The Distribution of Mutational Effects on Fitness in a Simple Circadian Clock

Laurence Loewe[1] and Jane Hillston[1,2]

[1] Centre for System Biology at Edinburgh,
The University of Edinburgh, Edinburgh EH9 3JU - Scotland
Laurence.Loewe@ed.ac.uk
[2] Laboratory for Foundations of Computer Science,
The University of Edinburgh, Edinburgh EH8 9AB, Scotland
jeh@inf.ed.ac.uk

Abstract. The distribution of mutational effects on fitness (DME^F) is of fundamental importance for many questions in biology. Previously, wet-lab experiments and population genetic methods have been used to infer the sizes of effects of mutations. Both approaches have important limitations. Here we propose a new framework for estimating the DME^F by constructing fitness correlates in molecular systems biology models. This new framework can complement the other approaches in estimating small effects on fitness. We present a notation for the various DMEs that can be present in a molecular systems biology model. Then we apply this new framework to a simple circadian clock model and estimate various DMEs in that system. Circadian clocks are responsible for the daily rhythms of activity in a wide range of organisms. Mutations in the corresponding genes can have large effects on fitness by changing survival or fecundity. We define potential fitness correlates, describe methods for automatically measuring them from simulations and implement a simple clock using the Gillespie stochastic simulation algorithm within StochKit. We determine what fraction of examined mutations with small effects on the rates of the reactions involved in this system are advantageous or deleterious for emerging features of the system like a fitness correlate, cycle length and cycle amplitude. We find that the DME can depend on the wild type reference used in its construction. Analyzing many models with our new approach will open up a third source of information about the distribution of mutational effects, one of the fundamental quantities that shape life.

1 Introduction

Evolutionary theory has been very successful in predicting the fate of mutations in various settings, assuming that the mutational effect on fitness is known. Determining the actual effects of mutations is difficult. While many biological wet-lab experiments have been conducted with the aim of determining the effects of new mutations, these have been particularly successful for mutations with large effects, as experimental noise obscures small effects [1].

This is unfortunate, as the evolutionary fate of mutations with big effects on fitness is rather simple to understand: many advantageous ones become fixed in the population,

M. Heiner and A.M. Uhrmacher (Eds.): CMSB 2008, LNBI 5307, pp. 156–175, 2008.

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

so that in some future generation all individuals will have inherited a copy, while deleterious (harmful) ones are removed rather quickly. The most interesting questions are currently posed by mutations of "small", but not "too small" effects on fitness. Here the difference between "small" and "too small" depends on a threshold set by the effective population size, where mutations in the "too small" category are behaving as if they had no effect on fitness and thus exhibit simple neutral dynamics. Since all organisms have a genome with a large number of opportunities to mutate in different ways, it has become custom to summarise these possibilities in the form of a distribution of mutational effects on fitness (DME^F), which associates a frequency of the occurrence of mutational changes with each mutational effect and abstracts the various molecular causes that determine the size of that effect. Various evolutionary theories make varying assumptions about this distribution [1] and their quality as a predictive tool often depends on the underlying DME^F. These theories are important for understanding the evolution of genomic sequences and thus play a crucial role in efforts to interpret the sequence of the human genome [1]. Because of the paramount importance of DME^Fs, recent work in population genetics has started to estimate DME^Fs directly from sequence data [1,2]. Such methods are not limited by the lack of sensitivity seen in wet-lab experiments, as they exploit the sensitivity of the evolutionary process on DME^Fs to infer the location and shape of a given type of DME^F in systems that evolve according to well understood forces. The drawbacks include:

Limited mechanistic details. Current population genetic estimates of DME^Fs from DNA sequence data are descriptions of observations that lack a rigorous underpinning in the form of a mechanistic model of mutational effects. There is little information for distinguishing various types of distributions (e.g. gamma, lognormal), once certain broad criteria are met.

Limited applicability. Each distribution is only a snapshot of a specific DME^F for a specific organism. While comparing such snapshots helps to ascertain common features of DME^Fs, such descriptive results do not help with further explorations.

