) z-axis is replaced by $L_3' := L_3/D_6$. 38

4.6 Genetic network for biological clock of *Neuropora crassa* with consideration of light. 3 more species are introduced: *Phot*, *WCCL* and $frq^1 L$, and consequently E_1, E_2, B, $\bar{B}$, C_c and Q-reactions are included. 38

4.7 Comparison to experiments. A model ensemble for the genetic network in Figure 4.6, with cooperativities $n = m = 4$, predicts *CCG* data on the biological clock of *Neurospora crassa*. In each panel, predictions of the model ensemble for the lg of concentration (in model concentration units "*cu*") of *CCG* protein are shown with +/− 2 ensemble standard deviations (shaded areas) about the ensemble mean (continuous lines). (A) 6+6 hours LD cycle; (B) 9+9 hours LD cycle; (C) 18+18 hours LD cycle. The *CCG* protein data are obtained from [58]. 39

4.8 Comparison of three different fitting methods for *CCG* and *FRQ* in the dark. (A) χ^2 is computed based on *wc-1*^{r1}. (B) χ^2 is computed based on *wc-1*^{r1}+*wc-1*^{r0}. The inset for *CCG* shows the lg-concentration at very early time. The peak value is greater than the average of later oscillation about 4 fold at lg-scale. (C) χ^2 is computed based on *wc-1*^{r1}+*wc-1*^{r0} together with the light dependent data from [58] (6+6, 9+9 and 18+18 hours LD cycle). 40

5.1 Area and vectors. 48

5.2 MINE prediction for different light intensity experiments. Three criteria are used, with LD cycle being 6+6 hours, the first measurement taking place at 0.5 hours, the spacing between consecutive measurements being 4 hours, and the total time points being 13. Görl's light intensity corresponds to 100 in the model units. x-axis shows different light intensity and y-axis shows three different scores for three different methods. (A) Trace method indicates as large light intensity as possible. (B) Determinant method shows a peak at 200, which is about 40 μmol photons/m^2/s. (C) Normalized determinant method clearly identifies 200 as the best choice. 49

5.3 MINE prediction for different measurement time experiments. x-axis, t_s, denotes the spacing between two consecutive measurements, and y-axis, t_L, denotes the first measurement time. The furthest time a microarray chip can measure is about 60 hours. Different contour color shows scores of three different maximization criteria. (A) Trace method indicates that we want to delay the start of measurement as late as possible and maximize the spacing at the same time. (B) Determinant method indicates that we want to make the spacing as large as possible while start the measurements as early as possible. (C) Normalized determinant method gives the same results as in the second method. 49

5.4 Definition of r ratio. Each 2 consecutive measurements are combined to define $2t\_s$. The ratio r is explored by moving the middle bar between its left barrier and right barrier, from 0 to 1. 50

5.5 MINE prediction for different spacing ratio experiments. x-axis is t_s, where $2t\_s$ is defined as total spacing for two consecutive measurements, and y-axis, r, denotes the spacing ratio as define in Figure 5.4. Again, three contour plots are shown for three different criteria. (A) The dependent trace method again predicts to delay the even-numbered measurement to be as late as possible, which is equivalent to make a replicate. (B) Determinant method enforces independence and the even-numbered measurements are forced to be in the middle of odd-numbered measurements for any fixed t_s. (C) The normalized determinant method gives similar results as the second method. 50

6.1 The Hierarchical Structure of ENSSOLVER. 52

6.2 Example of output. In this example, a plot of χ^2 versus Monte Carlo sweep is displayed together with caption of run name and model name. Some statistics about χ^2 is shown next. You can also replot the figure based on your needs and download the column file for further manipulation. This accumulation run explores a region with all the χ^2 being equally good (near-minimal values of χ^2) although some of them do not yield sustained oscillation. 53

6.3 Example of output. This is a 2-dimensional histogram plot for two Monte Carlo parameters: one is χ^2 and the other is the rate coefficient D_7, followed by some statistics of the two parameters. The rate coefficient D_7 is very well constrained in the range of 0.05 and 0.07. 54

List of Tables

4.1 Rate coefficients in the genetic network model of the biological clock ($n = m = 4$) predicting its observed oscillations. Ensemble mean $\langle X \rangle$ and ensemble standard deviation $\sigma(X) := [\langle X^2 \rangle - \langle X \rangle^2]^{1/2}$ for rate coefficients (X) in the $n = m = 4$ biological clock model. For a k^{th} order reaction (with k = 1,2 or 5), the rate coefficient is given in units of $1/(hour \times cu^{k-1})$ where "cu" represents the arbitrary, but common model unit of concentration.Column (A): χ^2 is computed based on *wc-1*^{r1}. Column (B): χ^2 is computed based on *wc-1*^{r1}+*wc-1*^{r0}.Column (C): χ^2 is computed based on *wc-1*^{r1}+*wc-1*^{r0} together with the light dependent data from [58] (6+6, 9+9 and 18+18 hours LD cycle). '-' means this reaction is not included in the corresponding model. 41

4.2 This analysis corresponds to Column (A) in Table reftable1. As the life-time of the translationally active *wc-1* messenger RNA, *wc-1*^{r1}, decreases (or equivalently, its decay rate coefficient D_7 increases), the model ensemble predicts that the system becomes arrhythmic. The percentage of stable oscillators, p_{osc}, in the model ensemble from Figure 4.4 (for the $n = m = 4$ model in Figure 2.2 with *WC-2* constant) perturbed by the decay rate coefficient D_7 for each ensemble member being varied from 0.05 ($\sim$ the mean estimate in Table 4.1) to 1.50 while keeping all other rate coefficients fixed at ensemble-generated values. This percentage monotonically decreases as the life-time of the translationally active *wc-1* mRNA decreases (or equivalently, as its decay rate coefficient D_7 increases). 42

Chapter 1

Introduction

With genome sequencing projects providing nearly complete inventory of the building blocks of life, functional genomics is now facing the challenge of "re-assembling the pieces" [1]. Time-dependent RNA [9] and protein profiling [10], protein-protein [11] and protein-DNA interaction mapping [12], and the *in vitro* reconstruction of biological reaction networks [13] are giving us detailed new insights into the make up and dynamics of a living cell's genetic and biochemical circuitry. Biological reaction network models or genetic networks provide a powerful theoretical and computational framework for integrating and summarizing such genomic, proteomic and metabolic information by allowing for a detailed, quantitative description of how the cell's molecular species (including genes, RNA, protein and other environmental molecules) interact with each other. In principle, genetic networks enable us to quantitatively describe the dynamics of a living cell. This may include, but not be limited to: how the cell evolves as a function of time, how the cell responds to the environmental change, how the cell behaves differently subject to alteration by genetic engineering.

The most fundamental question of life science is, what is a living system? From the quantitative biological circuit modelling's perspective, we can try to answer this question by asking another question: given the inventory of most molecular species and cellular compartments and sets of experimental data for the dynamic global response of the cell, *i.e.*, the time-dependent concentrations for the molecular species in a cell, how do you reconstruct the genetic network? This question basically consists of two parts: (1) the genetic network topology (which species is interacting with which species, in other words, how the

network is connected); (2) the model parameters (kinetic reaction rates and initial concentrations). The systems approach is now at the heart of functional genomics and seems to be the most promising approach to this fundamental question. A full understanding of fundamental processes like transcription, metabolism, development, biological clock, mating, aging and pathogenicity will be obtained when a hypothesized gene regulatory and biochemical reaction network can successfully predict the global response observed through genomics experiments.

Genetic networks can be partially identified for a few well-studied systems like the *lac* operon [2], *trp* operon [3, 4], *GAL* gene cluster [5], *qa* gene cluster [6], cell cycle [7], and biological clock [26]. These genetic networks display a diversity of dynamic behavior, including a transient response, switch-like behavior and oscillations. My research will focus on biological clock of *Neurospora crassa*. This particular genetic network has provided fundamental insights into how the clock functions in a variety of eukaryotes and provides an exciting example of a gene regulatory pathway with an oscillatory response.

Chemical reaction network [8] provides a simple, general framework for quantitative modeling of a genetic network's dynamic behavior in terms of the time-dependent concentrations of molecular species governed by a system of ordinary differential equations (ODEs). For the deterministic models, given all rates coefficients and initial concentrations, any genetic network's behavior will be determined. A general purpose kinetic simulator is developed armed with several standard ODE solvers [14]. But the problem we are facing here is that the rates coefficients and initial concentrations are mostly unknown, and the experimental data are typically sparse and noisy. The ensemble method has been proposed to step around this problem [27] by avoiding these traditional modeling attempts relying on educated guess of these unknown parameters. The basic idea is: instead of trying to identify one unique model parameter set, one should rather, at least initially, admit a random distribution of all possible parameter sets, *i.e.*, a statistical ensemble of candidate models according to the likelihood function based on the comparison of model prediction and the available experimental data.

Provided the ensemble of model parameter sets, we can then predict ensemble averages along with the uncertainties. Then one question follows naturally: how do we improve the prediction further? One quick answer is that we should collect more experimental data. The Maximally Informative Next Experiment (MINE) is designed and implemented to guide the future experiments, which will provide the maximally additional information so as to maximally constrain the model parameters. The basic idea is to compare the ensemble of model parameter sets with themselves by performing virtual experiments using the kinetics simulator.

In Chapter 2, I will discuss how to model an arbitrary genetic network and simulate it by computers. Chapter 3 contains the Monte Carlo ensemble method based on kinetics models. In Chapter 4, I will discuss the application of Monte Carlo ensemble method to the biological clock of *Neurospora crassa*. Chapter 5 will be focused on Maximally Informative Next Experiment (MINE). In Chapter 6, I will discuss the GUI software ENSSOLVER which is used to visualize the simulation results.

Chapter 2

Kinetics Model (KINSOLVER)

2.1 Introduction to kinetics model

A biological system can be viewed as a chemical reaction network [8]. Validating a genetic network depends upon our ability to simulate a particular reaction network and to predict how the network responds to various experimental perturbations. In order to refine and examine the behavior of a genetic network in a genomic context, an efficient general purpose simulator, KINSOLVER [14], is designed and implemented to simulate genetics networks represented by coupled nonlinear differential equations, *i.e.* compute the time dependent concentrations of each species from a simple interface for specifying and refining the target reaction network.

The system of Ordinary Differential Equations (ODEs) is a very basic and useful mathematical model in many areas, such as engineering, economics, physics, biology, etc.. The following is the canonical first order initial value problem (IVP) in general normal form (any higher order ODEs can be constructed by first order ODEs.):

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(t, \mathbf{y})$$

$$\mathbf{y}(t_0) = \mathbf{y_0}$$

where

t is the time-like independent variable,

$\mathbf{y}$ is the column N-vector of dependent variables,

N is the size of the vector,

$\frac{d}{dt}$ denotes differentiation with respect to t,

t_0 is the initial time and $\mathbf{y_0}$ is the N-vector initial condition, and

$\mathbf{f}$ is a N-vector valued function of t and $\mathbf{y}$.

A kinetics model is a specification of reactions between hypothesized molecular participants. The diagrammatic representation is like Figure 2.1. The species are represented as boxes. The reactions are represented by circles. The arrows indicate the directions of reactions. Since any living system can be viewed as a chemical reaction network, a thorough understanding of how these models behave is important.

For one particular equation: $S_1 + S_2 \rightleftharpoons S_3 + S_4$, there are 4 species, and 2 reactions: forward and backward, indicated by $\rightleftharpoons$. Suppose the forward reaction rate is k_f and backward reaction rate is k_b. We want to write out the differential equation for the species, *e.g.* S_1. In the forward reaction the species is consumed, therefore: $-k_f[S_1][S_2]$. For the backward reaction, S_1 is generated: $+k_b[S_3][S_4]$. Therefore we have:

$$\frac{dS_1}{dt} = -k_f[S_1][S_2] + k_b[S_3][S_4]$$

Similar equations for the other 3 species can be constructed. Usually the backward reaction rate is smaller, sometimes just 0.

For this simple Hydrogen Combustion Model, the 3 reactions are:

$$H_2 + O \rightleftharpoons H + OH$$

$$H + O_2 \rightleftharpoons O + OH$$

$$H_2 + OH \rightleftharpoons H_2O + H.$$

Based on the method just introduced, we can obtain the full multiplicative mass balance kinetics, *i.e.* the complete set of ODEs:

$$\frac{d[H_2]}{dt} = -k_{f1}[H_2][O] + k_{b1}[H][OH] - k_{f3}[H_2][OH] + k_{b3}[H_2O][OH]$$

$$
\begin{aligned}
\frac{d[O]}{dt} &= -k_{f1}[H_2][O] + k_{b1}[H][OH] + k_{f2}[H][O_2] - k_{b2}[O][OH] \\
\frac{d[O_2]}{dt} &= -k_{f2}[H][O_2] + k_{b2}[O][OH] \\
\frac{d[H]}{dt} &= +k_{f1}[H_2][O] - k_{b1}[H][OH] - k_{f2}[H][O_2] \\
&\quad +k_{b2}[O][OH] + k_{f3}[H_2][OH] - k_{b3}[H_2O][OH] \\
\frac{d[OH]}{dt} &= +k_{f1}[H_2][O] - k_{b1}[H][OH] + k_{f2}[H][O_2] \\
&\quad -k_{b2}[O][OH] - k_{f3}[H_2][OH] + k_{b3}[H_2O][OH] \\
\frac{d[H_2O]}{dt} &= k_{f3}[H_2][OH] - k_{b3}[H_2O][OH]
\end{aligned}
$$

where k_{fi} and k_{bi} are the forward and backward reaction rates respectively for reaction number i.

Now, we have formalized the kinetics modeling. The routine can be applied to any deterministic genetic network regardless how complex they are. The genetic network for biological clock of *Neuropora crassa* is shown as in Figure 2.2.

2.2 Numerical methods for kinetics model

Kinetics model is just a system of ODEs. So the simulation is just the numerical integration of ODEs.

The Euler method [15] is the simplest numerical method for solving ODEs which utilizes the first order of Taylor series expansion:

$$y_{n+1} = y_n + hf(t_n, y_n)$$

where y_n is the solution at time t_n, and y_{n+1} is the estimate of the solution of time t_{n+1} based on y_n and slope at t_n. h is the step size of discretization.

The RK method [16] is a classical higher order method. Here is one example for 4^{th} order RK:

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

where

$$\begin{aligned} k_1 &= hf(t_n, y_n) \\ k_2 &= hf(t_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1) \\ k_3 &= hf(t_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1) \\ k_4 &= hf(t_n + h, y_n + k_3) \end{aligned}$$

These two conventional methods are efficient for non-stiff problems. But for many genetic networks, they actually exhibit the property of stiffness and make conventional methods to be very inefficient [17].

A formal definition of stiffness was given by L. F. Shampine and C. W. Gear [18]:

> "By a stiff problem we mean one for which no solution component is unstable (no eigenvalue has a real part which is at all large and positive) and at least some component is very stable (at least one eigenvalue has a real part which is large and negative). Further, we will not call a problem stiff unless its solution is slowly varying with respect to the most negative real part of the eigenvalues."

In other words, stiffness is defined in relation to the most negative mode and the time span we are interested in.

Let us have a look at a simple example. A typical stiff differential equation can be given by:

$$\begin{aligned} \frac{dy}{dt} &= -10^3[y - \exp(-t)] - \exp(-t) \\ y(0) &= 0 \end{aligned}$$

where y is a scalar for simplicity. The exact solution is:

$$y(t) = \exp(-t) - \exp(-10^3 t)$$

which consists of two components: $\exp(-t)$ and $\exp(-10^3 t)$. Obviously, $\exp(-10^3 t)$ varies more rapidly than $\exp(-t)$. This system consists of two normal modes: $\tau_1 = 1$ and $\tau_2 = 10^{-3}$. In order to reach equilibrium, we have to consider the time scale τ_1, while we also have to consider τ_2 for the sake of step size of time. Apparently, if we make the step size too big, it will invalidate the conventional numerical methods. Hence the step size has to be small, about the same scale as τ_2. But the total integration time needed to reach equilibrium is much bigger than the step size. As a consequence, we have to wait an unbearably long time. A more general analysis using Jacobian matrix and various numerical methods for stiff problems are discussed in [17].

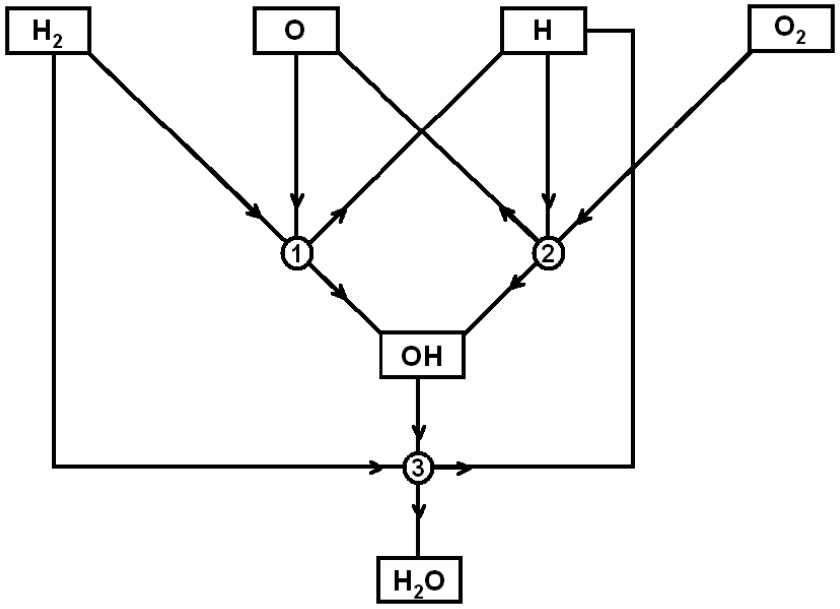


Figure 2.1: Simple Hydrogen Combustion Model.

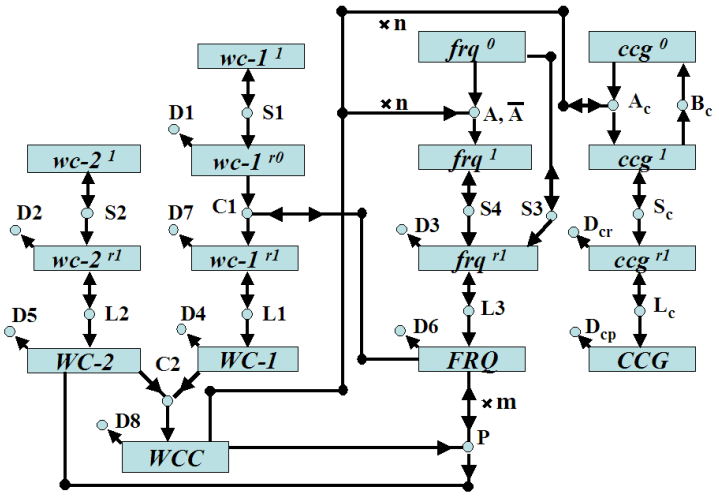


Figure 2.2: A genetic network for the biological clock. The *white-collar-1* (*wc-1*), *white-collar-2* (*wc-2*), *frequency* (*frq*), and *clock controlled gene* (*ccg*) gene symbols can be superscripted 0, 1, r0, r1, indicating, respectively, a transcriptionally inactive (0) or active (1) gene or a translationally inactive (r0) or active (r1) mRNA. Associated protein species are denoted by capitals. Reactions in the network are represented by circles. Arrows entering circles identify reactants; arrows leaving circles identify products; and bi-directional arrows identify catalysts. The labels on each reaction, such as S_4 , also serve to denote the rate coefficients for each reaction. Reactions without products, such as D_8 , are decay reactions. Reactions A and P have cooperative kinetics: (A) $nWCC + frq^0 \rightarrow frq^1$ and (P) $WCC + mFRQ \rightarrow WC\text{-}2 + \mathrm{m}FRQ$. The n and m are Hill coefficients or cooperativities. Only for the A reaction, a backward reaction, $(\bar{A})$ $frq^1 \rightarrow nWCC + frq^0$, is included, with non-zero rate coefficient $\bar{A}$.

