Computational Systems Biology for the Biological Clock of *Neurospora crassa*

by

Xiaojia Tang

(Under the direction of Heinz-Bernd Schüttler)

Abstract

Genetic networks have been applied to describe biological systems, *e.g.*, the biological clock, from a systems biology perspective. A model-driven discovery process, Computing Life, is developed and used to identify an ensemble of genetic networks to describe quantitatively the biological clock of the lowly bread mould *Neurospora crassa* for its light-responsive behavior through iterative cycles combining both experiments and computational simulations. Central to this discovery process is a new methodology for the rational design of a Maximally Informative Next Experiment(MINE) based on the genetic network ensemble. In each cycle, the MINE approach is used to design the most informative new experiment for the biological goal of discovering *clock-controlled genes* which is the outputs of the clock. The new experimental results are then added back to the data pool to provide more information to improve the estimates and predictions made by the genetic network ensemble. The identified ensemble of light-responsive genetic networks is expanded trying to describe the temperature response of the *N. crassa* and has been proved to be sufficient to explain the wild type data under different temperatures.

Index words: genetic networks, biological clock, ensemble approach, maximally informative next experiment

Computational Systems Biology for the Biological Clock of *Neurospora crassa*

by

Xiaojia Tang

B.E., Tsinghua University, 2002

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

2009

Computational Systems Biology for the Biological Clock of *Neurospora crassa*

by

Xiaojia Tang

Approved:

Major Professor: Heinz-Bernd Schüttler

Committee: Jonathan Arnold
David P. Landau

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2009

Dedication

This dissertation is dedicated to my parents,

who gave me life, love me and support me always.

Acknowledgments

I would like to thank 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. David P. Landau 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. Shan-Ho 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. Yihai Yu, who implemented the first version of MINE and gave me a lot of help for my further work. My appreciation also goes to Dr. Wubei Dong, James Griffith and Rosemary Kim for their significant input for my research. I also greatly appreciate the inspiring discussion with Dr. Jaxk Reeves 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 researc assistantship. This work is supported by US National Science Foundation under NSF DBI-0243754 and BES-0425762.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Genetic Network for the Biological Clock

For species with fully sequenced genomes, it is presently feasible to obtain time-dependent profiling of each biomolecule, *i.e.*, DNAs, RNAs and proteins through technologies including DNA microarrays and quantitative proteomics [1–3]. The development in genomics and measuring biotechnologies provides us the genomic, proteomic and metabolic information, bringing new detailed insights into the genetic and biochemical circuitry of a living cell. Grounded in an understanding at the molecular level, it is now time to step up to the system-level to understand a biological system's structure and dynamics from a more comprehensive view. Genetic networks, as the central part of the system approach to biology, provide a powerful theoretical and computational framework to integrate and summarize our current knowledge of genes and proteins, and mostly important, their interactions and interconnections [4]. More than just a static diagram, genetic networks can incorporate biochemical reactions and time-dependent profiling data of biomolecules to reveal how living systems function dynamically, how they react to perturbations, such as environmental changes and genetic engineering, etc [5, 6].

Many genetic networks can be partially identified for experimentally well-studied systems from available knowledge in the literature. For example, the biological clock is known to be a complex trait that can influence a large number of phenotypes from only a few regulatory macromolecules. Particularly the biological clock of *Neurospora crassa* [7], a filamentous bread mould, is chosen as our biological system of interest. Its biological clock is easy to observe and to manipulate(Figure 1.1). As an experimentally well studied model organism,

the entire genome of *Neurospora crassa* has has been completely sequenced [8]. Thus it has been possible to identify three molecular building blocks of the clock: the genes *white-collar-1* (*wc-1*), *white-collar-2* (*wc-2*), and frequency (*frq*) and their products. The genes *wc-1* and *wc-2* encode PAS-domain containing transcription factors that turn on the clock oscillator. The WC-1 protein also acts as a blue-light receptor. The gene *frq* encodes the clock oscillator FRQ and is activated by the WHITE-COLLAR transcription factor protein complex WCC=WC-1/WC-2. The FRQ protein in turn appears to function as a cyclin to recruit an as yet to be identified kinase/phosphatase pair for the phosphorylation-dependent inactivation of WCC [9–12].

From the above information, a detailed genetic network has been constructed to explain how the clock functions as shown in Figure 1.2 [14]. In the network model, the WCC protein activates the oscillator gene *frq*. The activated frq^1 gene is then transcribed into its cognate mRNA frq^{r1}, which in turn is translated into its cognate protein FRQ. The FRQ protein, in turn, deactivates the WCC in the P reaction. It thus forms a loop of dynamical frustration, *i.e.*, WCC turns on the oscillator gene (*frq*) whose product shuts down the activator WCC. This negative feedback loop between WCC and FRQ explains in part how clock oscillations arise. As the controlled outputs of the clock mechanism, WCC activates a large group of *clock-controlled genes* (*ccgs*), the number of which in the genome was largely unknown prior to the work reported in this dissertation. Hence, the extent of clock control over metabolism is largely unknown [15].

1.2 The Computing Life Paradigm

Beyond the capability of quantitatively describing a living biological system and answering some fundamental questions, *e.g.*, how a complex trait like the biological clock works, another important feature of a genetic network is making computational predictions about the corresponding biological systems. Researchers have proposed such an iterative process involving an interaction of modeling and experimentation to identify and validate genetic networks [4,16].

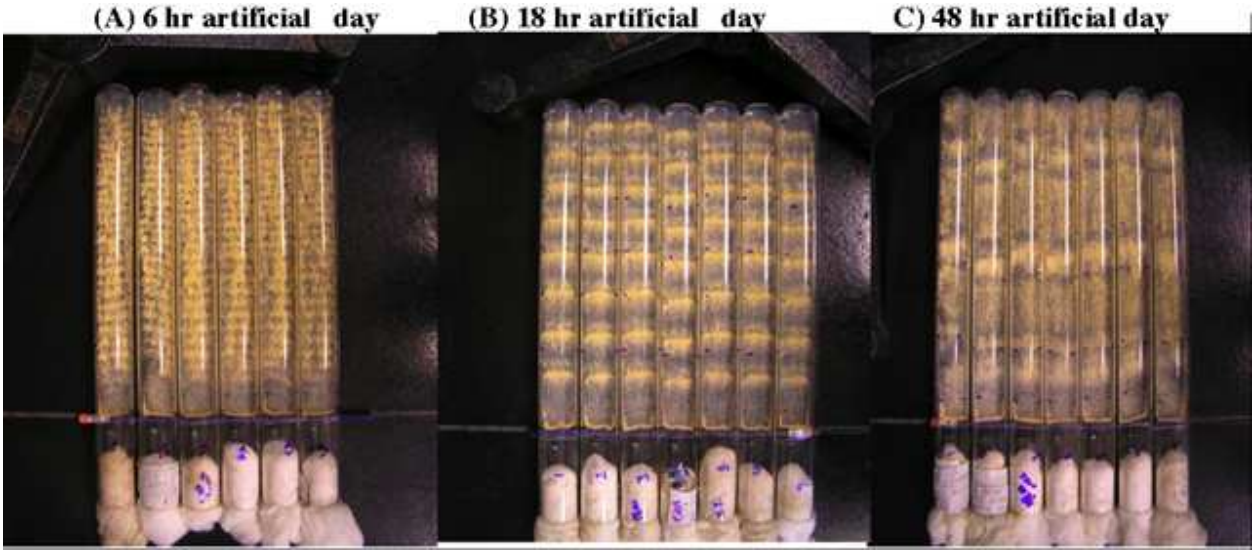

Figure 1.1: **The clock of *N. crassa* is remarkably adaptive in its entrainment to varied artificial days.** Replicate race tubes (as shown in the figures) are inoculated at one end and subject to a 6 hr, 18 hr, and 48 hr artificial day over 7 ordinary days. The clock is manifested by the appearance of orange bands (i.e., asexual production of spores) as the culture grows to the other end of the tube. The term "artificial days" means that the culture is grown in alternating dark and light periods with respective equal amount of time. In each artificial day the race tubes experienced (A) 3 hrs light following 3 hrs dark, (B) 9 hrs light following 9 hrs dark, or (C) 24 hrs light following 24 hrs dark. It can be seen that the number of conidial bands tracked the number of artificial days experienced. The tube with bands, marked with time points on them, are then photographed and analyzed the intensity change with respect to time [13].

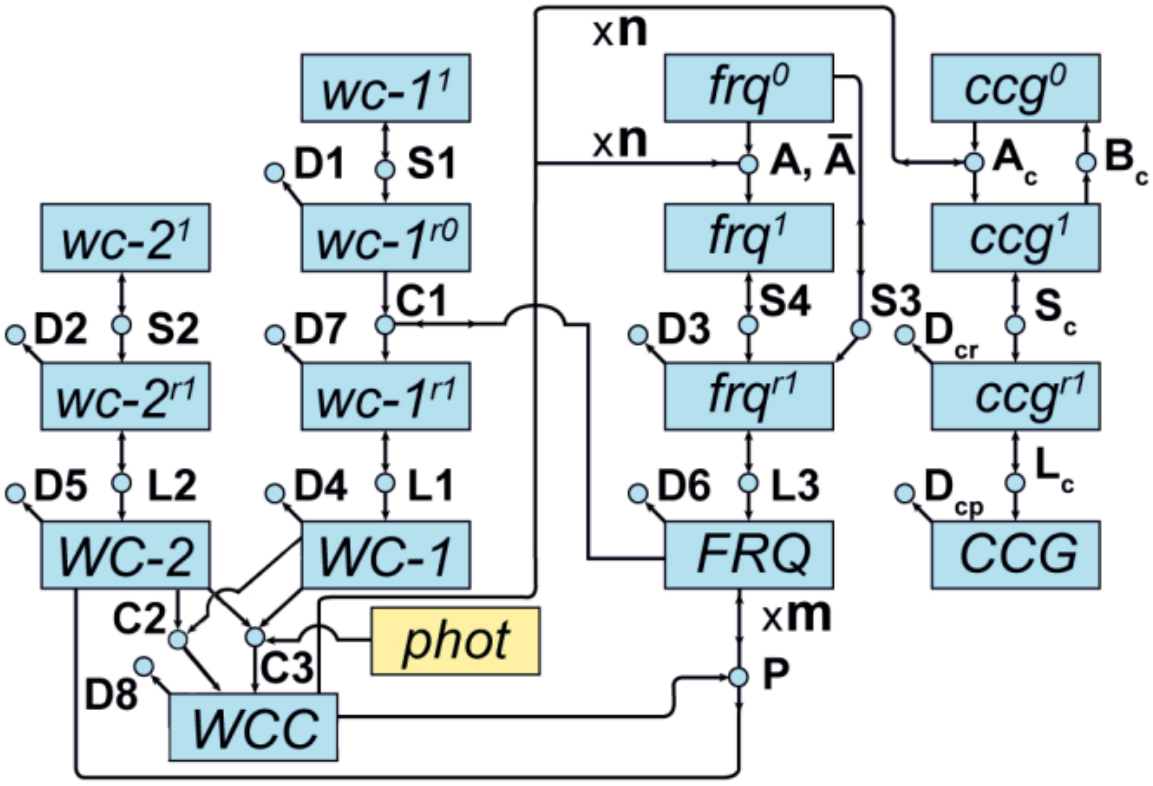

Figure 1.2: **Model diagram of the biological clock of *Neurospora crassa.*** Molecular species (*i.e.*, reactants or products) in the network are represented by boxes. The *white-collar-1* (*wc-1*), *white-collar-2 (wc-2)*, *frequency (frq)*, and *clock controlled gene (ccg)* gene symbols can be superscripted 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. A phot (in yellow) denotes a photon species. 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 S4, also serve to denote the rate coefficients for each reaction. Reactions labeled with an S, L, or D denote transcription, translation, or degradation reactions, respectively. Reactions without products, such as D8, are decay reactions. Reactions, such as A and P, have cooperative kinetics: (A) $n\mathrm{WCC} + frq^0 \rightarrow frq^1$ and (P) $\mathrm{WCC} + m\mathrm{FRQ} \rightarrow \mathrm{WC{-}2} + m\mathrm{FRQ}$. The n and m are Hill coefficients or cooperativities. Only one reaction, the ''A'' reaction, has a back reaction, ($\bar{\mathrm{A}}$), $frq^1 \rightarrow n\mathrm{WCC} + frq^0$, included, with nonzero rate. The rate constants specify the right hand side of the kinetics model in equation (1) through the Law of Mass Action in Materials and Methods. The figure has been used in [13]

.

Along these lines, we introduce a model-driven discovery process called Computing Life as shown in Figure 1.3. In this paradigm, a cycle consisting of computational modeling and genomics experiments is used to identify a genetic network model for the biological system of interest by iteratively tightening the estimates on model parameters and model predictions with cumulative experimental data in each cycle.

A preliminary genetic network model is first built to describe the biological system of interest according to known genetic and biochemical information. The biological system is then perturbed, and the outcomes are observed, i.e., measured by RNA and protein profiling. An ensemble fitting approach is applied with the constraint of the experimental data [14]. From the generated ensemble of model parameters, predictions are made and compared with available data to evaluate goodness of fit of the ensemble. The next step is then to choose a new perturbation to obtain more information to revise and improve the model. The difficulty here is: what is the best perturbation to be done next to improve maximally our knowledge of the genetic network, thus to make our model as close to the truth as possible [13]?

One proposed approach to find out the most "informative" next experiment is to design the next perturbation based on the assumption that genetic networks are in steady state and/or are linear [17, 18]. This is not applicable here since the biological clock is usually not in a steady state but is rather approaching a stable limit cycle. Also, the steady-state approach discards most information contained in observations on network dynamics, *i.e.*, its time-dependent behavior. Another approach is to generate an entire compendium of profiling experiments for varied genetic and environmental perturbations [19]. However, such profiling experiments nowadays are still very costly. In many situations, it is not a feasible choice, considering the available budget. We, therefore, developed a process of choosing the Maximally Informative Next Experiment (or MINEing) which can be guided by the continuously refined network model in an intelligent and cost-effective way while fully exploiting the information contained in the observed network dynamics.

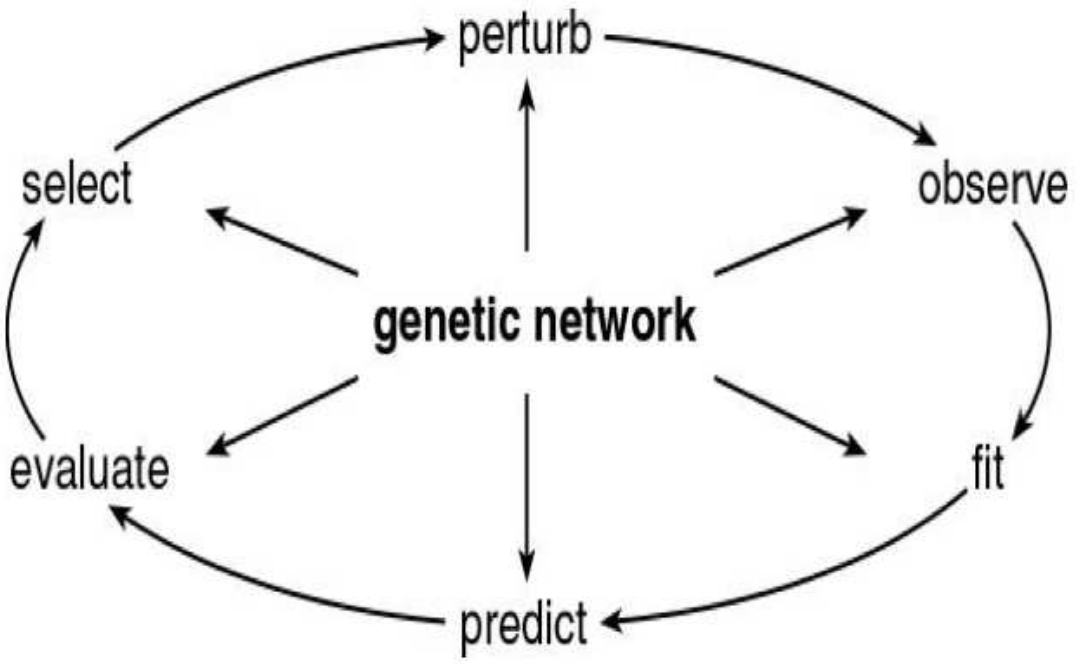

Figure 1.3: **Computing Life Paradigm.** The ‘‘perturb’’ and ‘‘observe’’ steps represent the experimentation phase; the ‘‘fit’’, ‘‘predict’’ and ‘‘evaluate’’ steps are the main components of the genetic network ensemble simulation phase; and the ‘‘select’’ step is the MINE design phase which closes the Computing Life workflow cycle. The figure has been used in [13]

In Chapter 2, I will discuss the kinetics model and the ensemble approach we used to describe and identify the genetic network for a light-responsive biological clock. Several examples of fitting, predicting and evaluating ensembles of genetic networks will also be presented. In Chapter 3, the methodology of selecting an optimal perturbation, "Maximally Informative Next Experiment" (MINE) will be discussed. In Chapter 4, the Computing Life process will be applied to the light-responsive genetic network. To illustrate this iterative process, we have traced three cycles through the Computing Life paradigm in the context of refining our network model for the biological clock's mechanism as well as mining for *clock-controlled genes* in *Neurospora crassa*. In Chapter 5, the light-responsive model will be expanded with a temperature-responsive feature, and applied to explain the so-called *temperature compensation*, i.e., the stability of a biological clock's period over the more than a 10℃ physiological temperature range.

Chapter 2

Ensemble Method of Genetic Network Identification and its Applications to the Light-Responsive Biological Clock Model

2.1 Introduction to the Kinetics Model

A biological system can be viewed as a chemical reaction network [20]. The diagrammatic represents is as shown in Fig 1.2. The species are represented as boxes. The reactions are represented by circles. The arrows indicate the directions of reactions. Then the time-dependence of the molecular species concentrations in the chemical reaction network model can be represented by coupled nonlinear ordinary differential equations (ODEs) based on mass-action kinetics.

Consider a general single-reaction process of type: $\alpha A + \beta B \rightleftharpoons \sigma S + \tau T$ with 4 species and 2 reactions (forward and backward). Suppose the forward reaction rate is k_f and backward reaction rate is k_b. The net reaction rate for this process is given by mass action kinetics according to:

$$\frac{1}{\alpha}\frac{d[A]}{dt} = -k_f[A]^{\alpha}[B]^{\beta} + k_b[S]^{\sigma}[T]^{\tau}$$

where

t is the time variable,

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

$[A]$, $[B]$, $[S]$ and $[T]$ are the concentration of the 4 species, respectively,

and and α, β, σ, and τ are the stoichiometric coefficients of the respective species.

That is, both for the forward and the backward reaction, the contribution to the net rate of production is proportional to the product of the reactant concentrations.

Similar equations for the other 3 species can be constructed. Usually the backward reaction rate is smaller, sometimes just 0. For a chemical reaction network model with multiple reactions, we can obtain the net rate of production for each species and construct the full multiplicative mass action kinetics, *i.e.*, the complete set of ODEs.

More generally, these ODEs have the form:

$$\frac{d[s]_t}{dt} = \mathrm{G}(...[s']_t..., \mathrm{t}) = \sum_r g_{r,s'} k_r \prod_{s'} [s']_t^{n_{r,s'}}, \tag{2.1}$$

where $[s]_t$ is the concentration of molecular species s at time t; $\mathrm{G}(...[s']_t..., \mathrm{t})$ is the net rate of production of species s at time t given all related species' concentration $[s']$ in related reactions r; $g_{r,s'}$is the stoichiometry, *i.e.*, the net number of molecules of species *s'* produced (if $g_{r,s'} > 0$) or consumed (if $g_{r,s'} < 0$) by the occurrence of reaction r; $n_{r,s'}$ is the cooperativity, *i.e.*, the number of molecules of species *s'* entering into reaction r as reactants. Both $g_{r,s'}$ and $n_{r,s'}$ are the parameters representing the topology of the genetic network; the rate coefficient k_r of reaction r, together with the initial concentration $[s]_{t=0}$ of species s, are the parameters that describe the dynamics of the genetics network.

This kinetic modeling method can be applied to any deterministic genetic networks regardless how complex they are. For example, the genetic network model for the biological clock of *Neurospora crassa* shown as in Figure 1.2 uniquely specifies a system of 16 ODEs that describes the temporal profiles of genes and their products shown in Figure 1.2 [14].

Since our kinetic model is now a system of ODEs, the task of simulating to the system is reduced to the numerical integration of ODEs [21]. An efficient general purpose simulator, KINSOLVER , was designed and implemented to compute the time dependent concentrations of each species with 5 standard methods (Euler, Modified Euler, Runge Kutta (RK), Adaptive RK-Fehlberg, and LSODES) [22]. With the kinetic network model and the numerical tools we are just ready to start our fitting for the experimental data. However, this approach can be implemented on a real system only if all topological ($g_{r,s'}$, $n_{r,s'}$) and dynamical (k_r, $[s]_{t=0}$) are known.

2.2 Ensemble Approach of Model Identification

2.2.1 Motivation of the Ensemble Approach

Before starting to model a genetic network, it is necessary to consider the quality and quantity of the experimental data that we are going to fit. As an example that is representative for the general case, let's take a look at the experimental data for the light responsive biological clock of *Neurospora crassa* shown in Figure 1.2.

The experimental data are time dependent microarray profiling concentration data of RNAs and their products, and the conidiation intensity in race tubes representing the concentration of CCG proteins, coming from the experiments done by our collaborators in the Department of Genetics (see dots in Figure 2.3) [13] and the digitization of published literature graphs [9, 14, 23–25].

One would notice that for our situation, and more commonly in systems biology research, the available experimental data generated by the profiling experiments are usually sparse, incomplete and noisy [26]. The RNA profiling data are available for only a few molecular species at a limited number of time points. For example, in our situation, when we started to fit the model shown in Figure 1.2, we had initially only data for 5 species (the FRQ protein, the WC-1 protein, the *wc1* RNA, the *frq* RNA and the conidiation density, *i.e.* the concentration of CCG protein) yielding 183 data points altogether [14].

Another issue to be noticed is that in our simulation, all species concentrations are to be measured and represented in a common but unknown "model unit" of concentration (cu), and all rate coefficients are in units of $1/(hour \times cu^{k-1})$ for a reaction of kth order (*i.e.*, having k reactants). This brings up the unknown unit conversion factors by which the experimental data are converted into model units. These factors should be counted as part of the model parameters which need to be inferred during our study. For the subsets of experimental data points measured under identical conditions, the unit conversion factors would be the same.

For example, there are 5 independent unknown unit conversion factors required for the data we used to fit the light-response model shown in Figure 1.2 .

Now when we start to model a genetic network such as shown in Figure 1.2, what we have is a kinetics model that is rich in unknown parameters. Essentially, all of the kinetic model parameters are unknown, including, *e.g.*, the initial concentration of the molecular species, the reaction rate coefficients and the unit conversion factors. In the above model to fit the experimental data set requires 16 initial species concentrations, 26 rate coefficients, and 5 unknown concentration unit conversion factors, which add up to 47 parameters. However, as we have seen above, the sparse, incomplete, and noisy data set we have is only a poor constraint for these 47 unknown parameters. In fact this difficulty is quite fundamental and ubiquitous in systems biology [27].

To solve the problem and obtain a meaningful comparison of the model to the data, we have used a novel ensemble method [28, 29] of genetic network 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 [30]. Instead of trying to identify one unique model parameter set, our goal in this ensemble method is to generate a large, random sample, *i.e.,* an ensemble, of models that are consistent with the available RNA and protein profiling data. In the ensemble method implemented as a Monte Carlo (MC) simulation technique [28], 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 [14]. The model ensemble is a probability distribution on the parameter space of rate coefficients and initial concentrations. The Q value in this context is the likelihood that the genetic network model in Figure 1.2 could have given rise to the observed profiling data, calculated as a function of the model parameters, *e.g.*, k_r, $[s]_{t=0}$ and others we have explained above. This approach has now been applied to several different genetic networks [31].Below is a more detailed, formal description of the ensemble approach.