Sensitive to evolutionary process. All methods that estimate DME^Fs from DNA sequence data require a set of assumptions about the evolutionary process that led to the sample of DNA sequences used for the inference. These assumptions can be difficult to test and may cast doubts on estimates of a DME^F. Since several evolutionary processes can lead to similar features in a set of sequences, it can be challenging to disentangle their effects from those of the underlying DME^F [3].

In this paper we propose a third approach to the study of distributions of mutational effects on fitness, besides the direct experiments and the population genetical methods mentioned above. Our main contribution is to describe the approach and to demonstrate how it works in principle in a simple circadian clock model. We suggest that molecular systems biological models can be used to obtain much of the evolutionary interesting properties of a DME^F for a particular limited model system. Combining the *in silico* experimental techniques of molecular systems biology with knowledge of the study system from the wet-lab experiments allows the following improvements:

More mechanistic details. Molecular systems biology allows the construction of rigorous mechanistic models of biological systems that maintain a close link to biological reality. This reduces errors in estimates that are caused by biologically

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

misleading abstractions. In addition, such computational models allow the further exploration of parameter space at the low cost of simulations as opposed to the often prohibitive costs or difficulties of performing equivalent wet-lab experiments.

Precision. The precise control over every aspect of a model that comes with *in silico* models makes it possible in principle to compute emerging properties with a very high degree of precision. Depending on the stochastic nature and computational complexity of the model, it may still be too costly for some models to achieve the level of precision that some evolutionary questions require. However, we anticipate that many useful models can be analysed without such problems. In addition, advances that reduce the cost of computing and improve the speed of algorithms can be translated into increasingly precise estimates of DME^{F}s.

To demonstrate the feasibility of our new approach we deliberately choose a simple model to make it easier to focus on the fundamental challenges that arise from this new perspective. Such challenges include (i) the construction of computable fitness correlates that can be used as surrogates for biological fitness in the wild, (ii) the accuracy with which such fitness correlates need to be (and can be) computed and (iii) fundamental biological questions about distributions of mutational effects. For example, how often will a change of an underlying reaction rate improve or degrade overall functionality? Will the relative size of effects on the emerging properties of the system be larger or smaller than the relative size of mutational effects on reaction rates?

Results demonstrate that our new approach can be used in principle to infer interesting properties of distributions of mutational effects, where details strongly depend on the model under focus. The rest of the paper is structured as follows. In Section 2 we present some background on key ideas from evolutionary biology which we will use in the remainder of the paper. Section 3 outlines our framework for taking a systems biology approach to the study of the distribution of mutational effects. The model we consider in this paper is presented in Section 4 whilst its analysis is described in Section 5. A discussion of the results is given in Section 6 and conclusions in Section 7.

2 Background

Instead of obtaining DME^{F}s directly, our basic strategy is to (i) build a mechanistic model of how the phenotype changes depending on lower level changes in reaction rates that are ultimately caused by DNA changes, (ii) define a function that computes fitness from the phenotype and (iii) use random perturbations together with (i) and (ii) to determine the DME^F. While some commonly used models in quantitative genetics also compute a phenotype as an intermediate step towards computing the distribution of mutational effects [4], the approach presented here can include much more mechanistic detail by building on molecular systems biological data. DME^{F}s can be used to quantify *robustness*. Understanding robustness [5] is important for drug design [6].

2.1 A Nomenclature of Distributions of Mutational Effects (DMEs)

In this subsection we explain precisely what we mean by *distributions of mutational effects*. This is necessary to avoid confusion when discussing the various distributions. We consider each of the terms in turn (for examples, see Figures 8+9):

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

Effects. The *effects* are the changes in the emerging high-level systemic property under focus in the investigated system. DME^Y is used to denote a DME of the emergent system property Y. All Y are high-level properties, so a superscript is used.

We denote DMEs that describe the effects on fitness in the wild by DME^F, where the *fitness* can be easily linked to a trait like survival rate or fecundity that can be observed in its natural environment. In the more limited example of our circadian clock DME^L describes variations in the length of a cycle and DME^A variations in the amplitude of the oscillations. *Effects* are changes in *phenotype* properties.