Chapter 3

Monte Carlo (MC) Method and Ensemble of Models

3.1 Monte Carlo (MC) method

Many systems in physics have very high degrees of freedom, which makes conventional direct numerical integration methods not applicable due to the fact that their estimate of error is $O(N^{-2/d})$, where N is the total number of points and d is the dimension of the system. But MC method's estimate of error is $O(N^{-2})$, it does not depend on the dimension of the system.

Modern Monte Carlo methods have their recent roots in the 1940s, when Fermi, Ulam, von Neumann, Metropolis began considering the use of random numbers to exam different problems in physics from a stochastic perspective [19]. A MC method allows us to follow the time dependence of a model for which the change or growth is not well defined. The basic idea is to provide approximate solutions to a problem by performing statistical sampling experiments.

The beauty of a MC method indicates that the scope of applications is enormous. It is now widely used in physics, biology, finance and other disciplines [21] [22].

3.1.1 Simple sampling

Suppose, we wish to calculate the one-dimensional definite integral:

$$I = \int_a^b f(\theta)\, d\theta. \tag{3.1}$$

The simple sampling Monte Carlo method gives an estimate value of this integral by choosing n points θ_i randomly from the interval $[a, b]$ with a uniform distribution:

$$I \cong \frac{b-a}{n} \sum_{i=1}^{n} f(\theta_i). \tag{3.2}$$

Here we actually approximate the average of the function $f(\theta)$ to be:

$$\langle f(\theta) \rangle \cong \frac{1}{n} \sum_{i=1}^{n} f(\theta_i). \tag{3.3}$$

Consider $f(\theta_i)$ as the random response variable, use the Law of Large Numbers when n is large, we have:

$$\sigma^2 = \frac{1}{n}[\frac{1}{n} \sum_{i=1}^{n} f(\theta_i)^2 - [\frac{1}{n} \sum_{i=1}^{n} f(\theta_i)]^2]. \tag{3.4}$$

The error for this estimate is $O(\frac{1}{\sqrt{n}})$, the convergence is very slow. Also, simple sampling will only be approximate for functions that are relatively smooth. Any sharp peak in the function f will probably be missed by a simple MC method. One simple, quick fix is to divide the interval into a set of unequal sub-intervals and perform the simple MC integration for each sub-interval. But the more general approach is to sample the function according to the its shape, which means the number of random points for each interval $d\theta$ should be selected proportional to $f(\theta)d\theta$. This leads to the development of importance sampling as a Monte Carlo method.

3.1.2 Importance sampling

We can rewrite the general definite integral 3.1 as:

$$I = \int_a^b f(\theta)\, d\theta = \int_a^b \frac{f(\theta)}{p(\theta)} p(\theta)\, d\theta, \tag{3.5}$$

where the density function $p(\theta)$ satisfies:

$$\int_a^b p(\theta)\, d\theta = 1. \tag{3.6}$$

Also, define:

$$y(\theta) = \int_0^\theta p(\theta')\, d\theta', \tag{3.7}$$

which gives:

$$\frac{dy}{d\theta} = p(\theta);\ y(\theta = a) = a;\ y(\theta = b) = b.$$

Now, the integration becomes:

$$I = \int_a^b \frac{f(\theta(y))}{p(\theta(y))}\, dy. \tag{3.8}$$

Apply the same technique as simple sampling here gives the *average* of the function of $f(\theta)$ as:

$$\langle f(\theta)\rangle \cong \frac{1}{n}\sum_{i=1}^{n} \frac{f(\theta(y_i))}{p(\theta(y_i))}, \tag{3.9}$$

with the variance to be:

$$\sigma^2 = \int_a^b [\frac{f(\theta)}{p(\theta)}]^2 p(\theta)\, d\theta - [\int_a^b \frac{f(\theta)}{p(\theta)} p(\theta)\, d\theta]^2. \tag{3.10}$$

If we can select the p to have similar behavior as f, then f/p will be very smooth. The reason for this major improvement is: the distribution of the points of x is now $dy/d\theta = p(\theta)$ based on the uniform distribution of y, which means we indeed allocate more points to the more *important* places.

Let's look at a real high-dimensional thermal system, which will be directly adopted to the ensemble method for genetic network identification. The expectation value of an observable $A(\mathbf{\Theta})$, where $\mathbf{\Theta}$ is the phase space vector governed by the Hamiltonian function $H(\mathbf{\Theta})$, can be expressed as:

$$\langle A(\mathbf{\Theta})\rangle = \frac{1}{Z}\int_{-\infty}^{\infty} A(\mathbf{\Theta}) e^{-\beta H(\mathbf{\Theta})}\, d\mathbf{\Theta}, \tag{3.11}$$

where $\beta = \frac{1}{kT}$, k is *Boltzman* constant and T is temperature. Z is called *partition function* which is the normalization factor for the canonical ensemble:

$$Z = \int_{-\infty}^{\infty} e^{-\beta H(\mathbf{\Theta})}\, d\mathbf{\Theta} \tag{3.12}$$

Comparing 3.12 with 3.5, we let $A(\mathbf{\Theta}) = \frac{f(\Theta)}{p(\Theta)}$ and $p(\mathbf{\Theta}) = \frac{1}{Z} e^{-\beta H(\mathbf{\Theta})}$. According to the results of 3.9, the expectation value is:

$$\langle A(\mathbf{\Theta})\rangle \cong \frac{1}{n}\sum_{i=1}^{n} A(\mathbf{\Theta}_i). \tag{3.13}$$

Now the question is how we can actually generate random numbers according to $p(\mathbf{\Theta})$. In general, there is no efficient analytical algorithm to generate a sequence of random numbers for arbitrary weight function. It will be even more complicated when the Hamiltonian is a general non-linear function, the *blackbox energy criterion*, depending on the vector $\mathbf{\Theta}$. One approach comes from *Metropolis Algorithm*.

3.1.3 *Metropolis Algorithm*

The idea of using *Markov Chain* proposed by Metropolis [20] is that one starts from an initial state $\mathbf{\Theta_0}$ and then further states are generated which are ultimately distributed according to equilibrium $p(\mathbf{\Theta})$ when n goes to infinity. In order that the Markov process converges to $p(\mathbf{\Theta})$, the transition rate (the probability of transition from $\mathbf{\Theta}_i$ to $\mathbf{\Theta}_j$ in phase space) $T(\mathbf{\Theta}_i \rightarrow \mathbf{\Theta}_j)$ must satisfy the detailed balance condition:

$$p(\mathbf{\Theta}_i)T(\mathbf{\Theta}_i \rightarrow \mathbf{\Theta}_j) = p(\mathbf{\Theta}_j)T(\mathbf{\Theta}_j \rightarrow \mathbf{\Theta}_i) \tag{3.14}$$

Then we get:

$$\frac{p(\mathbf{\Theta}_i)}{p(\mathbf{\Theta}_j)} = \frac{T(\mathbf{\Theta}_j \rightarrow \mathbf{\Theta}_i)}{T(\mathbf{\Theta}_i \rightarrow \mathbf{\Theta}_j)}, \tag{3.15}$$

where

$$\frac{p(\mathbf{\Theta}_i)}{p(\mathbf{\Theta}_j)} = \frac{W(\mathbf{\Theta}_i)}{W(\mathbf{\Theta}_j)} \tag{3.16}$$

This effectively allows us to rely on the weight function instead of the equilibrium density function.

The following is the overall flow of *Metropolis Algorithm*:

1. Specify an initial state $\mathbf{\Theta}_0$ in phase space.
2. Propose a new state $\mathbf{\Theta}'$ with $T_p(\mathbf{\Theta} \rightarrow \mathbf{\Theta}')$, where $T_p(\mathbf{\Theta} \rightarrow \mathbf{\Theta}')$ is a user defined perturbation function
3. Accept the new state $\mathbf{\Theta}'$ with $T_a(\mathbf{\Theta} \rightarrow \mathbf{\Theta}')$ and reject it with $1 - T_a(\mathbf{\Theta} \rightarrow \mathbf{\Theta}')$, where $T_a(\mathbf{\Theta} \rightarrow \mathbf{\Theta}') = \min(1, \frac{W(\mathbf{\Theta}')}{W(\mathbf{\Theta})})$

4. Go back to step (2)

For the previous thermal system example:

$$\frac{W(\mathbf{\Theta}')}{W(\mathbf{\Theta})} = e^{-\beta \triangle H}, \tag{3.17}$$

where $\triangle H$ is the energy difference between the proposed new state and the old state. This random walk approach enables us to get the required distribution regardless of the complexity of the Hamiltonian. In an equilibrium state, the knowledge of un-normalized weight function is enough to calculate the expectation value of corresponding observables, since only ratios of probabilities, $\frac{Q(\mathbf{\Theta}')}{Q(\mathbf{\Theta})} = \frac{W(\mathbf{\Theta}')}{W(\mathbf{\Theta})}$, enter into the Metropolis acceptance probability T_a.

3.2 Ensemble of Models

In trying to model the genetic network such as shown in Figure 2.2, we are faced with a fundamental and ubiquitous difficulty of systems biology: essentially all the relevant model parameters (including, *e.g.*, molecular species initial concentrations and reaction rate coefficients) are unknown and there are a large number of such unknown parameters while the available experimental data are sparse and noisy. Even a relatively simple genetic network can require many unknown model parameters. For example, 47 parameters (including 16 initial species concentrations, 26 rate coefficients and 5 unknown concentration unit conversion factors) are required to model the above simple genetic network for the biological clock in Figure 2.2. The unknown parameters are typically poorly constrained only by a sparse set of noisy profiling data, available only for a limited number of molecular species (*e.g.* 183 data points for altogether 5 different species in this biological clock system). To obtain a meaningful comparison of the model to the available data, we have employed a novel ensemble method of circuit identification which was developed for the context of sparse, noisy time-dependent profiling data without requiring, *e.g.*, any stationary state assumption concerning the reactants and products in the genetic network. Instead of trying to identify one unique model parameter set, our goal in this ensemble method is to generate a large, random

sample of models, *i.e.*, an ensemble of models, consistent with the available RNA and protein profiling data, implemented as a Monte Carlo (MC) simulation technique. In the ensemble method, a random walk is initiated in the 47-dimensional space of model parameters, and a likelihood function Q is used to guide the walk into a parameter region of near-maximum Q-values. The Q-value in this context is the likelihood that the genetic network model in Figure 2.2 could have given rise to the observed profiling data, calculated as a function of the model parameters (*i.e.*, the rate coefficients, initial concentration values of all species, and unit conversion factors of observed species in the genetic network). We now give a more detailed, formal description of the ensemble approach.

Let the M-dimensional vector $\mathbf{\Theta} := (\Theta_1, \cdots, \Theta_M)$ denote the unknown parameters, comprising the natural logarithms of the rate coefficients, of the initial species concentrations and of any unknown unit conversion factors in the model. All species concentrations are measured and given here in a common, but unknown "model unit" of concentration and all rate coefficients in unites of $1/(hour \times cu^{k-1})$ for reactions of k^{th}-order (*i.e.*, having k reactants). The ensemble of models is then formally described in terms of a probability distribution, the likelihood function $Q(\mathbf{\Theta})$, on the "model space" of all "model vectors" $\mathbf{\Theta}$. To construct such a $Q(\mathbf{\Theta})$, suppose that in a series of M_e experiments, labeled by $e = 1, \cdots, M_e$ in each experiment the concentrations $[s]$ of certain species s, labeled by $s = 1, \cdots, M_s$ (Note: M_s may just be a subset of the total species in this genetic network), are measured at time points t, labeled by $t = 1, \cdots, M_t$, let $Y_l := \ln([s]^x_{t,\Theta})$ for each data point labeled by $l := (t, s, e)$. Here, the superscript (x) in $[s]^x_{t,\Theta}$ denotes that concentration is measured in some experimental units of concentration, such as photon or radioactive decay count units or ratio of induction units. Next, let $\mathbf{Y} := (Y_1, \cdots, Y_D)$ denote the D-dimensional vector of all those Y_l, where $D = M_s M_T M_E$. Likewise, let $\mathbf{F}(\mathbf{\Theta}) := (F_1(\mathbf{\Theta}), \cdots, F_D(\mathbf{\Theta}))$ denote the corresponding predicted values for these observables $\mathbf{Y}$ for a given model $\mathbf{\Theta}$. For the above described set of observables $\mathbf{Y}$, the predicted values $F_l(\mathbf{\Theta})$ are calculated from $\mathbf{\Theta}$ by numerically solving the network's system of rate equations with the rate coefficients and

initial conditions comprised by $\mathbf{\Theta}$ and then calculating from that solution the predicted ln-concentration $F_l(\mathbf{\Theta}) := \ln(\phi_{s,t,e}[s]_{t,e})$ for each observed species s at each observation time point t in each experiment e. Here, $[s]_{t,e}$ denotes the predicted species concentrations, given in the model unit "cu", and $\phi_{s,t,e}$ denotes the unknown unit conversion factor from the model unit to the various detector units used to represent the experimental data. Subsets of experimental data points (s,t,e) which have been measured under identical conditions in the same experiment with the same experimental detection method share the same $\phi_{s,t,e}$-parameters are required.

It is reasonable to assume (but not fundamental to our ensemble method!) that the probability distribution $P(\mathbf{Y}|\mu)$ of the data $\mathbf{Y}$, given their corresponding mean values $\mu = \mu_1, \cdots, \mu_D$, is representable as a multivariate Gaussian, without error correlations between different data points Y_l. Hence, we will use in the following:

$$P(\mathbf{Y}|\mu) = \text{const} \times e^{-\chi^2/2}$$

with

$$\chi^2(\mathbf{Y};\mu) := \sum_{l=1}^{D} (Y_l - \mu_l)^2/\sigma_l^2$$

where μ_l and σ_l denote the mean and standard deviation of the observable Y_l. σ_l is an estimated value for all log-concentration data points Y_l. If multiple realizations of each profiling experiment are performed, then the full variance-covariance matrix for the experimental data can be estimated and used in the previous equation.

A given $P(\mathbf{Y}|\mu)$ does of course not uniquely determine the model ensemble. There is an infinite manifold of $Q(\mathbf{\Theta})$ which is consistent with the data distribution $P(\mathbf{Y}|\mu)$, and we have to make "reasonable" choices. The simplest choice which we have adopted here is to take $P(\mathbf{Y}|\mu)$ as the likelihood (in which the experimental data $\mathbf{Y}$ are viewed as fixed) to determine the ensemble $Q(\mathbf{\Theta})$. Thus the parameters $\mathbf{\Theta}$ are distributed according to the following likelihood function:

$$Q(\mathbf{\Theta}) = P(\mathbf{Y};\mathbf{F}(\mathbf{\Theta})) := \Omega^{-1} W(\mathbf{\Theta}) := \Omega^{-1} e^{-H(\mathbf{\Theta})}$$

where $\Omega := \sum_{\mathbf{\Theta}}$ is the normalization factor. Here the *Hamiltonian* or *energy function* is introduced to emphasize the analogy to the Boltzmann factor as in the previous thermal system example, $H(\mathbf{\Theta}) := -\ln W(\mathbf{\Theta})$.

In standard data-fitting methods, such as maximum likelihood, least-squared fitting and maximum entropy approaches, one would attempt to construct the correct model by finding a unique $\mathbf{\Theta}$ which minimize $Q(\mathbf{\Theta})$. Due to the large number of unknown model parameters and sparsity and noise of the experimental data, such approaches are bound to fail in the present context. The basic philosophy here is that one should not attempt to find a unique $\mathbf{\Theta}$, unless it is warranted by the quantity and quality of the underlying data. Rather, one should admit all $\mathbf{\Theta}$ as possible candidates for the correct model with a probability distribution which reasonably reflects a $\mathbf{\Theta}$'s degree of consistency with the data.

For any ensemble of the general form $Q(\mathbf{\Theta}) := \Omega^{-1} W(\mathbf{\Theta})$ with an analytically known or numerically calculable weight function $W(\mathbf{\Theta})$, we can evaluate the ensemble average of any quantity $G(\mathbf{\Theta})$,

$$\langle G(.) \rangle := \sum_{\mathbf{\Theta}} G(\mathbf{\Theta}) Q(\mathbf{\Theta}) = [\sum_{\mathbf{\Theta}} G(\mathbf{\Theta}) W(\mathbf{\Theta})] / [\sum_{\mathbf{\Theta}} W(\mathbf{\Theta})]$$

Of course we actually cannot explore all the possible $\mathbf{\Theta}$ space. For the Monte Carlo ensemble method,

$$\langle G(.) \rangle_{MC} := \frac{1}{I} \sum_{i=1}^{I} G(\mathbf{\Theta}^{(i)})$$

is actually used to calculate ensemble mean from Monte Carlo sample $\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(I)}$. The standard deviation is calculated based on the same Monte Carlo sample, where $G(.)$ could be squared of any unknown parameter.

The *Metropolis Algorithm* described in the previous chapter is perfect for this problem. All the crucial components needed in *Metropolis Algorithm* now have been defined here, to generate random samples of $\mathbf{\Theta}$ according to $Q(\mathbf{\Theta})$.

Chapter 4

Monte Carlo (MC) Studies for the Biological Clock of *Neurospora crassa*

4.1 Introduction

Biological rhythmicity and the clock mechanisms that drive biological rhythms are fundamental properties of many groups of cellular life, ranging from prokaryotes to humans. Circadian clocks function to control daily rhythms in cellular activities and behavior. A detailed understanding of the molecular and biochemical basis for circadian rhythmicity is essential to human physiology, including endocrine function, sleep/wake cycles, psychiatric illness, as well as drug tolerances and effectiveness [23] [24]. Simple eukaryotes provide appropriate experimental systems to investigate the clock because clock mechanisms are evolutionarily conserved, such as *Neurospora crassa*, a well-defined model organism with one of the most highly characterized clocks. The biological clock [25] provides a prototypical and biologically ubiquitous example of how a complex trait can emerge from the interaction of even a small number of gene regulatory elements.

In the lowly bread mould, *Neurospora crassa*, biomolecular reactions involving the *white-collar-1*(*wc-1*), *white-collar-2*(*wc-2*) and *frequency*(*frq*) genes and their products constitute building blocks of a biological clock [26]. A central, open question of systems biology is whether these building blocks are necessary and sufficient to define a genetic network that oscillates and how, in quantitative detail, such oscillations emerge from the interactions between these building blocks. A novel method of genetic network identification [27] is used to find an ensemble of oscillating network models, constituted from *wc-1*, *wc-2* and *frq* and their products, and which is quantitatively consistent with available RNA and protein profiling data on the *Neurospora crassa* biological clock. The use of genetic networks to integrate

diverse experimental information and to predict the behavior of a complex trait, such as the biological clock, provides a new paradigm for quantitative genetics at the molecular level [28].