2.2.2 A Formal Description of the Ensemble Approach

Let the M-dimensional vector $\boldsymbol{\theta} := (\theta_1,\ldots, \theta_M)$ denote the natural logarithms (ln) of unknown parameters that determine the model, including unknown reaction rate coefficients, unknown initial concentrations of species, and unknown unit conversion factors we just described. For example, for the full model in Figure 1.2 we will have M=47. And for short, $\boldsymbol{\theta}$ is referred to as "the model". Our ensemble of models is then formally described in terms of a probability distribution on the model space of all model $\boldsymbol{\theta}$s with the likelihood function $Q(\boldsymbol{\theta})$.

Suppose that in a series of M_e experiments labeled by $e = 1, \ldots, M_e$, the concentrations $[s]$ of species s are measured at time points t. Define $Y_l := ln([s]^{(x)}_{t,e})$ where $l := (t, s, e)$ labels the measured concentration of species s in experiment e at time t in logarithms, and $[s]^{(x)}_{t,e}$ denotes by the superscript *(x)* that the concentrations here are the measured values from the M_e experiments in various experimental or detector units, such as microarray reading , or photon or radioactive decay count units. Then let$\mathbf{Y} := (Y_1, \ldots, Y_D)$ denote the D-dimensional vector of all Y_l values. Similarly let $\mathbf{F}(\boldsymbol{\theta}) := (F_1(\boldsymbol{\theta}), \ldots, F_D(\boldsymbol{\theta}))$ denote the vector of corresponding predicted values of the observables in $\mathbf{Y}$ for a given model $\boldsymbol{\theta}$, which are calculated by numerically solving the system of rate equations of the network using the model parameters (*i.e.*, the initial concentrations and the rate coefficients) given by $\boldsymbol{\theta}$ and then calculating the predicted log-concentration

$$F_l(\boldsymbol{\theta}) := ln(\phi_{s,t,e}[s]_{t,e}), \tag{2.2}$$

for all D observables [22]. $[s]_{t,e}$ here without the superscript denotes the predicted species concentrations in the model unit cu which was described in 2.2.1. $\phi_{s,t,e}$ denotes the unknown unit conversion factor from the model unit to the various units of the experimental data.

It is reasonable to assume, but not fundamental to the ensemble method, that the probability distribution $P(\boldsymbol{Y})$ of the data $\mathbf{Y}$ is representable as a multivariate Gaussian, without error correlations between different data points Y_l, as

$$P(\boldsymbol{Y}; \mu) = const \times e^{-\chi^2/2}, \tag{2.3}$$

with

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

where μ_l denotes the mean values and σ_l denotes the standard deviation. σ_l is estimated to be $\sigma_l \cong 0.14$ for all log-concentration data points Y_l [14].

A given $P(\boldsymbol{Y};\mu)$ of course would not uniquely determine the model. There is then an infinite manifold of $\boldsymbol{\theta}$ which is consistent with the data distribution $P(\boldsymbol{Y})$, and our choice here is simply to take $P(\boldsymbol{Y};\mu)$ as the likelihood (in which the experimental data $\mathbf{Y}$ are viewed as fixed) to determine a model ensemble $Q(\boldsymbol{\theta})$. Thus the parameters $\boldsymbol{\theta}$ are distributed according to the following likelihood:

$$Q(\boldsymbol{\theta}) = P(\boldsymbol{Y}; F(\boldsymbol{\theta})) = \Omega^{-1}W(\boldsymbol{\theta}) = \Omega^{-1}\exp[-\chi^2(\boldsymbol{Y}; F(\boldsymbol{\theta}))/2], \tag{2.4}$$

with normalization factor $\Omega := \Sigma_{\boldsymbol{\theta}} W(\boldsymbol{\theta})$ where $\Sigma_{\boldsymbol{\theta}}$ denotes the integration over all M components of $\boldsymbol{\theta}$.

With the likelihood distribution function $Q(\boldsymbol{\theta})$ constructed, standard data-fitting methods, such as maximum likelihood, least-squared fitting and maximum entropy approaches, could be used to construct the correct model by finding the unique $\boldsymbol{\theta}$ which minimizes $Q(\boldsymbol{\theta})$. However, such approaches are bound to fail in our situation due to the large number of unknown model parameters and sparsity and noise of the experimental data here. Thus instead of attempting to find a unique $\boldsymbol{\theta}$ which is not warranted by the quantity and quality of the underlying data, we take all $\boldsymbol{\theta}$ as possible candidates for the correct model which reasonably reflects a $\boldsymbol{\theta}$'s degree of consistency with the data. This collection of models defines a probability distribution $Q(\boldsymbol{\theta})$

If the weight function $W(\boldsymbol{\theta})$ is analytically known or numerically calculable, we can then evaluate the ensemble average of any quantity $G(\boldsymbol{\theta})$,

$$\mathrm{E}[G(.)] := \sum_{\boldsymbol{\theta}} G(\boldsymbol{\theta})Q(\boldsymbol{\theta}) = \left[\sum_{\boldsymbol{\theta}} G(\boldsymbol{\theta})W(\boldsymbol{\theta})\right] / \left[\sum_{\boldsymbol{\theta}} W(\boldsymbol{\theta})\right],$$

where E[...] denotes the mean over the ensemble probability $Q(\boldsymbol{\theta})$. And the ensemble standard deviation can be evaluated as

$$\sigma_G = (\mathrm{E}[G(.)^2] - \mathrm{E}[G(.)]^2)^{1/2}$$

But it is actually impossible to explore all the possible $\boldsymbol{\theta}$ space. In the practical implementation of the ensemble method, a random sample of $\boldsymbol{\theta}$ is generated by the Monte Carlo (MC) method [32] using a *Metropolis algorithm* as we will describe in next subsection to construct $\boldsymbol{\theta}$, and the averages $\mathrm{E}[G(.)]$ are then approximated by averages over the Monte Carlo sample.

2.2.3 Monte Carlo implementation of the Ensemble method

The ensemble approach is implemented, as we mentioned at the end of last session, by a Monte Carlo simulation method. To generate a random sample of $\boldsymbol{\theta}$, a standard *Metropolis algorithm* is used: starting from some initial $\boldsymbol{\theta}^{(i)}$, a Markovian random walk is generated through $\boldsymbol{\theta}$-space. For each step of the walk, a random change is proposed to either one randomly selected $\boldsymbol{\theta}$-component ("local update") or simultaneously to all $\boldsymbol{\theta}$-components ("global update"). For both local and global updates, the proposed new θ' is generated so that it is randomly distributed according to a proposal probability $T_p(\theta \to \theta')$ which is symmetric, *i.e.*, $T_p(\theta \to \theta') = T_p(\theta' \to \theta)$. The proposed $\boldsymbol{\theta'}$ is then probabilistically either accepted or rejected so that the random walk would either walk to the proposed $\boldsymbol{\theta'}$(accepted) or stay at the old $\boldsymbol{\theta}$(rejected). The random walk will converge to a terminal equilibrium distribution after a large number of such updating steps, which is designed to be the desired ensemble distribution $Q(\boldsymbol{\theta})$. In order to converge, the transition rate $T(\boldsymbol{\theta} \to \boldsymbol{\theta'})$ from $\boldsymbol{\theta}$ to $\boldsymbol{\theta'}$ then must satisfy the detailed balance condition:

$$Q(\boldsymbol{\theta})T(\boldsymbol{\theta} \to \boldsymbol{\theta'}) = Q(\boldsymbol{\theta'})T(\boldsymbol{\theta'} \to \boldsymbol{\theta})$$

Then

$$\frac{Q(\boldsymbol{\theta})}{Q(\boldsymbol{\theta'})} = \frac{T(\boldsymbol{\theta'} \to \boldsymbol{\theta})}{T(\boldsymbol{\theta} \to \boldsymbol{\theta'})}$$

where

$$\frac{Q(\boldsymbol{\theta})}{Q(\boldsymbol{\theta'})} = \frac{W(\boldsymbol{\theta})}{W(\boldsymbol{\theta'})}.$$

Thus we don't have to evaluate the normalization factor $\boldsymbol{\Omega}$ since only the ratios of probabilities $\frac{Q(\boldsymbol{\theta})}{Q(\boldsymbol{\theta'})} = \frac{W(\boldsymbol{\theta})}{W(\boldsymbol{\theta'})}$ enters the calculation. The new $\boldsymbol{\theta'}$ is accepted with $T_a(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta'})$ and rejected with $1 - T_a(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta'})$, where $T_a(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta'}) = \min(1, \frac{W(\boldsymbol{\theta'})}{W(\boldsymbol{\theta})})$, where T_a is the Metropolis acceptance probability [32]. Combining $T_a(\theta \rightarrow \theta')$ with the proposal probability $T_p(\theta \rightarrow \theta')$ will then obey the detailed balance condition, due to the fact that $T_p(\theta \rightarrow \theta')$ is symmetric in θ and θ'.

In the actual simulation run, not all $\boldsymbol{\theta}$-components are updated with the above procedure. The unknown, independent unit conversion factors in the logarithm ($ln(\phi_{s,t,e})$) are separately chosen to maximize $Q(\boldsymbol{\theta})$. Only the remaining $\boldsymbol{\theta}$-components (unknown initial species concentrations and unknown reaction rate coefficients) are subjected to the random Metropolis updating procedure, using the maximized $Q(\boldsymbol{\theta})$ as the terminal distribution [14]. From eq. 2.4 and eq. 2.2, it is easy to see that the original $Q(\boldsymbol{\theta})$ is Gaussian dependent on $ln(\phi_{s,t,e})$, therefore the reduced MC procedure which only carries the ratio of probabilities is mathematically equivalent to the full MC procedure of updating all $\boldsymbol{\theta}$-components including all the unit conversion factors. Correspondingly, χ^2 used in the actual run is also a reduced value which is minimized with respect to the independent $ln(\phi_{s,t,e})$. This is the situation for all the χ^2 shown in Figure 2.1 and 2.4. An example of a typical minimizing process is shown in Figure 2.1. It can bee seen in the figure that the χ^2 first decreases very quickly with progressive Monte Carlo (MC) sweeps in the parameter space. It is stuck at about 3600 for a while, and then the MC process takes a risk and permits the χ^2 to increase, thereby allowing the walk to escape from local minimum of χ^2, which is just the situation illustrated here.

After the Markov chain converges, the generated random sample of $\boldsymbol{\theta}$ has a equilibrium distribution which is approximately the ensemble distribution. By the Monte Carlo importance sampling method, when the sample size I is very large, the ensemble average calculated

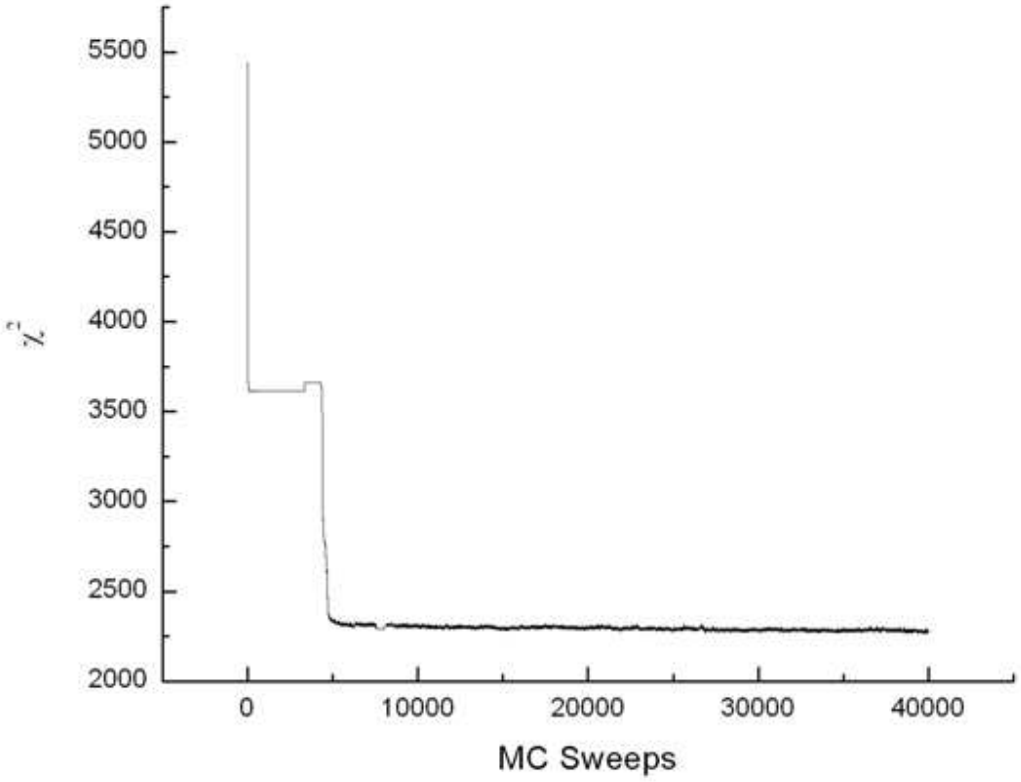

Figure 2.1: **A typical example of Monte Carlo random walk equilibration in the parameter (θ) space of the models in actual simulation.**Progress toward equilibrium is monitored by $\chi^2 = -2\ln Q + const$, which is a measure of the departure between the data and the model prediction. It can bee seen in the figure that the χ^2 decreases with progressive Monte Carlo (MC) sweeps in the parameter space. It is stuck at about 3600 for a while, and then the MC process takes a risk and permits the χ^2 to increase, thereby allowing the walk to escape from local minimum of χ^2, which is just the situation illustrated here.

from the Monte Carlo sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(I)}$ is then

$$\mathrm{E}[G(.)]_{MC} := \frac{1}{I} \sum_{i=1}^{I} G(\boldsymbol{\theta}^{(i)}),$$

where $G(.)$ could be the square of any unknown parameter. The ensemble standard deviation is calculated based on the same Monte Carlo sample,

$$\sigma[G(.)]_{MC} = (\mathrm{E}[G(.)^2]_{MC} - \mathrm{E}[G(.)]_{MC}^2)^{1/2}$$

2.3 Results and discussions for the light-responsive biological clock of *Neurospora crassa*

In terms of the Computing Life paradigm shown in Figure 1.3, the ensemble approach of genetic network identification is the stages that "fits" the genetic network models to the experimental data (*i.e.*, generates model ensemble consistent with the data), makes "predictions" with the generated model ensemble using ensemble averages, and "evaluates" the goodness of fitting by comparing the predictions with the experimental data using χ^2 as the merit function.

The genetic networks shown in this thesis are based on Yu's identification for a genetic network of the biological clock of *N. crassa* [14,21] which successfully explained the literature data both in the dark and in artificial days of different periods [9, 23–25]. However, due to the very limited knowledge about exactly how the biological clock functions at the molecular level, there exist different hypotheses, *i.e.*, alternative models, for the biological clock of *N. crassa*. In the work shown in this thesis, we have worked with different models. The genetic network shown in Figure 2.2 is one of the alternative models other that the one in Figure 1.2. This one has served as the main genetic network model throughout the three cycles of Computing Life process, which we will see in Chapter 4. In contrast with the model in Figure 1.2, this model features two forms of WCC: a dark version (not light responsive, denoted as $\mathrm{WCC}^{\mathrm{D}}$) and a light version (capable of photon absorption, denoted as $\mathrm{WCC}^{\mathrm{L}}$). The dark/light forms then result in dark and light forms of *frq* and *ccg* when WCC binds

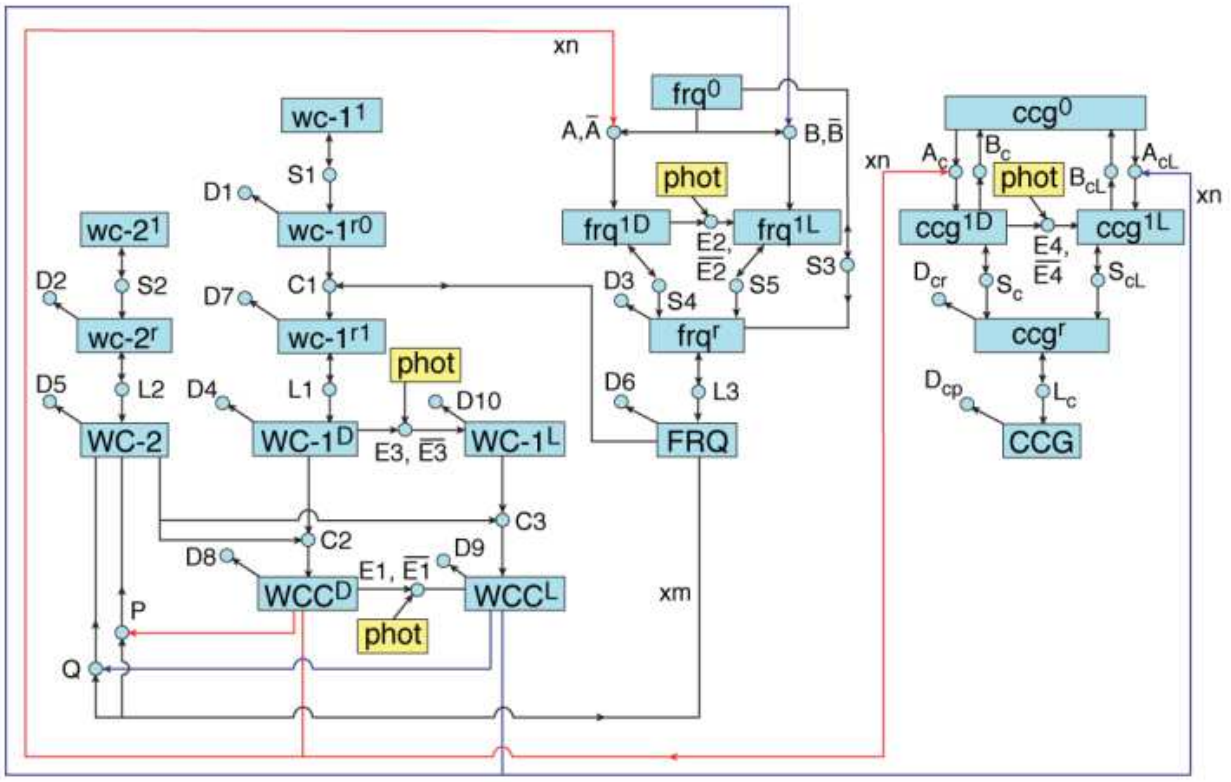

Figure 2.2: **Alternate genetic network for the biological clock**. Molecular species (*i.e.*, reactants or products) in the network are represented by boxes. The terms are the same as described in the legend of Figure 1.2. The main difference is that the WCC has a light and dark form denoted $\mathrm{WCC^D}$ and $\mathrm{WCC^L}$. When these two forms bind upstream of *frq* and *ccg* genes, it leads to two different transcriptionally active forms of the gene, such as frq^{1D} and frq^{1L}. In addition, photons (in yellow) can enter the system to interact with WC-1 in four ways, depending on the bound state of the WCC, in the reactions E1, E2, E3, and E4. All four of these reactions have been given nonzero back reaction rates. The final difference is that the two forms of WCC lead to two deactivation reactions of WCC by FRQ, labeled P and Q. Reactions, such as A and P, have cooperative kinetics: (A) $n\mathrm{WCC^D}+frq^0 \rightarrow frq^1$ and (P) $\mathrm{WCC^D}+m\mathrm{FRQ} \rightarrow \mathrm{WC}-2+m\mathrm{FRQ}$. The n and m in these two reactions are called Hill coefficients or cooperativities. Only 6 reactions, such as the reaction A, has a back reaction with non-zero rate, *e.g.* $(\bar{\mathrm{A}})frq^{1D} \rightarrow n\mathrm{WCC^D}+frq^0$. This figure has been used in [13]

to the upstream region of *frq* and *ccg* genes, respectively. In addition, it is believed that the photon absorbing feature of WCC comes from WC-1. Therefore, photons can enter the system to interact with WC-1 in four ways, depending on the bound state of the WCC, in the reactions labeled E1, E2, E3, and E4 in the network diagram. The final distinction is that the two forms of WCC lead to two deactivation reactions of WCC by FRQ, labeled P and Q. The model selection between this model and the one in Figure 1.2 will be discussed in subsection 2.3.2.

2.3.1 Experimental Data Fitting of Data in Three Cycles

As mentioned in the introduction and as will be discussed in detail in Chapter 4, our Computing Life paradigm was designed with the biological goal of searching for *clock controlled genes* (*ccg*) in the genome of *Neurospora crassa*, and three series of experiments have been performed, which are labeled cycle 1, cycle 2 and cycle 3, respectively. In each cycle, the concentrations of *wc-1* (in fact the total of $wc-1^{r0}$ and $wc-1^{r1}$), *wc-2*, *frq* mRNA were measured by microarray profiling. In cycle 1 the organism was grown in the dark for 48 hours; in cycle 2, a 48hr culturing was done with an artificial day of 24 hrs in the dark followed with 24 hrs in the light; cycle 3 involved a knock-down of *wc-1* genes in which *wc-1* was reduced to only 30% activity of the wild type.

The ensemble fitting, as we have mentioned at the beginning of this section, used the genetic network model in Figure 2.2. Note that although not displayed here, the literature data [9,23–25] were also included as well as data from the three cycles to keep the consistency with previous work [14]. This rule was followed in each cycle, *i.e.*, the data in each cycle were cumulative. As shown in Figure 2.3, the dots are the experimental data with at least 5 duplicates at the same time points. The curve is the mean prediction of the ensemble (ensemble average), and the gray band shows the 2 times ensemble standard deviation about the ensemble mean. The bars with either gray or white alternatively represent the light exposure to which the growing organism was subject.

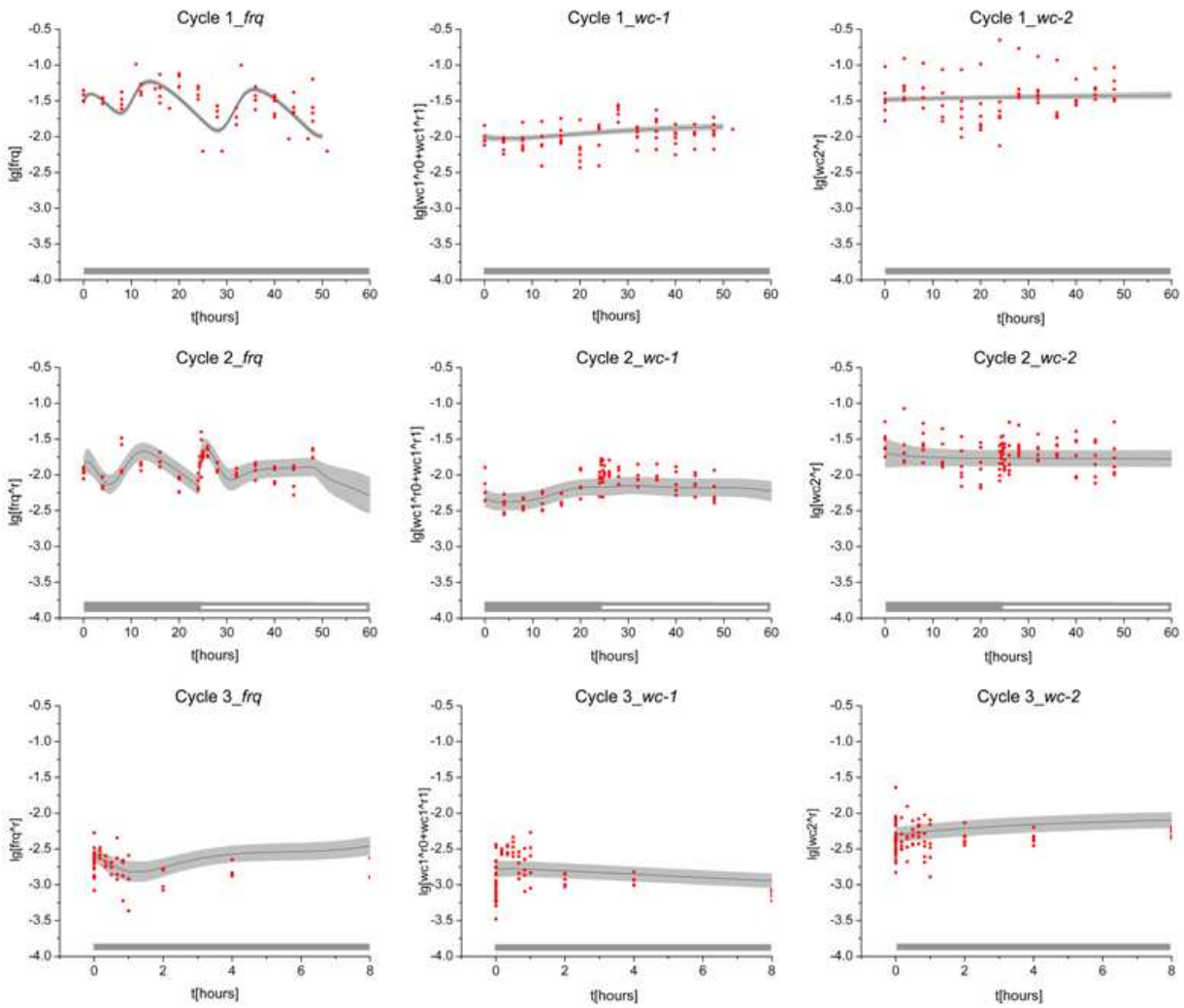

Figure 2.3: **An ensemble of genetic networks predicts the mRNA levels of wc-1, wc-2, and frq for cycles 1–3 by the model shown in Figure 2.2**. The decadic log (lg) of each gene's mRNA level is measured at least 5 times on an array for each time point. Some data points are from the literature [9, 25]. The curves represent the mean prediction of the ensemble of genetic networks in Figure 2.2 $\pm$ 2 ensemble standard deviations about the ensemble mean. Grey bars denote lights off; white bars denote lights on. This graph has been published in [13].