Mutations. The *mutations* are low-level genotype changes that perturb the wild type reference system and cause phenotypic effects to change. At the lowest level *mutations* are DNA changes. In the absence of a mechanistic model for predicting enzymatic reaction rates from DNA, *mutations* can also be introduced as reaction rate changes, as the mechanistic chain of causality that links DNA changes and fitness changes must pass through the corresponding reaction rates at some point. $\mathrm{DM}_X\mathrm{E}$ is used to denote the genotypic perturbations that are introduced into property X to measure a DME. All X are low-level properties, so a subscript is used.

If *mutations* are a representative sample of naturally occurring DNA changes, we omit X, as this is the most natural and most important DME. In the more limited example of our circadian clock we can only change the reaction rates listed in Table 2. For example $\mathrm{DM}_{v_d}\mathrm{E}^L$ denotes the *distribution of mutational changes in protein degradation rate v_d that have effects on the length of clock cycles L*.

Distribution sign. One may want to focus only on increases or decreases of the values of a DME. For example, advantageous mutations in the DNA that increase fitness could be analyzed separately from survival compromising mutations that decrease fitness. Here we denote an increase and decrease with the additional letter 'I' and 'D', respectively. If these occur in a high-level emerging property of the system, the letters are superscript, if in low-level mutational changes, the letters are subscript. Specifying nothing is equivalent to 'DI'.

Thus a *distribution of increasing mutational changes in protein production rate k_s that have only decreasing effects on fitness* is denoted by $\mathrm{D}_I^D\mathrm{M}_{k_s}\mathrm{E}^F$.

If we wish to be very general, we simply specify DME. If we want to be more specific, we include the additional information according to the notation introduced above. Since all DMEs describe how the emergent properties of complex systems change in response to changes in lower level components, some generalities may emerge from their study.

2.2 Fitness and Selection Coefficients

Fitness is the highest level function of any biological system. As such it is difficult to define rigorously [7]. Fitness correlates have been used successfully in the study of life-history evolution [8]. We propose that it is possible to define meaningful fitness correlates that are computable from molecular systems biological models. For simplicity, we will assume that W, the absolute fitness in the wild, can be estimated by observing a high level organismic fitness correlate in wet-lab experiments and that this is proportional to the fitness correlate F that we compute *in silico*. Thus we can define:

$$F_M = F_{WT}(1 + s) = \frac{W_M}{W_{WT}},$$

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

where the subscripts M and WT denote the mutant and the wild type. Here the wild type is considered to be relatively 'mutation free' and s is the selection coefficient, commonly used in population genetics to denote the effects of a mutation on fitness. Using this approach we can compute DMEs for F and any emerging property of our model if we specify an underlying distribution of how reaction rates are affected by DNA changes. Expressing our results as s allows direct comparison with population genetics results.

2.3 Circadian Clocks

Circadian clocks are the internal molecular clocks that govern large parts of the molecular machinery of life. They frequently have a huge impact on the behaviour of organisms. They are responsible for waking us up in the morning and they make us feel tired in the evening. Such clocks are of paramount importance for the vast majority of organisms from Cyanobacteria [9] through fruitflies to humans [10]. Much recent work has focused on elucidating the various molecular components that perform the chemical reactions that oscillate with a daily rhythm. This has resulted in a series of models with increasing numbers of interlocking feedback loops [11]. In this paper we focus on one of the simplest models for a circadian clock that exists. This decision is motivated by a desire to focus the reader's attention on the basic principles of our new approach and on fundamental aspects of observing DMEs in clocks. We also wanted to apply our new approach to simple systems first to collect experience before analyzing complex ones.