Key features of the genetic network that permit oscillations are:

1. The presence of functional *wc-1*, *wc-2*, and *frq* genes, generating protein products *WC-1*, *WC-2* and *FRQ*, and the *white collar complex* (*WCC*) formed by *WC-1* and *WC-2*;

2. A closed feedback loop of the biomolecular reactions in the genetic network with: *WCC* activating the *frq* gene $\rightarrow$ the activated *frq* gene producing *frq* mRNA $\rightarrow$ *frq* mRNA producing *FRQ* protein, and $\rightarrow$ *FRQ* deactivating *WCC*;

3. Dynamical frustration arising in the feedback loop due to *WCC*'s stimulating the production of *FRQ* while *FRQ* induces the deactivation of *WCC*;

4. A minimal level of cooperativity in the activity of *WCC* in activating the *frq* gene and/or in the activity of the frequency protein *FRQ* in deactivating *WCC*.

4.2 Genetic Network Model for the Biological Clock

A genetic network for the biological clock, consisting of 25 reactions and 16 participating biomolecular species, is shown in Figure 2.2. The experimental basis for each reaction in the network will now be described. There is strong evidence that the proteins *WC-1* and *WC-2* in this network form a complex (*WCC*) which acts as a transcription factor for the *frq* and *clock-controlled genes*, *ccg* [29][30]. In turn, the oscillator protein *FRQ* provides negative feedback by interacting with *WC-1* conditionally on *WC-2* [31] and positive feedback through the post-transcriptional control of *WC-1* synthesis [26]. The band (*bd*) gene is hypothesized to be one of these *ccg* genes in the circuit [29][32].

One alternative light-dependent biological clock model will also be introduced in the next sections, which includes 3 more species: *Phot*, *frq*1*L* and *WCCL*; and 10 more reactions. If I

say clock model, I actually mean the light-independent biological clock model as in Figure 2.2, unless otherwise specified.

The dynamical behavior of this network is then described in terms of kinetic rate equations [33], with assumed, standard mass action kinetics, forming a system of coupled ordinary differential equations (ODEs). A unique solution of these coupled ODEs which can be directly compared to experimental time-dependent profiling data requires as input a knowledge of the initial (starting) concentrations of all molecular species and of the rate coefficients of all reactions (such as those given in column (A) of Table 4.1, from an ensemble fit to the experimental data) which describe gene activation, transcription, protein synthesis, complex formation, and mRNA and protein decay. Some of these reactions (*i.e.* A, $\bar{A}$, C_1, P and A_c) involve the participation of the clock proteins *FRQ* and *WCC*. As uncovered by a mathematical analysis of this rate equation model (see Stability Analysis in the Appendix A), the genetic network in 2.2 can display a diversity of dynamical behaviors, including regular circadian oscillations and damped oscillatory transients to a stable stationary state.

4.3 Experimental Methods

The *wc-1* RNA data are digitized from Figure. 1 [26], and the *frq* RNA data are digitized from Figure. 1C [31]. The *WC-1* protein data are digitized from Figure. 1 [26], and the *FRQ* protein, from Figure. 1C [31]. The physiological *bd* data are from Figure. 2 [29]. The *CCG* data subject to different light entrainment experiments are taken from [58].

4.4 *Metropolis Algorithm* for Biological Clock

Here is the description how the *Metropolis Algorithm* is used for identification of a genetic network for the biological clock.

In our actual simulation runs, we did not update all $\mathbf{\Theta}$-components, according to the foregoing procedure. Rather, we chose the ln of unknowns, independent unit conversion factors $\ln(\phi_{s,t,e})$ so as to maximize $Q(\mathbf{\Theta})$, given the $M' = 42$ remaining (non-unit-conversion-factor)

$\mathbf{\Theta}$-components. Only the remaining $\mathbf{\Theta}$-components were subjected to the random Metropolis updating steps described above, using the so-maximized Q($\mathbf{\Theta}$) as the terminal distribution. Due to the Gaussian dependence of the original $Q(\mathbf{\Theta})$ on the ln unit conversion factors, $\ln(\phi_{s,t,e})$, this "reduced" MC procedure is mathematically equivalent to the "full" MC procedure of subjecting all M $\mathbf{\Theta}$-components, including all $\ln(\phi_{s,t,e})$, to random Metropolis updates. The corresponding "reduced" values of χ^2, minimized with respect to the independent $\ln(\phi_{s,t,e})$, are what is shown in Figures 4.1 and 4.3. $\sigma_l \cong 0.14$ is assumed for all ln-concentration data points Y_l.

For the model in Figure 2.2 with $n = m = 4$, we first chose some set of rate coefficients and initial concentrations to give us a (weakly damped or undamped) oscillatory solution. We then re-scaled the rate coefficients and initial concentrations and shifted the initial time value so that the period, maximal amplitude and phase of the oscillation for the $[CCG]$ protein species in the model roughly matched those of the experimental $[CCG]$ data. The resulting model parameter vector served as the initial $\mathbf{\Theta}$ in our MC equilibration run for the $n = m = 4$ model. For MC equilibration runs with other cooperativity exponents n and m, a $\mathbf{\Theta}$ from a fully equilibrated $n = m = 4$ run served as the MC initial. We used a $1 : 1$ random mixture of local and global updating moves, with the maximum proposed step width automatically adjusted after every 20^{th} sweep (where 1 sweep = M' Metropolis updating steps) so as to keep the average Metropolis acceptance probability in both local and global updating steps around 50%, *e.g.*, between 0.34 and 0.66 for results reported in Figures 4.2 and 4.3. After about $4 \sim 6 \times 10^4$ equilibration MC sweeps, about 4×10^4 accumulation MC sweeps were performed and the components of the resulting $\mathbf{\Theta}$ at the end of each accumulation sweep were included into our MC random sample.

In Figure 4.1 the progress of such a MC random walk towards its "equilibrium state" is shown. This "equilibrium" is reached when the probability for a given parameter set to be visited equals the likelihood, Q, and, consequently, when the walk mainly explores regions of near-maximal Q-values or, equivalently, near-minimal values of χ^2. The "model ensemble" is

then the collection of models "$\mathbf{\Theta}$" which are visited after the random walk has settled into its equilibrium state.

In conventional maximum-likelihood methods, one seeks to identify a unique model $\mathbf{\Theta}^{(opt)}$ by maximization of some likelihood function $Q(\mathbf{\Theta})$. This is then sometimes complemented by a sensitivity analysis, based on the local behavior of $Q(\mathbf{\Theta})$ in close proximity to $\mathbf{\Theta}^{(opt)}$, or based on an *ad hoc*, *brute force* exploration of a few wider, but dimensionally limited parameter regions. Such an approach is justified if experimental data are abundant, available for essentially all molecular species, and low in noise, resulting in a $Q(\mathbf{\Theta})$ sharply peaked at $\mathbf{\Theta}^{(opt)}$. By contrast, in our current situation, experimental data are sparse, noisy and available for only a few of the many potentially relevant molecular species. As a consequence, there may then exist vast expanses of $\mathbf{\Theta}$-space where $Q(\mathbf{\Theta})$ is maximal, or nearly so, and any unique, "optimal" choice of (if one exists, by whatever choice of likelihood!) may seriously misrepresent the information actually contained in the data. The crucial advantage of the ensemble method is that it systematically explores those expanses of $\mathbf{\Theta}$-space. In doing so, it allows us to get a more complete and systematic understanding of what *can* be known, inferred or predicted on the basis of the existing data and, of equal importance, what is *not* known and *can not* be predicted. Thus, the method allows us to make some quite definitive, experimentally testable model predictions for some model parameters and some observable properties, even though many other parameters and properties may be very poorly constrained. Furthermore, the presently most poorly constrained properties are those whose future measurement will provide the most stringent additional constraints. Hence the ensemble can systematically guide the design of maximally informative "new" experiments, based on the available "old" data.

4.5 Results for light-independent biological clock model

4.5.1 Comparison to profiling experiments

Predictions by the model ensemble using the ensemble averages +/− 2 ensemble standard deviations are shown in Figure. 4.2 and are quite in accord with the experimental data. The conclusion is that the genetic network in Figure 2.2 is sufficient to explain published profiling data on the biological clock. The ensemble means and standard deviations for the 26 rate coefficients of the model are given in column (A) of Table 4.1. While a plethora of models have been proposed to explain biological rhythms [34], there needs to be a tighter linkage between theory and experiment, as noted by [35]. In Figure 4.2 we present detailed experimental support for the model in Figure 2.2.

The model ensemble makes a number of predictions consistent with experimental observations. The *WC-1* protein is predicted to lag the *FRQ* protein with close to an 8-hour phase difference in Figure 4.2C, consistent with the experimentally observed 8-hour phase difference [26]. The *FRQ* protein and *frq* RNA are predicted to oscillate with a (4-6)-hour phase difference in Figure 4.2B and Figure 4.2C, as observed [32]. The de-repression of *FRQ* takes 14-19 hours of the circadian cycle, as observed in Figure 4.2C [36]. The range of ln[$wc\text{-}1^{r1}$] oscillations implies a less than 2-fold induction of the $wc\text{-}1^{r1}$ mRNA (compared to over 12-fold induction of *frq* mRNA) during the cycle in the model in Figure 4.2B; indeed, only limited oscillations in wc-1 mRNA are observed, if any [26]. The level of *WC-2* (presumptively in the nucleus) is predicted to be in great excess of other proteins, as observed [31]. Finally, the rate of translational synthesis of *FRQ* (L_3 in Figure 2.2) is relatively rapid (with translation coefficient on the order of $L_3 \sim 4/hour$ in column (A) of Table 4.1) compared to the post-translational degradation of *WCC* mediated by *FRQ* in the decay reaction p (with an average cycle-maximum decay coefficient on the order of $P \times [FRQ]^m \sim 1.2/hour$) [31]. The model also trivially concurs with experiments in that knocking out either the *wc-1*, *wc-2*, or *frq* genes is predicted to eliminate oscillations, as observed [29]. Our estimated value

of 5 hours is consistent with the *FRQ* protein life-time of $\approx 4-7$ hours, obtained directly from the *FRQ*-decay data of Liu et al [37] , independent of our model ensemble.

4.5.2 Minimal requirements for a ticking clock

Having identified a circuit in Figure 2.2 that is simple and explains observations on the biological clock in the absence of external environmental stimuli, it is natural to ask what features of the clock are essential for oscillations. A key feature to obtain oscillations in the genetic network has been the introduction of cooperative kinetics in the activation of *frq* (A) and/or the deactivation of *WCC* (P), with cooperativity exponents or *Hill coefficients* n and m, respectively, as defined in Figure 2.2. From the mathematical stability analysis in the Appendix A (for a slightly simplified version of the model, with *WC-2* set to a time-independent constant), it can be seen that some minimal amount of cooperativity, namely $nm > 4$, is required for the model to exhibit undamped oscillations regardless of initial conditions. There is some evidence that *FRQ* acts as a dimer [38]. Four model ensembles were identified with varying *Hill coefficients*, with n=m and n=4, 3, 2, or 1 [33]. From Figure 4.3A, it can be seen that the ensemble without cooperativity ($n = m = 1$) has χ^2-values substantially larger than those of the three ensembles with cooperativity. The χ^2-values of the remaining three model ensembles, with cooperativity, significantly overlap, and the best fits are achieved with *Hill coefficients* of $n = m = 3$, substantially less than postulated in some previous models [39] and in correspondence to the most robust version of a simplified stochastic model with the same *Hill Coefficient* $n = 3$ [40]. On the basis of the limited-duration data available, we cannot at present discriminate between truly oscillatory (undamped) models and weakly damped oscillatory models, such as the $n = m = 2$ model shown in Figure 4.3A. The exact mechanism by which the *FRQ* protein deactivates *WCC* complex is unknown. Smolen et al. [41] propose that the *FRQ* protein simply sequesters *WCC* complex in contrast to the model in Figure 2.2 in which *FRQ* degrades *WC-1* in the P reaction. Is it necessary that the P reaction be a degradation reaction? To answer this ques-

tion the ensemble method was used to reconstruct the likelihood functions under 4 distinct hypotheses about *WCC* deactivation with 3 slight modifications in the P reaction, defined in Figure 4.3B. Two of these hypotheses are variants on Smolen's sequestering hypothesis supplemented with cooperativity; and another simply assumes that *FRQ* catalytically triggers *WCC* complex falling apart into its constituents, *WC-1* and *WC-2*. As can be seen in Figure 4.3B, these alternative deactivation mechanisms are reasonably competitive with the proposed degradation mechanism in Figure 2.2. At this point in time the data do not strongly support a particular deactivation mechanism, although the original deactivation mechanism in Figure 2.2 appears to outperform the 3 other mechanisms proposed in Figure 4.3B. Reaction networks with both positive and negative feedback elements have been proposed to explain the dynamics of the biological clock [42][43]. The network in Figure 2.2 has both kinds of elements. One positive feedback element in the clock appears to be the post-transcriptional control of *WC-1* synthesis [26] by *FRQ* in reaction C_1 of Figure 2.2. In Figure 4.3A, we also show the results for a model without this positive feedback of *FRQ* on *WC-1* synthesis, *i.e.*, assuming a modified C_1 reaction, $wc\text{-}1^{r0} \rightarrow wc\text{-}1^{r1}$, without participation of *FRQ*. As can be seen, the two ensembles overlap substantially in their likelihood χ^2 values and they do not differ significantly with regard to fit to available profiling data on the *Neurospora crassa* clock. The *FRQ* positive feedback in reaction C_1 is evidently not an essential element of the network topology for the biological clock to function, as concluded elsewhere [41][38].

So, why is there post-transcriptional control of *WC-1* by *FRQ* [26][38][44]? A possible answer is suggested by the predicted life-time of the translationally active $wc\text{-}1^{r1}$ mRNA, $\langle D_7 \rangle^{-1} \approx 20$ *hours* (from Table 4.1), which is comparable to a full ($\sim$ 24-hour) oscillation period and about 30 times longer than the predicted life-time of the inactive $wc\text{-}1^{r0}$ species, $\langle D_1 \rangle^{-1} \approx 0.65$ *hour*. We thus hypothesize that the primary function of the C_1 reaction is not to control *WC-1* production, but simply to enable it by conferring sufficient longevity to the *wc-1* mRNA. Without this mRNA stabilization, the clock system would be relegated

to a non-oscillatory region of its parameter space, *i.e.*, the *wc-1* mRNA would decay too fast for the clock to tick. One might ask how a fit to the data can be achieved without the positive feedback by *FRQ* on *wc-1* mRNA, as shown by the light-blue curve in Figure 4.3A. The answer is that, in our circuit without posttranscriptional regulation of *wc-1* mRNA, the modified C_1 reaction (without *FRQ* participation) still serves to lengthen the life-time of the *wc-1* mRNA, *i.e.*, $\langle D_7 \rangle$ is still substantially reduced relative to $\langle D_1 \rangle$. The problem is that, without *FRQ* participation, we are lacking a biochemical explanation for the life-time increase. The existence of the *FRQ*-induced mRNA stabilization and its detailed biochemical mechanism needs to be explored further experimentally. The fact that, compared to its *FRQ* regulator (Figure 4.2C), $wc\text{-}1^{r1}$ has a much weaker oscillation amplitude is an immediate consequence and already a direct experimental confirmation of the *FRQ*-induced mRNA stabilization. For life-times comparable to the oscillation period, $[wc\text{-}1^{r1}]$ tends to "average out" the oscillations in its *FRQ*-controlled production rate. If the $wc\text{-}1^{r1}$ mRNA life-time were much shorter than the oscillation period it would be difficult to reconcile the two experimental observations that, on the one hand, *FRQ* is a critical translational activator; yet, on the other hand, the resulting activation, as measured by $[wc\text{-}1^{r1}]$, oscillates much more weakly than the activator itself.

A central feature of the genetic network representing the biological clock in Figure 2.2 is its closed, dynamically frustrated feedback loop [45] where, on the one hand, *WCC* activates the *frq* gene and, on the other hand, the *FRQ* protein deactivates *WCC*. Visualization of the model ensemble provides insights into how clock oscillations emerge in the genetic network in a way consistent with the data in Figure 4.2. Key parameters of this feedback loop in Figure 2.2 are the rates of activation (A) and deactivation ($\bar{A}$ of *frq* by *WCC* and the rate of deactivation of *WCC* by *FRQ* (P). In the mathematical stability analysis given in the Appendix A, a function R has been identified which partitions the 47-dimensional parameter space into one domain where only undamped oscillations occur (red) and another where damped oscillations are possible (blue) as shown in Figure 4.4.

In Figure 4.4, the $n = m = 4$ model ensemble is projected into the $(A', P', \bar{A}')$ volume, where the rate constants $(A, P, \bar{A})$ have been re-scaled to be dimensionless quantities. As can be seen, these re-scaled ensemble rates represent a cylinder (red) containing about 82% of the ensemble that satisfy the Routh-Hurwitz Criterion for instability ($R > 0$ at all fix points, see below) which is necessary and sufficient for the model to exhibit only sustained oscillations. The remaining 18% of the ensemble (blue), which are scattered, do not have sustained oscillations. If this subset of damped oscillators (blue) is trimmed from the ensemble, the remaining models (red) form a tight droplet of re-scaled model parameters. Similar plots in which the z-axis $\bar{A}$ is replaced with the other re-scaled rate coefficients that control transcription and translation of *frq* in the closed feedback loop (S_4, L_3) are not as constrained and can take a broader spectrum of values on the vertical axis (as shown in Figure 4.5). The values of the re-scaled rates $(A', P', \bar{A}')$ of activation and inactivation of *frq* and decay of *WCC* are thus key quantitative elements for sustained oscillations. The plot in Figure 4.4 emphasizes that the data in Figure 4.2 are consistent with a genetic network with sustained oscillations, but do not eliminate some genetic networks with damped oscillations (in blue).

4.5.3 Robustness of the biological clock

Perturbation of the ensemble shown in Figure 4.4 also allows examination of the robustness of the biological clock. One of the key predictions of the model ensemble is that the life-time $(1/D_7)$ of the translationally active *wc-1* mRNA is long ($\sim$ 20 hours). To examine robustness of the model, we varied this key parameter for each ensemble member from 14 hours in the reduced model with *WC-2* constant ($\langle D_7 \rangle = 0.05/hour$ in column (A) of Table 4.1) down to 1.0 hour (or $D_7 = 1.0/hour$), keeping all other rate coefficients fixed at their ensemble-generated values. As indicated in Table 4.2 by the percentage of stable oscillators in the so-perturbed ensemble, the cyclical dynamics is robust against an about 15-fold decrease of life-time, but then the ensemble precipitously becomes arrhythmic at a life-time of $\sim$ 1.33 hours (or $D_7 = 0.75/hour$) or shorter. The actual distribution of

D_7-values in the unperturbed ensemble of Figure 4.4 imposes a much tighter constraint of $D_7 < 0.20/hour$, *i.e.*, the experimental data, through the ensemble likelihood, only support a life-time of translationally active *wc-1* mRNA longer than 5 hours.