In cycle 1 where the *frq* gene showed an obvious oscillation rhythm in the biological clock, the predicted oscillation by the model ensemble displayed great consistency. For *wc-1* and *wc-2* mRNAs, the prediction also showed consistency with much reduced oscillations comparing with *frq* mRNA. In cycle 2, the light was turned on after 24 hrs of dark. Both the microarray data and the ensemble mean prediction reflected a coordinated response to the light in the *frq* genes, which can be interpreted as a reset of the biological clock caused by turning on the light. In cycle 3 which involved inhibition of the *wc-1* gene to 30% of its original activity, the ensemble prediction successfully displays the slow decay of the concentration level of *wc-1*, while this decay observed over 8 hrs in the microarray data also supported a predicted lifetime of 7.4 hrs by the ensemble. More detailed discussion of the 3 cycles procedure in the point of view of an iterative Computing Life paradigm, including the refining progress of the genetic network model, can be found in Chapter 4.

2.3.2 Model Selection Between Alternative Genetic Networks

χ^2 defined in subsection 2.2.2 has been used as the merit function to evaluate the goodness of fit to the data by the ensemble of genetic networks. Therefore according to the distribution of χ^2 obtained under same conditions (*i.e.*, based on the same data sets), we can compare between different models and make selections [14].

The first comparison shown here is between the genetic networks model shown in Figure 1.2 and in Figure 2.2. Although the latter is the genetic network we have been used in all 3 cycles of ensemble fitting throughout the whole Computing Life paradigm for finding *ccg* genes (refer to Chapter 4), later development of the former one in Figure 1.2 did slight better with a minimum chi-square of 4240 versus that of 4474 for the network in Figure 2.2. The comparison is shown in Figure 2.4 (it is $\chi^2/2$ that is displayed in the figure), based on the same data set. The model in Figure 1.2 also contains fewer parameters. Therefore by both chi-square and Occam's Razor, the simpler network with fewer parameters in Figure 1.2 is more preferred and is used in later work.

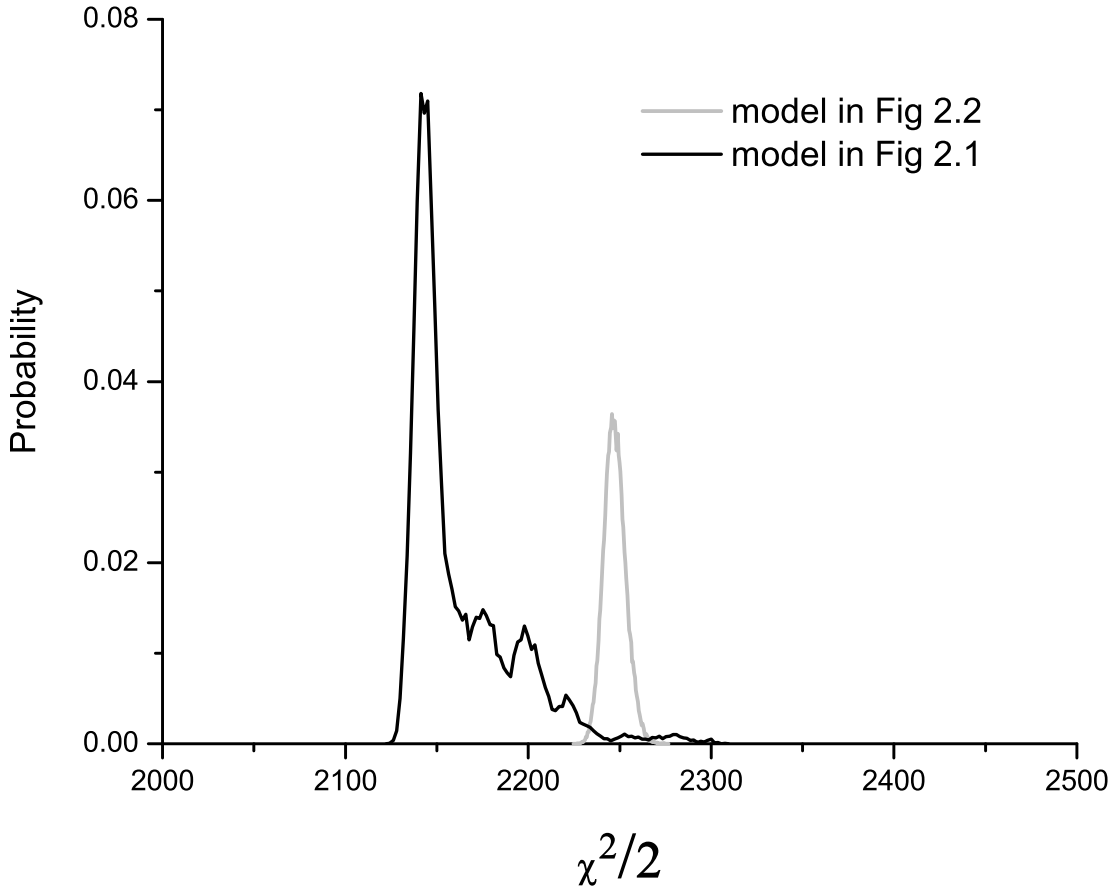

Figure 2.4: **Comparison of model fits.** The histogram of values of χ^2 (as defined in Eq. 2.4)is shown for model ensembles of genetic network shown in Figure 2.2 and Figure 1.2 respectively. The model in Figure 1.2 does have a smaller chi square and also contains less parameters than the model in Figure 2.2, which makes us finally choose it as our model in future work.

The second comparison is then based on the selected model in Figure 1.2. The details of this selected model, especially the details of the P-reaction, are still subject to debate. It has been proved mathematically that to keep the clock ticking, *i.e.*, to obtain oscillations in the genetic network, the minimum requirement is that the choice of the Hill coefficients, n and m, must satisfy $nm > 4$ [14]. There is also evidence reported that FRQ acts as a dimer [33], which would suggest $m = 2$. Therefore, four model ensembles had been identified with genetic network models of varying Hill coefficients, with $n = m$ and $n = 4, 3, 2, or 1$ based on literature data set only in [14]. It turned out to be that $n = m = 3$ was substantially less in χ^2 than other models. However with limited duration data at that time, the comparison of chi-square statistics failed to discriminate between undamped oscillatory ($n = m = 4$) and weakly damped oscillatory models (*e.g.*, $n = m = 2$). With cumulative data shown in Figure 2.3, we can now test again the appropriate Hill coefficients sets. As shown in the upper part of Figure 2.5, the most obvious result is that the $n = m = 2$ model is outperformed clearly with a huge χ^2. The $n = m = 3$ keeps doing well and shows low chi-square statistics, supporting the corresponding most robust version of a simplified stochastic model of $n = m = 3$ [34]. However $n = 4, m = 2$ slightly beats the $n = m = 3$ set with the lowest chi-square on average by far. The current model $n = m = 4$ shows a chi-square distribution very close to that of the $n = m = 3$ model. All three models have a large overlap with each other thus are not discriminated by current data set. Although currently not used in the simulation, the results do suggest to revisit the $n = 4, m = 2$ or $n = m = 3$ cases in future work.

A similar comparison also has been applied to possible mechanisms for the P-reaction. There are different proposed versions of P-reactions about how WCC is deactivated by FRQ [12, 35]. To test these 4 distinct hypotheses about the deactivation of WCC (as shown in Figure 2.5B), the likelihood function under each hypothesis was reconstructed using the ensemble method. Previous analysis with model ensemble generated by only the literature data supported our current choice of using 4FRQ + WCC $\rightarrow$ 4FRQ + WC-2 as our P-reaction, but could not distinguish it from 4FRQ + WCC $\rightarrow$WCC/FRQ, and 4FRQ +

WCC →3FRQ +WCC/FRQ [14]. With more data from cycle 1 to cycle 3 added into the simulation, it is very interesting to see that 4FRQ + WCC→4FRQ + WC-2 is still our best choice, outperforming all others. The other two hypotheses 4FRQ + WCC →WCC/FRQ, and 4FRQ + WCC→3FRQ +WCC/FRQ are suggested to fail by a much larger chi-square. And surprisingly 4FRQ + WCC→4FRQ + WC-2+ WC-1, which previously had slightly worse chi-squares, now has a second lowest set of chi-squares quite close to our best choice. The addition of data does bring us a clearer view and to successfully discriminate model under different hypotheses.

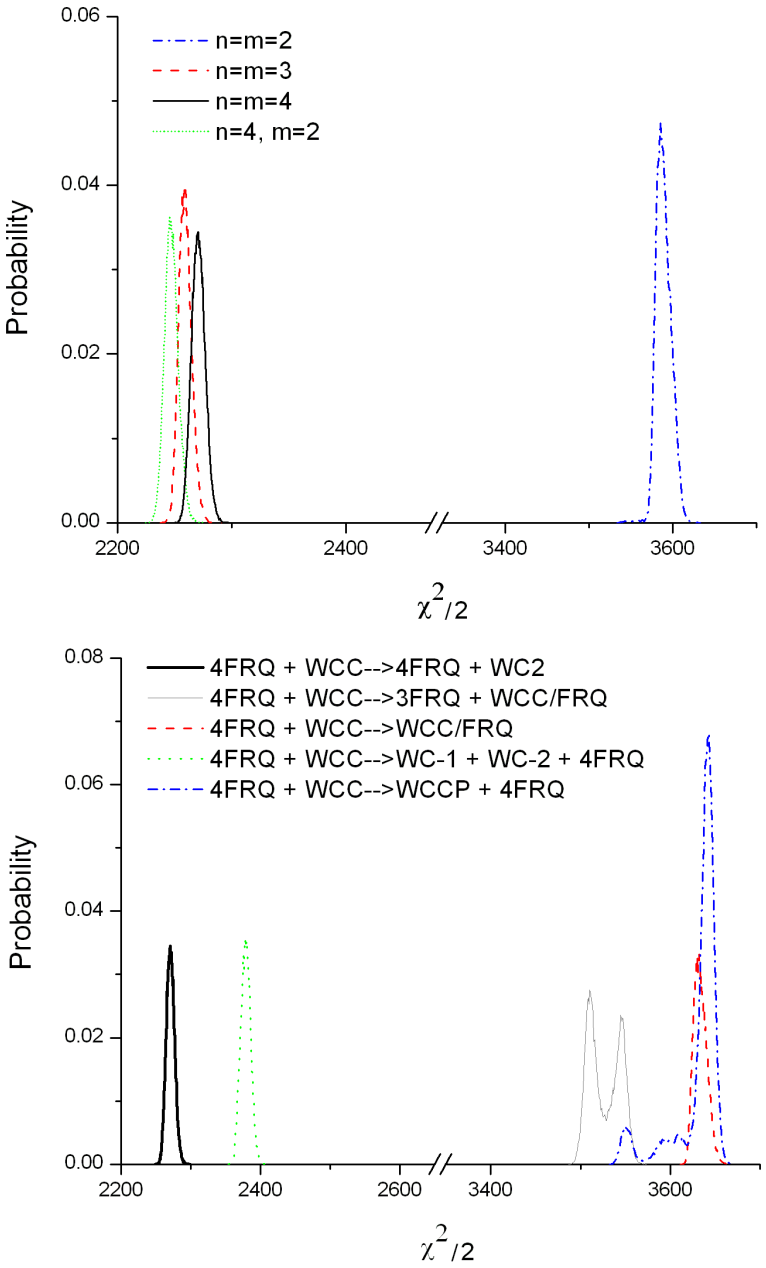

Figure 2.5: **Another comparison of model fits.** A histogram of values of χ^2 is shown for varying model ensembles. In the upper panel models with different Hill coefficients are compared. In the bottom panel, model with different P-reactions are compared. A smaller χ^2 is preferred.

Chapter 3

Maximally Informative Next Experimental (MINE) Design

3.1 Introduction

The model ensemble method described in Chapter 2 summarizes what we know, and equally importantly, what we do not know about the biological network, given the prior or "old" experimental data. The next step in our Computing Life paradigm as shown in Figure 1.3 is to select a perturbation as the next round of experiments. Since the profiling experiments are costly, we would like the designed perturbation experiments to be maximally informative to gain the maximal information about the genetic network. In order to select such an optimal perturbation, we introduce the novel method of evaluating the Maximally Informative Next Experiment (MINE) [13].

In Chapter 2, the kinetics model is represented by a system of ODEs which reveals the topological and dynamical details of the genetic network providing the *prior* experimental data. If we change the point of view and include the experiment design together with the model parameter selection, the system of ODEs in fact gives the model prediction if the model parameters and experimental conditions are provided, eq. 2.1 now becomes:

$$\frac{d\mathbf{S}_t}{d\mathrm{t}} = \mathbf{G}(\mathbf{S}_t; \boldsymbol{\theta}, \mathbf{u}), \tag{3.1}$$

where $\mathbf{S}_t = ([s]_{t,1}, \ldots [s]_{t,N})^{\mathrm{T}}$is a N×1 vector of species concentrations, with N denoting the number of molecular species evolving according to the kinetics rate equations; $\mathbf{G} = [\mathrm{G}_1, ...\mathrm{G}_\mathrm{N}]^{\mathrm{T}}$ specifies the kinetics, *i.e.*, $\mathrm{G_n}(\mathbf{S_t}; \boldsymbol{\theta}, \boldsymbol{u})$ is the net rate of production of species n at time t, given the species concentration $\mathbf{S}_t$. The model parameter vector $\boldsymbol{\theta} \equiv [\theta_1, ...\theta_M]^T$

in $\mathbf{G}$ is the M×1 vector of all $\boldsymbol{\theta}$ model parameter variables as we have discussed in previous chapters, and for short, is referred to as "the model". The rate functions $\mathbf{G}$ now also explicitly depend on an array of control variables, $\mathbf{u}$ of unspecified array dimension, which are known and can be varied by the experimenter. These control variables specify, for example, the nature of the perturbations and external conditions to be applied to the biological system. They comprise all control variables defining the specific data point to be measured, including, for example, the choice of molecular species to be observed and the time of observation. More generally, $\mathbf{u}$ specifies the experiment to be done.So for short it will be referred to as "the experiment" in the following.

Then for a given choice of model $\boldsymbol{\theta}$, let $f(\boldsymbol{\theta}, \boldsymbol{u})$ denote the prediction for a single species log-concentration $\log(y)$ to be measured for a single time point by the next profiling experiment, where y is one of the elements of $\mathbf{S}_t$ to be measured at some specific observation time t. The vector $\boldsymbol{u}$, as explained above, would specify *all* control variables defining the specific date point y to be measured. For the planned next experiment measuring multiple variables $y_1, ...y_d$, let $F(\boldsymbol{\theta}, \mathbf{U}) := [f(\boldsymbol{\theta}, \boldsymbol{u}_1), ...f(\boldsymbol{\theta}, \boldsymbol{u}_d)]^T$ denote a $d \times 1$ vector of the corresponding predicted outcomes in logarithms, and $\mathbf{U} := [\boldsymbol{u}_1, ...\boldsymbol{u}_d]$ denote the supervector of corresponding control parameter vectors $\mathbf{u}_i$ where $\mathbf{u}_i$ specifies the control variables for the measurement of the data point y_i for $i = 1, ...d$. That is to say, $F(\boldsymbol{\theta}, \mathbf{U})$ is "the prediction" for "the observables" $\log(y_i)$ in "the next experiment" $\mathbf{U}$. We are trying to find a maximally informative perturbation in the next experiments.

Clearly, it is not a mathematically well-defined problem: which next experiment $\mathbf{U}$ is "maximally informative". We have to make an *ad hoc* choice for a design criterion and then try it out in real-life applications. The basic conceptual ideas underlying this *ad hoc* construction of a MINE criterion are borrowed from microscopy in which we need to "image" the inner workings of the cell. A microscope generates images of the cell's material components in the three-dimensional *physical space*. By analogy the profiling experiments generate images of the cell's high-dimensional *kinetics parameter space*.

Just as we always want to obtain a sharp and clear image through the microscope, ideally we would like to obtain "images" with high resolution to determine accurately a genetic network's "location" in kinetics parameter space, which is specified by a unique choice of parameter vector $\boldsymbol{\theta}$. However from the present-day profiling experiments the "images" we *do* obtain do not allow us to completely re-construct the $\boldsymbol{\theta}$: our "vision" in $\boldsymbol{\theta}$-space is seriously blurred so that we cannot "locate" exactly a unique $\boldsymbol{\theta}$. As we explained in Chapter 2, given the *prior* experimental data, we are able to find the model ensemble $Q(\boldsymbol{\theta})$ from *what we know* by imposing constraints on $\boldsymbol{\theta}$; however the $Q(\boldsymbol{\theta})$ is spread out within those constraints in $\boldsymbol{\theta}$-space, which also presents the blurring caused by *what we do not know*. Our goal is therefore to reduce this blurring as much as possible in the *next* experiment to be performed: we want to tune our "microscope" on model parameters to get a view of $\boldsymbol{\theta}$-space different with what we have seen with the maximum possible resolution.

The imaging procedure of a microscope is, from the model point of view, to relate, or say to map the observed image (F) to the underlying object ($\boldsymbol{\theta}$) through a mathematical model and a mapping function $F(\boldsymbol{\theta}, \mathbf{U})$ then captures this image model. It is important to have such an imaging model; otherwise, we cannot, for example, reconstruct the shape, size and location ($\boldsymbol{\theta}$) of a cellular organelle from the light intensity pattern (F) of the cell's magnified image produced by an optical microscope. This model for the optical microscopy is now well-established, highly reliable: it is simply the physical optics. Our "imaging procedure" for RNA profiling experiments, by analogy relates the observables F to the underlying $\boldsymbol{\theta}$-vector. The appropriate model framework may well be mass action kinetics as we have introduced in eq. 2.1. But the details of the kinetics model are still very much subject to debate.

As we have seen in Chapter 2, the profiling experiment's data are sparse and noisy. Thus, unlike in the microscope in which it is easy to look in all different directions, we do not have sufficient experimental data with sufficient diversity to look in to all possible directions of the kinetic parameter space. Each experiment only yields a (in general non-linear) projection of the object points $\boldsymbol{\theta}$ in the M-dimensional kinetics parameter space onto the image points

F in d-dimensional image space. It is implied by the sparsity and noise of the experimental data that typically only a lower dimensional image sub-space, of dimension $\mathrm{d_{eff}} < \mathrm{M}$, can be actually resolved by the experiment. The MINE experimental design procedure cannot eliminate the blurring of the vision, but it can help to minimize the blur.

3.2 The First Criterion : MINE by maximal distance in image space

In order to develop a quantitative MINE criterion, first consider the simplest case: the maximally informative next experiment involves measuring only one single data point y. Suppose from the model ensemble Q deduced from the *prior* experimental data, two possible choices of models are randomly drawn, denoted by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Both of the two models would give predictions consistent with the prior experiments within the experimental uncertainties. A next experiment with control vector $\boldsymbol{u}$ is planned in order to distinguish between these two choices. The predicted outcomes for the two models would then be $f(\boldsymbol{\theta}, \boldsymbol{u})$ and $f(\boldsymbol{\theta}', \boldsymbol{u})$, respectively. In our analogy to microscopy, the resolving power of a microscope is the ability of the microscope to measure the angular *separation* of images close together. The further the separation is, the higher the resolving power that the microscope would have. The crucial point here in our MINE to reduce the "blurring" for a better resolution in the parameter space is: the more the two predictions *differ* from each other, the better the next experiment will allow us to discriminate between the two choices, *i.e.*, the more informative the next experiment will be.

It is necessary to choose a "metric" of the difference between the two model choices. One possible choice could be, for example, the square of the difference of the two predicted outcomes:

$$V_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{u}) = [f(\boldsymbol{\theta}, \boldsymbol{u}) - f(\boldsymbol{\theta}', \boldsymbol{u})]^2/2 \tag{3.2}$$

The Maximally Informative Next Experiment $\boldsymbol{u}$ is then the one that maximizes this difference metric.

Since the two $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are randomly drawn from Q, their joint distribution would be $Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = Q(\boldsymbol{\theta}) \times Q(\boldsymbol{\theta}')$. The general idea is to apply the criterion explained above to an ensemble of models by choosing $\boldsymbol{u}$ which maximizes the average of $V_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{u})$. For the proposed $V_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{u})$ in Eq. 3.2,

$$V(\boldsymbol{u}) := \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\theta}'} V_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{u}) Q(\boldsymbol{\theta}) Q(\boldsymbol{\theta}') = \mathrm{E}[f(., \boldsymbol{u})^2] - \mathrm{E}[f(., \boldsymbol{u})]^2, \tag{3.3}$$

where $\int_{\boldsymbol{\theta}}$ denotes integration or summation over all θ-components and E[...] denotes the mean over the ensemble probability distribution $Q(\boldsymbol{\theta})$. The second equality follows immediate from $\sum_{\theta} Q(\boldsymbol{\theta}) = 1$. $V(\mathbf{u})$ is then just the variance in our prediction within the ensemble that can be evaluated by Monte Carlo methods described in the previous chapter.

In the general case of the next experiment measuring multiple variables $y_1, ...y_d$, we will replace the square of the predicted difference of the two one-dimensional outcomes, $\Delta f(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{u}) := f(\boldsymbol{\theta}, \mathbf{u}) - f(\boldsymbol{\theta}', \mathbf{u})$ in Eq. 3.2 by the corresponding squared length of the difference vector of the two d-dimensional outcome vectors in the image space:

$$\Delta F(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{U}) := F(\boldsymbol{\theta}, \mathbf{U}) - F(\boldsymbol{\theta}', \mathbf{U}). \tag{3.4}$$

The $V_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{u})$ in Eq. 3.2 is then replaced by

$$V_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{U}) = |\Delta F(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{U})^2|/2 \tag{3.5}$$

where $|...|$ denotes the Euclidean norm, *i.e.*, $|\boldsymbol{\Phi}| := (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{1/2}$ for $\boldsymbol{\Phi} = [\Phi_1, ...\Phi_d]^T$. Inserting Eq. 3.5 into Eq. 3.3, we get a MINE design criterion in the d-variable next experiment

$$V(\mathbf{U}) = \mathrm{E}[|F(., \boldsymbol{U})|^2] - |\mathrm{E}[F(., \boldsymbol{U})]|^2 = \sum_{i=1,...d} \left(\mathrm{E}[|f(., \boldsymbol{u_i})|^2] - (\mathrm{E}|f(., \boldsymbol{u_i})|)^2\right) \tag{3.6}$$

It is easy to see from Eq. 3.6 that each $\mathbf{u_i}$ could be independently chosen to make $\mathrm{E}[|f(., \boldsymbol{u_i})|^2] - |\mathrm{E}[f(., \boldsymbol{u_i})]|^2$ the largest, which would result in all $\mathbf{u_i}$ collapsing to the same $\boldsymbol{u}$ point. That is to say, this MINE criterion tries to guide us to observe exactly the same

y-variable for d times, instead of observing d independent y-variables. Clearly, this criterion lacks the ability to enforce independence of multiple observables. We need to search for a better criterion.