3 Evolutionary Systems Biology

Evolutionary genetics and molecular biology have both been very successful in furthering our understanding of the natural world. However, after decades of research some familiar simplifying assumptions are now reaching their limits and evolutionary biologists are getting increasingly interested in the molecular details of their systems. At the same time molecular biologists progressively recognize the merit of quantitative modelling. Growing genomics and systems biology datasets provide a strong motivation for exploring realistic models at the interface (e.g. [12]). Increasingly detailed models of intracellular processes could help understand evolution by deriving DME^Fs *ab initio* by computing fitness correlates. Below we define one possible fitness correlate for circadian clocks. The following procedure can estimate a DME in a particular system:

1. Define a wildtype for use as a fixed 'mutation-free' reference point.
2. Treat each protein like a system in order to link DNA changes to changes in protein function by assuming a DME^{rate}, which denotes a realistic distribution of mutational effects on reaction *rate*s for DNA changes within protein-coding and regulatory sequences. If necessary, scale the frequencies of mutations for a given *rate* change by an estimate of the number of base pairs in the DNA that influence this *rate* to reflect the varying mutational target sizes in the system under investigation.
3. Compute enough samples [13] to obtain a $\mathrm{DM}_{rate}\mathrm{E}^F$, which denotes a distribution of the changes in the fitness correlate that emerges from the lower level distribution(s) of the underlying reaction rate changes.

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

4. Plot the differences to the mutation free reference as a $DM_{rate}E^F$, on one logscale for decreasing and on another logscale for increasing effects to visualise changes in the frequency of small effects from potential random noise expectations. Compare the number of fitness increases and decreases with the increases and decreases of the underlying reaction rate distributions in order to establish whether the molecular structure within biomolecules, or the network structure of biochemical reaction systems, has a larger influence on the DME of the fitness correlate.

The particular importance of small mutational effects in long-term evolution emphasizes the need for a careful analysis of numerical issues while computing fitness correlates. In plotting DME^Fs, logscales were found to be more helpful than linear scales due to their ability to visualize very small differences.

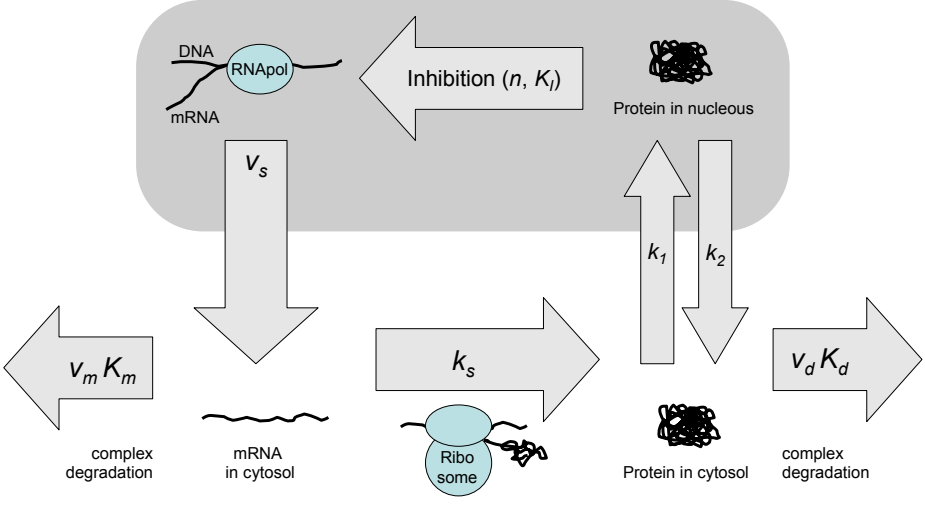

Fig. 1. Overview over the basic negative transcriptional feedback system that implements the simple clock analysed here

4 Model

The very simple model of a circadian clock that we use has been described elsewhere [14] and is closely related to the elementary transcriptional feedback oscillator described by Goodwin [15].

The basic reaction scheme is found in Figure 1. Briefly, the RNA polymerase complex transcribes a gene into mRNA, which is exported into the cytosol, where it accumulates at constant rate v_s. The ribosome translates the mRNA into a protein which accumulates at rate k_s. The protein can migrate between the nucleus and the cytosol, where the rates of transport are k_1 and k_2. Transportation into the nucleus is assumed here to be equivalent to turning the protein into a repressor, so potentially more than one reaction might be subsumed here. If enough copies of the protein have accumulated in the nucleus, they can cooperatively bind to the DNA and thus prohibit the

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

Table 1. Stochastic simulation implementation of the simple clock model. The kinetics of the chemical reactions shown here is governed by the propensity functions that determine which reaction out of this list will occur next. Once it has occurred, the species counts are adjusted according to the transition entry.