4.5.4 Stability analysis of genetic network

The model in Figure 2.2 can be translated into a system of 16 differential equations describing the rate of change of each of the 16 species in the genetic network as a function of time t. The 16 species concentrations [wc-1^1], [wc-1^{r0}], [wc-1^{r1}], [WC-1], [wc-2^1], [wc-2^r], [WC-2], [WCC], [frq^0], [frq^1], [frq^{r1}], [FRQ], [ccg^0], [ccg^1], [ccg^{r1}], and [CCG] are abbreviated here to $u_1, u_{r0}, u_{r1}, u_p, v_1, v_r, v_p, w, f_0, f_1, f_r, f_p, g_0, g_1, g_r$ and g_p, respectively, with constant total *frq*-gene concentration $f_G := f_0 + f_1$. The reaction labels in Figure 2.2 double as the rate coefficients in the reaction network. This 16-dimensional model can be reduced to a 7-dimensional one by several simplifications. The clock-controlled gene and its products (g_0, g_1, g_r and g_p) can be dropped from the rate equations because their dynamics are driven entirely by the clock genes (*wc-1, wc-2* and *frq*) and their products, and the *ccg* products do not feed back on the clock genes in Figure 2.2. The *WC-2* protein is in 5-fold molar excess over *FRQ* and *WC-1* in the nucleus [31], and hence *wc-2* and its products (v_1, v_r, v_p) can be treated approximately as constants. The total amount of each gene, *e.g.*, $f_0 + f_1 =: f_G$ is constant, allowing us to eliminate f_0. Likewise, the concentration of the unregulated *wc-1* gene, u_1, is a constant [26]. These simplifications leads to a reduced model with a "dynamical vector" obeying the following 7 rate equations, of the general form $\mathbf{y} := (f_1, f_r, f_p, w, u_p, u_{r1}, u_{r0})$, with the 7 components of the "reaction rate vector" given by the right-hand sides of the rate equations: $\dot{\mathbf{y}} = \mathbf{\Gamma}(\mathbf{y})$, with the 7 components of the "reaction rate vector" given by the right-hand sides of the rate equations:

$$\begin{aligned} \dot{f}_1 &= A(f_G - f_1)w^n - \bar{A}f_1 \\ \dot{f}_r &= S_3(f_G - f_1) - S_4 f_1 - D_3 f_r \end{aligned}$$

$$
\begin{aligned}
\dot{f}_p &= L_3 f_r - D_6 f_p \\
\dot{w} &= E_2 u_p - D_8 w - nA(f_G - f_1)w^n + n\bar{A} f_1 - P w f_p^m \\
\dot{u}_p &= L_1 u_{r1} - D_4 u_p - E_2 u_p \\
\dot{u}_{r1} &= C_1 u_{r0} f_p - D_7 u_{r1} \\
\dot{u}_{r0} &= V_1 - D_1 u_{r0} - C_1 u_{r0} f_p
\end{aligned}
$$

Here, *e.g.*, $\dot{w} := dw/dt$, denotes the time derivative of $w(t)$; $E_2 := C_2 v_p =$ constant, and $V_1 := S_1 u_1 =$ constant. The *Hill coefficients* n and m are, respectively, the number of *WCC* molecules needed to cooperatively activate *frq* and *ccg*; and the number of *FRQ* molecules needed to degrade cooperatively *WCC*.

To explore the long-time dynamics of our biological clock model, we analyze its stationary states or "fixed points" (FP), denoted by $\mathbf{y}^*$, where all species' time derivatives would vanish, *i.e.*, the solution(s) of the 7 coupled equations $\mathbf{\Gamma}(\mathbf{y}^*) = 0$. By quasi-static approximation, all species can be eliminated except f_p and w. The FPs are given by the following 2 functions:

$$
\begin{aligned}
w &= \frac{C_2 L_1 S_1 u_{r0} f_p v_p}{D_7(D_8 + P f_P^m)(D_1 + C_1 f_p)(D_4 + C_2 v_p)} \\
f_p &= \frac{L_3(S_3 \bar{A} f_G + S_4 P w^n)}{D_6 D_3(\bar{A} + A w^n)}
\end{aligned}
$$

By linearizing the rate equations near the FP, we can find out whether or not the FP is stable (*i.e.*, for slight departures the system returns to the FP) [46][47]. If all FPs of the model are unstable, then a variety of non-trivial dynamical behaviors are possible, including oscillations. So, a necessary and sufficient condition for the model to exhibit only sustained oscillations, regardless of initial conditions, is that all its FPs be unstable [46]. Stability or instability of a FP is governed by the "stability matrix" $\mathbf{J}$, the Jacobian of $\mathbf{\Gamma}(\mathbf{y})$ with matrix elements $J_{ij} := \partial \Gamma_i / \partial y_j$ evaluated at which, for our 7-dimensional model, has the general form:

$$\begin{pmatrix} -\gamma_1 & 0 & 0 & d_1 & 0 & 0 & 0 \\ \bar{b}_1 & -\gamma_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \bar{b}_2 & -\gamma_3 & 0 & 0 & 0 & 0 \\ \bar{d}_1 & 0 & \bar{b}_3 & -\gamma_4 & b_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\gamma_5 & b_5 & 0 \\ 0 & 0 & \bar{d}_3 & 0 & 0 & -\gamma_6 & b_6 \\ 0 & 0 & \bar{e}_3 & 0 & 0 & 0 & -\gamma_7 \end{pmatrix}$$

The non-zero **J**-matrix elements are given by $\gamma_1 = Aw^n + \bar{A}$, $\gamma_2 = D_3$, $\gamma_3 = D_6$, $\gamma_4 = D_8 + n^2 A w^{n-1}(f_G - f_1) + P f_p^m$, $\gamma_5 = D_4 + E_2$, $\gamma_6 = D_7$, $\gamma_7 = D_1 + C_1 f_p$, $\bar{b}_1 = S_4 - S_3$, $\bar{b}_2 = L_3$, $\bar{b}_3 = -mPw f_p^{m-1}$, $\bar{d}_1 = n(Aw^n + \bar{A})$, $\bar{d}_3 = -\bar{e}_3 = C_1 u_{r0}$, $b_4 = E_2$, $b_5 = L_1$, $b_6 = C_1 f_p$ and $d_1 = nAw^{n-1}(f_G - f_1)$, with all concentrations set to their respective FP values (*e.g.* $f_p = f_p^*$). The sparseness and regularity of this matrix **J** is due to the closed feedback loop in the genetic network in 2.2 and mathematically resembles the linearized system of the synthetic oscillator known as the repressilator [45].

A FP is unstable if and only if **J** at least one of the (in general complex) eigenvalues of acquires a positive real part. The eigenvalues of **J**, denoted by λ are the roots of the 7^{th} order characteristic polynomial $\Phi(\lambda) := \det(\mathbf{J} - \lambda \mathbf{E})$ where **E** denotes the unit matrix [48]. By factorization of $\Phi(\lambda)$ into lower-order sub-polynomials and a Routh-Hurwitz analysis [49] of these sub-polynomials, we can prove that an FP is unstable (*i.e.*, a complex λ exists with $\Phi(\lambda) = 0$ and $Re(\lambda) > 0$) if and only if

$$R := A_3^2 + A_1^2 A_4 - A_1 A_2 A_3 > 0$$

where the a_n are coefficients of a 4^{th} order sub-polynomial of $\Phi(\lambda)$ given in terms of the **J**-matrix elements by

$$A_1 = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4,$$

$$
\begin{aligned}
A_2 &= (\gamma_1+\gamma_2)(\gamma_3+\gamma_4)+\gamma_1\gamma_2+\gamma_3\gamma_4-d_1\bar{d}_1,\\
A_3 &= \gamma_1\gamma_2(\gamma_3+\gamma_4)+\gamma_3\gamma_4(\gamma_1+\gamma_2)-d_1\bar{d}_1(\gamma_2+\gamma_3),\\
A_4 &= \gamma_1\gamma_2\gamma_3\gamma_4-d_1\bar{b}_1\bar{b}_2\bar{b}_3-d_1\bar{d}_1\gamma_2\gamma_3.
\end{aligned}
$$

The more straightforward alternative is to calculate the eigenvalues directly and check the sign of the real parts. The results of the two methods are identical.

Figure 4.4 shows a projection of a Monte-Carlo-generated model ensemble into a 3D parameter subspace. Different colors indicate whether the model FPs are all unstable ($R > 0$, in red) or whether at least one FP of the model is stable ($R < 0$, in blue) according to the Routh-Hurwitz analysis. It can also be proved that the foregoing FP instability criterion ($R > 0$) can be satisfied if and only if the level of cooperativity in the model exceeds a threshold given by

$$nm > 4.$$

as in Appendix A.

4.6 Light-dependent biological clock model and results

The previous model describes the circadian clock of *Neurospora crassa* in the absence of environmental cues. But the biological clock is indeed entrainable by exogenous stimuli such as light, temperature and nutrition. Entrainment synchronizes the endogenous oscillations with the exogenous time and allows behavioral flexibility. The blue light photoreceptor, *WC-1*, mediates light input to the circadian system through direct binding (with *WC-2* to form *WCC* by heterodimerizing via PAS domains) to the *frq* promoter [30]. A modified clock model is proposed with inclusion of extra light-dependent species and reactions as shown in Figure 4.6.

Phot is the species representing externally controlled light, which transforms *WCC* and *frq* to *WCCL* and *frqL* respectively. The behavior of the light dependent species are analogical to the original species in dark. This model effectively includes the light signaling pathway.

In Figure 4.7, 3 more profiling data sets of *CCG* are present [58](40 data points for 6+6 hours LD cycle, 31 data points for 9+9 hours LD cycle and 31 data points for 18+18 hours LD cycle). The light intensity is 20 μmol photons/m^2/s. The fitting also includes the original data used for the previous light-independent model. The subtitles of duration indicate the light-dark cycle (*e.g.* 12 hours means 6 hours dark + 6 hours light). The initial $\mathbf{\Theta}$ used in these model identification runs is taken from the original light-independent $n = m = 4$ model parameter vector. The predicted conidiation rhythm by Monte Carlo average fits the experimental data very well, which gives the conclusion that the genetic network given in Figure 4.6 is sufficient to explain these available clock data. The circadian rhythm is synchronized by the external light entrainments, which shows the adaptiveness of the biological system. The endogenous clock anticipates the day cycle for the whole system to optimize its various physiological processes. 12-hour LD cycle gives about 17 cycles, 18-hour LD cycle gives 11 cycles, and 36-hour LD cycle gives 5 cycles. So Figure 4.7 indeed shows the virtual daily life of *Neuropora crassa*.

Based on the prediction of column (C) of Table 4.1, same statement as before can be made to address the necessity of post-transcriptional control of *WC-1* by *FRQ*. $\langle D_7\rangle^{-1} \approx 16.7$ *hour*, while $\langle D_1\rangle^{-1} \approx 0.136$ *hour*. The life time of *wc-1*^{r1} ($\langle D_7\rangle^{-1}$) is longer that of *wc-1*^{r0} ($\langle D_1\rangle^{-1}$) by a magnitude of 100. At the same time, the range of D_7 values is again constrained to be significantly above 5 hours. Also, the estimated value of D_6 is about 0.193 (the *FRQ* protein life time $\sim$ 5 hours), which is again consistent with the experimental observation [37].

When we introduce the light-dependent species and reactions, what happens to the fitting to the original dark data and does inclusion of more light-dependent data indeed improve the prediction? Figure 4.8 shows the comparison to the original dark data. This figure contains 3 columns for different χ^2 calculation. As you may notice that in Table 4.1, the difference between D_1 ($\sim 0.141/hour$) and D_7 ($\sim 0.0717/hour$) for column (B) is not as big as the other two (One of our key predictions is that life time of *wc-1*^{r1} is significantly longer). The

explanation for this is that for column (B) fitting, the predicted behavior of *CCG* protein at initial time is unreasonably high as shown in the inset of Figure 4.8B. When the light data is included, values of D_1 ($\sim 7.33/hour$) and D_7 ($\sim 0.0632/hour$) return to the *appropriate* ratio (115). The fitting to the dark data by the light-dependent clock model still gives accurate prediction for *FRQ* protein, and even better outcome for *CCG* (without the unrealistic early time concentration). The light-dependent model not only interprets the light entrainment experimental data, but also maintains the prediction for the dark data. This also warns us that this high dimensional unknown parameters space could give very different, sometimes biologically wrong prediction as column (B) runs into a wrong parameters space. Then more experimental data and more realistic model guides the exploration back (or closer) to correctness.

4.7 Discussion

The usual modus operandi for quantitative genetics is to narrow progressively the search for quantitative trait loci (QTLs) to explain a complex trait in terms of a position on a chromosome [50]. The ultimate expression of this approach is the Human HapMap [51]. Once there with the QTL in hand from the use of the HapMap, the story ends with the question of what the QTL does. Here we have introduced a different complementary paradigm for explaining a complex trait. Two genetic networks are introduced as a precise hypothesis to explain how genes and their products control the biological clock (as in Figure 4.2 and 4.7). The resulting genetic networks provide quantitative and testable predictions about how biomolecules interact to determine such a complex trait. Analysis of the biological clock in *Neurospora crassa* has yielded predictions as to the necessary and sufficient conditions for sustained endogenous biological rhythms. Recently Elowitz and Leibler demonstrated that a synthetic oscillator called the repressilator could be constructed in *E. coli* [45]. Successful engineering of the repressilator depended critically on using the associated genetic network describing the repressilator. The ultimate test of our genetic network herein for the biological

clock in *Neurospora crassa* will be its successful exploitation to synthesize a biological clock in a strain without a timepiece.

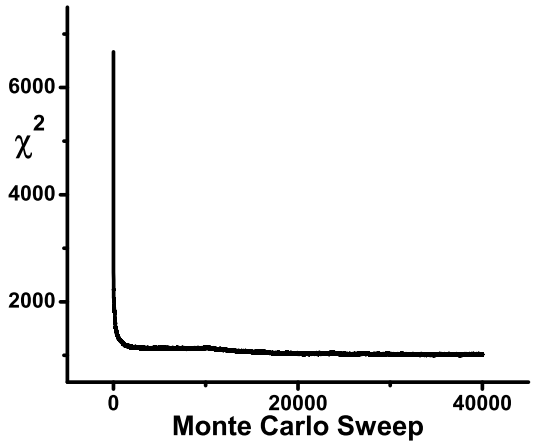


Figure 4.1: Monte Carlo random walk equilibration in the parameter ($\mathbf{\Theta}$) space of the models. Progress towards equilibrium is monitored by $\chi^2 = -2 \ln Q$+const which is a measure of the departure between the data and the model prediction of the genetic network in Figure 2.2, for model cooperativities (*i.e.*, Hill Coefficients) $n = m = 4$. As the fit is refined, χ^2 , on average, decreases with progressive Monte Carlo sweeps in the parameter space. One "sweep" comprises one visit, on average, to all M unknown parameter values, $\Theta_1, ...\Theta_M$ (where $M = 47$ in this model), with a consequent random decision whether or not to change the visited parameter, based on the Metropolis updating rule. This random search method for finding a parameter region near a minimum of χ^2 occasionally takes a risk and permits the χ^2 to increase (worsen the fit), thereby allowing the walk to escape from local minima in the χ^2-surface, as seen in the figure.

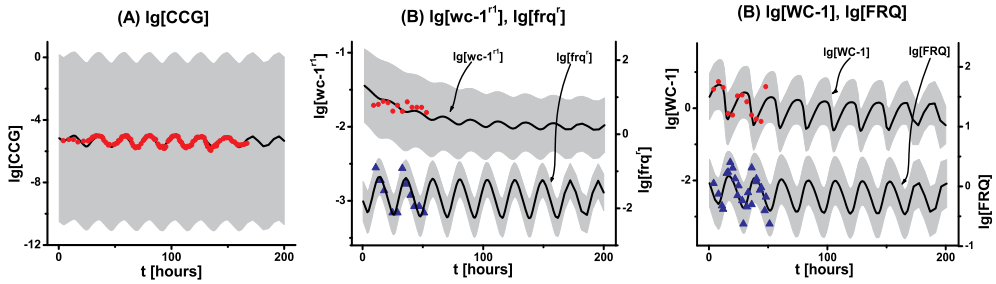


Figure 4.2: Comparison to experiments. A model ensemble for the genetic network in Figure 2.2, with cooperativities $n = m = 4$, predicts various profiling data on the biological clock of *Neurospora crassa*. In each panel, predictions of the model ensemble for the lg of concentration (in model concentration units "*cu*") are shown with $+/-$ 2 ensemble standard deviations (shaded areas) about the ensemble mean (continuous lines). The points are the experimental lg data. (A) levels of conidiation for a *bd* mutant, hypothesized to be a measure of a *CCG* protein, over a 7-day interval; (B) levels of *wc-1*^{r1} and *frq*^{r} RNA over time t; (C) levels of *FRQ* and *WC-1* protein over time t.

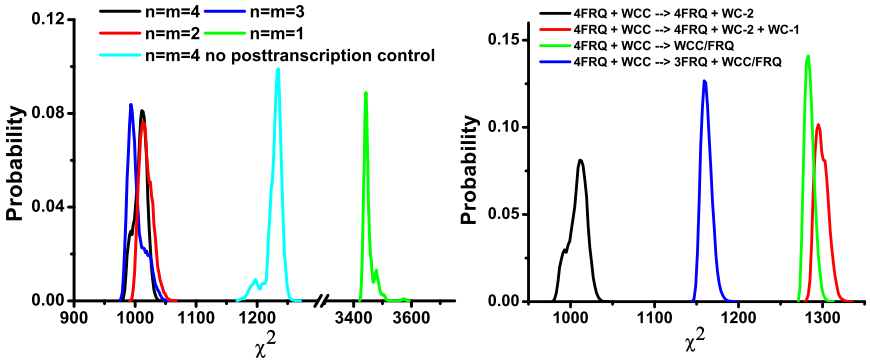


Figure 4.3: Comparison of model fits. Some model ensembles, using modified or simplified versions of the model in Figure 2.2, outperform the original model in terms of χ^2-values. A histogram of values of $\chi^2 = -2 \ln Q + \text{const}$ is shown for varying model ensembles. (A) The number of molecules of *WCC* (n) or *FRQ* (m) working together cooperatively (*i.e.*, the *Hill coefficients*) in reactions A and P are varied with $n = m$. Also the histogram of χ^2-values for a model ensemble without post-transcriptional regulation of *wc-1* by *FRQ* is reported. Some models with less cooperativity in the action of *WCC* or *FRQ* (*e.g.* $n = m = 3$, dark blue curve) or without post-transcriptional regulation (light blue curve) have smaller χ^2-values on average than the $n = m = 4$ model. We have explored in detail models with $n \neq m$ (for $n, m = 2, \cdots, 4$) with additional Monte Carlo runs as in Figure 4.3; it was unnecessary to consider $n = 1$ and $m = 1, \cdots, 4$ or $m = 1$ and $n = 1, \cdots, 4$ because these cases do not satisfy the condition $nm > 4$ necessary for oscillations (see Appendix A). The best model ensemble (histogram most shifted to the left) to date remained $n = m = 3$. (B) The deactivation reaction (P) is varied, allowing by 4 different deactivation mechanisms (including the one in Figure 2.2, in black) to be compared for their goodness of fit. Each deactivation reaction is defined in the Figure legend.