3.3 Criterion 2: MINE by maximal volume in image space

Now we need to construct a more useful MINE criterion, which enforces a measure of independence of the observables. Instead of the *Euclidean length* in Eq. 3.6, we consider the *volume* swept out by the image difference vector ΔF. The idea comes again from the microscopy analogy. Suppose we are trying to observe a certain volume $\boldsymbol{v_o}$ in the *object space* through our "microscope". Hence from the $\boldsymbol{v_o}$, the mapping function generates an "image difference volume" $\boldsymbol{v_\Delta}$ in d-dimensional image difference space. $\boldsymbol{v_\Delta}$ is the volume swept out by the image difference vector $\Delta F(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{U})$ for all pairs of object points $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in $\boldsymbol{v_o} \times \boldsymbol{v_o}$; or formally, $\boldsymbol{v_\Delta}(\boldsymbol{v_o}, \mathbf{U}) := \Delta F(\boldsymbol{v_o}, \boldsymbol{v_o}, \mathbf{U})$. From the notation, it is clear that $\boldsymbol{v_\Delta}$ depends on the choice of the control vector $\mathbf{U}$, as well as on $\boldsymbol{v_o}$.

The basic idea here inspired by microscopy is: the greater the volume amount contained in $\boldsymbol{v_\Delta}(\boldsymbol{v_o}, \mathbf{U})$, the more detail we should be able to discern in $\boldsymbol{v_o}$. That is to say, in order to gain more information about the contents of $\boldsymbol{v_o}$, we need to tune U, our microscope's control vector, so as to increase $|\boldsymbol{v_\Delta}(\boldsymbol{v_o}, \mathbf{U})|$, the d-dimensional image difference volume amount. Compared with the Euclidean distance criterion described in 3.1.1 which guides the y-variables to be the same, the requirement of sweeping out a higher dimensional volume $\boldsymbol{v_\Delta}$ will naturally enforce a certain degree of independence of the observables. The Euclidean norm measures just the length of the ΔF-vector and can be maximized even if ΔF sweeps out only a 1-dimensional sub-manifold (*i.e.*, the same $\boldsymbol{u}$ point). However, $\boldsymbol{v_\Delta}$ is constructed to be a higher-dimensional manifold with a dimensionality of d or M, whichever is less.

Then we need to find out how to choose an appropriate $\boldsymbol{v_o}$ or the corresponding $\boldsymbol{v_\Delta}$, in terms of the ensemble pair distribution $Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = Q(\boldsymbol{\theta})Q(\boldsymbol{\theta}')$. However, what we explained above of constructing a $\boldsymbol{v_\Delta}$ from an underlying $\boldsymbol{v_o}$ in object ($\boldsymbol{\theta}$-) space, is not a practical

solution to the question here and should only be regarded as what inspired us to introduce the $\boldsymbol{v_\Delta}$. In our practical research, we simplify the MINE with $\boldsymbol{v_\Delta}$ criterion that we do not try to construct such a $\boldsymbol{v_\Delta}$ from a given $\boldsymbol{v_o}$. Instead, a "representative" $\boldsymbol{v_\Delta}$ is defined as what is swept out by $\Delta F(\boldsymbol{\theta}, \boldsymbol{\theta'}, \mathbf{U})$ where $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$ are drawn from typical values prescribed by the ensemble pair distribution $Q(\boldsymbol{\theta}, \boldsymbol{\theta'}) = Q(\boldsymbol{\theta})Q(\boldsymbol{\theta'})$. This $\boldsymbol{v_\Delta}$ will be constructed from the characteristic variance/co-variance ellipsoid of ΔF and then it will be a dependent of the control vector U.

First let's define the ensemble distribution of ΔF as :

$$Q_\Delta(\boldsymbol{\Phi}, \mathbf{U}) := \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\theta'}} \delta(\boldsymbol{\Phi} - \Delta F(\boldsymbol{\theta}, \boldsymbol{\theta'}, \mathbf{U}))Q(\boldsymbol{\theta})Q(\boldsymbol{\theta'}) \tag{3.7}$$

where $\boldsymbol{\Phi} := [\Phi_1, ...\Phi_d]^T$ is any point in ΔF-space and $\delta(...)$ is the Dirac delta-function in d-dimensions. $Q_\Delta(\boldsymbol{\Phi}, \mathbf{U})$ is the probability density for $\Delta F(\boldsymbol{\theta}, \boldsymbol{\theta'}, \mathbf{U})$ to take on the value $\boldsymbol{\Phi}$, and $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$ are independently distributed according to $Q(\boldsymbol{\theta})$ and $Q(\boldsymbol{\theta'})$, respectively. Note that the characteristic ellipsoid of ΔF's $d \times d$ variance/co-variance matrix $D(\mathbf{U})$ is given by

$$D_{ik}(\boldsymbol{U}) := \int_{\boldsymbol{\Phi}} \boldsymbol{\Phi}_i \boldsymbol{\Phi}_k Q_\Delta(\boldsymbol{\Phi}, \boldsymbol{U})/2. \tag{3.8}$$

One can see that $Q_\Delta(\boldsymbol{\Phi}, \mathbf{U})$ then defines an effective $\boldsymbol{v_\Delta}(\mathbf{U})$ in the image difference (ΔF) space. And since $\Delta F(\boldsymbol{\theta'}, \boldsymbol{\theta}, \mathbf{U}) = -\Delta F(\boldsymbol{\theta}, \boldsymbol{\theta'}, \mathbf{U})$, $Q_\Delta(\boldsymbol{\Phi}, \mathbf{U})$ defined in Eq. 3.7 is even in $\boldsymbol{\Phi}$, *i.e.*, $Q_\Delta(-\boldsymbol{\Phi}, \mathbf{U}) = Q_\Delta(\boldsymbol{\Phi}, \mathbf{U})$. Thus, this characteristic variance/co-variance ellipsoid of ΔF is centered at the origin, $\boldsymbol{\Phi} = 0$. The squared half-axis lengths of the ellipsoid are the eigenvalues of the D-matrix, corresponding to eigenvectors that defines the respective half-axis orientations. These half-axes are orthogonal to each other, and they define a rectangular parallelotope (i.e., a high dimensional parallelepiped prism) in ΔF-space. The volume of this parallelotope is proportional to that of the ellipsoid by a universal constant prefactor. It is then natural to choose this parallelotope constructed from the variance/covariance matrix D of ΔF as our image difference volume $\boldsymbol{v_\Delta}(\mathbf{U})$, instead of the ellipsoid itself. The determinant

of $D(\mathbf{U})$ is the square of the parallelotope's volume amount, $|\boldsymbol{v}_{\boldsymbol{\Delta}}(\boldsymbol{U})|^2$ and could be used as a possible MINE criterion to be maximazed:

$$V(\boldsymbol{U}) := \det(D(\mathrm{U})) = |\boldsymbol{v}_{\boldsymbol{\Delta}}(\boldsymbol{U})|^2. \tag{3.9}$$

The above Eq. 3.9 is sometimes referred to as the *generalized variance*, whose distribution is known exactly if the distribution of the prediction is Gaussian over the ensembles [36]. However, it should be strongly emphasized here that it is *not* implied or required that the distribution $Q_{\Delta}(\boldsymbol{\Phi}, \mathbf{U})$ of ΔF is Gaussian by invoking the variance/covariance ellipsoid of ΔF. Such an ellipsoid can be constructed for any $Q_{\Delta}(\boldsymbol{\Phi}, \mathbf{U})$ from the matrix $D(\mathbf{U})$ in Eq. 3.8. Although the whole MINE approach here is *ad-hoc*, the main advantage of this constructed $\boldsymbol{v}_{\boldsymbol{\Delta}}$ is that $D_{ik}(\boldsymbol{U})$ could be deduced from Eq. 3.4, Eq. 3.7 and Eq. 3.8 to be

$$D_{ik}(\mathbf{U}) = \mathrm{E}[f(.,u_i)f(.,u_k)] - \mathrm{E}[f(.,u_i)]\mathrm{E}[f(.,u_k)]$$

in which the ensemble means E[...] could be calculated by ensemble Monte Carlo evaluation. Also it should be noticed that since in the calculation of the matrix $D(\mathbf{U})$ we only use $\Delta F(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{U})$, the difference between log-concentrations, *i.e.*, the logarithm of concentration ratio, is independent of the choice of model concentration units and scale-free. Hence our MINE criteria is also scale-free.

3.3.1 A Hilbert Space picture of MINE formalism

In terms of Hilbert Space (HS) formalism, we could define again the MINE approach constructed above using a Hilbert Space of functions that are defined on the model parameter ($\boldsymbol{\theta}$) space. The variance/covariance is the HS inner product, which for any pair of functions defined in the $\boldsymbol{\theta}$- space is defined as:

$$(g|h) := \mathrm{E}[g(.)h(.)] - \mathrm{E}[g(.)]\mathrm{E}[h(.)]. \tag{3.10}$$

The observables $\log(y_i)$ are now represented by a HS vector $f_i := f(\boldsymbol{\theta}, u_i)$, for i=1,...d; and the element in the variance/covariance matrix, D_{ik}, is then the inner product of two HS vectors f_i and f_k,

$$D_{ik} = (f_i \mid f_k).$$

The predicted ensemble standard deviation of the observable f_i is the length in HS, *i.e.*, the HS vector norm $\|f_i\|$, where $\|f_i\| := (f \mid f)^{1/2}$.

Now the independence of the HS vector set $f_1, ...f_d$ naturally represents the independence of the corresponding observables $\log(y_i), ... \log(y_d)$. If the d HS vectors $f_1, ...f_d$ are linearly independent, the finite-dimensional subspace spanned by these HS vectors has a dimension of d; otherwise the subspace has a dimension less than d. Thus a linearly independent HS vector set $f_1, ...f_d$ spans a parallelotope of d-dimensional in HS. Since $D_{ik} = (f_i \mid f_k)$, one can see that $\det(D)$ in the MINE criterion Eq. 3.9 is the *Gramian determinant* of the vector set $f_1, ...f_d$, which is the square of the volume of this HS parallelotope [37]. Hence $\det(D)$ is then an alternative representation of the characteristic variance/covariance parallelotope $\boldsymbol{v_\Delta}$ in the d-dimensional ΔF space. Notice that $\boldsymbol{v_\Delta}$ in ΔF space is a rectangular parallelotope as we have described in 3.2.2., while the HS parallelotope spanned by the HS vectors $f_1, ...f_d$ is not rectangular in general, since $f_1, ...f_d$ are not required to be mutually orthogonal with respect to their HS inner product Eq.3.10.

If the $f_1, ...f_d$ are linearly dependent, $\det(D)$ is zero and the HS prism spanned by these HS vectors collapses to a lower-dimensional one. In the other extreme situation, if all observables are uncorrelated, the corresponding HS vectors $f_1, ...f_d$ are maximally independent and mutually orthogonal in terms of HS inner product and the volume of the HS parallelotope is simply the product of the HS vector length $\|f_i\|$ and $\det(D) = (\|f_1\| \cdot \ldots \cdot \|f_d\|)^2$. In general cases, the observables are correlated and the HS vectors $f_1, ...f_d$ are non-orthogonal, hence $\det(D) < (\|f_1\| \cdot \ldots \cdot \|f_d\|)^2$. The ratio $\det(D)/(\|f_1\| \cdot \ldots \cdot \|f_d\|)^2$ with a range of (0, 1), indicates the degree of independence of the observables. This ratio depends only on the HS

angles between pairs of HS vectors, but not on their individual lengths $\|f_i\|$. Therefore we can write the MINE criterion as:

$$V(\boldsymbol{U}) = \det(\mathrm{D}(\boldsymbol{U})) = (\|f_1\| \cdot \ldots \cdot \|f_d\|)^2 \cdot \frac{\det(D)}{(\|f_1\| \cdot \ldots \cdot \|f_d\|)^2}. \tag{3.11}$$

As we have seen in 3.2.1, if the experimental design tries to maximize only the individual variances, it tends to also maximize the correlation thus makes the ratio vanish to zero. Therefore, in order to maximize this MINE criterion, we must compromise between maximal mutual independence of all variables and maximal variance of each individual observable.

With one more look into the matrix $D(\mathbf{U})$ and the Euclidean distance criterion Eq. 3.6, it is easy to see that the right hand side of it is just the trace of the matrix $D(\mathbf{U})$: $V(\boldsymbol{U}) = trace(D(\boldsymbol{U})) = \|f_1\|^2 + \ldots + \|f_d\|^2$. For maximizing the $V(\mathbf{U})$, what the Euclidean distance criterion suggests is to maximize each individual HS vector length, *i.e.*, to maximize the variance of each individual variable, even if that results in high co-variances between variables. This is consistent with our discussion in 3.2.1..

3.4 Criterion 3: MINE by Maximal Observational Independence

We have seen that when we maximizing the MINE criterion Eq. 3.11, simply increasing the variances of individual observables only results in decreasing the independence between the observables and vise versa. This inspires us to propose another MINE criterion that emphasizes even more the independence of the observables comparing with Eq. 3.11. We substitute the original HS vectors f_i with the normalized unit HS vectors g_i where

$$g_i(\boldsymbol{\theta}) := f_i(\boldsymbol{\theta}) / \|f_i\| \qquad \text{for } i = 1, ...d; \tag{3.12}$$

and define a new "normalized" variance/covariance matrix similar with the one defined in Eq. 3.11,

$$\begin{aligned} E_{ik}(\boldsymbol{U}) &= (g_i \mid g_k) = D_{ik}(\boldsymbol{U})/(\|f_i\| \cdot \|f_k\|) \\ &= \mathrm{E}[g(., u_i)g(., u_k)] - \mathrm{E}[g(., u_i)]\mathrm{E}[g(., u_k)] \end{aligned} \tag{3.13}$$

This is, in fact, the well known correlation matrix between the predictions [36]. And the proposed MINE criterion is then to maximize

$$V(\boldsymbol{U}) := \det(E(\boldsymbol{U})) = \det(D(\boldsymbol{U}))/(\|f_1\| \cdot \ldots \cdot \|f_d\|)^2 \tag{3.14}$$

Comparing with Eq. 3.11, it is easy to see that the lengthes of the individual HS vector, *i.e.*, the variances of the observables do not affect $\det(E(\mathbf{U}))$, only the angles between pairs of the HS vector affect it, which measures the independence of the observables. Geometrically, in the Hilbert Space, $\det(E(\mathbf{U}))$ is the square of the volume of the parallelotope formed by the d HS unit vectors $g_1, ...g_d$ with fixed length $\|g_i\| = 1$. It is the angles between pairs of the unit vectors that determine the volume instead of their individual length, which then emphasizes only the independence of the observables when maximizing this MINE criterion.

Such a MINE criterion tends to design a next experiment such that the maximized information is obtained by improving the independence of the observables, rather than maximizing the individual variances. It is more advantageous in applications where the HS vectors are likely to be linearly dependent, *i.e.*, the observables are easy to be highly correlated with each other. This scenario is in fact what we have encountered consistently in our MINE calculations for our three Computing Life cycles. Therefore the third MINE criterion Eq. 3.14 to maximize $\det(E(\mathbf{U}))$ is what we have applied in our implementation of the three cycles.

There are some additional advantages of choosing $\det(E(\mathbf{U}))$, the correlation matrix, as the MINE criterion. It is well known in statistics that the correlation matrix measures the linear independence between the predictions of the observables. From our discussion in 3.2.3, $\det(E(\mathbf{U}))$ is known to have a range of 0 to 1, where the value 0 means perfect linear dependence of the predictions and the value of 1 means perfect linear independence. Furthermore, if the predictions are Gaussian over the ensemble, the correlation matrix measures also the stochastic independence of the predictions. In this case the value 0 means perfect stochastic dependence and 1 means perfect stochastic independence. That is to say, by applying the third MINE criterion Eq. 3.14, we can easily interpret the measures. And finally the dis-

tribution properties of $\det(E(\mathbf{U}))$ are well known, particularly when the distribution of the predictions are Gaussian.

3.4.1 Volume collapse pathology

In the actual MINE calculation, we have encountered consistently the situation that the observables are highly correlated, which results in the "almost" linearly dependent HS vector set $g_1, ...g_d$ (or, equivalently, the $f_1, ...f_d$ set). It is easy to see this lack of sufficient linear independence by calculating the d eigenvalues of the D matrix. Denote the d eigenvalues as $\lambda_n = \lambda_n(\boldsymbol{U})$ and enumerate by $n = 1, ..., d$ in descending order, with corresponding complete, orthonormal $d \times 1$ eigenvectors e_n. Note that the variance/covariance matrix D is non-negative, so the d eigenvalues λ_n are also non-negative. Thus we can decompose D into its eigenvector representation

$$D(\boldsymbol{U}) = \sum_{n=1,...d} \lambda_n(\boldsymbol{U}) e_n(\boldsymbol{U}) e_n(\boldsymbol{U})^T,$$

and $\det(D)$ is simply the product of all d eigenvalues. However, a common numerical difficulty we encounter is that the ratio of a smaller eigenvalue λ_n and the largest one λ_1 becomes of order or smaller than the *machine precision* ϵ_{mp} because the HS vector set is "almost" linear dependent. Then all such eigenvalues λ_n that are comparable to or small than $\lambda_1 \epsilon_{mp}$ are dominated by rounding errors, *i.e.*, they are too small to be numerically calculable. Thus $\det(D)$ is not numerically calculable for the exact value either. Geometrically this simply means the near "collapse" of the characteristic ellipsoid of the matrix D: the half-axis along the direction of the corresponding e_n with a length of $\sqrt{\lambda_n}$ has shrunk very close to, and not numerically distinguishable from zero. However, there is still useful information contained in such almost collapsed ellipsoid about the range swept out by the image difference vector ΔF. We need to remedy this numerical pathology to obtain such information.

Thus an ellipsoid volume collapse pathology is proposed to introduce a numerically stable lower cut-off into the eigenvalue spectrum of the matrix D so that if the eigenvalue λ_n is

comparable or even smaller than $\lambda_1 \epsilon_{mp}$, the product of the largest eigenvalue and the machine precision, it is replaced by a modified eigenvalue:

$$\mu_n(\boldsymbol{U}) = \max(\lambda_n(\boldsymbol{U}), \lambda_1 \epsilon_{cut}),$$

where the cut-off ratio ϵ_{cut} is fixed to be 10^{-10}, which is typically at least 2 or 3 orders of magnitude larger than the machine precision ϵ_{mp}. Thus in the ellipsoid every half-axis is at least $\sqrt{\lambda_1 \epsilon_{cut}}$ along the corresponding eigenvector, which keeps the ellipsoid from collapsing. Now the numerically incalculable exact value of $\det(E(\mathbf{U}))$ is then replaced by a numerically stable $\det(D(\mathbf{U}))$ of a modified D matrix such that

$$D^{(cut)}(\mathbf{U}) := \sum_{n=1,\ldots d} \mu_{\mathrm{n}}(\mathbf{U}) \mathrm{e}_{\mathrm{n}}(\boldsymbol{U}) \mathrm{e}_{\mathrm{n}}(\boldsymbol{U})^{\mathrm{T}},$$

and

$$\det(D^{(cut)}(\boldsymbol{U})) = \mu_1(\boldsymbol{U}) \cdot \ldots \cdot \mu_d(\boldsymbol{U}).$$

In our actual MINE calculation, the MINE criterion Eq. 3.14 is used instead of Eq. 3.11. The same lower cut-off procedure is then applied to generate a modified matrix E which is numerically stable, and the determinant becomes the corresponding $\det(E^{(cut)}(\boldsymbol{U}))$. Finally, note that this ellipsoid volume collapse does not numerically affect the MINE criterion Eq. 3.6 using Euclidean distance, since $V(\mathbf{U}) = \mathrm{trace}(\mathrm{D}(\mathbf{U}))$ is the sum of the eigenvalues of D matrix instead of the product of them. The sum is dominated by the largest eigenvalue and thus is numerically well controlled.

3.5 Summary of the Maximally Informative Next Experiment Approach

The goal of the MINE approach described above is to develop a quantitative criterion (or measure) to answer the question of how much information about the genetic network can be obtained from the next experiment to be performed, and then maximize this measure of information which is denoted by $V(\mathbf{U})$ with respect to the experimental design, *i.e.*, the control parameters of the next experiment, denoted by the control vector $\mathbf{U}$. The vector $\mathbf{U}$

contains all of the parameters that are controllable by the experimenter and that characterize the measurements to be performed, including the external conditions and the perturbations and mutations to the species.

As input to the MINE calculations, it is critical to have both the underlying kinetic rate coefficient model of the genetic network and all available *prior* or *old* experimental data. These two inputs first serve as the inputs of the ensemble simulation approach to constrain the unknown kinetics model parameters and generate an ensemble of fitted parameters, which is then used to predict the likely information content $V(\mathbf{U})$ for the next experiment with respect to $\mathbf{U}$. The mine criterion Eq. 3.14, $V(\boldsymbol{U}) = \det(E(\boldsymbol{U}))$, is chosen to guide the design of the next experiment as what we have discussed above. This criterion is the determinant of the *ensemble correlation matrix* $E(\mathbf{U})$ between predictions. For example, in our application of the MINE calculation, an ensemble simulation was performed for the kinetic model for the genetic network in Eq. 1.2 to generate a representative MC sample of 40,000 random θ-vectors, drawn from the respective ensemble distribution $Q(\theta)$ for that cycle. A subset of 200 random θ-vector from this sample is then used to calculate MC estimates for the ensemble expectation values E[...] for the matrix E via Eq. 3.13.

If two predictions in the fitted ensemble, given by two different models with different control vectors, are highly correlated, it will be difficult to distinguish these two models in the next experiment. If the two predictions in the fitted ensemble are less correlated, it will be easier to distinguish the two corresponding models. Hence by applying the criterion Eq. 3.14 that emphasizes the correlations between predictions, it is suggested to choose the next experiment U such that it gives predictions in the fitted ensemble more uncorrelated to each other and then more distinguishable. We are going to see a practical example of the application of the MINE method in the next chapter.

Chapter 4

Application of the Computing Life Paradigm on the Biological Clock of *Neurospora crassa*

After discussion of the ensemble method of genetic network identification and the Maximally Informative Next Experiment design respectively, it is now time to see how these two approaches are combined with the experiments to work together in the Computing Life paradigm shown in Figure 1.3. In the general context of refining the identification of the genetic network model of the biological clock in *N. crassa*, a particular biological goal is set to specify the direction of advance that we want to discover the *clock-controlled genes*(*ccgs*) in the *N. crassa* genome.

These so-called *clock-controlled genes*, in another word, are the outputs of the biological clock. Therefore, from the genetic network model shown in either Figure 1.2 or Figure 2.2, it is predicted that a *ccg* gene should have three characterizations: (1) maintain an endogenous circadian rhythm when the organism is grown in the dark; (2) be light-responsive when the organism is moved from dark to light (D/L); (3) change its expression when the level of the transcription factor WCC is knocked down. These three predictions are going to be used as criteria for *ccg* selection.

It should be noted that, as we mentioned in section 2.3, the genetic network model used throughout the Computing Life process shown in this thesis is the one shown in Figure 2.2 instead of that shown in Figure 1.2 for historical reasons.

4.1 Cycle 1 - Which genes are circadian?

After fitting the data in the literature and building the biological clock model [14], the first series of microarray experiments to be designed is to determine how many genes are under clock control. Such genes, as shown in both Figure 1.2 and Figure 2.2, are the output of the clock mechanism, thus should have an endogenous rhythm of ∼22hrs in the dark. So the first experiment to be done was to culture a strain for about 48 hrs in the dark to observe the clock's endogenous rhythm. The initial MINE task is to find out the appropriate starting time of observation and the optimal interval between observations.