Number	Reaction	Propensity function	Transition
1	gene → gene + mRNA	$(v_s\Omega)\frac{(K_I\Omega)^n}{(K_I\Omega)^n + P_N^n}$	$M \to M + 1$
2	mRNA → ∅	$(v_m\Omega)\frac{M}{(K_m\Omega) + M}$	$M \to M - 1$
3	mRNA → mRNA + protein	$k_s M$	$P_C \to P_C + 1$
4	protein → ∅	$(v_d\Omega)\frac{P_C}{(K_d\Omega) + P_C}$	$P_C \to P_C - 1$
5	protein → repressor	$k_1 P_C$	$P_C \to P_C - 1$ $P_N \to P_N + 1$
6	repressor → protein	$k_2 P_N$	$P_N \to P_N - 1$ $P_C \to P_C + 1$

Table 2. The parameters of our basic clock model and their assumed values for the two 'wild types' explored here

Parameter	Meaning	*Neurospora*	24h-clock
Ω	Size of system	10^5	10^5
n	Degree of Hill-type cooperativity	4	4
K_I	Threshold for Hill-type repression	1	1
v_s	Effective rate of mRNA accumulation in cytosol	1.6	1.6
v_m	Maximal effective turnover of mRNA degradation	0.505	0.505
K_m	Michaelis-Menten constant for mRNA degradation	0.5	0.5
k_s	Effective rate of protein production in cytosol	0.5	0.5
v_d	Maximal effective turnover of protein degradation	1.4	1.4
K_d	Michaelis-Menten constant for mRNA degradation	0.13	0.13
k_1	Effective rate of repressor accumulation in nucleus	0.5	0.4623
k_2	Effective rate of repressor movement out of nucleus	0.6	1.2

binding of the RNA polymerase complex, effectively shutting down the production of mRNA. This cooperative binding is described by kinetics of the Hill type with a given degree of cooperativity, n, and a threshold constant for repression, K_I. To allow transcription to start again, mRNA and the protein are constantly degraded by reactions of the Michaelis-Menten type. Here v_m denotes the maximal effective turnover rate of the

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

mRNA degradation complex (with Michaelis-Menten constant K_m). The corresponding reaction for the protein is described by v_d and K_d.

This model can be described by the following ordinary differential equations (ODEs), where M, P_C and P_N denote the concentrations of mRNA, cyctosolic protein and nuclear repressor, respectively. The change in the concentration of mRNA is given by

$$\frac{\mathrm{d}M}{\mathrm{d}t} = v_s \frac{K_I^n}{K_I^n + P_N^n} - v_m \frac{M}{K_m + M}, \tag{1}$$

the change in concentration of the cytosolic protein is given by

$$\frac{\mathrm{d}P_c}{\mathrm{d}t} = k_s M - v_d \frac{P_C}{K_d + P_C} - k_1 P_C + k_2 P_N \tag{2}$$

and the change in the concentration of the repressor form of the protein in the nucleus is given by

$$\frac{\mathrm{d}P_N}{\mathrm{d}t} = k_1 P_C - k_2 P_N. \tag{3}$$

To translate these ODEs into chemical reaction equations, we followed the scheme described in [16]. To this end all molecular concentrations in the ODEs are turned into actual molecule counts by multiplying them by Ω, the parameter that describes the scale of the system. Table 1 gives the important quantities that were used to compute the propensity functions and the stoichiometry matrix in the stochastic simulations of the system.

Such a model will have a degree of approximation due to the presence of the reaction with Hill kinetics, since it has been shown that a direct application of Gillespie's algorithm to implement Hill's kinetic law can lead to an overestimate of the variance when compared to a more faithful low-level representation of the actual elementary reactions [17]. Mass action and Michaelis-Menten reactions do not suffer from this problem [18,19].