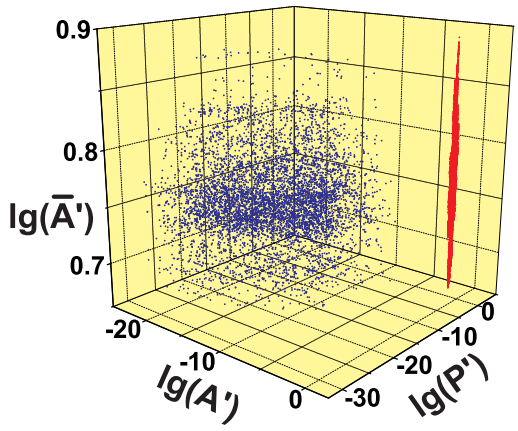


Figure 4.4: 3D projection of the model ensemble. The 3D parameter sub-space is spanned by re-scaled, dimensionless rate coefficients P', A' and $\bar{A}'$. Points in the ensemble allowing only sustained, undamped oscillations are shown in red; points in the ensemble that allow for damped oscillations (from the Stability Analysis in Appendix A) are shown in blue, for the model of Figure 2.2 with $n = m = 4$. To construct P', A' and $\bar{A}'$, define the "maximal" concentrations of *WCC* and *FRQ*, $w_x := L_1 V_1/(D_7 D_8)$ and $f_{px} := f_G S_4 L_3/(D_3 D_6)$, where $V_1 := S_1 u_1 := S_1[wc\text{-}1^1]$ is the rate of $[wc\text{-}1^{r0}]$-production and $f_G := f_0 + f_1 := [frq^0] + [frq^1]$ is the total, constant *frq*-gene concentration. Using the *FRQ* protein life-time $1/D_6$ as the "unit of time", then define dimensionless, re-scaled rate coefficients, such as P', A', $\bar{A}'$, L_3', S_4' and D_8' , respectively, for the P, A, $\bar{A}$, L_3, S_4 and D_8-reactions as follows: $P' := P(f_{px})^m/D_6$, $A' := A(w_x)^n/D_6$, $\bar{A}' := \bar{A}/D_6$, $L_3' := L_3/D_6$, $S_4' := S_4/D_6$ and $D_8' := D_8/D_6$. The re-scaling permits us to collapse higher-dimensional parameter sub-spaces, of dynamically equivalent models, into lower-dimensional ones.

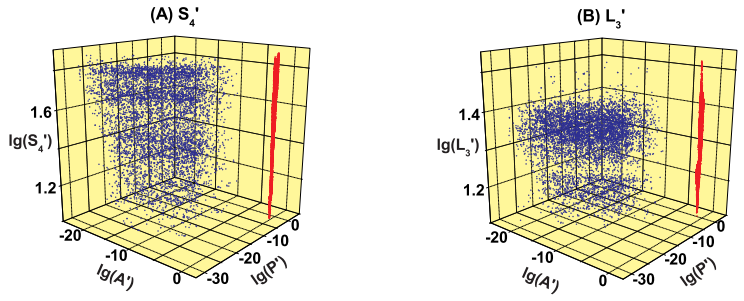


Figure 4.5: 3D projection of the model ensemble. (A) z-axis is replaced by $S_4' := S_4/D_6$. (B) z-axis is replaced by $L_3' := L_3/D_6$.

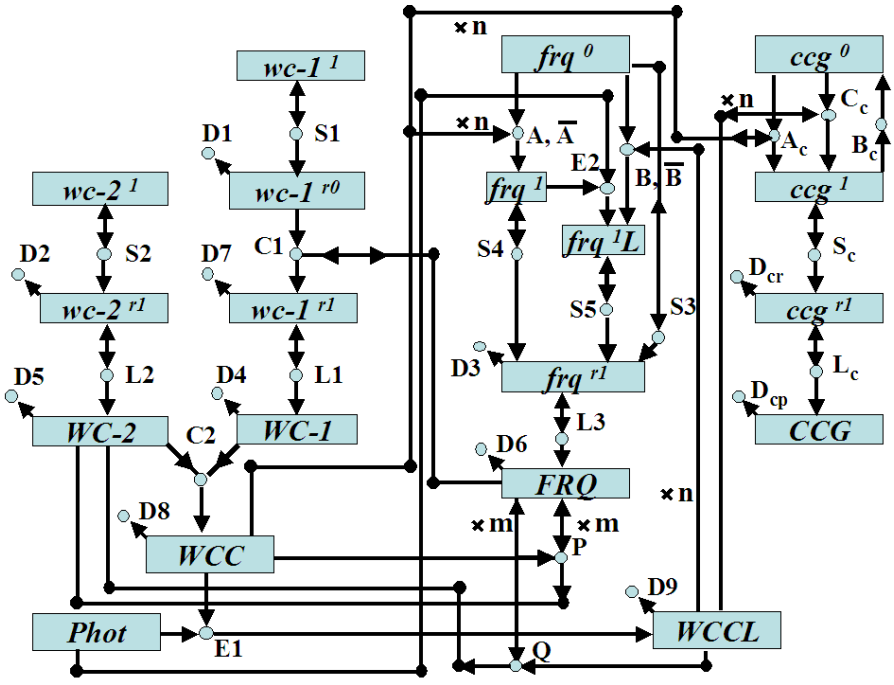


Figure 4.6: Genetic network for biological clock of *Neuropora crassa* with consideration of light. 3 more species are introduced: *Phot*, *WCCL* and frq^1L, and consequently E_1, E_2, B, $\bar{B}$, C_c and Q-reactions are included.

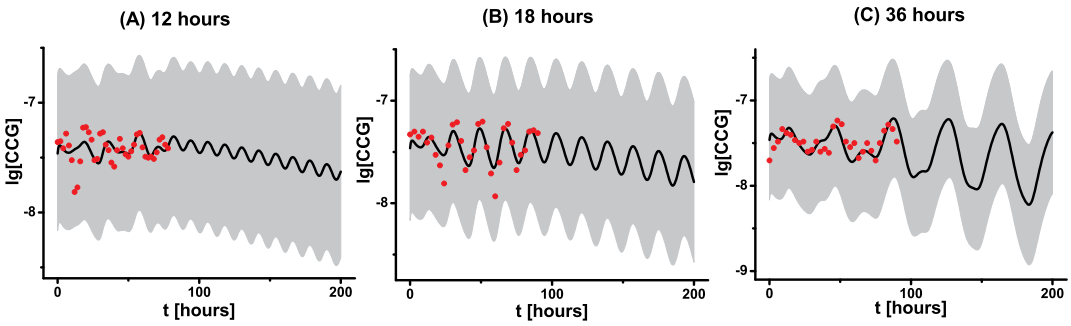


Figure 4.7: Comparison to experiments. A model ensemble for the genetic network in Figure 4.6, with cooperativities $n = m = 4$, predicts *CCG* data on the biological clock of *Neurospora crassa*. In each panel, predictions of the model ensemble for the lg of concentration (in model concentration units "*cu*") of *CCG* protein are shown with $+/-$ 2 ensemble standard deviations (shaded areas) about the ensemble mean (continuous lines). (A) 6+6 hours LD cycle; (B) 9+9 hours LD cycle; (C) 18+18 hours LD cycle. The *CCG* protein data are obtained from [58].

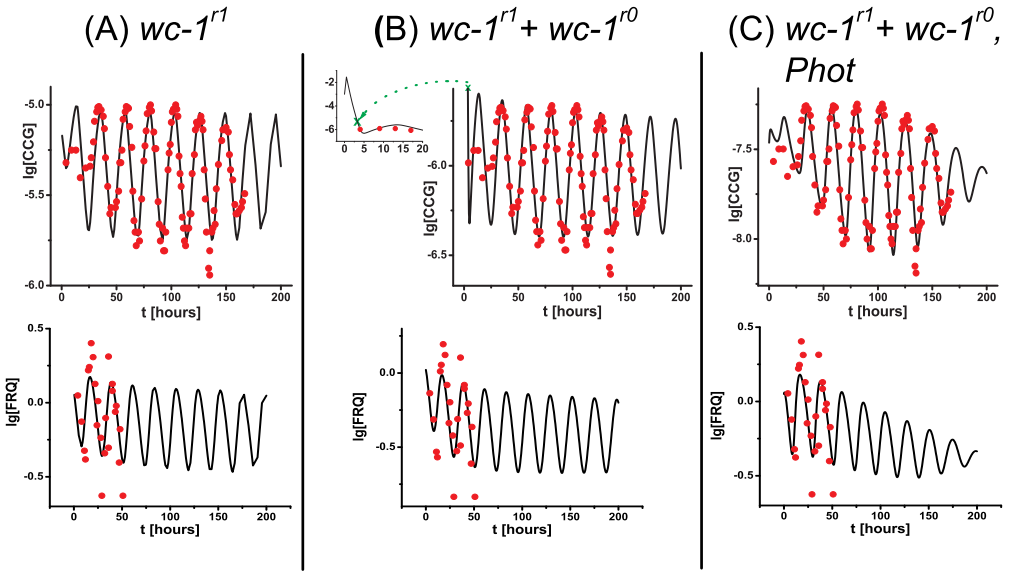


Figure 4.8: Comparison of three different fitting methods for *CCG* and *FRQ* in the dark. (A) χ^2 is computed based on *wc-1*^{r1}. (B) χ^2 is computed based on *wc-1*^{r1}+*wc-1*^{r0}. The inset for *CCG* shows the lg-concentration at very early time. The peak value is greater than the average of later oscillation about 4 fold at lg-scale. (C) χ^2 is computed based on *wc-1*^{r1}+*wc-1*^{r0} together with the light dependent data from [58] (6+6, 9+9 and 18+18 hours LD cycle).

$\boldsymbol{X}$	$\boldsymbol{k}$	$\langle\ \boldsymbol{X}\rangle$			$\sigma(\boldsymbol{X})$		
		(**A**)	(**B**)	(**C**)	(**A**)	(**B**)	(**C**)
A	5	0.0126	0.572	0.260	0.112	0.145	0.103
$\bar{A}_{bar}$	1	0.297	0.338	0.293	0.457	0.0323	0.051
S_1	1	10.4	24.4	2.00E-04	3.53	9.85	4.36E-05
S_2	1	0.247	0.077	0.333	0.740	0.535	1.20
S_3	1	2.77E-3	5.29E-4	1.91E-04	0.0103	1.59E-3	6.10E-04
S_4	1	6.05	8.34	5.41	2.11	0.989	2.57
D_1	1	1.54	0.141	7.33	0.485	0.0608	1.36
D_2	1	8.87E-4	1.73E-3	0.572	6.18E-4	5.60E-4	1.57
D_3	1	0.547	0.439	0.448	0.0584	0.0489	0.0622
C_1	2	0.0127	0.0519	5.08	8.44E-3	0.0316	1.85
L_1	1	63.0	79.2	71.2	12.0	12.8	8.29
L_2	1	15.6	20.2	1.49	1.34	2.49	8.01
L_3	1	6.56	5.76	62.5	2.68	2.08	14.9
D_4	1	0.318	0.352	0.600	0.0812	0.0599	0.277
D_5	1	0.368	0.349	0.0977	0.0384	0.0356	0.0124
D_6	1	0.208	0.241	0.193	0.0123	0.0200	9.52E-3
D_7	1	0.0496	0.0717	0.0632	0.0102	0.0158	9.62E-3
D_8	1	1.17E-3	6.09E-3	0.0362	2.59E-3	7.5E-3	9.27E-3
C_2	2	2.16	7.34	6.77	0.749	1.27	1.96
P	5	4.83	10.5	3.69	9.66	10.2	3.15
A_c	5	1.83	2.26	6.95	1.70	2.02	1.90
B_c	1	5.26	0.186	0.455	2.43	8.27E-3	0.0957
S_c	1	0.467	1.92E-3	2.56E-05	1.35	3.05E-3	2.09E-05
L_c	1	2.42	9.21	0.376	7.35	10.2	0.462
D_{cr}	1	3.69	4.83	0.487	2.40	2.64	0.0921
D_{cp}	1	0.185	5.66	0.453	7.80E-3	2.25	0.0832
E_1	2	-	-	1.67E-3	-	-	1.12E-4
$\bar{E}_{1bar}$	1	-	-	7.54E-3	-	-	7.47E-4
E_2	2	-	-	8.02E-4	-	-	2.71E-4
$\bar{E}_{2bar}$	1	-	-	2.32E-05	-	-	3.23E-05
B	2	-	-	1.08E-05	-	-	2.95E-06
$\bar{B}_{bar}$	1	-	-	4.9E-3	-	-	2.33E-4
S_5	1	-	-	5.12E-4	-	-	6.91E-4
D_9	1	-	-	2.00E-05	-	-	2.14E-05
C_c	5	-	-	3.24E-05	-	-	1.00E-05
Q	5	-	-	3.49E-07	-	-	5.67E-07

Table 4.1: Rate coefficients in the genetic network model of the biological clock ($n = m = 4$) predicting its observed oscillations. Ensemble mean $\langle X\rangle$ and ensemble standard deviation $\sigma(X) := [\langle X^2\rangle - \langle X\rangle^2]^{1/2}$ for rate coefficients (X) in the $n = m = 4$ biological clock model. For a k^{th} order reaction (with k = 1,2 or 5), the rate coefficient is given in units of $1/(hour \times cu^{k-1})$ where "cu" represents the arbitrary, but common model unit of concentration.Column (A): χ^2 is computed based on $wc\text{-}1^{r1}$. Column (B): χ^2 is computed based on $wc\text{-}1^{r1}$+$wc\text{-}1^{r0}$.Column (C): χ^2 is computed based on $wc\text{-}1^{r1}$+$wc\text{-}1^{r0}$ together with the light dependent data from [58] (6+6, 9+9 and 18+18 hours LD cycle). '-' means this reaction is not included in the corresponding model.

D_7 $[1/hours]$	p_{osc} [%]
0.05	82
0.10	81
0.15	79
0.20	78
0.25	77
0.50	73
0.75	68
0.80	58
0.85	41
0.90	26
0.95	14
1.00	6
1.25	0
1.50	0

Table 4.2: This analysis corresponds to Column (A) in Table reftable1. As the life-time of the translationally active *wc-1* messenger RNA, *wc-1*^{r1}, decreases (or equivalently, its decay rate coefficient D_7 increases), the model ensemble predicts that the system becomes arrhythmic. The percentage of stable oscillators, p_{osc}, in the model ensemble from Figure 4.4 (for the $n = m = 4$ model in Figure 2.2 with *WC-2* constant) perturbed by the decay rate coefficient D_7 for each ensemble member being varied from 0.05 ($\sim$ the mean estimate in Table 4.1) to 1.50 while keeping all other rate coefficients fixed at ensemble-generated values. This percentage monotonically decreases as the life-time of the translationally active *wc-1* mRNA decreases (or equivalently, as its decay rate coefficient D_7 increases).

Chapter 5

Maximally Informative Next Experiment (MINE)

5.1 Introduction

To gain complete information about a cell reaction network's topology and rate constants, so as to predict system response, it is crucial to perform time-dependent profiling experiments on the system under a wide variety of externally controlled perturbations, including varying combinations of gene knock-outs, enzyme inhibition and environmental control parameters, such as amino acid availability, carbon source availability, light intensity, and their time dependencies. However, not all experiments will be equally informative. Given the data from the "old" experiments already performed, which "new" experiment should we perform next, in order to gain the maximal information about the genetic network?

To formalize this question, let the design of any individual profiling experiment be completely described in terms of an L-tuple of "control variables" $\mathbf{u} := (u_1, \cdots, u_L)$ which comprises all those parameters, both continuous and discrete, whose values are known and, to some extent, controllable by the experimenter (in contrast to the a priori unknown model parameters $\mathbf{\Theta}$). This "control vector" $\mathbf{u}$ comprises, but is not limited to:

1. discrete (binary) variables specifying those species for which measurements are taken;
2. discrete variables specifying for each measured species (mRNA, protein, or other) the number of time-points;
3. the continuous t-values for each of these measurement time-points;

4. binary variables describing, respectively, the functional presence of or absence of known molecular species (including genes and proteins) and known reactions, which are controllable, *e.g.* by gene knockout or enzyme inhibition;

5. continuous variables describing the (in general) time-dependent values of extracellular environmental parameters, e.g, time-dependent carbon source and amino acid availability.

Thus, $\mathbf{u}$ comprises all experimental design parameters which we need to adjust or choose so as to gain the maximal amount of information from the "new" experiment. In return, the new experimental results can be fed into the ensemble Monte Carlo program to further shrink the size of the uncertainty.

Let $Q(\mathbf{\Theta}, \mathbf{u})$ denotes the ensemble likelihood, based on the prior, old experiments, with observed "old" data vector $\mathbf{Y}$ and corresponding model prediction $\mathbf{F}(\mathbf{\Theta}, \mathbf{u})$, with respect to a proposed, new profiling experiment with control vector $\mathbf{u}$. Clearly, the question of which new experiment is "maximally informative" is not a mathematically well-defined problem. We have to make a reasonable *ad hoc* choice for a design criterion and then try it out in real-life applications. To motivate the choice of design criterion, suppose, for now, that we are given only two possible choices of models, $\mathbf{\Theta}$(such as a *Hill Coefficient* of $n = 4$ in 4.1) and $\mathbf{\Theta}^*$ (such as $n = 3$ in Fig 2), which both give predictions, $\mathbf{F}(\mathbf{\Theta}, \mathbf{u})$ and $\mathbf{F}(\mathbf{\Theta}^*, \mathbf{u})$, consistent with the "old" experimental data (within the experimental uncertainties). In order to distinguish between these two choices, we want to perform a new experiment, with control vector $\mathbf{u}$. The predicted outcomes for this new experiment would be, respectively, $\mathbf{F}(\mathbf{\Theta}, \mathbf{u})$ and $\mathbf{F}(\mathbf{\Theta}^*, \mathbf{u})$. The crucial point to notice here is this: the more these two predicted outcomes $\mathbf{F}(\mathbf{\Theta}, \mathbf{u})$ and $\mathbf{F}(\mathbf{\Theta}^*, \mathbf{u})$ differ from each other, the "better" the new experiment will allow us to discriminate between the two model choices.

As a "metric" of the difference between the two outcomes we could choose, *e.g.*, $V_{\mathbf{\Theta},\mathbf{\Theta}^*}(\mathbf{u}) := |\mathbf{F}(\mathbf{\Theta}, \mathbf{u}) - \mathbf{F}(\mathbf{\Theta}^*, \mathbf{u})|^2$, with $|\cdots|$ denoting the Euclidean norm. We would thus arrive at a reasonable design criterion for a "maximally informative" new experiment:

choose your $\mathbf{u}$ so that it maximizes $V_{\mathbf{\Theta},\mathbf{\Theta}^*}(\mathbf{u})$, given $\mathbf{\Theta}$ and $\mathbf{\Theta}^*$. Different criteria have been proposed.

5.2 Maximization Criteria

In our ensemble method for circuit identification, we are of course given a large set, or even an entire continuum, of choices for our $\mathbf{\Theta}$ and $\mathbf{\Theta}^*$ which are consistent with the "old" experimental data and distributed according to their joint distribution $Q(\mathbf{\Theta}, \mathbf{\Theta}^*) = Q(\mathbf{\Theta}) \times Q(\mathbf{\Theta}^*)$. The generalized criterion for a maximally informative experiment is to choose $\mathbf{u}$ such that it maximizes the average of $V_{\mathbf{\Theta},\mathbf{\Theta}^*}(\mathbf{u})$ over all possible choices of $(\mathbf{\Theta}, \mathbf{\Theta}^*)$:

$$V(\mathbf{u}) := \sum_{\mathbf{\Theta}} \sum_{\mathbf{\Theta}^*} V_{\mathbf{\Theta},\mathbf{\Theta}^*} Q(\mathbf{\Theta}) Q(\mathbf{\Theta}^*) = 2[\langle |\mathbf{F}(., \mathbf{u})|^2 \rangle_Q - |\langle \mathbf{F}(., \mathbf{u}) \rangle_Q|^2]$$

where the second equality follows immediate from $\sum_\Theta Q(\mathbf{\Theta}) = 1$. Note that our proposed approach does not require that the genetic network reaches a stationary state or equilibrium [55]. Standard non-linear minimization methods [56] can be used to minimize $V(\mathbf{u})$.