In each of our MINE calculations, the predictions are of the log concentration of *wc-1*, *wc-2* and *frq* mRNAs over time in the next experiment, *i.e.*, $\log([wc-1^r])$, $\log([wc-2^r])$ and $\log([frq^r])$, where $[wc-1^r]$ is, in fact, the summed total of the two version of *wc-1* mRNA, $wc-1^{r1}$ and $wc-1^{r2}$, as shown in Figure 1.2 and Figure 2.2, *i.e.*, $[wc-1^r] = [wc-1^{r1}]+[wc-1^{r2}]$. Each MINE calculation is constrained to have a fixed budget of 13 microarray chips per experimental cycle, which means that there are 13 time points to be sampled in each cycle. This is a more realistic design since presently microarray profiling is still quite expensive to perform. Hence the dimension d of the correlation matrix $E(\mathbf{U})$ is determined by the fixed number of time points such that $d = 3 \times 13 = 39$, which is fixed by our budget.

Denote the starting time point, *i.e.*, the time lag till the first observation time by t_L, and the spacing between measurements by t_S. As what we have clarified above, the maximum in spacing is limited by the total culture time (48 hrs) and the cost constraint of 13 microarray chips. In the next experiment we are going to measure 3 mRNA species at 13 time points, t_j with j=1, ...13; and there are d=39 data points, $y_1, ...y_d$, to be observed in the next experiment. The observation times t_j are chosen to have equidistant spacing t_S after an initiation time lag of t_L, which is counted from the experiment starting time t=0. So

$$t_j = t_L + t_S(j-1) \qquad for\ j = 1, ..., 13.$$

In the cycle 1 ensemble simulation we constructed $Q(\theta)$ from the prior experimental input data taken from literature [9, 23–25]. An ensemble of Monte Carlo samples of 40,000 random θ-vectors was generated according the ensemble distribution $Q(\theta)$. From this ensemble, $V(\mathbf{U})$=det$(E(\mathbf{U}))$ was maximized with respect to $t_S$and t_L, for the next experiment designed to measure the three clock RNA species during the 48 hrs in the dark. The resulting optimal MINE values (as shown in Figure 4.1) appeared to be $t_s = 3.5$hrs and $t_L = 0$. In the real experiment, the theoretical optimal values were slightly modified to $t_s = 4$hrs and $t_L = 0$, so as to cover the behavior at the end of culture (42-48hrs) which is missed by the optimal design here but still important to know for the behavior at the later part of the strain culture. In the latter cycle2 and cycle3 microarray profiling and corresponding MINE calculations, t_S and t_L are then set at the same values used here without further adjustment.

The experimental results of microarray RNA profiling experiments following the MINE design above can be found in [13]. Biological and statistical analysis showed that, in the first cycle of the Computing Life process, 2436 genes showed circadian rhythmicity when grown in the dark, behaving consistently with the first prediction of *ccgs*. Therefore these 2436 genes were selected out as the future candidates of *ccgs*

4.2 Cycle 2 - Which genes are light-responsive?

4.2.1 Design of Light Response Experiments

From cycle1, 2436 genes showing circadian rhythmicity in the dark were identified [13] as the candidate of *ccgs* according to the first criterion. However not all of them are "real" *ccg* genes. The possible controller of the circadian rhythm shown in the dark could be either WCC, or a different oscillator [38], or even multiple oscillators [39]. There is also a false positive probability of 18% [13] for the selected 2436 genes. To discriminate the *ccg* genes from the other possibilities, it is necessary to design the next experiment.

As shown in Figure 1.2 and Figure 2.2, an important feature of the clock is light-entrainment. That is to say, the biological clock of the organism will be speeded up or

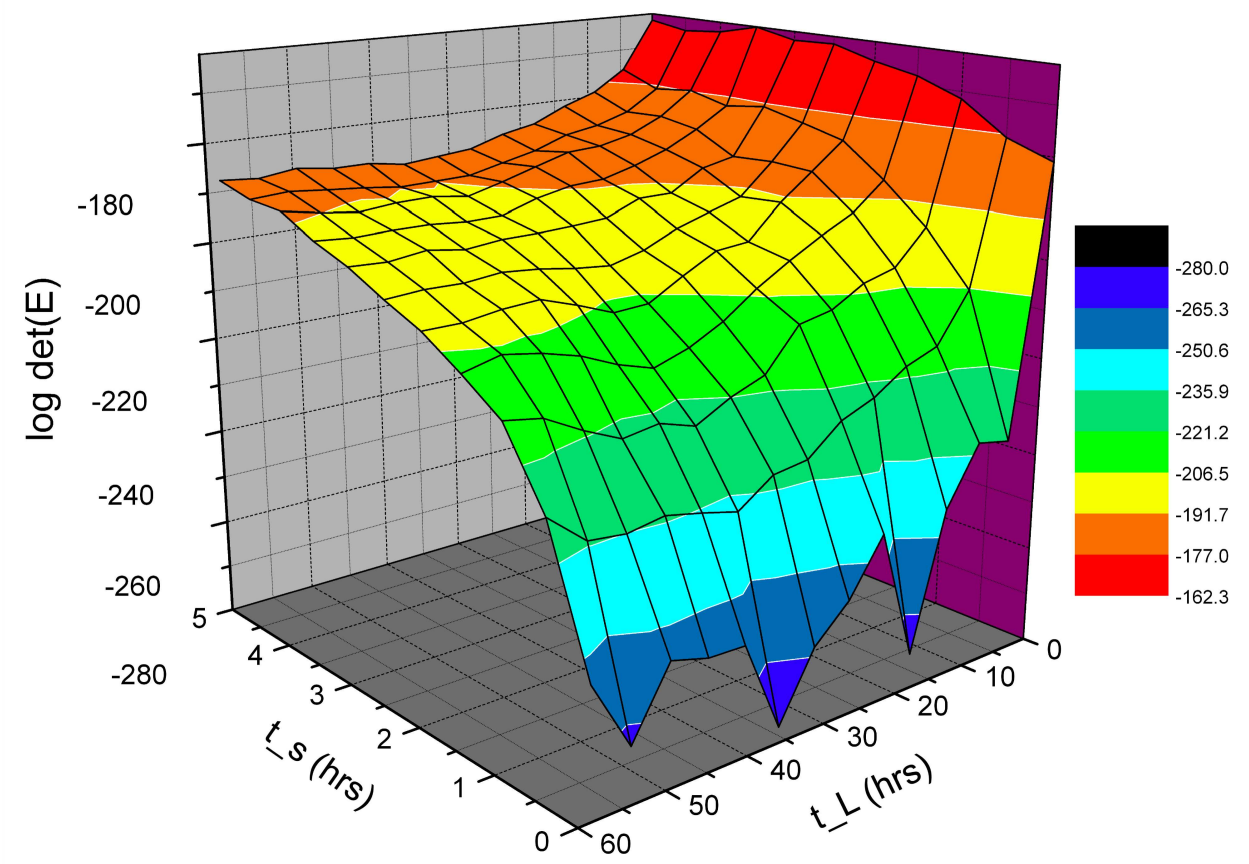

Figure 4.1: **MINE calculation to determine when to start sampling (t_L) and how often (t_S)**. The MINE surface is plotted as function of the lag $t_L$in hrs and spacing t_S in hrs. Higher values on the MINE surface suggest the preferred design points (t_L, t_S). Color contours of the log of the MINE criterion det(E) are overlayed as a function of the lag(t_L) and spacing(t_S) to show points on the surface of similar MINE values. The MINE surface suggests to start sampling immediately (small t_L) and to make the spacing (t_S) between observations as large as possible. The maximum permissible spacing (t_S) between observations is 5 hrs, as determined by two constraints: the cost constraint of 13 microarray chips per cycle; and beyond a 50hr experiment in cycle 1 stable oscillations in liquid culture are not guaranteed. This graph has been published in [13].

slowed down to accommodate the artificial days that it is exposed to, which consist of alternating light exposure and dark exposure. If the gene is a *ccg*, *i.e.*, under the control of WCC, it should also be light-responsive. Then the question is what artificial day should be used, *i.e.*, how long should the alternating dark and light exposure be.

In all *new* experiments we have performed, whenever light exposure was applied, the light intensity at the sample location was about 70 μmole(photons)/(s $\cdot$ m^2) in Einsteinian units, or about 15 W/m^2 in radiometric units, or about 5,300 lux/490 ft-candles in photometric units, assuming a "cool white" spectrum, as given in [13, 40]. By contrast, for the light exposure experimental data we have taken from the literature [9, 23–25], the light intensity was report to be only 20 μmole(photons)/(s $\cdot$ m^2), with unspecified spectral distribution. In the kinetics ensemble for cycle2 (also in cycle 3 below) the kinetics ensemble simulation needs to incorporate the new experimental data from cycle 1 and the literature data [9,13,23–25]into the distribution $Q(\boldsymbol{\theta})$. So, here we have assumed that both the literature data and our experimental data were generated with the same photon spectral distribution. Therefore we have modeled all light exposure experiments in terms of the photon concentration to be proportional to the respective light intensities when light enters the reaction rate coefficient. For all light-exposed experimental data throughout our simulation, the light exposure is periodic with 50% duty cycle having the same dark and light duration, in a phase of either D/L (*i.e.*, dark first, then light) or L/D.

After the new experimental data from cycle 1 has been incorporated into the kinetic ensemble simulation and fitted(see Figure 2.3), based on the new cumulative $Q(\boldsymbol{\theta})$, $V(\mathbf{U}) = \det(E(\mathbf{U}))$ was maximized with respect to the duration $t_{p/2}$ of the light exposure with D/L and phasing 50% duty cycle (*i.e.*, the dark duration is also $t_{p/2}$). The result is shown in Figure 4.2, suggesting a long artificial day with a half-period of daylight of between 19 and 24 hrs. The half period $t_{p/2}$ was chosen to be 24 hrs for the actual cycle2 experiments. A second cycle of microarray experiments was therefore performed with an artificial day of 24 hrs dark and 24 hrs light in a 48 hrs observation time.

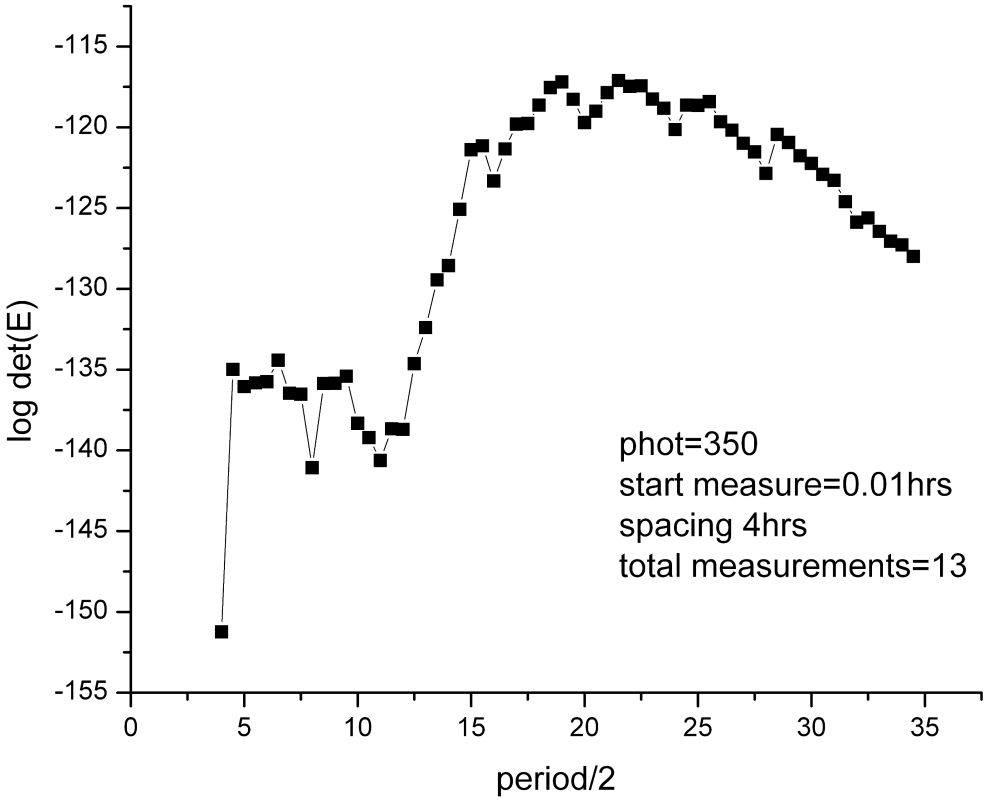

Figure 4.2: **MINE calculation to determine what artificial day to use in cycle 2**. Graph of the decadic log of the MINE criterion det(E) as a function of the half period of the artificial day in hrs. The calculation suggests trying a long artificial day and the half-period of daylight is between 19 and 24 hrs. Repetition runs were performed with different subsets of θ vectors drawn from the 40,000 model ensemble Monte Carlo sample. The log(det(E)) values from repetition runs differ typically by no more than $\pm$ 1 from the results shown here and give the same results: the half period of the artificial day should be in the range of 19-24 hrs. The inset gives: (1) the photon concentration of micromoles per second per meter squared ($\mu M/s \cdot m^2$); (2) the starting time (t_L), which was selected to be close to zero but not zero to assist in the computation of the MINE criterion det(E); (3) spacing (t_S) in hrs between observations; and (4) the total number of time-points, at which mRNA levels were measured (the number of arrays used). This graph has been published in [13].

Statistical and biological analysis of the experimental data from cycle 2 showed that, a total of 768 genes were both circadian and light-responsive, with a very low false positive rate of 0.03 [13]. These genes then passed the first and second criterions and remain as candidates of *ccgs* in Figure 1.2 and Figure 2.2. Also, as a control, other experiments with designs very similar to our cycle 2 reported in [20] and [41] independently obtained results in good agreement with us.

4.2.2 Possible Evidence for a Modified Model with Auto-Feedback Loops Activating *wc-1* and *wc-2*

With a second look at the MINE calculation of cycle 2 in Figure 4.2, a minor peak was found at the shorter period end of the graph, suggesting possible informative experiments with a very short artificial day ranging 3 to 6hrs. At the same time, when looking at the profiling experimental results of the three cycles, we noticed that there was evidence of the presence of auto-feedback loops for WCC activating *wc-1* and *wc2*. In cycle 1 searching for circadian genes, *wc-1* and *wc-2* mRNA showed fast oscillations with a period much shorter than the circadian genes. In cycle 2, a fast light-response less than one hour was also noticed. In cycle3, *wc-2* and, of course, *wc-1*, also showed a quick response signal right after the knocking down of *wc-1* genes [13]. For the modified genetic network shown in Figure 4.3, it is reasonable to predict that the auto-feedback loops added should permit entrainment to short artificial days. Thus motivated by the MINE prediction, a series of experiments were performed with extra short 3+3 artificial day that was usually avoided by researchers.

The result showed that the wild type strain did entrain into the 3+3 artificial day, and banding patterns were observed similar with that shown in Figure 1.1 but much denser [13], showing the direct evidence of the clock did oscillate and get entrained in to this very short artificial day. And for a *frq* mutation, which disabled the *frq* gene and thus eliminated FRQ from the genetic network in Figure 2.2, the entrainment still happened and the banding pattern was shown. If the FRQ involved P-reaction is the only feedback loop in the biological

clock, the oscillation and entrainment is impossible to happen. Further more, mutations that disabled either *wc-1* or *wc-2*, *i.e.*, resulted in no bandings, indicating no entrainment to the short artificial days with the *frq* gene being kept.

The above then suggested strong evidence of a modified model with auto-feedback loops that activate *wc-1* and *wc-2* respectively shown in Figure 4.3. However, failure in further ensemble fitting to the above quick response data stopped us from validating this model and making further use of this model, although revisiting this model is possible in the future. In conclusion it is a very interesting trial and suggests the high informativeness of the Computing Life paradigm.

4.3 Cycle 3 - Which genes are under WCC control?

By cycle 2, the number of genes as candidates for *ccgs* was limited to 768. To be confirmed as *clock controlled genes*, they need to agree also with a third prediction of the genetic networks in Figure 1.2 and 2.2, namely that a gene under the control of WCC should experience a sudden change in its mRNA level if WCC were knocked down (By saying "knock down" here, it means that the amount of WCC in the system would be decreased to a lower level compared to wild type strain.) It is easy to see from the model that interfering with the WCC concentration level can be realized by reducing either one of *wc-1* or *wc-2* or *frq* mRNA level. Therefore, to test the third prediction with a gene knock-down experiment, we need to first know which gene should be perturbed to gain maximum information about the genetic network in Figure 2.2.

The cycle 3 MINE calculation was done using the $Q(\boldsymbol{\theta})$ distribution generated from the literature data [9, 13, 23–25] plus our experimental data in cycle 1 and cycle 2. The mine criterion $V(\mathbf{U}) = \det(E(\mathbf{U}))$ was maximized with respect to the transcriptional ratio (TR) of the selected clock gene species to be knocked down. TR=1 corresponds to full expression, while a non-zero TR value less than 1 simply means the selected gene species would have its transcription rate coefficient S reduced to TR*S, where TR $\leqslant 1$.

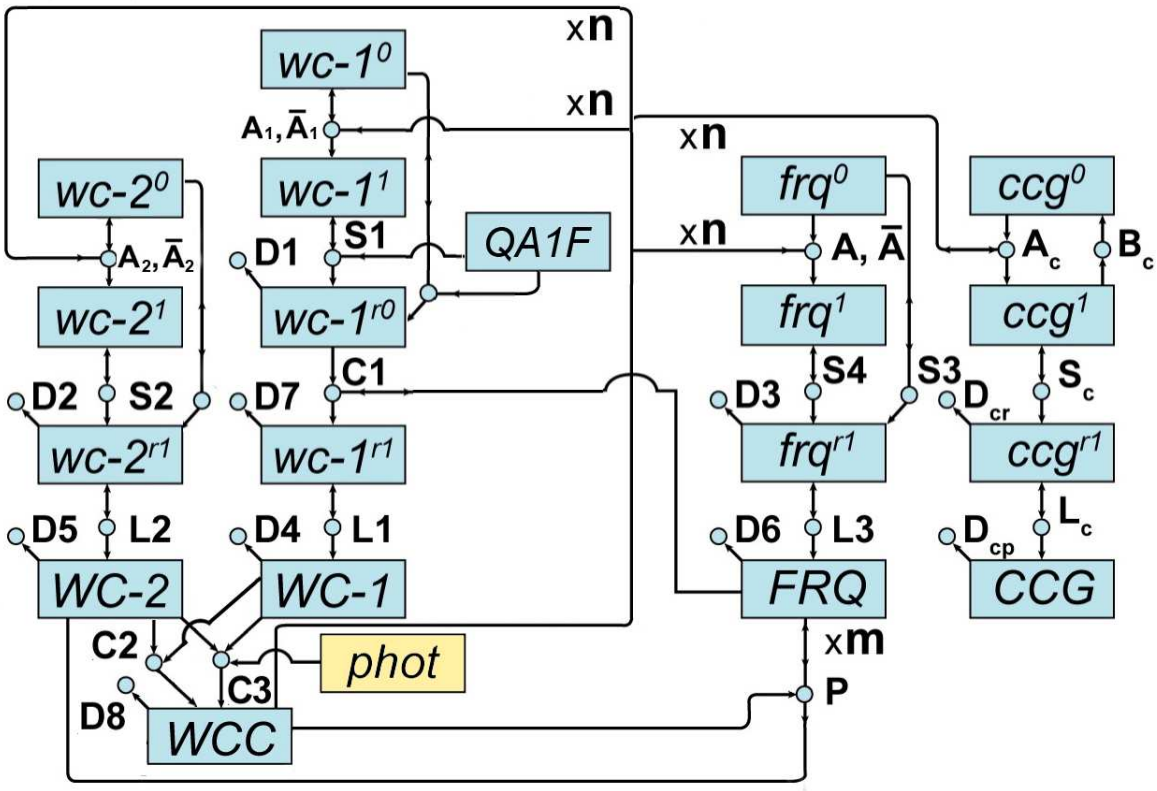

Figure 4.3: **A modified model with auto-feedback loops of *wc-1* and *wc-2*.** It is reasonable to predict that with the presence of these two auto-feedback loops the system can be entrained into very short artificial days, which is not a property of the *frq*-centered biological clock. This figure was used in [13].

The maximally informative gene to be knocked down was found to be *wc-1*, with a TR=0.1 (shown in Figure 4.4). It suggested reducing the transcription rate coefficient to 10% of that of wild type. The actual experiment was done by engineering a mutation strain of wild type *wc-1* and a quinic acid inducible copy of $wc-1^{+}$ producing a knock-down to 30% of its original activity when being transferred into quinic acid.

An additional MINE calculation is shown in the right bottom plot in Figure 4.4. The low $\log(\det(E))$ values suggested that over-expression of *wc-1* will not be very informative. Lewis *et al.* reported the over-expression of *wc-1* was "not sufficient to induce most light-regulated gene expression", which is consistent with our MINE prediction [41].

Statistical analysis of the data from this "gene knock-down" cycle 3 experiment found 328 *clock-controlled genes* supported by all three cycles of microarray experiment series and also biologically reasonable (*ccgs* must have a WCC binding site to interact with WCC) [13]. The possibility of a false positive was 0.0067, providing all three series were done independently. These 328 genes therefore satisfy the three predictions of the genetic network and constitute the *clock controlled genes*.

4.4 Identifying an ensemble of genetic networks for the biological clock of *Neurospora crassa*

Besides accomplishing the biological goal of finding *clock-controlled genes*, the three cycles of the Computing Life paradigm also have identified an ensemble of genetic networks of the biological clock with the cumulative data from 3 cycles of microarray experiments and the initial published literature data [9, 13, 23–25].

The estimation of rate coefficients has been improved after the three cycles, comparing with our previous results for an in-the-dark model of the biological clock of *Neurospora crassa* (Table 1 in [13]), which is the starting point of our simulation. The results of fitted rate coefficients are summarized in Table 4.1. For 69% of the rate coefficients in common (*i.e.*, 18 out of 26) of these two models, ensemble standard deviations were reduced by the

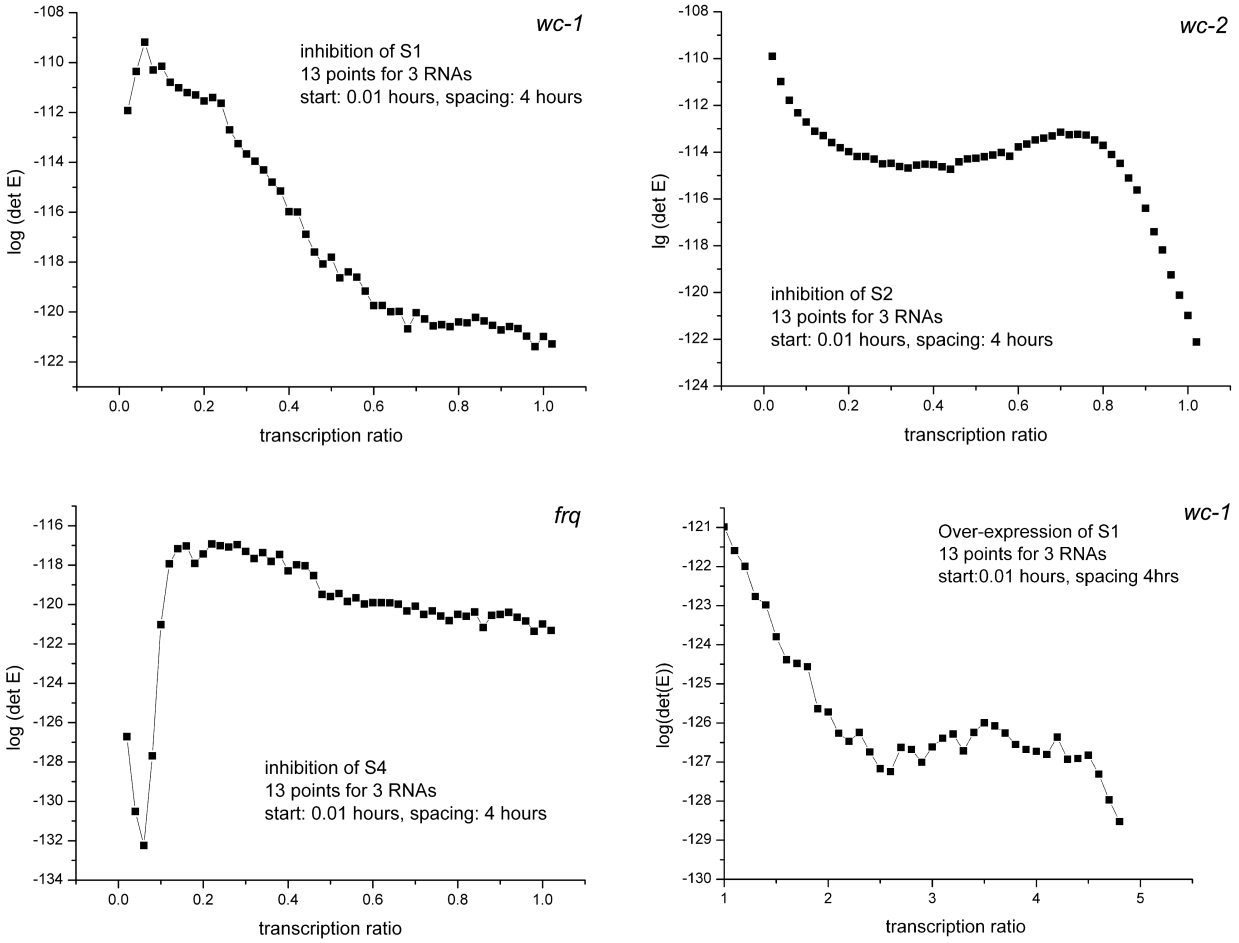

Figure 4.4: **A 90% knock-down of the *wc-1* gene is the MINE experiment**. The decadic log of the MINE criterion $\det(E)$ is displayed as a function of percent remaining activity of the three clock genes *wc-1*, *wc-2*, and *frq*. The matrix E is the correlation matrix of the predictions, emphasizing independence of predicted data points $f(., u_i)$. The predictions are for the mRNA levels of *wc-1*, *wc-2*, and *frq* over time. The right bottom figure shows the MINE prediction that over-expressing *wc-1* is not very informative, which was proved to be true by the experimental results reported by Lewis *et al.* [41]. Repetition runs were performed with different subsets of θ vectors drawn from the 40,000 model ensemble Monte Carlo sample. The $\log(\det(E))$ values from repetition runs differ typically by no more than $\pm$ 0.5 from the results shown here and give the same results: the maximal $\det(E)$ appears at transcription ratio 0.1 for *wc-1*. Part of this graph has been published in [13].