We used two sets of reaction rates as the starting points for our simulations: (i) the original set of parameters that Leloup *et al.* [14] used to describe a simple model of the 22h cycle circadian clock in *Neurospora crassa* and (ii) a modification of their parameter combination which we introduced to approximate a 24h cycle with the same set of reactions. Table 2 summarises the corresponding parameters.

5 Model Analysis

5.1 Simulations

We employed stochastic simulations to measure the emerging features of our model. To allow for flexibility in the analysis and speed of computation, we employed StochKit 1.0 (http://www.engineering.ucsb.edu/ cse/StochKit/), which implements a variety of algorithms that speed up Gillespie's Direct Method algorithm for stochastic simulation under particular sets of circumstances. For example, when large numbers of molecules are in the system, the library can choose to use a tau-leaping algorithm. It then no longer simulates every single reaction but rather estimates the number of reactions that will

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

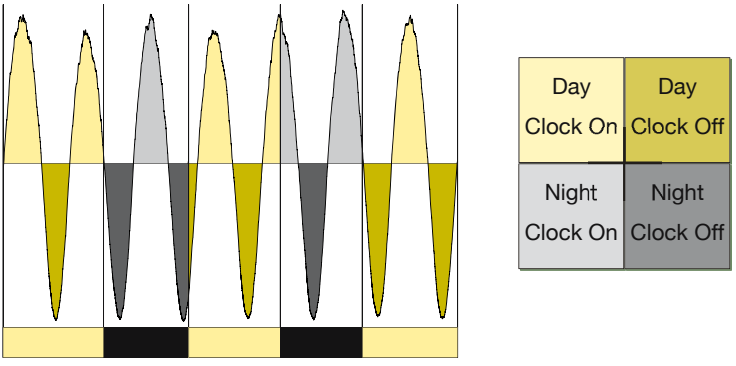

Fig. 2. The four states of a system with external and internal oscillations. It is possible to unambiguously assign one of the four states given in the table on the right to every point on a time course. Assignment to one of the four states is indicated by different shades in the time course on the right. We estimate the threshold from the observations as half the distance between the minima and maxima.

happen in a particular period of time. Our implementation of StochKit automatically switched between adaptive tau leaping [20] and fully detailed stochastic simulations. An overview of the corresponding methods has been presented elsewhere [20,21] and is also included in the StochKit manual.

5.2 Measuring Fitness Correlates

In order to explore the construction and behaviour of fitness correlates, we defined a simple biologically credible measure that we expect to be linked to fitness in many realistic situations. A schematic overview of the core principle is given in Figure 2.

Basically the existence of an external cycle (day or night) and an internal cycle (molecule count high or low) allows the definition of four states that describe all states that such a system can be in. Either it is 'in phase' (which can mean day or night) or it is 'anti-cyclic' (molecule count high, while external light is low and vice versa). If the internal system oscillates with a 24-hour period, then selection will favour mutations that help the cell to organise its patterns of gene activity around that cycle. However, if the internal oscillations are faster or slower, then the benefit of mutations that link particular genes with a particular state of the clock will be very limited, if existent at all: the genes that are in phase today will be out of phase in a few weeks and thus the long-term expectation of such a clock is probably not very different from random noise. From such considerations we can derive two measures of fitness, the cyclical fitness correlate F_C, defined as:

$$F_C = \frac{T_{1D} + T_{0N}}{T_{tot}}, \tag{4}$$

and the anti-cyclical fitness correlate F_A defined as;

$$F_A = \frac{T_{0D} + T_{1N}}{T_{tot}}, \tag{5}$$

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

where T_{1D}, T_{0N}, T_{0D} and T_{1N} sum over all time when the system is "On" during "Day", "Off" during "Night", "Off" during "Day" and "On" during "Night", respectively. All these quantities scale with the total time that the system has been under observation, T_{tot}. Based on such a definition, these fitness correlates can never be larger than 1. While any of the two measures can become zero under some special circumstances, we argue that more often the minimal value is 0.5, based on the random expectation of the complete absence of an internal cycle. For our 24h-clock we found F_A to be high and F_C to be low. Therefore we report only F_A below.