Generally speaking, $\mathbf{F}(\mathbf{\Theta}, \mathbf{u})$ is a vector. Suppose it is a 2-dimensional vector

$$\mathbf{F}(\mathbf{\Theta}, \mathbf{u}) = \begin{pmatrix} F_1(\mathbf{\Theta}, \mathbf{u}) \\ F_2(\mathbf{\Theta}, \mathbf{u}) \end{pmatrix}$$

where $F_1(\mathbf{\Theta}, \mathbf{u})$ and $F_2(\mathbf{\Theta}, \mathbf{u})$ correspond to two measurements. The score function can be re-written as:

$$V(\mathbf{u}) = 2(\langle [\Delta F_1(\mathbf{\Theta}, \mathbf{u})]^2 \rangle_Q + \langle [\Delta F_2(\mathbf{\Theta}, \mathbf{u})]^2 \rangle_Q)$$

where $\Delta F_i = F_i - \langle F_i \rangle$. The variances will get minimized independently. Hence, the selection of time points for these two measurements are not independent, which means this criterion guides the next experiment to be measured at 2 same time points.

Figure 5.1 shows the basic idea of the solution. We consider not only the magnitude, but also the relative direction between them in a fictitious space. Now we call them vectors: $\mathbf{A_i}$.In the case of 2 vectors, we want to maximize $|\mathbf{A_1} \times \mathbf{A_2}|$, *i.e.*, the area enclosed by these 2

vectors. Obviously, the maximum is achieved when the 2 vectors are orthogonal. Same idea can be applied to higher dimensional cases.

Let the number of species we want to measure is denoted by M_s, and for each species, M_t time points are measured for each species. Therefore, we have $N = M_s M_t$ unknown parameters for vector $\mathbf{u}$, where u_i corresponds to (s_i, t_i) for $i = 1 \cdots N$. Also define

$$\mathbf{d}_i(\mathbf{\Theta}, \mathbf{\Theta}^*) = \mathbf{d}(\mathbf{\Theta}, \mathbf{\Theta}^*, u_i) = \mathbf{F}(\mathbf{\Theta}, u_i) - \mathbf{F}(\mathbf{\Theta}^*, u_i).$$

Here we want to choose $u_1, \cdots, u_N$ so that all the $|\mathbf{d}_i(\mathbf{\Theta}, \mathbf{\Theta}^*)|^2$ are maximized, with the constraint that $\mathbf{d}_i(\mathbf{\Theta}, \mathbf{\Theta}^*)$ is independent of $\mathbf{d}_j(\mathbf{\Theta}, \mathbf{\Theta}^*)$.

Define a finite subspace $\mathcal{D} = \{\mathbf{a}(\mathbf{\Theta}, \mathbf{\Theta}^*)\}$, where $\mathbf{a}(\mathbf{\Theta}, \mathbf{\Theta}^*) = \sum_i a_i \mathbf{d}_i(\mathbf{\Theta}, \mathbf{\Theta}^*)$. Also define the orthornormal system (ONS) in this subspace $\mathcal{D}$ as $\mathbf{e}_1, \cdots, \mathbf{e}_N \in \mathcal{D}$, where $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$. Hence, $\mathbf{d}_i = \sum_{l=1}^{n} \mathbf{e}_l (\mathbf{e}_l \cdot \mathbf{d}_i)$, where $d_i^l = \mathbf{e}_l \cdot \mathbf{d}_i$. As a natural consequence, the determinant is thus well defined as

$$\det(d_i^l) \text{ where } i = 1 \cdots N \text{ and } l = 1 \cdots N.$$

In 2-dimensional case, we have

$$\det \begin{pmatrix} d_1^1 & d_1^2 \\ d_2^1 & d_2^2 \end{pmatrix} := \text{ area } .$$

The score function which we want to maximize can now be re-written as $V(\mathbf{u}) = |\det(d_i^l)|^2$. $V(\mathbf{u})$ is independent of the choice of $\mathbf{e}_1, \cdots, \mathbf{e}_N$. Therefore, we get

$$|\det(d_i^l)|^2 = |\det(D_{ij})|^2$$

where $D_{ij} = D_{ji} = \mathbf{d}_i \cdot \mathbf{d}_j$, and $i, j = 1, \cdots, n$. Hence,

$$D_{ij} = \sum_{\mathbf{\Theta}} \sum_{\mathbf{\Theta}^*} (\mathbf{d}_i \cdot \mathbf{d}_j) Q(\mathbf{\Theta}, \mathbf{\Theta}^*) = 2\langle \Delta F_i \Delta F_j \rangle_Q = 2 \sum_{\mathbf{\Theta}} \Delta F_i \Delta F_j Q(\mathbf{\Theta}) = 2[\langle F_i F_j \rangle_Q - \langle F_i \rangle_Q \langle F_j \rangle_Q].$$

In summary, 3 criteria have been proposed. The first one is to use the trace of the variance-covariance matrix of proposed measurements, $\mathrm{Tr}(D_{ij})$. The second method is to use

the determinant of this variance-covariance matrix to enforce the independence, $\det(D_{ij})$. The third one follows the second method directly by normalization, $\mathrm{Tr}(D_{ij})\det(D_{ij})/\prod_i D_{ii}$.

The foregoing approach is the last link to complete the "computing life paradigm" and is being incorporated into an integrated workflow system [57] wherein new experiments $\mathbf{u}$ are designed by the simulations and minimization procedures outlined above, with an ensemble Q based on the pre-existing "old" data. The results from the new experiments are then merged into the pre-existing data vector $\mathbf{Y}$ and its distribution $P(\mathbf{Y})$ and are thereby incorporated into new ensemble Q, to be used in the design of the next new experiment. With the inclusion of each new experimental data set, the ensemble Q will become more refined, until, at last, it identifies the "true" underlying kinetics model of the system under study. The approach thus provides a rational way to design new experiments.

5.3 MINE for Biological Clock

Microarrays are the powerful tools which are widely used for surveying the expression levels of thousands of genes simultaneously. Each microarray experiment to be designed involves obtaining mRNA levels on all $11,000$ genes in the Neurospora genome for 13 time points. Some of the genes are duplicated on the chip so that we have several replicate expression profiles ($12,544$ spots in total). For example, all of the clock genes and *qa* cluster genes are measured 5 times by each chip.

In the MINE calculations, 3 mRNA species, ccg^r, frq^r and $wc\text{-}1^r$, are virtually measured under different experimental perturbation for each Θ on the ln scale. For a typical Monte Carlo accumulation run, there are $40,000$ different Θ values. For every 200 Θ values, one is picked to run the virtual experiment. So the ensemble average is taken at level of 200. The dimension of matrix D is 39 (3 mRNA times 13 time points).

To run experiments for the light-dependent clock model as in Figure 4.6, the first question to ask is of course at what level of light intensity is most informative. In Figure 4.7, the light intensity was set at 20 μmol photons/m^2/s. This is simply a one-dimensional optimal search.

The microarray experiments are very expensive although this technology indeed enables us to measure thousands of mRNA levels simultaneously. This solves the "spatial" problem, but not the "temporal" problem. In other words, which time points should be measured remains an "expensive" question. Figure 5.3 shows the MINE results for different time points. This is a two-dimensional optimization: one direction for first time point, and the other direction for spacing between two consecutive points. The virtual experiments are run under constant dark. Variation grows as time goes on for any prediction, so late measurements will be dominant. In other words, all 13 time points are condensed to the latest possible time for microarray measurements. This is a reasonable, but not useful result due to the fact of dependence between different time points. The second method, determinant criterion, gives different results. Subject to the constraint of 60-hour microarray measurement, 5-hour spacing and immediate start of measurement is the guide for the next experiment. The third normalized determinant method yields the same results as the second one. Obviously, the independence enforcement pushes all the time points away from each other to maximize the criterion.

Another interesting dimension we can explore is the the different spacing ratio. Define r to be a ratio as in Figure 5.4. This virtual experiment is run under constant dark again.

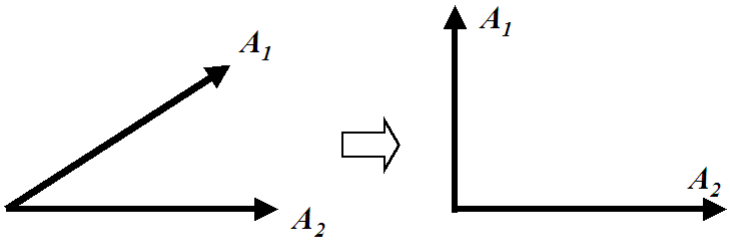


Figure 5.1: Area and vectors.

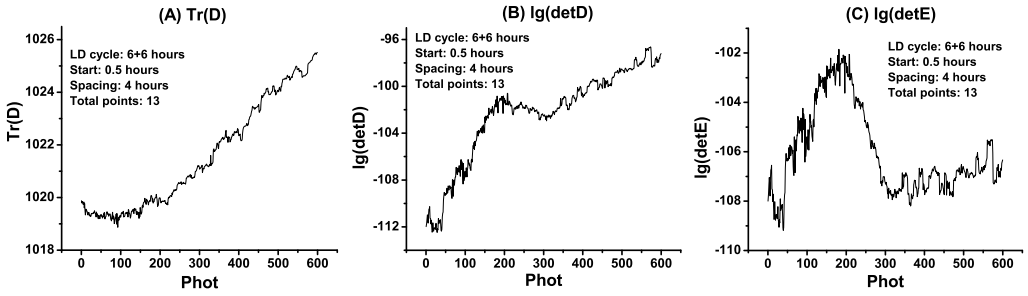


Figure 5.2: MINE prediction for different light intensity experiments. Three criteria are used, with LD cycle being 6+6 hours, the first measurement taking place at 0.5 hours, the spacing between consecutive measurements being 4 hours, and the total time points being 13. Görl's light intensity corresponds to 100 in the model units. x-axis shows different light intensity and y-axis shows three different scores for three different methods. (A) Trace method indicates as large light intensity as possible. (B) Determinant method shows a peak at 200, which is about 40 μmol photons/m^2/s. (C) Normalized determinant method clearly identifies 200 as the best choice.

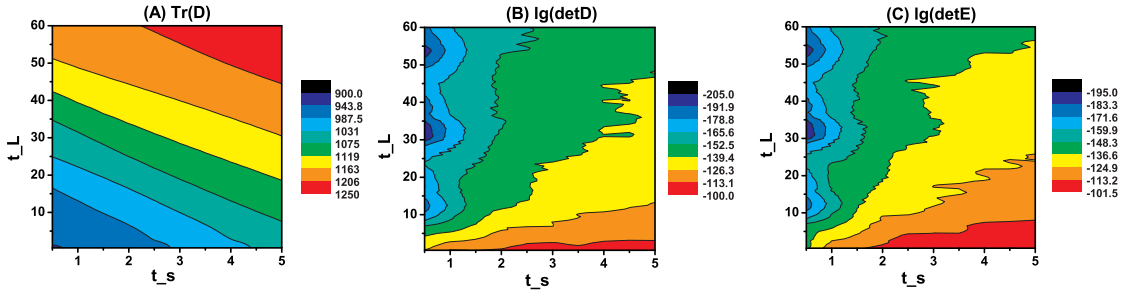


Figure 5.3: MINE prediction for different measurement time experiments. x-axis, t_s, denotes the spacing between two consecutive measurements, and y-axis, t_L, denotes the first measurement time. The furthest time a microarray chip can measure is about 60 hours. Different contour color shows scores of three different maximization criteria. (A) Trace method indicates that we want to delay the start of measurement as late as possible and maximize the spacing at the same time. (B) Determinant method indicates that we want to make the spacing as large as possible while start the measurements as early as possible. (C) Normalized determinant method gives the same results as in the second method.

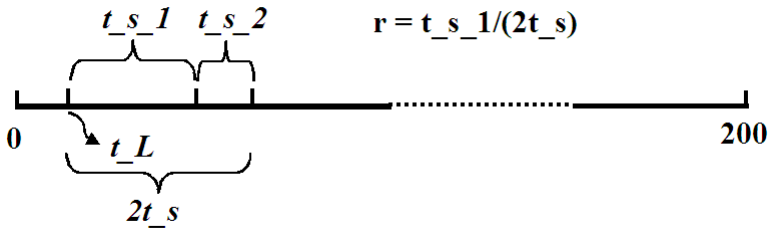


Figure 5.4: Definition of r ratio. Each 2 consecutive measurements are combined to define $2t\_s$. The ratio r is explored by moving the middle bar between its left barrier and right barrier, from 0 to 1.

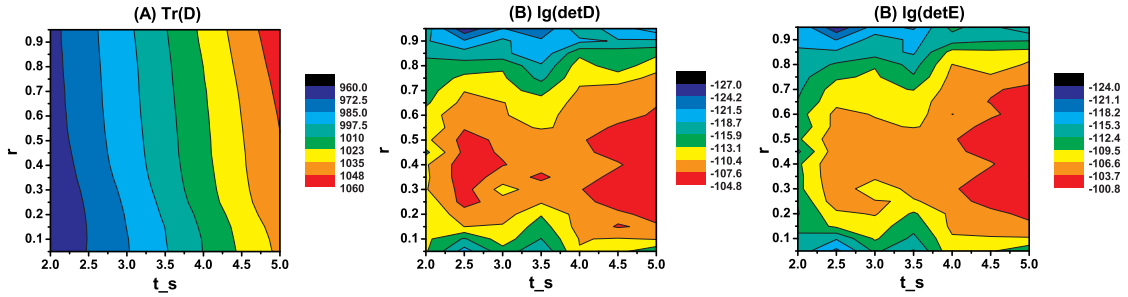


Figure 5.5: MINE prediction for different spacing ratio experiments. x-axis is t_s, where $2t\_s$ is defined as total spacing for two consecutive measurements, and y-axis, r, denotes the spacing ratio as define in Figure 5.4. Again, three contour plots are shown for three different criteria. (A) The dependent trace method again predicts to delay the even-numbered measurement to be as late as possible, which is equivalent to make a replicate. (B) Determinant method enforces independence and the even-numbered measurements are forced to be in the middle of odd-numbered measurements for any fixed t_s. (C) The normalized determinant method gives similar results as the second method.

Chapter 6

ENSSOLVER

The simulation results from the statistical ensemble code ens.for are consisted of tremendous amount of raw data in plain text files. It's very difficult to draw any inference from raw data files directly. A visualization tool, ENSSOLVER, has been designed and implemented to help the ensemble code users to debug the program and analyze the simulation results. This web-based Graphical User Interface (GUI) can be accessed via http://gene.csp.uga.edu (or http://gene2.csp.uga.edu as public demonstration with limited functions).

Figure 6.1 shows the hierarchical structure of ENSSOLVER. After you log onto gene.csp.uga.edu, you get to the root. Then different models shall be created, *e.g.* biological clock. Under each model, you can upload different runs, which is ens.o01 file. Then ens.o01 is processed to generate all the appropriate modules for further plotting and analysis. ENSSOLVER is based on Java Servlet [52]. GNUPLOT [53] and GIFSICLE [54] are used for plotting.

For each run, there are 2 frames. The left frame displays 5 different tasks you can perform:

- Monte Carlo Average vs Real Time: plot the ensemble average of each species against real time;
- Monte Carlo Average Projection: plot one species versus another species against real time and animation of the trajectory movement;
- Monte Carlo Parameter: plot Monte Carlo parameters, such as rate coefficients, against the virtual time (Monte Carlo Step);

- ENSSOLVER for one MC Step: plot the kinetic simulation according to the selection of one particular parameters set;
- Manage Files: manage the data files.

The right frame displays the results from server, such as the gif figure, statistical calculation, downloadable column data files.

The following Figures 6.2 and 6.3 give examples of using ENSSOLVER for analysis of biological clock.

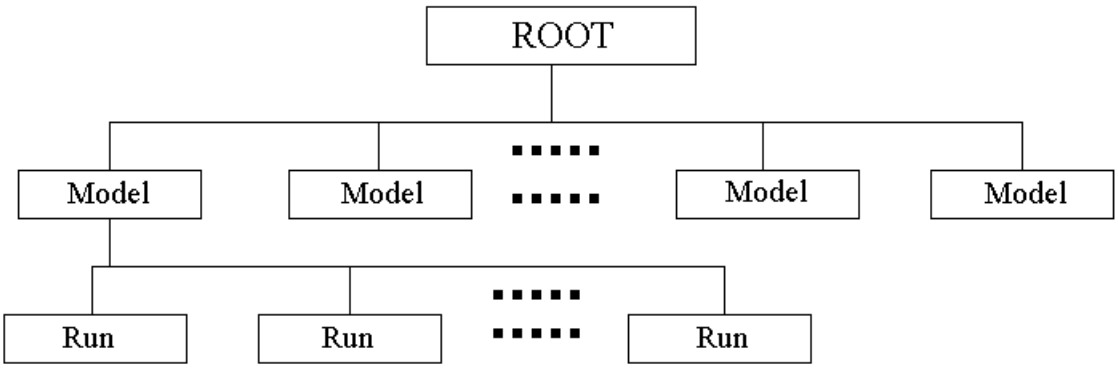
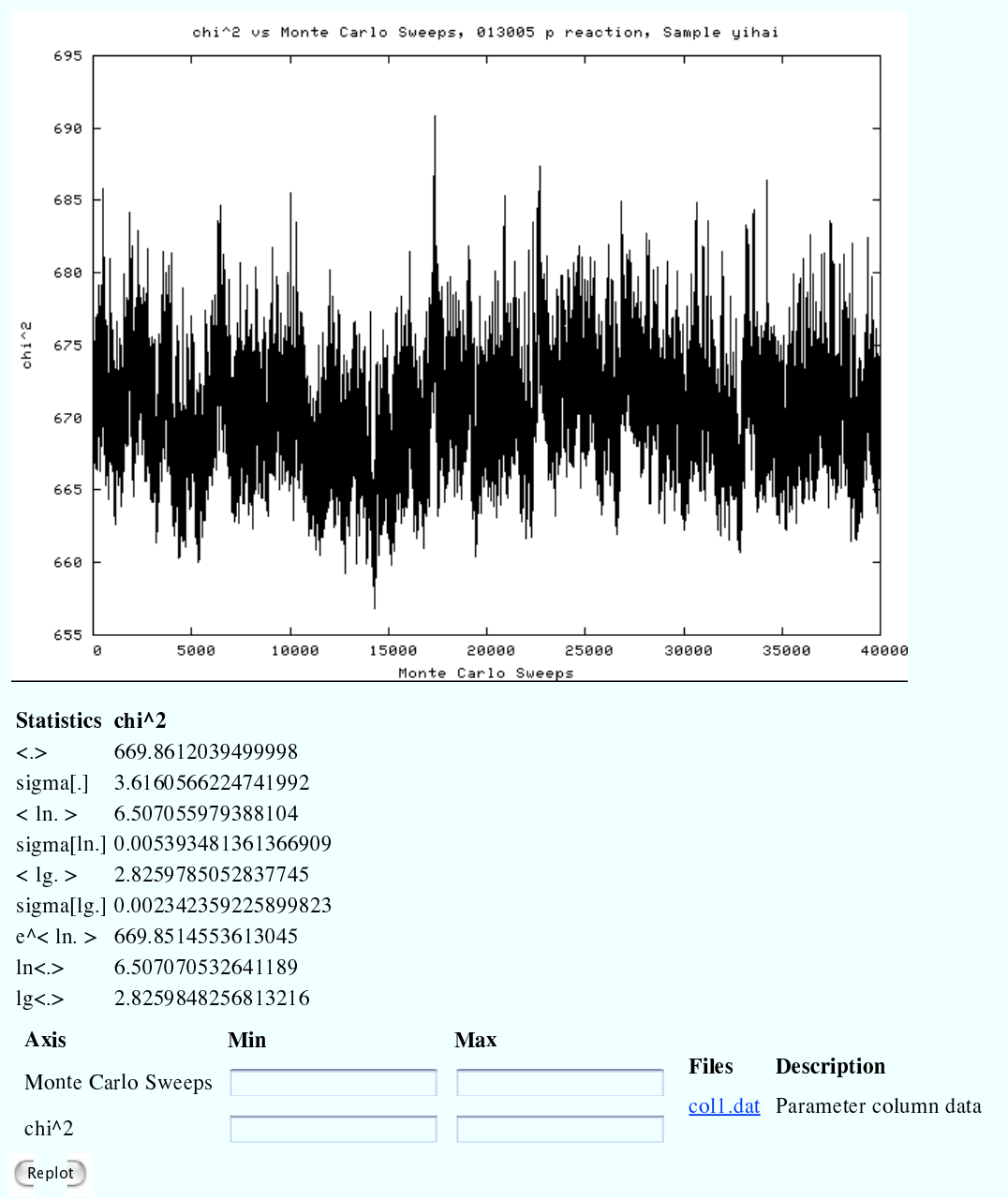


Figure 6.1: The Hierarchical Structure of ENSSOLVER.