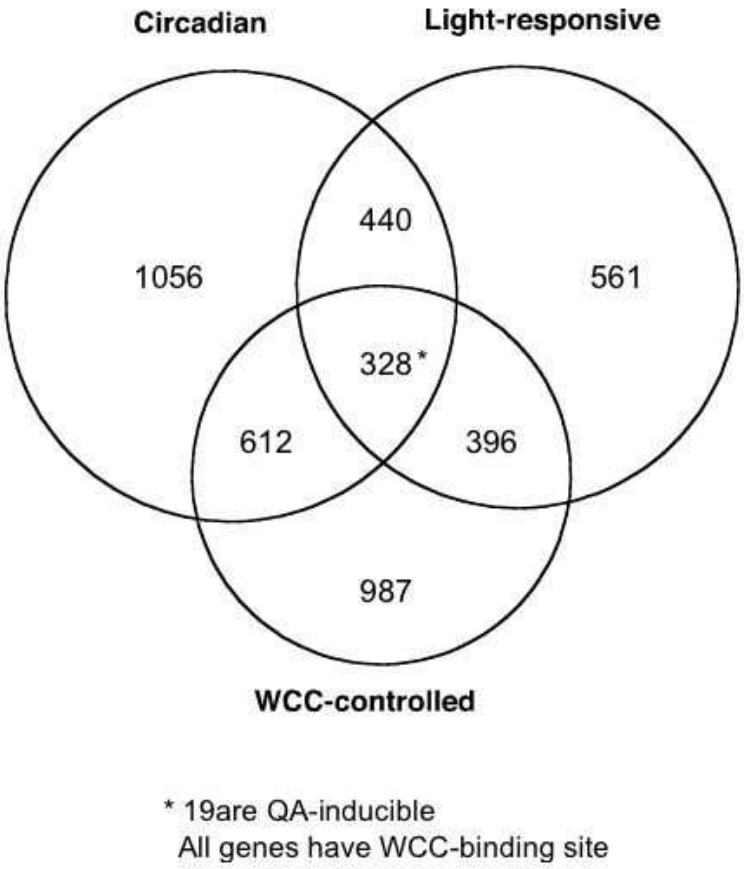

Figure 4.5: **Classification of 4380 *N. crassa* genes with upstream LREs in a Venn Diagram by their response in each of the three microarray experiments: (1) cycle 1 (assay for circadian rhythm); (2) cycle 2 (assay for light response); and (3) cycle 3 (assay for response to changing levels of WCC)**. The diagram summarizes the microarray experiments in cycles 1–3 of the Computing Life Paradigm. This graph has been used in [13].

addition of data from cycle 1-3 from earlier published results in [14]. The estimated lifetimes of the *wc-1* mRNA and the FRQ protein remain in good agreement with the measured values with data increasing by an order of magnitude(Table 4.3). The *wc-1* mRNA life time estimate of $7.4hrs = D_7^{-1}$ continues to be supported by experiments in cycle 3, in which the transcription level of *wc-1* was knocked down to a low level and the degradation time of the *wc-1* concentration was observed to be about 8 hrs (see Figure 2.3). The lifetime of FRQ protein is estimated to be $1/\langle D_6 \rangle \approx 5hrs$, being consistent with both our previous results and the estimation of 4-7 hrs from published data by others [33]. Transcription rates of *frq* (A and $\bar{\text{A}}$) and the deactivation rate of WCC (P), which were critical parameters for maintaining oscillations [14], are now more sharply defined than before. As described in 2.3.2, the model discussed above was replaced by the one in Figure 1.2 which has a close chi-square statistics (Figure 2.4) but fewer model parameters. The summary of the rate coefficients of the model in Figure 1.2 is shown in Table 4.2. There is no obvious improvement in the parameter estimation.

To assess progress in the Computing Life paradigm and also to compare to other models in the future, the error per observation σ^2, or the error variance, is estimated. This has served as a standard approach in linear and nonlinear models to estimate the precision of an experiment [42], as we have introduced in Chapter 2. It is also illustrated by simulation and data analysis that such a common variance components can be extracted from each of a variety of microarray experiments and used to compare different experiments [43]. Under the multivariate Gaussian assumption that we used in the likelihood in Chapter 2, a simple estimator for the error per observation can be constructed for our successive cycles of the Computing Life paradigm:

$$\hat{\sigma}^2 = \frac{1}{n}\chi^2_{min}\sigma_0^2$$

where n reports the number of data used in fitting, which is cumulative across cycles; χ^2 is the goodness of fit measure, and χ^2_{min} is the minimum chi-squared statistic over the ensemble; σ_0^2 is the error per observation in the multivariate Gaussian likelihood, which is allowed to

vary across observations. In the initial data drawn from the literature [9, 23–25], σ_0^2 is 0.02. In calculating χ^2_{min} , preliminary estimates of σ_0^2 , $4\sigma_0^2$ and $36\sigma_0^2$ were used respectively for literature data, microarray data and conidiation data, respectively, in order to give equal weight per time to different experiments in the ensemble fitting progress.

In Table 4.3, the progress can be seen in reducing the error variance in successive cycles. In the fourth cycle, the model was switched from the one in Figure 2.2 to the one in Figure 1.2. An additional 842 data points of conidial banding data (as shown in Figure 1.1) were collected under a 48 hr artificial day as in cycle 2. A reduction in the estimated error variance with a simpler model confirms the switch of the model is reasonable, as we have discussed in Chapter 2. Thus the advantage of this estimated error variance also offer a diagnosis of whether or not further experiments will refine the model ensemble. The downward trend in the estimated error variance suggests that further cycles could be profitable.

Table 4.1: **Rate coefficients in the genetic network model (Figure 2.2) of the biological clock (n=m=4) based on data from cycles 1-3 predicting the clock's observed oscillations, light response, and wc-1 perturbation.** Ensemble mean $< X >$ and ensemble standard deviation $\sigma(\mathrm{X}) := [< \mathrm{X}^2 > - < \mathrm{X} >^2]^{1/2}$ for rate coefficients (X) in the n=m=4 biological clock model of Figure 2.2. 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 for all species, except for the photon species where $1cu(photons) = 0.20\mu\mathrm{mole(photons)}/(\mathrm{s} \cdot \mathrm{m}^2)$.

X	k	$< X >$	σ(X)	X	k	$< X >$	σ(X)
A	5	0.0313	0.00974	A_c	5	0.1293	0.0826
$\bar{A}$	1	0.1108	0.00498	B_c	1	0.6091	0.1718
B	5	4.010E-4	1.020E-4	S_c	1	2.572	2.757
$\bar{B}$	1	0.382	0.0412	L_c	1	3.664	8.993
S1	1	4.20E-4	0.048E-4	D_{cr}	1	0.579	0.137
S2	1	0.0220	0.00838	D_{cp}	1	0.5536	0.1173
S3	1	5.474E-5	1.597E-4	E_1	2	0.003125	9.865E-4
S4	1	1.252	0.286	$\bar{E}_1$	1	0.0965	0.0104
D1	1	6.607	1.399	E_2	2	2.614	2.607
D2	1	0.153	0.0247	$\bar{E}_2$	1	0.0128	0.0298
D3	1	0.798	0.134	S5	1	8.924	0.696
C1	2	1.047	0.220	D9	1	1.234E-4	3.259E-4
L1	1	94.39	4.346	A_{cL}	5	0.0524	0.0156
L2	1	0.3698	0.2207	Q	5	4.812E-4	6.111E-4
L3	1	63.93	21.50	D10	1	2.865E-4	9.257E-4
D4	1	0.00451	0.0118	C3	2	5.559	1.794
D5	1	0.00890	0.00242	B_{cL}	1	0.00576	0.00633
D6	1	0.205	0.00899	S_{cL}	1	0.07454	0.1344
D7	1	0.135	0.0148	E_3	2	0.00974	0.00298
D8	1	0.0122	0.00304	$\bar{E}_3$	1	5.42E-4	0.00188
C2	2	3.322	0.912	E_4	2	1.335E-5	3.456E-5
P	5	0.2233	0.2701	$\bar{E}_4$	1	0.0121	0.00682

Table 4.2: **The estimates of rate coefficients after 3 cycles of Computing Life for the genetic network shown in Figure 1.2.** All symbols are similar with Table 4.1.

X	k	$< X >$	σ(X)	X	k	$< X >$	σ(X)
A	5	5.60481E-4	2.97967E-4	A_c	5	0.57962	0.37478
$\bar{A}$	1	0.12207	0.00571	B_c	1	52.19490	25.41051
S1	1	0.11210	0.18048	S_c	1	0.37529	0.70020
S2	1	86.98142	8.23094	L_c	1	7.30599	14.4736
S3	1	0.01720	0.06524	D_{cr}	1	0.58116	0.26025
S4	1	39.02247	6.10560	D_{cp}	1	0.36204	0.16057
D1	1	1.70002E-4	5.04858E-4	D2	1	0.01101	0.00162
D3	1	3.43327	0.60242	C1	2	2.37212	1.50706
C2	2	36.17778	11.09423	C3	2	0.05305	0.1832
L1	1	90.45906	7.23261	L2	1	1.66063	0.21721
L3	1	8.87033	2.46278	D4	1	0.13855	0.01669
D5	1	20.49211	7.75688	D6	1	0.48731	0.03427
D7	1	0.01759	0.00115	D8	1	2.61376E-5	6.46480E-5
P	5	0.01270	0.01688				

Table 4.3: **The quality of fit of the model usually improves in successive cycles through the Computing Life paradigm.** The column n reports the number of data used in fitting, which is cumulative across cycles. χ^2 is the goodness of fit measure, and χ^2_{min} is the minimum chi-squared statistic over the ensemble. $\hat{\sigma}^2$ is the estimates of the error variance σ^2, calculated using the formula $\hat{\sigma}^2 = \frac{1}{n}\chi^2_{min}\sigma_0^2$, where $\sigma_0^2 = 0.02$ is the error per observation in the multivariate Gaussian likelihood. The genetic network fitted is shown in Fig.2.2, except cycle 4. This table was used in [13]

profiling Experiment	n	χ^2_{min}	$\hat{\sigma}^2$
data from literature (cycle 0)	333	1188	0.0714
circadian cycle in the dark (cycle 1)	553	2918	0.1055
light response in D/L artificial day (cycle 2)	1927	3938	0.0409
WCC response by turning of WCC (cycle3)	2165	5528	0.0511
genetic network of model in Fig. 2 (cycle4)	3007	4640	0.0309

Chapter 5

A Temperature Responsive Model of the Biological Clock

5.1 Temperature Compensation of *Neurospora Crassa*

The biological clock network model that we have described and discussed by now has a circadian daily rhythm and also can adapt to external light stimuli. Besides light, the biological clock is also affected by other stimuli such as temperature and chemicals. Similar to the response to the external light entrainment, the biological clock can be reset by temperature pulses or steps, and it also can adapt to temperature entrainment, which confirms temperature as another important entrainment factor [44]. However, there is another interesting feature that makes temperature stimuli quite special: the period of the biological clock remains stable over a relatively broad temperature range. This mechanism is commonly called "temperature compensation" which is the key feature to make the clock tick with accurate time in spite of the ambient temperature variation in the natural environment [45, 46]. Also, there are physiological temperature limits for clock function, *i.e.*, the biological clock only works rhythmically within a certain temperature range. Outside this range the clock will stop ticking and freeze in a certain phase [47]. All of the above features motivate us to add a temperature response feature to our light responsive model.

However, how the biological clock works to produce such a temperature compensation remains a puzzle. A famous law for the temperature dependence of an elementary chemical reaction rate is the Arrhenius equation [48]:

$$k = k_0 \exp(-\Delta E_a / RT) \tag{5.1}$$

which means that the reaction rates can rise very rapidly as the temperature T (in Kelvin) rises. Obviously if the Arrhenius equation is the only temperature dependence that works in our genetic network model, one would naively expect that the oscillation period would be shorter for a higher temperature since all reaction would simply speed up and vice versa. There must be some mechanism that allows the genetic network to keep the period constant when T is raised.

Recent experimental research shows that the *frq* gene plays an essential role in the temperature response of the biological clock of *Neurospora crassa*. While the wild type strain of *Neurospora crassa* could be entrained to external ambient temperature cycles, the *frq*-null mutant strain, in which the *frq* gene is either deleted or disabled, could not truly entrain to the same cycles. The *frq* gene and its derivatives, *i.e.*, *frq* mRNA and FRQ protein, are then believed to be the central components of the temperature response mechanism in the Neurospora circadian clock [49].

Liu and Dunlap *et al.* [47] first proposed a complementary response of two FRQ isoforms that give temperature stability at different temperatures separately. They suggested that there exists temperature sensitive translational control of the production of the main oscillator protein FRQ, which results in the two isoforms (*i.e.*, different proteins produced from the same gene and the same RNA) of FRQ: $FRQ^{1-989}$and $FRQ^{100-989}$. There is experimental evidence showing that they have similar function but different behavior. Either form of FRQ is able to support circadian rhythmicity independently. But the longer isoform (FRQ^{1-989}) generates a shorter period than the shorter isoform $FRQ^{100-989}$ does. And the long FRQ^{1-989} is relatively more abundantly produced at a higher temperature, while the short $FRQ^{100-989}$ is relatively more abundantly produced at lower temperature [47]. They are produced through alternative in-frame initiation of translation. This competing initiation translational control then forms the proposed mechanism of temperature compensation. Liu's report, coming with rich experimental data, described this hypothesis only qualitatively. Akman *et al.* [50] created a toy model to represent this hypothesis, but they did not seriously reconcile their model

with the data. We then decided to combine this interesting mechanism with our identified model in Figure 1.2. The modified genetic network is shown in Figure 5.1.

Different hypotheses have also been proposed to explain how the *frq* gene and products are related with the temperature compensation mechanism. A second hypothesis is given by Ruoff and Dunlap [51] saying that there is a careful balancing of rate coefficients to yield temperature compensation, which is then a property of the clock network and involves an antagonistic balance of different reactions. A third hypothesis proposed by Hong and Tyson [52] on the contrary considers antagonistic balance extremely unlikely. Rather they develop a simple model in which some of the rate coefficients are temperature-dependent by analogy to a mechanism they proposed for the cell cycle. Brunner and Dunlap *et al.* [53] have developed a fourth hypothesis that the splicing in Introl 6 of FRQ during the transcription of *frq* mRNA is temperature dependent and may contribute to temperature compensation. This is an alternative mechanism to Liu's original hypothesis of the generation of two FRQ isoforms.

5.2 A Primary Model of the Alternative Initiation Translational Control

The proposed competing initiation of the translational control process are illustrated in Figure 5.2. Multiple ribosomal subunits, including 40S and elF2.GTP.Met-tRNAi (for short denoted by "E" below), must assemble on the *frq* mRNA to form a fully functional ribosome, before translation can begin at translation initiation sites (TIS) AUG#1 or AUG#3, respectively. First, 40S, from the solution in the cytoplasm attaches itself at the binding site A and scans along the *frq* mRNA in the 5'-to-3' direction. If the 40S captures E from solution, with a certain probability, before reaching AUG#1, the long isoform FRQ^{1-989} will be produced by translation starting at AUG#1; if 40S captures E after passing AUG#1, but before reaching AUG#3, the short isoform $\mathrm{FRQ}^{100-989}$ gets produced; if 40S doesn't capture anything before reaching AUG#3, nothing gets produced.

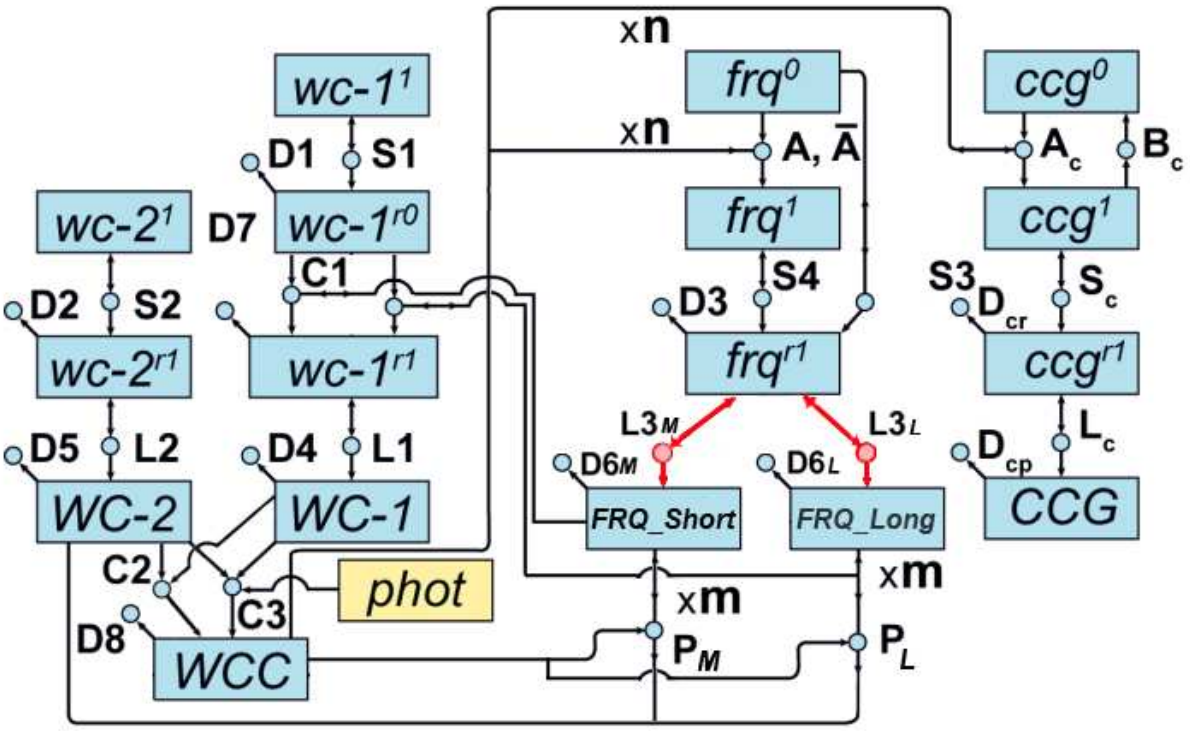

Figure 5.1: **Genetic network diagram of the alternative initiation translation control model.** Comparing with the original model in Figure 1.2, the only difference in the diagram in Figure 5.1 is the two isoforms of FRQ protein in parallel. The alternative translation reactions are emphasized by red arrows. "FRQ_Long" represents FRQ^{1-989} and "FRQ_Short" represents $\mathrm{FRQ}^{100-989}$. The subscript L or S, *e.g.*, in $\mathrm{P_S}$, mean the reaction is related to the "Long" or "Short" isoform respectively.

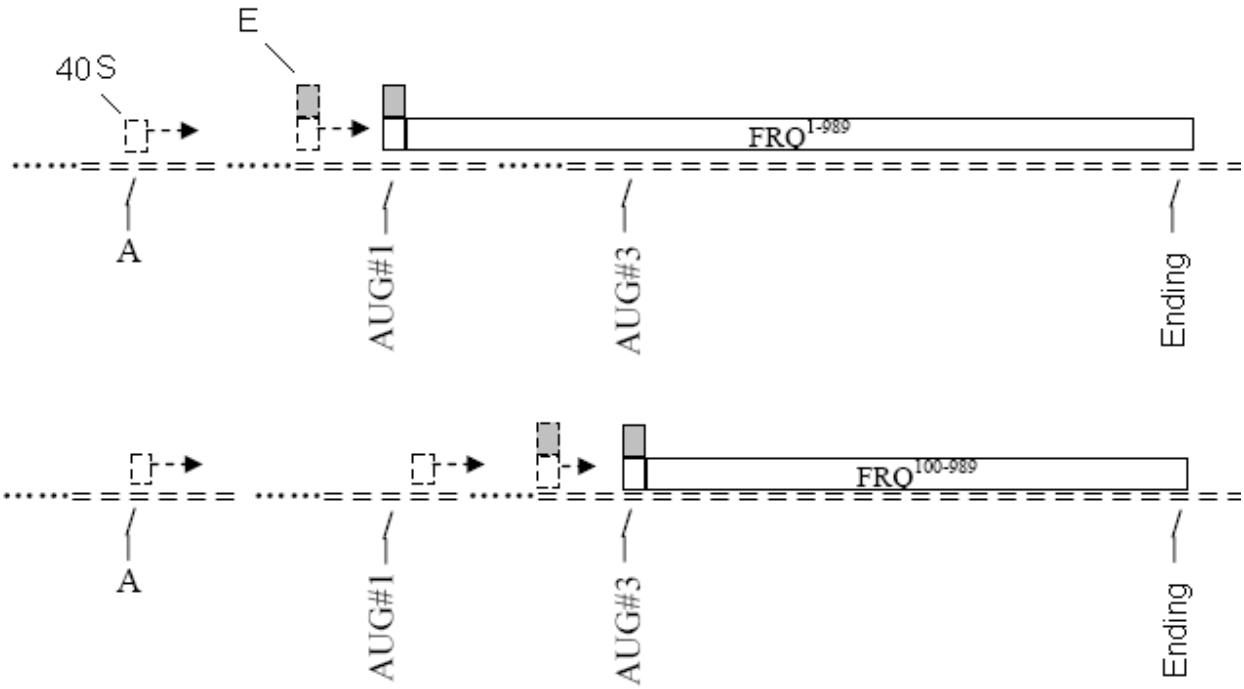

Figure 5.2: **Schematic of frq mRNA translation into two FRQ isoforms**. A is the 40S attachment site at the 5'-end of the mRNA strand; AUG#1 labels Long isoform translation initiation site; AUG #3 labels Short isoform translation initiation site; "Ending": Translation termination site for both isoforms at the 3'-end of mRNA strand. "=" in the graph means an mRNA site (base pair A, C, G, or U).