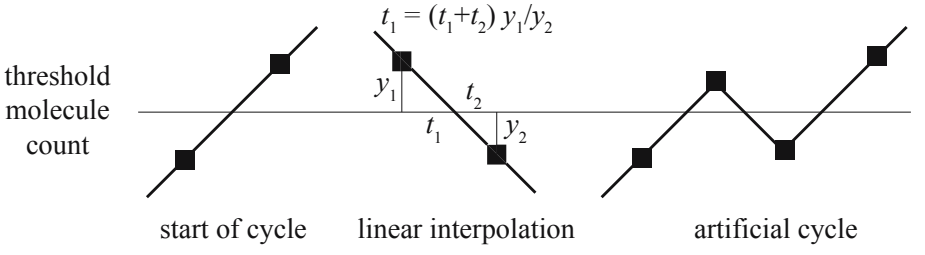

Fig. 3. Transitions of the clock. We use a threshold to distinguish the 'on' and 'off' states of the clock and keep track of the interpolated transition times to measure cycle length. In the presence of high levels of stochastic noise, artificial cycles can be generated. This was no problem above $\Omega > 10000$ in our system. Squares denote observations.

5.3 Measuring Cycle Length and Amplitude

To obtain a robust understanding of DMEs in a particular system it is preferable to investigate several of the emerging higher level properties of the system under investigation. In our case we also wanted to explore properties that could be determined more precisely than our present implementation of direct fitness correlates.

We decided to automatically observe the cycle length L and amplitude A, where the amplitude is the difference in molecule counts between the highest and the lowest point of a cycle. To define the beginning of a new cycle requires a threshold between the number of molecules at which the clock is considered to be 'off' and the same count in the 'on' state. We implemented this by using the same threshold required for our fitness measurements. Thus we stored the past state and determined for each current state, whether the threshold had been passed in upwards or downwards direction (Figure 3). If it had, the transition time was interpolated and cycle length was recorded. If it had not, it was checked whether the current value was a new extremum, facilitating the observation of cycle amplitudes. To get a good estimate of the true transition time we computed the intersection of the threshold with the line joining the two closest observations using the law of proportionality (see Figure 3, linear interpolation).

This system worked very well for large values of Ω. However, analyses at smaller values of Ω showed a sudden increase in the corresponding standard error estimates. Further scrutiny revealed that this was due to rare cases, where stochastic fluctuations had temporarily crossed the threshold, bucking the trend for just a moment and thereby triggering what the code considered a new, very short, cycle (Figure 3, right). This

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

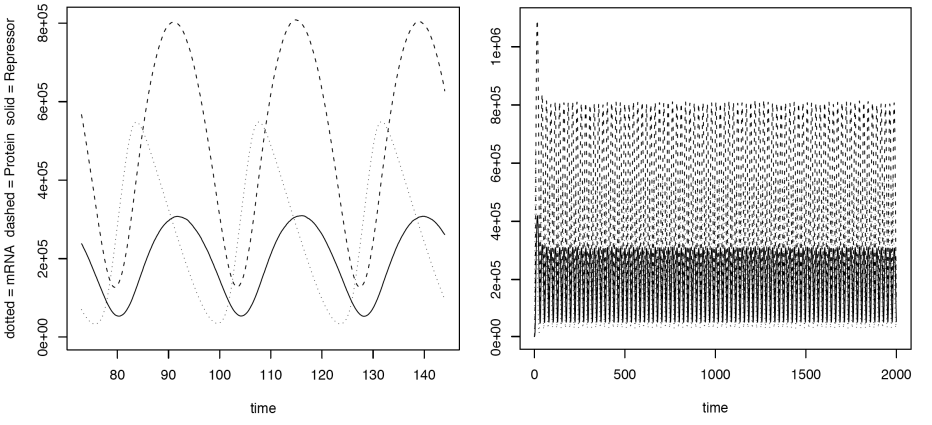

Fig. 4. Oscillations of our 24h-clock at $\Omega = 100000$. As in all our simulations the initial concentration for all the reactants was 1 molecule. To avoid any influence from initial concentrations, we allowed the clock to run for 50 hours before starting another 50 hour period, where merely maxima and minima were recorded to automatically estimate the threshold for fitness, cycle length and amplitude at half way between the two. Before actual observations started after calibration, two more cycles would be discarded, so that most observations would span the time from about 150 – 2000 hours.