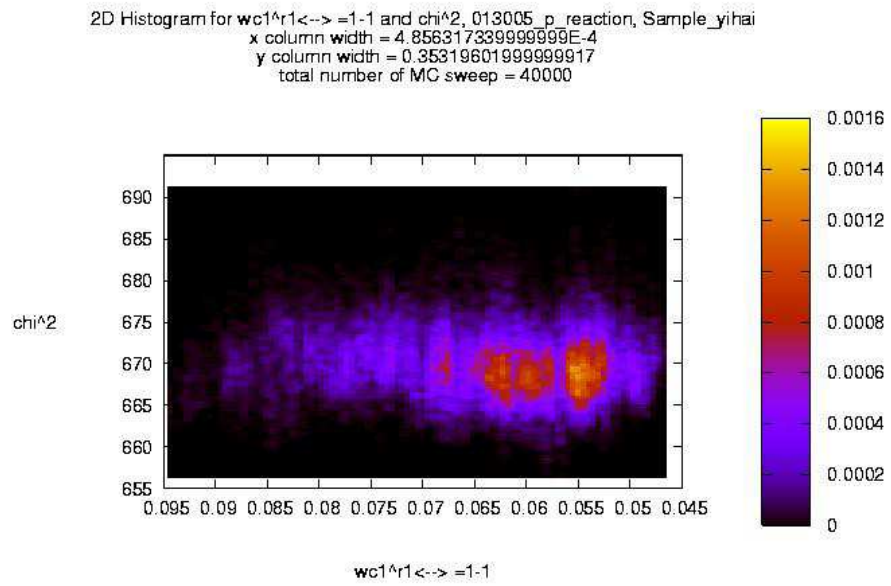
Statistics	chi^2
<.>	669.8612039499998
sigma[.]	3.6160566224741992
< ln. >	6.507055979388104
sigma[ln.]	0.005393481361366909
< lg. >	2.8259785052837745
sigma[lg.]	0.002342359225899823
e^< ln. >	669.8514553613045
ln<.>	6.507070532641189
lg<.>	2.8259848256813216

Axis	Min	Max
Monte Carlo Sweeps		
chi^2		

Files	Description
col1.dat	Parameter column data

Replot

Figure 6.2: Example of output. In this example, a plot of χ^2 versus Monte Carlo sweep is displayed together with caption of run name and model name. Some statistics about χ^2 is shown next. You can also replot the figure based on your needs and download the column file for further manipulation. This accumulation run explores a region with all the χ^2 being equally good (near-minimal values of χ^2) although some of them do not yield sustained oscillation.



Statistics	wc1^r1<--> =1-1	chi^2
<.>	0.06325827161250082	669.8612039499998
sigma[.]	0.009628240638613427	669.3526720011569
< ln. >	-2.7715921116664144	6.507055979388104
sigma[ln.]	0.14724962468594371	0.005393481361366909
< lg. >	-1.2036871601832912	2.8259785052837745
sigma[lg.]	0.06394969946359384	0.002342359225899823
e^< ln. >	0.06256231920966894	669.8514553613045
ln<.>	-2.7605293834427798	6.507070532641189
lg<.>	-1.1988826783609852	2.8259848256813216

Figure 6.3: Example of output. This is a 2-dimensional histogram plot for two Monte Carlo parameters: one is χ^2 and the other is the rate coefficient D_7, followed by some statistics of the two parameters. The rate coefficient D_7 is very well constrained in the range of 0.05 and 0.07.

Bibliography

[1] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001) Science 291, 1304C1351.

[2] Jacob, F.; and Monod, J. (1961) *Genetic regulatory mechanisms in the synthesis of proteins*. J. Mol. Biol. 3, 318-356.

[3] Yanofsky, C.; and Kolter, R. (1982) *Attenuation in amino acid bisynthesis operons*. Ann. Rev. genetic 16, 113-134.

[4] Yanofsky, C. (2001) *Advancing our knowledge in biochemical, genetics, and microbiology through studies of tryptophan metabolism*. Ann. Rev. Biochem. 70, 1-37.

[5] Johnston, M. (1987) *A model fungal regulatory mechanism: the GAL genes of Saccharomyces cerevisiae*. Microbiolog. Rev. 51, 458-476.

[6] Giles, N. H.; Case, M. E.; Baum, J.; Geever, R.; Huiet, L.; Patel, V.; and Tyler, B. (1985) *Gene organization and regulation in the qa (Quinic Acid) gene cluster of Neurospora crassa*. Microbiological Reviews 49, 338-358.

[7] Sveiczer, A.; Csikasz-Nagy, A.; Gyorffy, B.; Tyson, J.J.; and Novak, B. (2000) *Modeling the fission yeast cell cycle: quantized cycle times in wee1- cdc25D mutant cells*. Proc. Natl. Acad. Sci. USA 97, 7865-7870.

[8] Beadle, G. W.; and Tatum, E. I.(1941) *Genetic control of biochemical reactions in Neurospora*. Proc. Natl. Acad. Sci. USA 27, 499-506.

[9] DeRisi, J. L., Iyer, V.R., and Brown, P. O. (1997) *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science 278:680-686

[10] Gygi, S. P., RistB., Gerber, S. A., Trecek, F., Gelb M. H., and Aebersold, R. (1999) *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags*. Nature Biotechn. 17: 994-999

[11] Walhout, A.J.M., Sordella, R., Lu, R., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000) *Protein interaction mapping in C. elegans using proteins involved in vulval development*. Science 287: 116-122

[12] Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E., Simon, I., Zeitlinger, J., Schreiber, J., Hanett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000) Science 290: 2306-2309

[13] Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000) *Construction of a genetic toggle switch in Escherichia coli*. Nature 403: 339-342

[14] Aleman-Meza, B.; Yu, Yihai; Schüttler,H. B.; Arnold, Jonathan; and Taha, Thiab R.(2005) *KINSOLVER: A simulator for computing large ensembles of biochemical and gene regulatory networks* (Submitted to Software: Practice and Experience).

[15] Dormand, J. R. (1996) *Numerical Methods for Differential Equations – A Computational Approach*. CRC Press.

[16] Cheney, E. W.; and Kincaid, D.R. (1999) *Numerical Mathematics and Computing*. Fourth Edition. Brroks/Cole Publishing.

[17] Yu, Yihai (2004) *Stiff Problems in Numerical Simulation of Biochemical and Gene Regulatory Networks*. Master Thesis, University of Georgia, August 2004.

[18] Shampine, L.F.; and Gear, C.W. (1979) *A User's View of Solving Stiff Ordinary Differential Equations*. SIAM Review, Vol. 21, Issue 1, 1-17.

[19] Cooper, N. G. (1989) *From Cardinals to Chaos*. Cambridge Universtiy Press, Cambridge.

[20] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M. N., Teller, A. H., Teller E. (1953) J. Chem. Phys. 21, 1087.

[21] Landau, David P., Binder, Kurt (2000) *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press.

[22] Glasserman, Paul (2003) *Monte Carlo methods in financial engineering*. Springer-Verlag New York.

[23] Gorbacheva, Victoria Y., Kondratov, Roman V., Zhang, Renliang, Cherukuri, Srujana, Gudkov, Andrei V., Takahashi, Joseph S., and Antoch, Marina P. (2005) *Circadian sensitivity to the chemotherapeutic agent cyclophosphamide depends on the functional status of the CLOCK/BMAL1 transactivation complex* . PNAS 102: 3407-3412.

[24] Frédéric Gachon, Philippe Fonjallaz, Francesca Damiola, Pascal Gos, Tohru Kodama, Jozsef Zakany, Denis Duboule, Brice Petit, Mehdi Tafti, and Ueli Schibler (2004) *Genes and Development*, Jun 2004; 18: 1397 - 1412.

[25] Dunlap, J.C. (1999) *Molecular bases for circadian clocks*. Cell, 96, 271-290.

[26] Lee, K., Loros, J. J., and Dunlap, J. C. (2000) *Interconnected feedback loops in the Neurospora circadian system*. Science 289, 107-110

[27] Battogtokh, D., Asch, D. K., Case, M. E., Arnold, J., and Schüttler, H.-B. (2002) *An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa*. Proc. Natl. Acad. Sci. (USA) 99, 16904-16909

[28] Lynch, M. and Walsh, J. B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA

[29] Crosthwaite, S. K., Dunlap, J. C., and Loros, J. J. (1997) *Neurospora wc-1 and wc-2: transcription, photoresponses, and the origins of circadian rhythmicity*. Science 276, 763-769

[30] Froehlich, A. C., Liu, Y., Loros, J. J., and Dunlap, J. C. (2002) *White Collar-1, a circadian blue light photoreceptor, binding to the frequency promoter*. Science 297, 815-819

[31] Denault, D. L., Loros, J. J., and Dunlap, J. C. (2001) *WC-2 mediates WC-1-FRQ interaction with the PAS protein-linked circadian feedback loop of Neurospora*. EMBO. J. 20, 109-117

[32] Garceau, N. Y., Liu, Y, Loros, J. J., and Dunlap, J. C. (1997) *Alternative initiation of translation and time-specific phosphorylation yield multiple forms of the essential clock protein FREQUENCY*. Cell 89, 469-476

[33] Segel, I. H. (1975) *Enzyme Kinetics*. Wiley, NY

[34] Winfree, A. T. (2000) *Geometry of Biological Time*. Springer-Verlag. NY, NY

[35] Winfree, A. T. (2002) *On emerging coherence*. Science 298: 2336-2337

[36] Merrow, M. W. Garceau, N. Y., and Dunlap, J. C. (1997) *Dissection of a circadian oscillation into discrete domains*. Proc. Natl. Acad. Sci. (USA) 94, 3877-3882

[37] Liu, Y., N.Y. Garceau, J.J. Loros, J. J., and J.C. Dunlap, J. C. (1997) *Thermally regulated translational control of FRQ mediates aspects of temperature responses in the Neurospora Circadian Clock*. Cell 89: 477-486

[38] Cheng, P., Yang, Y., Heintzen, C., and Liu, Y. (2001) *Coiled-coil domain mediated FRQ-FRQ interaction is essential for its circadian clock function in Neurospora*. EMBO J. 20: 101-108

[39] Ruoff, P., Vinsjevik, M., Mahsenzadeh, and Rensing, L. (1999) *The Goodwin model: simulating the effect of cycloheximide and heat shock on the sporulation rhythm of Neurospora crassa*. J. Theor. Biol. 196: 483-494

[40] Gonze, D., Halloy, J., and Goldbeter, A. (2002) *Robustness of circadian rhythms with respect to molecular noise*. Proc. Natl. Acad. Sci. (USA) 99: 673-678

[41] Smolen, P., Baxter, D. A., Byrne, J. H. (2001) *Modeling circadian oscillations with interlocking positive and negative feedback loops*. J. Neurosci. 21: 6644-6656

[42] Tyson, J. J., Hong, C. I., Thron, C. D., and Novak, B. (1999) *A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. Biophys*. J. 77: 2411-2417

[43] Leloup, J.-C. and Goldbeter, A. (1998) *A model for circadian rhythms in Drosophila incorporating the formation of a complex between the PER and TIM proteins*. J. Biological Rhythms 13: 70-87

[44] Becskei, A. and Serrano, L. (2000) *Engineering stability in gene networks by autoregulation*. Nature 405: 590-593

[45] Elowitz, M. B. and S. Leibler. (2000) *A synthetic oscillatory network of transcriptional regulators*. Nature 403: 335-338

[46] Strogatz, S. H. (2001) *Exploring complex networks*. Nature 410, 268-276

[47] Minorsky, N. (1962) *Nonlinear Oscillations*. Van Nostrand, Princeton, NJ

[48] Noble, B. and J. W. Daniel. (1977) *Applied Linear Algebra*. 2nd Edition. Prentice-Hall, Inc. Englewood Cliffs, NJ

[49] Gantmacher, F.R. (1959) The Theory of Matrices II. New York, Chelsea Publishers

[50] Glazier, A. M., Nadeau, J. H., and Aitman, T. J. (2002) *Finding genes that underlie complex traits*. Science 298: 2345-2349

[51] The International HapMap Consortium. (2003) *The International HapMap Project*. Nature 426: 789-796

[52] Rossbach P., and Schreiber, H. (2000) *Java Server and Servlets: Building Portable Web Applications*. Addison-Wesley

[53] http://www.gnuplot.info/

[54] http://www.lcdf.org/ eddietwo/gifsicle/

[55] Gardner, T.S., Dibernardo, D, Lorenz, D., and Collins J.J. (2003) *Inferring genetic networks and identifying compound mode of action via expression profiling* Science 301: 102-104

[56] Press, W.H, Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1993) *Numerical Recipes* Cambridge University Press, NY, NY

[57] Kochut, KJ, Arnold J., Sheth, A., Miller, J., Kraemer, E., Arpinar, B., and Cordoso, J. (2003) *INTELLIGEN: a distributed workflow system for discovering protein-protein interactions*. Parallel and Distributed Databases 13: 43-72

[58] Margit Görl, Martha Merrow, Benedikt Huttner, Judy Johnson, Till Roenneberg, and Michael Brunner (2001) *A PEST-like element in FREQUENCY determines the length of the circadian period in Neurospora crassa*. The EMBO Journal 20, 24 pp.7074-7084

[59] Allan C. Froehlich, Jennifer J. Loros, and Jay C. Dunlap (2003) *Rhythmic binding of a WHITE COLLAR-containing complex to the frequency promoter is inhibited by FREQUENCY*. PNAS 100, 10 pp. 5914-5919

Appendix A

Stability Analysis for Biological Clock Model

A.1 Introduction to Routh-Hurwitz Analysis

The Routh-Hurwitz stability analysis is a method for determining whether or not a system is stable based upon the coefficients in the system's characteristic equation. It is particularly useful for higher-order systems because it does not require the polynomial expressions in the transfer function to be factored.

Consider the characteristic equation

$$\det(\mathbf{J} - \lambda\mathbf{E}) = A_0\lambda^n + A_1\lambda^{n-1} + \cdots + A_{n-1}\lambda + A_n = 0$$

determining the n eigenvalues λ of a real $n \times n$ square matrix $\mathbf{J}$, where $\mathbf{E}$ is the identity matrix. The following 2-step criterion can be used:

1. If any of the coefficients are zero or negative and at least one of the coefficients are positive, there is a root or roots that are imaginary or that have positive real parts. Therefore, the system is unstable.

2. If all coefficients are positive, arrange the coefficients in rows and columns in the following pattern:

$$\begin{array}{cccccc} s^n & A_0 & A_2 & A_4 & A_6 & \cdots \\ s^{n-1} & A_1 & A_3 & A_5 & A_7 & \cdots \\ s^{n-2} & B_1 & B_2 & B_3 & B_4 & \cdots \\ s^{n-3} & C_1 & C_2 & C_3 & C_4 & \cdots \\ s^{n-4} & D_1 & D_2 & D_3 & D_4 & \cdots , \\ \vdots & \vdots & \vdots & & & \\ s^2 & E_1 & E_2 & & & \\ s^1 & F_1 & & & & \\ s^0 & G_1 & & & & \end{array}$$

where

$$\begin{array}{cccc} B_1 = \frac{A_1A_2 - A_0A_3}{A_1} & C_1 = \frac{B_1A_3 - A_1B_2}{B_1} & \cdots \\ B_2 = \frac{A_1A_4 - A_0A_5}{A_1} & C_2 = \frac{B_1A_5 - A_1B_3}{B_1} & \cdots \\ B_3 = \frac{A_1A_6 - A_0A_7}{A_1} & C_1 = \frac{B_1A_7 - A_1B_4}{B_1} & \cdots \\ \vdots & \vdots & \vdots \end{array}$$

The Routh-Hurwitz stability criterion states that the number of roots with positive real parts is equal to the number of changes in sign of the coefficients in the $s^n, \cdots, s^0$ column. Note that the exact values are not required for the coefficients; only the sign matters.