A mathematical model describing this mechanism of alternative translation initiation can be set up by treating the formation of the 40S-E complex as a Poisson event. Suppose the 40S ribosomal subunit attaches on the *frq* mRNA strand at site A with a probability k_A (*i.e.*, the 40S-A-attachment reaction rate coefficient), and denote the probability of 40S grabbing a ribosomal subunit E while residing at some site on the mRNA strand as p, so the probability of 40S moving to the next site without capturing an E unit is $q = 1 - p$. Thus according to the Poisson process, the possibility of 40S capturing an E subunit after it moves exactly l steps, is

$$P_l = q^l(1-q),$$

where l = 0,1,..., counting from A. So the probability that the capture happens in no more than l steps (*i.e.*, exactly l steps or less) is

$$P_{k\leqslant l} = \sum_{k=0}^{l} P_k = (1-q)\sum_{k=0}^{l} q^k = 1 - q^{l+1} = 1 - (1-p)^{l+1},$$

where k is the number of steps that it takes 40S to capture an E unit. Similarly, the probability of 40S capturing the E unit after moving $l_1$steps but no more than l_2 steps ($l_1 \leqslant l_2$) is

$$P_{l_1\leqslant k\leqslant l_2} = \sum_{k=l_1}^{l_2} P_k = (1-q)\sum_{k=l_1}^{l_2} q^k = (1-q)\cdot q^{l_1}\sum_{k=0}^{(l_2-l_1)} q^k = q^{l_1+1} - q^{l_2+1}, \tag{5.2}$$

or, we can write Eq. 5.2 as

$$P_{l_1\leqslant k\leqslant l_2} = [1-(1-p)^{l_2+1}] - [1-(1-p)^{l_1+1}] = P_{k\leqslant l_2} - P_{k\leqslant l_1}. \tag{5.3}$$

Since the translation of the long isoform FRQ^{1-989} requires the 40S to capture the E unit before it arrives at initiation site AUG#1, the net production rate coefficient of FRQ^{1-989} is

$$k_L = k_A[1-(1-p)^L], \tag{5.4}$$

where k_A is the 40S-A-attachment reaction rate coefficient, and $L = 1420$ is the site number of AUG#1, counting site A as the first one. Similarly, the net production rate of the short isoform $\mathrm{FRQ}^{100-989}$, which requires the capture to happen between AUG#1 and AUG#3, is

$$k_S = k_A[(1-p)^L - (1-p)^S] \tag{5.5}$$

where $S = 1720$ similarly is the site number of AUG#3.

Both k_A and p describe elementary reaction steps and, therefore, likely follow the Arrhenius Law Eq. 5.1, *i.e.*, $k_A = k_{A0}\exp(-\Delta E_A/T)$ and $p = p_0\exp(-\Delta E_p/T)$. In this manner the temperature dependence of the complicated mRNA scanning and isoform production process can be modeled by only 4 additional unknown thermal parameters: the activation energies, ΔE_A and ΔE_p, and the pre-factors k_{A0} and p_0.

The same Poisson kinetics approach as described above can also be used to model the temperature dependence of other translation reactions and of the transcription reaction caused by alternative initiation, with or without isoforms. Other reactions shown in the network in Figure 5.1 are elementary and can be modeled by a conventional Arrhenius law. However, it is noted that we don't have all reactions modeled by Arrhenius law to keep this primary model as simple as possible for easy fitting. Therefore, for the genetic network shown in Figure 5.1 featured with the alternative initiation translational control, we only add the above 4 more unknown parameters to this model for the temperature dependence comparing with our previous light-responsive model in Figure 1.2. If this model works, the genetic network model shown in Figure 5.1 would be a complete biological clock model with three most important features of a biological clock: circadian rhythmicity, light response and temperature response.

5.3 Ensemble Fitting of the Temperature Response Model with Wild Type Data

An ensemble fitting has been performed using the genetic network in Figure 5.1 featuring the alternative transcription initiation mechanism above (shown in Figure 5.2). The experimental data are obtained from experiments performed on wild type strains of *Neurospora crassa* in the dark under 6 different temperatures (18℃, 20℃, 22℃, 25℃, 27℃ and 30℃) published

by Liu *et al.* [47]. At the same time, part of literature microarray data were also included to constrain the model to still keep the correct light-response behavior while developing the new temperature-response feature.

Since the genetic network model is a slightly modified version of our light-responsive model shown in Figure 1.2, we initialized the reaction rate coefficients based on our previous fitting results using this model [13]. The prefactor k_{A0} was initialized with the same value of L_3 in Figure 1.2, and the initial value of the activation energy ΔE_A, in units of Kelvin temperature (so is ΔE_p), was chosen to be the room temperature, 298.15K. This initial setup of k_{A0} and ΔE_A, together with other reaction rate coefficients directly from the old model in Figure 1.2, mathematically made the new model have an easy start in an equivalent point in the $\boldsymbol{\theta}$ space to the already well fitted light responsive model.

The initial value of another activation energies, ΔE_p, and the pre-factors p_0 were chosen such that the value of $p = p_0 \exp(-\Delta E_p/T)$ has the same order of magnitude as 1/L, *i.e.*,$1/1420 \sim 10^{-3}$. This therefore ensured the values of both $1-(1-p)^L$ and $(1-p)^L-(1-p)^S$ to be reasonable , *i.e.*, both values should not be too small to be numerically calculated. That is to say, physically, we want to be assured that the amount generated for both isoforms of FRQ would be at comparable levels. This has been verified by experimental data, that the ration of FRQ^{1-989} and $\text{FRQ}^{100-989}$ has the same order of magnitude of 1 [47].

The fitting results are shown in Figure 5.3. The discrete dots represent the experimental data, and the curve is the ensemble averages with ± 2 SD showing by the gray band, all expressed in decadic log-concentration. The predictions by the model ensemble are in pretty good agreement with the experimental data over the whole temperature range. So we can conclude that our new light- and temperature- responsive model is sufficient to explain these published profiling data generating by wild type strains.

The ensemble fitting results of the 4 newly added parameters are summarized in Table 5.1 with the ensemble averages and the ensemble SDs listed. It is easy to verify that $p = p_0 \exp(-\Delta E_p/T)$ has the order of magnitude of 10^{-3} over the temperature range used here

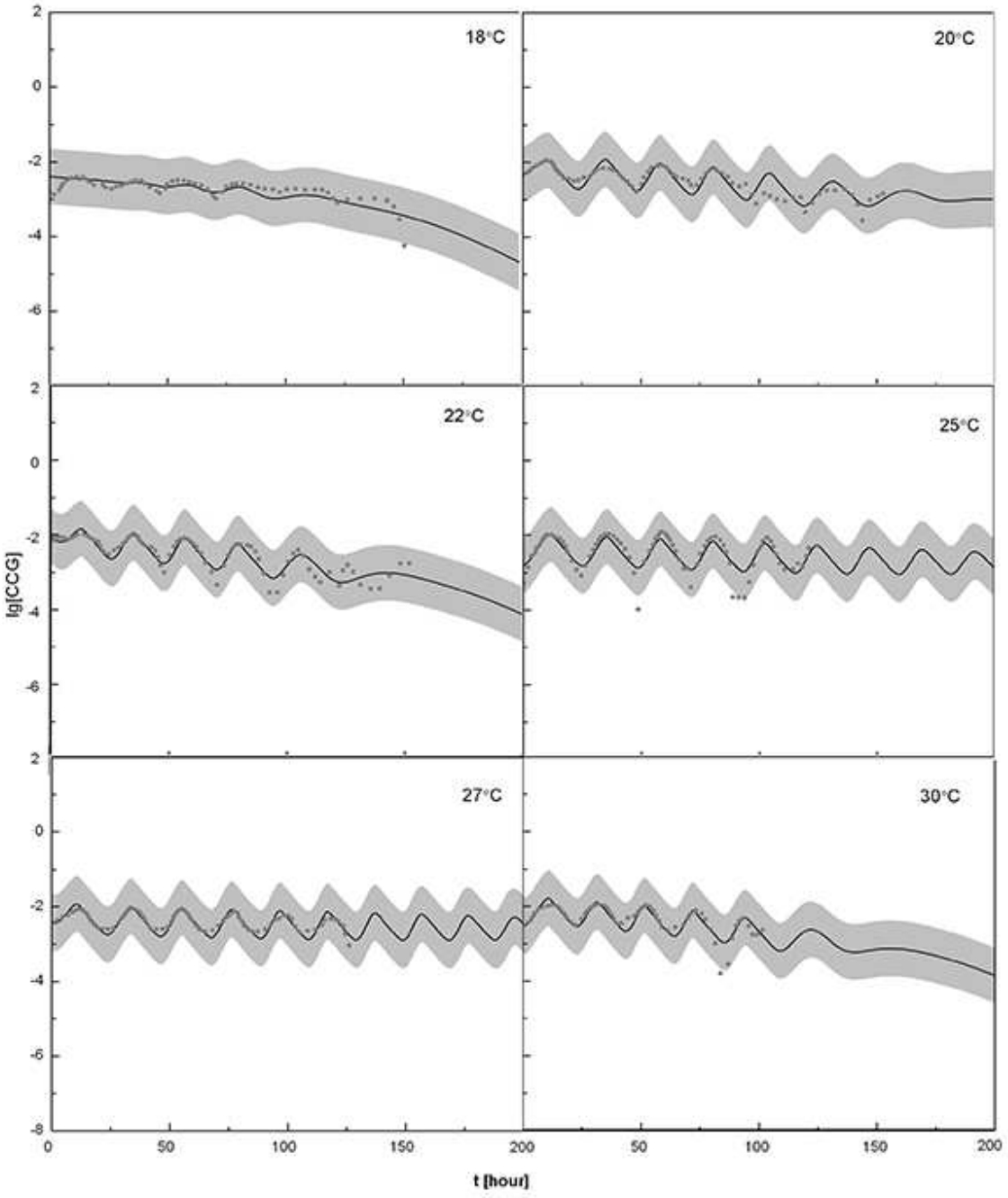

Figure 5.3: **A model ensemble for the genetic network in Figure 5.1 predicts the profiling data of wild type *N. crassa*.** The strains were grown in the dark under 6 different temperature. The discrete dots represent the experimental data, and the curve is the ensemble averages with ± 2 SD showing by the gray band, all expressed in decadic log-concentration. The predictions by the model ensemble are in pretty good agreement with the experimental data over the whole temperature range.

Table 5.1: **The model ensemble predictions for the 4 parameters key to the alternative initiation translation mechanism in Figure 5.2** $< X >$ is the ensemble average and $\sigma(X)$ is the ensemble SD defined same as before. The unit of ΔE_p and ΔE_A is Kelvin.

X	p_0	ΔE_p	ΔE_A	k_{A0}
$< X >$	6.3108094E-01	2.2084690E+03	4.0160124E+02	6.5413293E+00
$\sigma(X)$	2.9876921E-02	1.4351865E+01	9.0030012E+00	1.9426574E-01

(18-30℃). In Figure 5.4 (A), the predicted translation rate coefficients k_L Eq. 5.4 and k_S Eq. 5.5 are plotted against temperature. The plots of k_S have been moved vertically without changing the scale for easy comparison. Both k_L and k_S increases with temperature, but k_L increases in a much faster manner. The k_L/k_S ratio in the Figure 5.4(B) shows this trend more explicitly.

The translation ratio k_L/k_S alone cannot tell us how effectively each of these two isoforms functions in the genetic network in degrading the WCC activator. We need to find some more straightforward quantity to evaluate their performance in WCC degradation. It might be intuitive to use the respective FRQ protein concentration of these two isoforms, since [FRQ] is directly related to the P-reaction which results in the negative feedback loop. However it is not sufficient to use the concentration [FRQ], because the WCC degradation rate also depends on the P-reaction rate coefficients (P_L and P_L for the long isoform and short isoform respectively), and these P-reaction rate coefficients are randomly varied in the MC random walk, along with the random variations of the concentration [FRQ]. Further more, the MC random walk during the ensemble simulation randomly varies certain scale factors, *i.e.*, in effect it randomly changes the units in which concentration is measured. Both the FRQ concentration and P-coefficient values depend on that random choice of scale factors.

Thus we define a scale free quantity $\mathrm{P}_X[\mathrm{FRQ}^{\mathrm{X}}]^m[\mathrm{WCC}]$ (where X $=L$ or S, and the cooperativity m=4, refer to Figure 5.1). It is a quantity that is not affected by the random unit

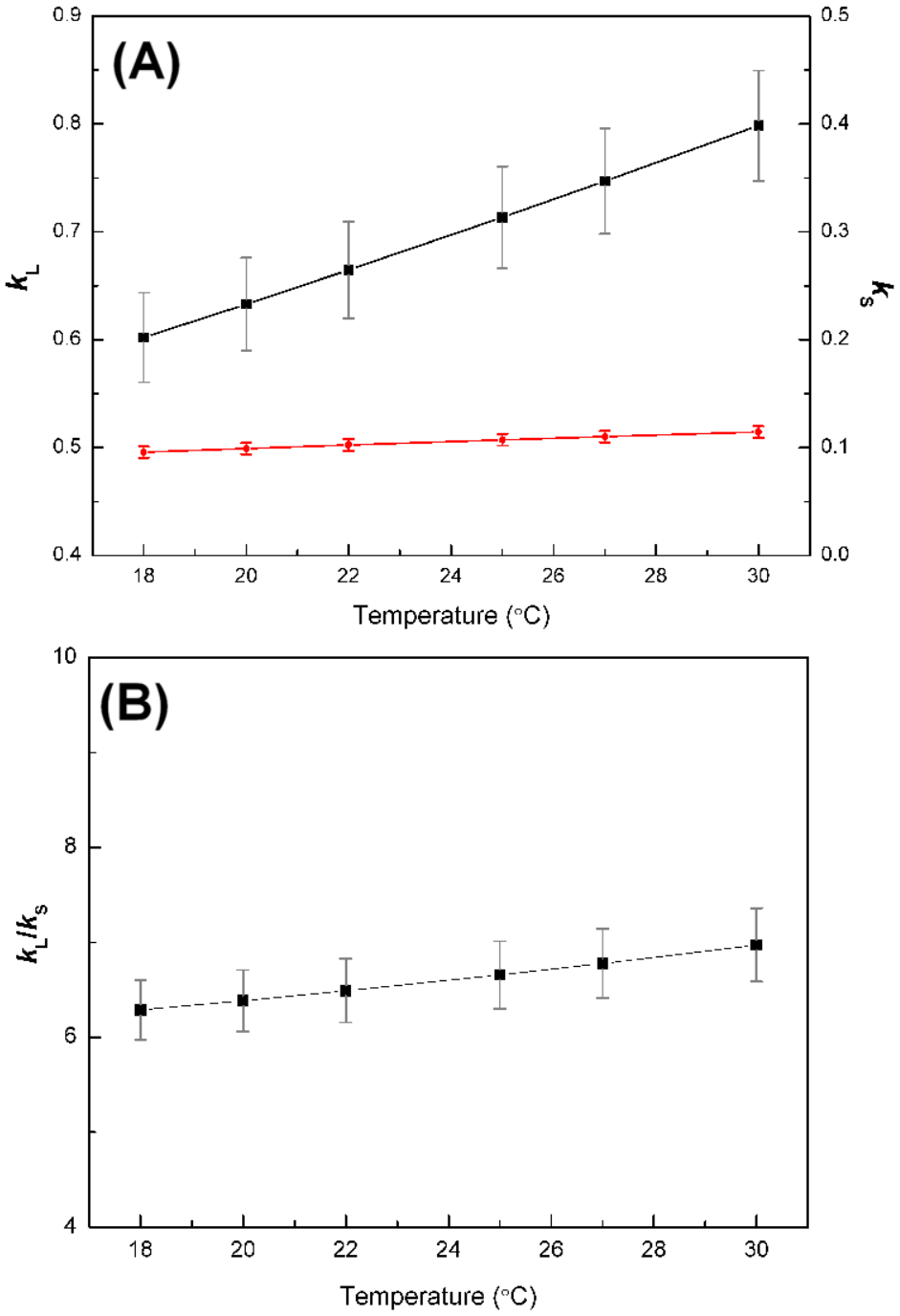

Figure 5.4: **The alternative translation mechanism makes $k_L$increases much faster than k_S with temperature.** (A) is the predicted value of k_L and k_S vs. temperature, respectively. (B) is the ratio k_L/k_S plotted against temperature. All values are calculated with the model ensemble predictions listed in Table 5.1. Variation is estimated assuming the 4 parameters are independent. The lines shown in the graph are auxiliary lines connecting data points.

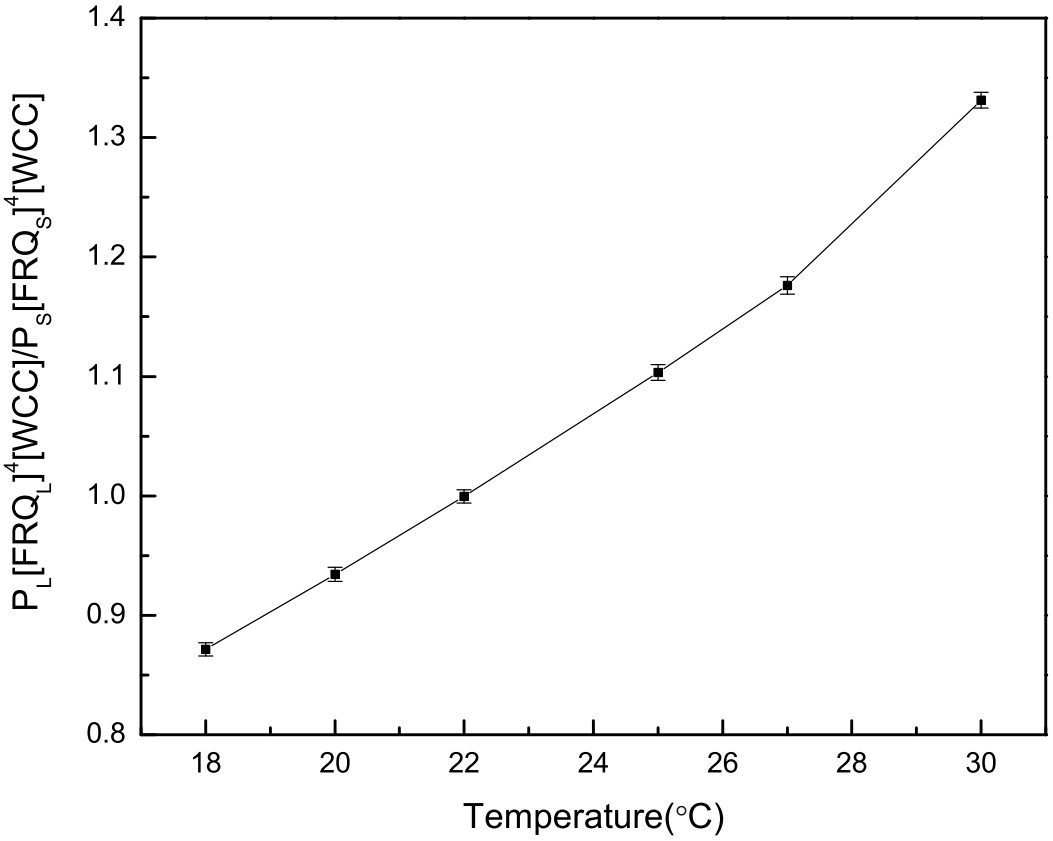

Figure 5.5: **The ratio of the scaling free quantity** $\mathrm{P_X[FRQ]}^m\mathrm{[WCC]}$ **of the two FRQ isoforms is plotted against the ambient temperature**. The model ensemble predicted scaling free quantity measures the efficiency of the FRQ protein in the P-reaction. Error bar is ensemble SD. At the low end of 18℃, the ratio is smaller than one, indicating that at lower temperature the shorter isoform is more efficient than the longer one. The lines shown in the graph are auxiliary lines connecting data points.

Table 5.2: **Temperature coefficient Q10 is calculated using 30℃ as reference temperature.** $Q10 := (\frac{P_1}{P_2})^{\frac{10}{T1-T2}}$, where T_2 here is the reference temperature 30℃ and P_2 is the corresponding period at 30℃. The unitless quantity Q10 here measures the rate of change of period in the biological clock of *N. crassa* for every 10℃ rise in the temperature. All Q10 values are closed to but smaller than 1, indicating the period is stable and only slightly decreased (i.e., shortened) with temperature rises.

T_1	18℃	20℃	22℃
Q10	0.9852 ± 0.0074	0.9058 ± 0.0043	0.9472 ± 0.0300
T_1	25℃	27℃	
Q10	0.8222 ± 0.0188	0.7662 ± 0.0283	

scaling used in the ensemble approach and thus allows meaningful comparisons. It directly measures the "power" or "efficiency" of FRQ^{1-989} and $\text{FRQ}^{100-989}$ in the P-reaction in Figure 5.1, respectively, by calculating the scale-independent reaction rate of the P-reaction. The plot of an ensemble average of this ratio of the two quantities against the temperature is shown in Figure 5.5. It proves again that the long isoform FRQ^{1-989} increases faster with temperature. However, it should be noted first of all that the ratio is close to 1 at all temperatures, which reveals a much smaller difference between the performance of the two isoforms than what is found by the translation ratio k_L/k_S. Furthermore, it is clear that at the low temperature end of 18℃, the ratio is smaller than 1, meaning that the shorter isoform does play a more important role in the biological clock at the lower temperature.

Figure 5.6 shows the model ensemble prediction of the period of the biological clock under different temperatures. It can be seen that they remain relatively stable and fluctuate in a narrow range of ∼21-24 hours, which is consistent with the concept of temperature compensation. This is more clear with the calculated temperature coefficient Q10 listed in Table 5.2. Q10 is a unitless quantity that measures the rate of change of a certain quantity (*e.g.*, reaction rate) in a biological or chemical system for every 10℃ rise in the temperature.

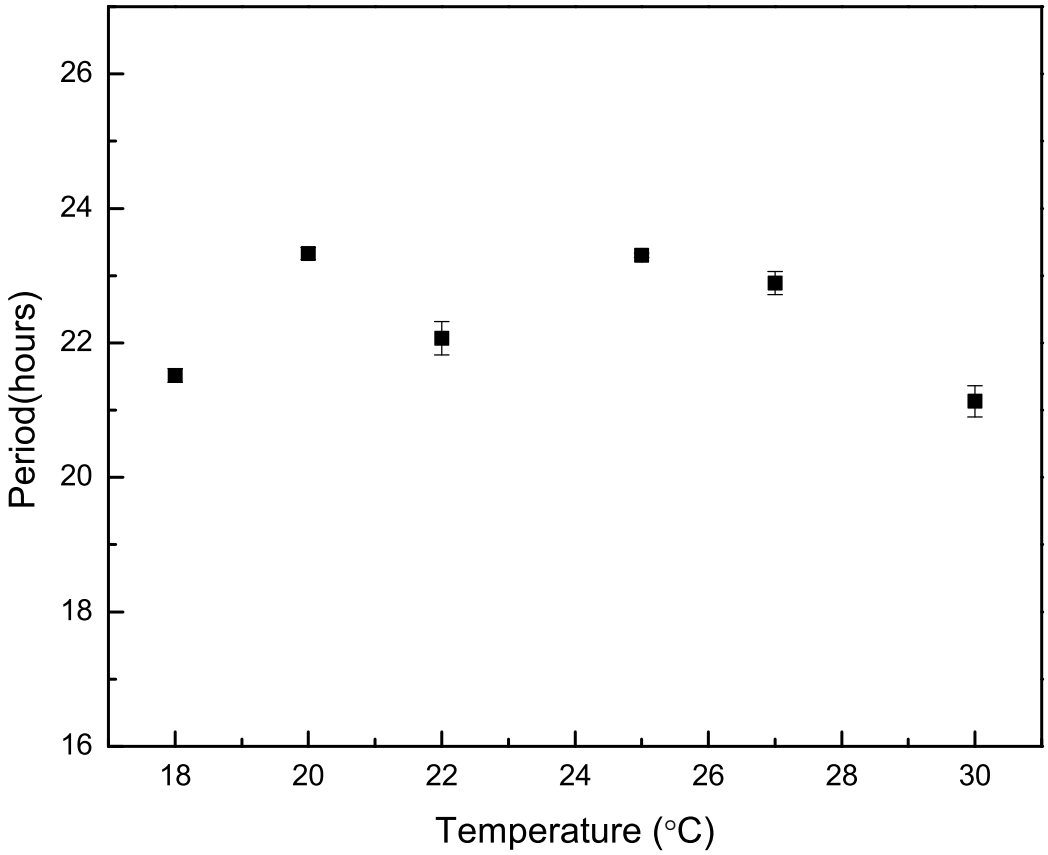

Figure 5.6: **The period of the biological clock at 6 different temperatures over a range of 10°C remains stable** The values of period are obtained from the fitted model ensemble average in Figure 5.3. They remain stable with slight fluctuation in the range of 21-24 hours.