A.2 Stability Criterion $R > 0$ for Biological Clock Model

According to 7-dimensional clock model's rate equations, we can have

$$\mathbf{D} = \mathbf{J} - \lambda\mathbf{E} =$$

$$\begin{pmatrix} -(\gamma_1+\lambda) & 0 & 0 & d_1 & 0 & 0 & 0 \\ \bar{b}_1 & -(\gamma_2+\lambda) & 0 & 0 & 0 & 0 & 0 \\ 0 & \bar{b}_2 & -(\gamma_3+\lambda) & 0 & 0 & 0 & 0 \\ \bar{d}_1 & 0 & \bar{b}_3 & -(\gamma_4+\lambda) & b_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & -(\gamma_5+\lambda) & b_5 & 0 \\ 0 & 0 & \bar{d}_3 & 0 & 0 & -(\gamma_6+\lambda) & b_6 \\ 0 & 0 & \bar{e}_3 & 0 & 0 & 0 & -(\gamma_7+\lambda) \end{pmatrix}.$$

Then, the characteristic polynomial can be written as

$$K(\lambda) = [Q(\lambda) - q(\lambda)]G(\lambda)$$

where

$$\begin{aligned} q(\lambda) &= \gamma_1 + \lambda, \\ Q(\lambda) &= \frac{d_1\bar{d}_1}{\gamma_4+\lambda} + \frac{d_1\bar{b}_3\bar{b}_2\bar{b}_1}{(\gamma_4+\lambda)(\gamma_3+\lambda)(\gamma_2+\lambda)}, \\ G(\lambda) &= (\gamma_2+\lambda)(\gamma_3+\lambda)\cdots(\gamma_7+\lambda). \end{aligned}$$

Obviously, $\lambda_5 = -\gamma_5 < 0, \lambda_6 = -\gamma_6 < 0, \lambda_7 = -\gamma_7 < 0$. They always yield stable FP. Therefore, we investigate the reduced characteristic polynomial instead as

$$P(\lambda) = (\gamma_1+\lambda)\cdots(\gamma_4+\lambda) - d_1\bar{d}_1(\gamma_2+\lambda)(\gamma_3+\lambda) - d_1\bar{b}_3\bar{b}_2\bar{b}_1.$$

Also define

$$\begin{aligned} g_1 &= \frac{\gamma_1+\gamma_4}{2} - [(\frac{\gamma_1-\gamma_4}{2})^2 + d_1\bar{d}_1]^{1/2}, \\ g_2 &= \gamma_2, \\ g_3 &= \gamma_3, \\ g_4 &= \frac{\gamma_1+\gamma_4}{2} - [(\frac{\gamma_1-\gamma_4}{2}) + d_1\bar{d}_1]^{1/2}. \end{aligned}$$

Then, $P(\lambda)$ can be expressed as

$$P(\lambda) = (g_1+\lambda)(g_2+\lambda)(g_3+\lambda)(g_4+\lambda) + b^4$$

where $b = -d_1\bar{b}_1\bar{b}_2\bar{b}_3$. The proof for g_2 and g_3 being positive is trivial. Since $d_1\bar{d}_1 < \gamma_4\gamma_1$, g_1 and g_4 are also positive. For real λ, since g_1, g_2, g_3 and g_4 are all larger than 0, no real unstable eigenvalues exist. Expand $P(\lambda)$ as

$$P(\lambda) = A_0\lambda^4 + A_1\lambda^3 + A_2\lambda^2 + A_3\lambda + A_4$$

where

$$\begin{aligned}
A_0 &= 1, \\
A_1 &= \sum_{i=1}^{4} g_i, \\
A_2 &= \frac{1}{2}(\prod_{i=1}^{4} g_i)\sum_{i\neq j}\frac{1}{g_i g_j}, \\
A_3 &= (\prod_{i=1}^{4} g_i)\sum_{i=1}^{4}\frac{1}{g_i}, \\
A_4 &= \prod_{i=1}^{4} g_i + b^4.
\end{aligned}$$

Since $a_i > 0$ for $i = 0, 1, 2, 3, 4$, second step of RH criterion is required. Thus the RH table is constructed as in the previous section:

$$\begin{array}{lll}
A_0 & A_2 & A_4 \\
A_1 & A_3 & \\
B_1 & B_2 & \\
C_1 & & \\
D_1 & &
\end{array}$$

where

$$\begin{aligned}
B_1 &= (\prod_{i=1}^{4} g_i)(\sum_{i\neq j}\frac{1}{g_i g_j} - \frac{\sum_{i=1}^{4}\frac{1}{g_i}}{\sum_{i=1}^{4} g_i}), \\
B_2 &= \prod_{i=1}^{4} g_i + b^4, \\
C_1 &= (A_1A_2A_3 - A_3^2 - A_1^2A_4)/(A_2A_1 - A_3), \\
D_1 &= B_2.
\end{aligned}$$

All the elements in the first column of the RH table is inherently positive except c_1. According to the RH criterion, if and only if $c_1 < 0$ can yield an eigenvalue with positive real part to make the system unstable, which gives the following criterion:

$$R := A_3^2 + A_1^2 A_4 - A_1 A_2 A_3 > 0.$$

A.3 Stability Criterion $nm > 4$ for Biological Clock Model

Before working on the necessary and sufficient condition for instability: $nm > 4$ criterion, let us summarize all the relevant stability parameters as the following:

$$\begin{aligned}
&(\mathbf{P1}) \quad \gamma_1 = Aw^n + \bar{A} \\
&(\mathbf{P2}) \quad \gamma_2 = D_3 \\
&(\mathbf{P3}) \quad \gamma_3 = D_6 \\
&(\mathbf{P4}) \quad \gamma_4 = n^2 A f_0 w^{n-1} + D_8 + P f_p^m = n d_1 + D_8 + P f_p^m \\
&(\mathbf{P5}) \quad \bar{b}_1 = S_4 - S_3 \\
&(\mathbf{P6}) \quad \bar{b}_2 = L_3 \\
&(\mathbf{P7}) \quad \bar{b}_3 = -mPw f_p^{m-1} \\
&(\mathbf{P8}) \quad d_1 = nAf_0 w^{n-1} = \frac{1}{n}(\gamma_4 - \eta) \text{ where } \eta := D_8 + P f_p^m \\
&(\mathbf{P9}) \quad \bar{d}_1 = n(Aw^n + \bar{A}) = n\gamma_1
\end{aligned}$$

The Stationary constraints in Section 4.5.4 can be expressed in terms of the stability parameters by setting all the differential equations to be zero:

$$\begin{aligned}
&(\mathbf{S1}) \quad A = \frac{\gamma_1}{w^n}\frac{f_1}{f_G} \\
&(\mathbf{S2}) \quad S_3 = -\bar{b}_1\frac{f_1}{f_G} + \frac{\gamma_2\gamma_3}{\bar{b}_2}\frac{f_p}{f_G} \\
&(\mathbf{S3}) \quad f_r = \frac{\gamma_3}{\bar{b}_2} f_p \\
&(\mathbf{S4}) \quad E_2 = \frac{1}{u_p}(D_8 w - \frac{\bar{b}_3 f_p}{m})
\end{aligned}$$

$$(\mathbf{S5}) \quad L_1 = (D_4 + E_2)\frac{u_p}{u_{r1}}$$

$$(\mathbf{S6}) \quad D_7 = C_1 f_p \frac{u_{r0}}{u_{r1}}$$

$$(\mathbf{S7}) \quad V_1 = (D_1 + C_1 f_p) u_{r0}$$

Lemma A.3.1

$$\bar{b}_1 \bar{b}_2 \leq \gamma_2 \gamma_3 \frac{f_p}{f_1}.$$

Proof From (**S2**), we get

$$S_3 = -\bar{b}_1 \frac{f_1}{f_G} + \frac{\gamma_2 \gamma_3}{\bar{b}_2} \frac{f_p}{f_G}$$

. Because S_3 has to be positive as a rate coefficient, hence, the result follows. ∎

Lemma A.3.2

$$d_1 \leq \frac{\gamma_4}{n} \frac{1}{1 + |\bar{b}_3|/b_{30}}$$

where

$$b_{30} := mn^2 (1 - \frac{f_1}{f_G}) \frac{\gamma_1}{f_p/f_1}.$$

Proof Put (**P1**) into (**S1**), we get

$$\bar{A} = \gamma_1 (1 - \frac{f_1}{f_G}) \text{ and } A = \frac{1}{w^n}(\gamma_1 - \bar{A}) = \frac{1}{w^n} \gamma_1 \frac{f_1}{f_G}.$$

The (**P8**) can be written as

$$w = \frac{n}{d_1} \gamma_1 f_1 (1 - \frac{f_1}{f_G}).$$

At the same time, (**P7**) indicates that

$$P f_p^m = \frac{|\bar{b}_3| f_p}{mw} = \frac{|\bar{b}_3|}{mn} \frac{d_1}{\gamma_1 (1 - f_1/f_G)} \frac{f_p}{f_1}.$$

Then, (**P4**) gives

$$D_8 = \gamma_4 - n d_1 - P f_p^m = \gamma_4 - (n + \frac{|\bar{b}_3|/\gamma_1}{mn(1 - f_1/f_G)} \frac{f_p}{f_1}) d_1.$$

D_8 is a rate coefficient, which means $D_8 > 0$. The result follows. ∎

Lemma A.3.3

$$d_1|\bar{b}_3| \leq \frac{1}{n}\gamma_4 b_{30}\varphi_3 < \frac{1}{n}\gamma_4 b_{30}$$

where

$$\varphi_3 := \frac{|\bar{b}_3|/b_{30}}{1+|\bar{b}_3|/b_{30}}.$$

Proof This is a direct result from Lemma A.3.2 since

$$\varphi_3 = \frac{|\bar{b}_3|/b_{30}}{1+|\bar{b}_3|/b_{30}} < 1.$$

∎

Lemma A.3.4

$$\eta \geq \eta_x \text{ where } \eta := \gamma_4 - nd_1 \text{ and } \eta_x = \gamma_4\varphi_3.$$

Proof From Lemma A.3.2, we have $\gamma_4 - nd_1 \geq \gamma_4\varphi_3$, which is $\eta \geq \eta_x$. ∎

Lemma A.3.5

$$b^4 \leq b_x^4\varphi_f\varphi_3 < b_x^4$$

where

$$b^4 := d_1\bar{b}_1\bar{b}_2|\bar{b}_3|, b_x^4 := \gamma_1\gamma_2\gamma_3\gamma_4 nm, \varphi_f := 1 - \frac{f_1}{f_G} \text{ and } \varphi_3 = \frac{|\bar{b}_3|/b_{30}}{1+|\bar{b}_3|/b_{30}}.$$

Proof Combine Lemma A.3.1 and A.3.3, we get

$$d_1\bar{b}_1\bar{b}_2|\bar{b}_3| \leq \gamma_2\gamma_3\frac{f_p}{f_1}\frac{1}{n}\gamma_4 b_{30}\varphi_3.$$

Since

$$b_{30} = mn^2(1-\frac{f_1}{f_G})\frac{\gamma_1}{f_p/f_1},$$

the result follows. Also note that $0 < \varphi_f < 1$ and $0 < \varphi_3 < 1$. ∎

The original characteristic polynomial coefficients can be written as:

$$\begin{aligned} A_0 &= 1, \\ A_1 &= \gamma_1+\gamma_2+\gamma_3+\gamma_4, \\ A_2 &= (\gamma_2+\gamma_3)(\gamma_1+\gamma_4)+\gamma_2\gamma_3+\gamma_1\eta, \\ A_3 &= \gamma_2\gamma_3(\gamma_1+\gamma_4)+(\gamma_2+\gamma_3)\gamma_1\eta, \\ A_4 &= \gamma_1\gamma_2\gamma_3\eta+b^4. \end{aligned}$$

Consider a special case such that: $\gamma_1=\gamma_2=\gamma_3=\gamma_4=\gamma$. We have:

$$\begin{aligned} A_1 &= 4\gamma, \\ A_2 &= 5\gamma^2+\gamma\eta, \\ A_3 &= 2\gamma^3+2\gamma^2\eta, \\ A_4 &= \gamma^3\eta+b^4. \end{aligned}$$

Put them into the R criterion:

$$R := A_3^2+A_1^2A_4-A_1A_2A_3 = 16\gamma^2b^4-36\gamma^6-24\gamma^5\eta-4\gamma^4\eta^2.$$

Use the lemmas just proved, we can get

$$R \le [16nm\varphi_f\varphi_3-(36+24\varphi_3+4\varphi_3^2)]\gamma^6.$$

The necessary condition for $R>0$ is that

$$16nm\varphi_f\varphi_3-(36+24\varphi_3+4\varphi_3^2)>0,$$

which means

$$16nm > \frac{1}{\varphi_f}(\frac{36}{\varphi_3}+4\varphi_3+24) =: \Gamma_1.$$

By checking the monotonicity of Γ_1 with respect to φ_3,

$$\frac{\partial\Gamma_1}{\partial\varphi_3} = \frac{1}{\varphi_f}(-\frac{36}{\varphi_3^2}+4) < 0 \text{ for } 0<\varphi_f<1 \text{ and } 0<\varphi_3<1.$$

Therefore, the necessary condition for instability is

$$nm > \frac{15}{4} = 3.75.$$

Define a new set of variables:

$$\begin{aligned} a &:= \frac{1}{2}(\gamma_2 + \gamma_3), \\ g &:= (\gamma_2\gamma_3)^{1/2} \leq a, \\ f &:= \gamma_1 + \gamma_4, \\ h &:= (\gamma_1\eta)^{1/2} \\ j &:= (\gamma_2 + \gamma_3)^{1/2} \leq \frac{1}{2}f. \end{aligned}$$

The characteristic polynomial coefficients can be re-written as:

$$\begin{aligned} A_1 &:= f + 2a, \\ A_2 &:= 2af + g^2 + h^2, \\ A_3 &:= g^2 f + 2ah^2, \\ A_4 &:= g^2h^2 + b^4. \end{aligned}$$

Because $j^2g^2 = \gamma_1\gamma_2\gamma_3\gamma_4$ and Lemma A.3.5, $b^4 \leq nm\varphi_f\varphi_3 j^2 g^2$. Therefore, $A_4 \leq g^2(h^2 + \frac{nm}{4}j^2)$ and $\varphi_3 j^2 \leq h^2 < j^2$.

Also define

$$\begin{aligned} X &:= A_1A_2A_3 \\ Y &:= A_3^2 + A_1^2A_4 \end{aligned}$$

Then

$$\begin{aligned} X &:= (f + 2a)(2af + j^2 + a^2)(g^2f + 2ah^2)(g^2f + 2ah^2) \\ Y &:= (g^2f + 2ah^2)^2 + (f + 2a)^2(g^2h^2 + b^4). \end{aligned}$$

Therefore,

$$\begin{aligned} R &= Y - X \\ &= (f+2a)^2 b^4 - 2af[(g^2-h^2)^2 + (g^2 f + 2ah^2)(f+2a)], \end{aligned}$$

where $R > 0$ indicates instability. Check R with respect to h^2 at fixed f, g, a, b:

$$\frac{\partial R}{\partial (h^2)} = -2af[2(h^2 - g^2) + 2a(f+2a)] = 0.$$

Hence, $h^2 = g^2 - a(f+2a) < g^2 - 2a^2 < -a^2 < 0$, which means the maximum location is on the negative part of h^2. For $h^2 > 0$, since $g^2 \leq a^2$,

$$\frac{\partial R}{\partial (h^2)} = -4af(h^2 + af + 2a^2 - g^2) < 0.$$

Therefore, R monotonically decreases with respect to h^2 for $h^2 \geq 0$. Based on the monotonicity, $h^2 \geq \varphi_3 j^2$ immediately concludes that $R(h^2) \leq R(\varphi_3 j^2)$. Also, note that

$$b^4 \leq \gamma_1\gamma_2\gamma_3\gamma_4 nm\varphi_f\varphi_3 = nm\varphi_f\varphi_3 j^2 g^2,$$

hence, $R \leq R_1$, where

$$R_1 := nm\varphi_f\varphi_3 j^2 g^2 (f+2a)^2 - 4a^2 f\varphi_3 j^2 (f+2a) - 2af^2 g^2 (f+2a) - 2af(g^2 - \varphi_3 j^2)^2.$$

Let $j = \frac{1}{2}\varphi_j f$, note $0 < \varphi_j \leq 1$. Then

$$\begin{aligned} R_1 = &- 2ag^4 f + (4p - 4a^2 g^2 f^2) \\ &+ (4pag^2 + \varphi_j^2 ag^2 - 2\varphi_3\varphi_j^2 a^3 - 2ag^2) f^3 \\ &+ (pg^2 - \varphi_3\varphi_j^2 a^2) f^4 - \frac{1}{8}\varphi_3^2\varphi_j^4 af^5 \end{aligned}$$

where

$$p := \frac{nm}{4}\varphi_f\varphi_3\varphi_j^2.$$

Consider the special test case again, $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma$. R_1 is reduced to:

$$R_1 = [64p - (36 + 24\varphi_3 + 4\varphi_3^2)]\gamma^6.$$

The necessary condition for instability requires $R_1 > 0$, which means

$$p > \frac{36 + 24\varphi_3 + 4\varphi_3^2}{64}.$$

Therefore, we have

$$nm > \frac{1}{4\varphi_f}(\frac{9}{\varphi_3} + 6 + \varphi_3) =: \Gamma_2.$$

Function Γ_2 is monotonically decreasing for $0 < \varphi_3 < 1$. Therefore the necessary condition for instability is the same as the previous conclusion:

$$nm > \frac{15}{4} = 3.75.$$

Let $x = \frac{f}{2a}$ and $\varphi_g = \frac{g}{a}$, R_1 can be re-written as:

$$\begin{aligned} R_1 = & - 4\varphi_g^4 a^6 x + 4(nm\varphi_3\varphi_f\varphi_j^2\varphi_g^2 - 4\varphi_g^2)a^6x^2 \\ & + 8(nm\varphi_3\varphi_f\varphi_j^2\varphi_g^2 - 2\varphi_3\varphi_j^2 - 2\varphi_g^2)a^6x^3 \\ & + 4(nm\varphi_3\varphi_f\varphi_j^2\varphi_g^2 - 4\varphi_3\varphi_j^2)a^6x^4 \\ & - 4\varphi_3\varphi_j^2 a^6 x^5. \end{aligned}$$

Since $0 < \varphi_f < 1$,$0 < \varphi_3 < 1$,$0 < \varphi_j \leq 1$, then

$$R_1 < R_2 := G_2(x)\varphi_f\varphi_3\varphi_j^2 a^6,$$

where

$$G_2(x) := -4\varphi_g^4 x + 4(k-4)\varphi_g^2 x^2 + 8[(k-1)\varphi_g^2 - 2]x^3 + 4(k\varphi_g^2 - 4)x^4 - 4x^5$$

Also let $\varphi_g^2 = 1 - y$, where $0 \leq y < 1$, then

$$G_2(x) = H_0(x) + H_1(x)$$

where

$$\begin{aligned} H_0(x) &:= -4x + 4(k-4)x^2 + 8(k-3)x^3 + 4(k-4)x^3 - 4x^5 \\ H(x,y) &:= 4(2y - y^2)x - 4(k-4)yx^2 - 8(k-1)yx^3 - 4kyx^4 \\ k &:= nm \end{aligned}$$

Note that $H_0(x)$ does not contain y.

$G_2(x)$ monotonically increases with respect to k. So, let us try the marginal case such that $k = 4$. $H_0(x)$ is greatly simplified,

$$H_0(x) = -4x(x^2 - 1)^2,$$

which indicates that

$$H_0(x) \leq 0 \ \forall x \geq 0.$$

At the same time,

$$H(x, y) = 4x[2(1 - 3x^2 - 2x^3)y - y^2].$$

Let $s(x) := 1 - 3x^2 - 2x^3$. There are 2 cases for analysis:

1. $s(x) \geq 0$. Maximize $H(x, y)$ with respect to y:

$$\frac{\partial H(x, y)}{\partial y} = 4x[2s(x) - 2y] = 0,$$

which gives $y_0(x) = s(x) \geq 0$. Obviously, $H(x, y) \leq H(x, y_0(x)) = 4x[s(x)]^2$. Therefore,

$$\begin{aligned} G_2(x) &= H_0(x) + H(x, y) \\ &\leq H_0(x) + H(x, y_0(x)) \\ &= 4x\{[s(x)]^2 - (1 - x^2)^2\} \\ &= 4x[s(x) - (1 - x^2)][s(x) + (1 - x^2)] \\ &= -4x(2x^2 + 2x^3)[s(x) + (1 - x^2)] \end{aligned}$$

Because $s(x) \geq 0$ and $0 \leq x < 1$, $G_2(x) < 0$.

2. $s(x) < 0$. Because $0 \leq y < 1$, $H(x, y) \leq 0$. Hence,

$$G_2(x) = H_0(x) + H(x, y) < 0.$$

Also note that R_2 is a monotonically increasing function with respect to k. Therefore for any $k \leq 4$, $R \leq R_1 < R - 2, 0$ is always satisfied, which means the instability is not possible.

Therefore, the necessary condition for existence of FP instability is that:

$$nm > 4.$$

Now let us prove that $nm > 4$ is also a sufficient condition for existence of FP instability.

Take $\varphi_j = 1$, $\varphi_g = 1$, $\varphi_3 \to 1 - 0^+$, $\varphi_f \to 1 - 0^+$, we can get $R \to R_1 - 0^+ \to R_2 - 0^+$, *i.e.* R can be arbitrarily close to R_1 and R_2. Let $\epsilon := k - 4 > 0$, and note that $y = 1 - \varphi_g^2 = 0$, we get

$$\begin{aligned} R_2(x) &= H_0(x) \\ &= -4x + 4\epsilon x^2 + 8(1+\epsilon)x^3 + 4\epsilon x^4 - 4x^5 \end{aligned}$$

In order to prove the "existence", we just need to choose the right point, *e.g.* $x = 1$, then

$$R_2(1) = +16\epsilon > 0,$$

which gives

$$R \to G_2(1)a^6 > 0.$$

So, the unstable FP is found at $x = 1$, $y = 0$. Hence, $nm > 4$ being the sufficient condition is proved.

In conclusion, the level of cooperativity in the clock model, $nm > 4$, is a necessary and sufficient condition for existence of FP instability ($R > 0$).