The closer the Q10 values are to 1, the stabler the measured quantity is with temperature change. In Table 5.2, Q10 is calculated with 30℃ as the reference temperature. All Q10 values are closed to but smaller than 1, indicating the period is stable and only slightly decreased (i.e., shortened) with temperature rises.

5.4 Mutant Type Data: Discrepancies and Possible Solutions

Although the Figure 5.3 shows that the ensemble of models identified for the genetic network in Figure 5.1 gives a good fit to the experimental data of *wild type* strain of *Neurospora crassa*, discrepancies show up when we move on to the mutant type experiments under the same 6 different temperatures. Liu's experimental work [47] shows that, if either one of the two initiation sites is disabled, the resultant mutant strains will behave differently than before. The AUG#1 mutant (*i.e.*, AUG#1 is disabled) generates only the short $\mathrm{FRQ}^{100-989}$, which results in the elimination of the rhythmicity at temperature near the high end of the physiological temperature range. In contrast, the AUG#3 mutant generating only the long FRQ^{1-989}, results in the elimination of the rhythmicity at the low end of the physiological temperature range. This at first seems quite easily explained by our model in Figure 5.1 and Figure 5.2, since the analysis above shows that the short isoform has a lower translation rate, but it is relatively more efficient at lower temperature. The longer isoform FRQ^{1-989} is less efficient at the lower end than the shorter one.

However, it must be noticed that, once one of the initiation sites is deleted or disabled, the alternative initiation translation mechanism at once degenerates into the same Poisson model with only one initiation site which is just located at different positions on the *frq* mRNA strain. What's more, the two sites are very close to each other, considering that L=1420 and S=1720 are the same order of magnitude. Therefore, the behavior with respect to the temperature change of the two mutant types should be very similar to each other. This makes it quite difficult to explain the different response that the short FRQ-protein would work at low temperature end but fails at the high end, while the long isoform works at

high temperature end and fails at the low end. Liu explained this by qualitatively and very briefly stating that there might be a threshold of the concentration of FRQ [47]. However, our continuous failures of fitting the two mutant experimental data disprove this possibility, since the genetic network in Figure 5.1 automatically contains the requirement of appropriate concentration of FRQ to serve as the input of P-reaction. There must be other reasons, or say mechanisms, to explain the different responses of the two isoforms of the FRQ proteins.

This motivates us to again look at the genetic network in Figure 5.1 more carefully. The only differences among those parameters that relate to the two FRQ isoforms, besides the site position of the translation reaction, are the decay reaction rate coefficients and the P-reaction coefficients. And as the key component of the genetic network that gives the negative feedback loop, the P-reaction affects the rhythmicity of the clock most directly. What if the P-reaction is also temperature dependent? The two isoforms of FRQ protein with different length are very possible to have different properties with respect to the ambient temperature changes, as temperature changes could affect their folding differently. In fact, a recent report from Mehra and Dunlap [54] also notices this problem. They suggest the temperature dependence of the P-reaction is an additional factor to the alternative translation mechanism that Liu has reported [47]. Based on the above idea, two possible modifications of the model in Figure 5.1 are proposed for future work. We are going to introduce them below, respectively.

5.4.1 Independent P-Reaction Rate Coefficients Model

This solution to modeling the mutant data is a very simple idea. It is more like an empirical solution which simply allows the rate coefficient of the P-reaction to be varied independently under each temperature for each isoform, instead of having a fixed rate coefficient for all temperature and for both isoforms as what we are doing right now. This gives extra degrees of freedom for each isoform and allows them to develop different behavior under different temperatures. No physical or mathematical model is considered at this stage, however. If all the mutant type experimental data could be fitted in this way, the two groups of ensemble

predictions of the P-reaction rate coefficients for each isoform can be listed out and plotted. It is natural to expect that they would show different behaviors with respect to temperature. Possible model fitting could be proposed afterwards.

Then we introduce 6 x 2=12 new parameters in total into the model. This fitting is still underway. Currently, the AUG#3 mutant with only long isoform has been fitted at higher temperature end of 25℃and 27℃. It is reasonable to have these two fitted since it is equivalent to our previous light responsive model in Figure 1.2.

5.4.2 Multi-State FRQ proteins Model

This model starts from the fact that different protein folding and shaping, corresponding to different energy status, should have different behavior in the translation reactions. Thus, suppose in order to take part in the P-reaction, there is an energy threshold for the protein. For the short isoform, its energy must be lower than this threshold (otherwise it changes folding shape and is deactivated); similarly for the long isoform, its energy should be higher than this threshold. Let's first concentrate on the situation of the shorter isoform. Suppose there are N_l possible energy levels below the threshold and N_h above it. And for simplification, we assume the N_l lower energy levels are degenerate with the energy E_l, and the other N_h higher status are degenerate and the energy is E_h. So, the probability ratio to find a FRQ protein in the lower energy level and in the higher energy level is

$$\frac{P_l}{P_h} = \frac{N_l}{N_h} \exp(-\frac{E_l - E_h}{RT}).$$

Notice that $P_l + P_h = 1$, let N denote $\frac{N_h}{N_l}$. It is easy to see $N > 1$; and $\Delta E = E_h - E_l$ where $\Delta E > 0$. The probability of finding a FRQ protein at the lower energy level is

$$P_l = \frac{1}{1 + N \exp(-\frac{\Delta E}{RT})}.$$

So for a shorter isoform FRQ, we need to take this P_l into account as a prefactor of the P-reaction. This will introduce two more parameters N and ΔE. Similarly, the probability of a long FRQ protein at higher energy level is

$$P_h = \frac{1}{1 + \frac{1}{N'} \exp(\frac{\Delta E'}{RT})},$$

where N' and $\Delta E'$are two more independent new parameters to be added in. It can be easily seen that P_h and P_l do bring different temperature dependency for the two isoforms for the FRQ proteins.

The work of ensemble identification of this model is beyond the scope of this dissertation. These two new models discussed in this section will become the focus of future work.

Chapter 6

Conclusion

Throughout the work in this dissertation, genetic networks consisting of bio-macromolecules and their interactions, have been used as a powerful framework for the quantitative description of the structure and dynamics of a biological system. Being built from a currently limited biology knowledge database, genetic network modeling usually starts with an incomplete topology as well as plenty of unknown parameters, even if the targeted biological systems have already been completely sequenced and well studied, *e.g.*, like the biological clock of *Neurospora crassa*. Competing multiple hypotheses of genetic networks for a biological system reflect the incompleteness of experimental information. To test these hypotheses and identify a precise model for a genetic network of high degree of parametric freedom is a real challenge, considering the limited availability, high cost and large quantitative uncertainties of the biological experimental data. There needs to be a decision-making algorithm to suggest theoretically a next experiment that reduces most effectively the ambiguity in the network hypotheses [4]. The Computing Life paradigm introduced in this dissertation is such a novel iterative model-driven discovery process. This new methodology combines experiments seamlessly with computational experimental data fitting and theoretically best, or more precisely, maximally informative next experiment (MINE) design. The iterative process accumulates maximal information in each cycle and thus improves the discrimination between competing models and the precision of the model thereby identified.

The computational experimental data fitting method used in the Computing Life paradigm is the ensemble approach which has successfully identified an ensemble of oscillating network models quantitatively consistent with available RNA and protein profiling

data on the *N. crassa* clock [14]. The philosophy of searching for an *ensemble* of models is the essential idea here, since it is not possible to identify a unique model. The Metropolis algorithm, a Markov chain Monte Carlo method, is used to generate a large number of data-compatible models, all of which are treated as possible model candidates to represent our ensemble. Model parameters are estimates over the whole ensemble as ensemble averages and ensemble standard deviations. This approach is of great advantage when dealing with systems of a high degree of parametric freedom that are poorly constrained due to a lack of information from the data. The goodness of fit is evaluated by chi-square statistics with respect to the data, which also provides an effective criterion to compare and discriminate between different network hypotheses. The effectiveness of this approach has been verified throughout the three cycles of Computing Life for the biological clock. Both the precision of prediction and the ability of model comparison increased with the cumulative information from the data.

The MINE design is built upon the model ensemble formalism and it is the central part of this Computing Life paradigm: it provides a practical systems biology decision-making algorithm. The criterion used in the MINE approach is to maximize the independence of any pair of data points in observations to be done, based on the predictions made by the current model ensemble. Therefore, it compares between different choices of the next experiment $\mathbf{U}$ and tries to increase as much as possible the *scope* of observation (V($\mathbf{U}$))in the next experiment instead of *precision* of a single observation. For the biological purpose of "mining clock-controlled genes in the genome of *N. crassa*", after three cycles,from 4380 *N. crass* genes 295 such genes were selected that show meaningful biological significance [13]. At the same time, as we have mentioned above, the precision of the model ensemble is also improved. In cycle 1, the MINE choice of the initial observation time and the observation time spacing conforms to the conventional experimental practice [11]. The second MINE application in cycle2, suggests light exposure experiments with a 48hr artificial day, and it also motivated us with a short 6hr-day of light response experiment, which is usually avoided

by researchers [25], to examine the additional auto-feedback loop. In cycle 3, experiments designed to perturb WCC protein levels, the MINE-recommended knocking down of *wc-1* does not seem to be an obvious conventional choice, but the subsequent *wc-1* knock-down experiment successfully identified 328 clock-controlled genes and thereby validated this choice.

It must be noted that the selection of the next experiment **U** in fact is delimited by the specified biological goal, *e.g.*, the *ccgs* mining, which ultimately leaves the choice of what kind of set of next experiments to be performed in the hands of the experimenter. The MINE approach guides the design of the next experiment; on the other hand, the particular biological goal defines which detailed specific set of possible experiments it should choose from. One could naively make the biological goal the completely unconstrained objective of learning the most about the genetic network to instruct the MINE criterion, but then the set of possible experiments would become too large to actually realize the optimal experiment computationally. By specifying a particular biological goal, the design questions are then put into a more detailed and confined context and this makes it possible to parameterize and optimize the next experiment.

We finally also utilized the ensemble approach to study a model expansion which describes temperature compensation behavior of the biological clock. This is the first time in systems biology that a complete biological clock model has been identified with all three basic functions: circadian rhythm in the dark, light response and temperature compensation. The main feature of this expanded model is to allow for alternative translation initiation control of the FRQ protein that results in two FRQ isoforms with similar function but different temperature response. The ongoing project has successfully found an ensemble of models that is consistent with the experimental data from the wild type organism. However further simulations with mutant data show the incompleteness of this expanded model. They suggest the necessity to make the activator-degrading P-reaction (in Figure 5.1) have different temperature dependence for the two isoforms. Latest experimental data are consistent with this

idea [54]. Two possible model modifications have been developed to capture this differential temperature response of the P-reaction. These new models will be the focus of future work.

In conclusion, we have developed Computing Life as a novel model-driven iterative workflow process which fully integrates experimentation with a powerful new decision-making and modeling algorithms for systems biology. For the first time in systems biology, a biological clock model with both light response and temperature compensation features has been identified with this new approach, based on the wild type data.

Bibliography

[1] J. DeRisi and P. Iyer, VR & Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale.," *Science*, vol. 278, pp. 680–686, 1997.

[2] S. Gygi, B. Rist, S. Gerber, F. Trecek, M. Gelb, and R. Aebersold, "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.," *Nature Biotechnology*, vol. 17, pp. 994–999, 1999.

[3] A. Walhout, R. Sordella, X. Lu, J. Hartley, M. Temple, GF Brasch, N. Thierry-Mieg, and M. Vidal, "Protein interaction mapping in c. elegans using proteins involved in vulval development," *Science*, vol. 287, pp. 116–122, 2000.

[4] H. Kitano, "Systems biology: A brief overview," *SCIENCE*, vol. 295, pp. 1662–1664, 2002.

[5] T. Ideker, V. Thorsson, J. Ranish, R. Christmas, J. Buhler, J. Eng, R. Bumgarner, D. Goodlett, R. Aebersold, and H. L, "Integrated genomic and proteomic analyses of a systematically pertrubed metabolic network," *Science*, vol. 292, pp. 929–934, 2001.

[6] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268–276, 2001.

[7] J. Dunlap, "Molecular bases for circadian clocks.," *Cell*, vol. 271-290, p. 1999, 96.

[8] J. E. Galagan, S. E. Calvo, K. A. Borkovich, E. U. Selker, N. D. Read, D. Jaffe, W. FitzHugh, L. Ma, S. Smirnov, S. Purcell, B. Rehman, T. Elkins, R. Engels, S. Wang, C. B. Nielsen, J. Butler, M. Endrizzi, D. Qui, P. Ianakiev, D. Bell-Pedersen, M. A. Nelson, M. Werner-Washburne, C. P. Selitrennikoff, J. A. Kinsey, E. L. Braun, A. Zelter, U. Schulte, G. O. Kothe, G. Jedd, W. Mewes, C. Staben, E. Marcotte, D. Greenberg,

A. Roy, K. Foley, J. Naylor, N. Stange-Thomann, R. Barrett, S. Gnerre, M. Kamal, M. Kamvyssells, E. Mauceli, C. Bielke, S. Rudd, D. Frishman, S. Krystofova, C. Rasmussen, R. L. Metzenberg, D. D. Perkins, S. Kroken, C. Cogoni, G. Macino, D. Catcheside, W. Li, R. J. Pratt, S. A. Osmani, C. P. DeSouza, L. Glass, M. J. Orbach, J. A. Berglund, R. Voelker, O. Yarden, M. Plamann, S. Seller, J. C. Dunlap, A. Radford, R. aramayo, D. O. Natvig, L. A. Alex, G. Mannhaupt, D. J. Ebbole, M. Freitag, I. Paulsen, M. S. Sachs, E. S. Lander, C. Nusbaum, and B. Birren, "The genome sequence of the filamentous fungus *Neurospora crassa.*," *Nature*, vol. 422, pp. 859–868, 2003.

[9] S. Crosthwaite, J. Dunlap, and J. Loros, "Neurospora wc-1 and wc-2: transcription, photoresponses, and the origins of circadian rhythmicity.," *Science*, vol. 276, pp. 763–769, 1997.

[10] A. Froehlich, Y. Liu, J. Loros, and J. Dunlap, "White collar-1, a circadian blue light photorecptor, binding to the frequency promoter.," *Science*, vol. 297, pp. 815–819, 2002.

[11] B. Aronson, K. Johnson, and J. Dunlap, "Circadian clock locus frequency: protein encoded by a single open reading frame defines period length and temperature compensation.," *PNAS USA*, vol. 91, pp. 7683–7687, 1994.

[12] T. Schafmeier, A. Haase, K. Kaldi, J. Scholz, M. Fuchs, and M. Brunner, "Transcriptional feedback of *Neurospora* circadian clock gene by phosphorylation-dependent inactivation of its transcription factor.," *Cell*, vol. 122, pp. 235–246, 2005.

[13] W. Dong, X. Tang, Y. Yu, R. Nilsen, R. Kim, J. Griffith, J. Arnold, and H. Schüttler, "Systems biology of the clock in *Neurospora crassa*," *PLoS ONE*, vol. 3, p. e3105, 2008.

[14] Y. Yu, W. Dong, C. Altimus, X. Tang, J. Griffith, M. Morello, L. Dudek, J. Arnold, and H.-B. Schuttler, "A genetic network for the biological clock of *Neurospora crassa.*," *PNAS USA*, vol. 104, pp. 2809–2814, 2007.

[15] A. Correa and B.-P. D, "Distinct signaling pathways from the circadian clock participate in regulation of rhythmic conidiophore development in *Neurospora crassa.*," *Eukaryotic Cell*, vol. 1, pp. 273–280, 2002.

[16] J. Locke, M. Southern, L. Kozma-Bognar, V. Hibbard, B. Brown, T. MS, and M. AJ, "Extension of a genetic network model by iterative experimentation and mathematical analysis.," *Molecular Systems Biology*, vol. 1, p. 10.1038, 2005.

[17] M. Yeung, J. Tegner, and C. JJ, "Reverse engineering gene networks using singular value decomposition and robust regression.," *PNAS USA*, vol. 99, pp. 6163–6168, 2002.

[18] T. Gardner, C. Cantor, and C. JJ., "Construction of a genetic toggle switch in escherichia coli.," *Nature*, vol. 403, pp. 339–342, 2000.

[19] T. Hughes, M. Marton, A. Jones, C. Robers, R. Stoughton, and et al, "Functional discovery via a compendium of expression profiles.," *Cell*, vol. 102, pp. 109–126, 2000.

[20] C. H. Chen, C. S. Ringelberg, R. H. Gross, J. C. Dunlap, and J. J. Loros, "Genome-wide analysis of light-inducible responses reveals hierarchical light signaling in *Neurospora*," *EMBO J.*, vol. 28, pp. 1029–1042, 2009.

[21] Y. Yu, *Monte Carlo studies of genetic networks with special reference to the biological clock of Neurospora crassa.* PhD thesis, University of Georgia, Athens, GA, 2005.

[22] B. Aleman-Meza, Y. Yu, H. Schuttler, J. Arnold, and T. Taha, "Kinsolver: a simulator for computing large ensembles of biochemical and gene regulatory networks.," *Computers and Mathematics with Applications*, vol. 57, pp. 420–435, 2009.

[23] N. Garceau, Y. Liu, J. Loros, and D. JC, "Alternative initiation of translation and time-specific phosphorylation yield multiple forms of the essential clock protein frequency.," *Cell*, vol. 89, pp. 469–476, 1997.

[24] K. Lee, J. Loros, and J. Dunlap, "Interconnected feedback loops in the *Neurospora* circadian system.," *Science*, vol. 289, pp. 107–110, 2000.

[25] M. Görl, M. Merrow, B. Huttner, J. Johnson, T. Roenneberg, and M. Brunner, "A pest-like element in frequency determines the length of the circadian period in *Neurospora crassa*," *EMBO J.*, vol. 20, pp. 7074–7084, 2001.

[26] H. de Jong and D. Ropers, "Strategies for dealing with incomplete information in the modeling of molecular interaction networks," *Briefings in Bioinformatics*, vol. 7, pp. 354–363, 2006.

[27] H. De Jong, "Modeling and simulation of genetic regulatory systems: a literature review.," *J Comput. Biol.*, vol. 9, pp. 67–103, 2002.

[28] D. Battogtokh, D. Asch, M. Case, J. Arnold, and H. Schuttler, "An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of *Neurospora crassa.*," *PNAS USA*, vol. 99, pp. 16904–16909, 2002.

[29] K. . J. S. Brown, "Statistical mechanical approaches to models with many poorly known parameters.," *Phys. Rev E*, vol. 68, pp. 021904–1–021904–9, 2003.

[30] T. Gardner, D. d. Bernardo, D. Lorenz, and J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling.," *Science*, vol. 301, pp. 102–104, 2003.

[31] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf, "Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems.," *J. R. Soc. Interface*, vol. 6, pp. 187–202, 2009.

[32] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics (2nd Edition)*. Cambridge University Press, 2005.

[33] P. Cheng, Y. Yang, C. Heintzen, and Y. Liu, "Coiled-coil domain mediated frq-frq interaction is essential for its circadian clock function in neurospora.," *EMBO J.*, vol. 20, pp. 101–108, 2001.

[34] D. Gonze, J. Halloy, and A. Goldbeter, "Robustness of circadian rhythms with respect to molecular noise.," *PNAS USA*, vol. 99, pp. 673–678., 2002.

[35] P. Smolen, D. Baxter, and J. Byrne, "Modeling circadian oscillations with interlocking positive and negative feedback loops.," *J. Neurosci.*, vol. 21, pp. 6644–6656, 2001.

[36] R. Muirhead, *Aspects of Multivariate Statistical Theory.* New York: Wiley, 1982.

[37] V. V. Prasolov and V. M. Tikhomirov, *Geometry.* American Mathematical Society, 2001.

[38] A. Correa, Z. Lewis, A. Greene, I. March, R. Gomer, and D. Bell-Pedersen, "Multiple oscillators regulate circadian gene expression in *Neurospora.*," *PNAS USA*, vol. 100, pp. 13597–13602, 2003.

[39] R. De Paula, Z. Lewis, A. Greene, K. Seo, L. Morgan, M. Vitalini, L. Bennett, R. Gomer, and D. Bell-Pedersen, "Two circadian timing circuits in *Neurospora crassa* cells share components and regulate distinct rhythmic processes.," *J. Biol. Rhythms*, vol. 21, pp. 159–168, 2006.

[40] "See http://www.intl-light.com/handbook/; and especially http://www.intl-light.com/handbook/ch05.html for spectral distributions and http://www.intl-light.com/handbook/ch07.html for unit conversions."

[41] Z. Lewis, A. Correa, C. Schwerdtfeger, K. Link, X. Xie, R. Gomer, T. Thomas, D. Ebbole, and D. Bell-Pedersen, "Overexpression of white collar-1 (wc-1) activates circadian clock-associated genes, but is not sufficient to induce most light-regulated gene expression in *Neurospora crassa.*," *Molecular Microbiology*, vol. 45, pp. 917–931, 2002.

[42] M. Davidian and G. DM, *Nonlinear Models for Repeated Measurement Data.* New York: Chapman and Hall, 1995.

[43] J. Townsend, "Resolution of large and small differences in gene expression using models for the bayesian analysis of gene expression levels and spotted dna microarrays.," *BMC bioinformatics*, vol. 5: 54. doi:10.1186, pp. 1471–2105–5–54, 2004.

[44] B. Sweeney and J. Hastings, "Effects of temperaure upon diurnal rhythms," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 25, pp. 87–104, 1960.

[45] J. Hastings and B. Sweeney, "On the mechanism of temperature independence in a biological clock.," *PNAS USA*, vol. 43, pp. 804–809, 1957.

[46] C. Pittendrigh, "Temporal organization: reflections of a darwinian clock-watcher.," *Annual Review of Physiology*, vol. 55, pp. 17–54, 1993.

[47] Y. Liu, N. Y. Garceau, J. J. Loros, and J. C. Dunlap, "Thermally regulated translational control of frq mediates aspects of temperature responses in the *Neurospora* circadian clock," *Cell*, vol. 89, pp. 477–486, 1997.

[48] K. J. Laidler, J. H. Meiser, and B. C. Sanctuary, *Physical Chemistry.* Brooks Cole, 2002.

[49] A. Preguerio, N. P. Lloyd, D. Bell-Pedersen, C. Heintzen, J. Loros, and J. Dunlap, "Assignment of an essential rold for the *Neurospora* frequency gene in circadian entrainment to temperature cycles," *PNAS USA*, vol. 102, pp. 2210–2215, 2005.

[50] O. Akman, J. C. Lock, S. Tang, I. Carre, A. Millar, and D. Rand, "Isoform switching facilitates period control in the the *Neurospora crassa* circadian clock.," *Molec. Systems Biolog.*, vol. 4, p. 164, 2008.

[51] P. Ruoff, J. Loros, and J. C. Dunlap, "The relationship between frq-protein stability and temperature compensation in the *Neurospora* circadian clock," *PNAS USA*, vol. 102, pp. 17681–17686, 2005.

[52] C. Hong, E. Conrad, and J. J. Tyson, "A proposal for robust temperature compensation of circadian rhythms," *PNAS USA*, vol. 104, pp. 1195–1200, 2007.

[53] A. Diernfellner, H. Colot, O. Dintsis, J. Loros, J. Dunlap, and M. Brunner, "Long and short isoforms of *Neurospora* clock protein frq support temperature compensated circadian rhythms," *FEBS Letters*, vol. 581, pp. 5759–5764, 2007.

[54] A. Mehra, MS, C. Baker, H. Colot, J. Loros, and J. Dunlap, "A role for casein kinase 2 in the mechanism underlying circadian temperature compensation," *Cell*, vol. 137, pp. 749–760, 2009.