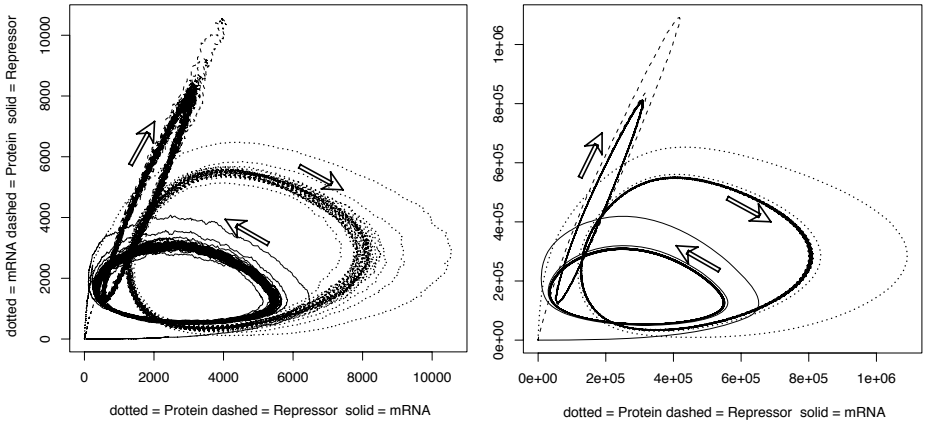

Fig. 5. Limit cycles for the 24h-clock at $\Omega = 1000$ (left) and 100 000 (right). Smaller Ω increase noise even more, but the general presence of oscillations is remarkably robust at this parameter combination, even if $\Omega = 10$.

phenomenon became prevalent at about $\Omega < 6000$ in the parameter combinations that we tested. To remedy this problem, two thresholds will have to be set up in such a way that a cycle is only recognised as such, if it has crossed both thresholds.

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

5.4 Basic Clock Behaviour

Our clock models do not behave differently from those analysed in the literature. Figure 4 demonstrates the extraordinary regularity and long-term stability of the oscillations at $\Omega = 100000$. If Ω is reduced, noise is increased, as can be seen in the limit cycles of Figure 5. To obtain solid estimates of the stochastic variability of the Neurospora and our 24h-clock, we observed 6380 and 6927 single simulations for 2000 hours (less the calibration period). The resulting distribution of anti-cyclic fitness, cycle length and amplitude can be found in Figures 6 and 7 (see the next section for an explanation of the DME plots in these figures).

5.5 Bootstraps and DME Estimates

Bootstraps. To obtain robust estimates of a DME^F is a statistical challenge. If an underlying $\mathrm{D}^D\mathrm{ME}^{k_1}$ or $\mathrm{D}^I\mathrm{ME}^{k_1}$ is assumed to map DNA sequence changes to repressor production rate changes, then one would like to know the effects on the distributions of emerging properties given by $\mathrm{D}_{DI}\mathrm{M}_{k_1}\mathrm{E}^{F_A}$, $\mathrm{D}_{DI}\mathrm{M}_{k_1}\mathrm{E}^{L}$ and $\mathrm{D}_{DI}\mathrm{M}_{k_1}\mathrm{E}^{A}$.

Here we propose to use a slight modification of the statistical bootstrap technique to achieve this. Bootstrapping in statistics was introduced to estimate the unknown distribution U of variates that are computed by a known function f from a known distribution D [13]. This is achieved by repeatedly sampling (with replacement) from the known distribution (or dataset). Then the function f is applied to each sample $\mathbf{x}$ to obtain samples from the unknown distribution:

$$U \sim f(\mathbf{x}), \text{ where } x \sim D \tag{6}$$

Thus U can be quantified rigorously if enough samples can be generated. Here we use as D the underlying $\mathrm{D}_D\mathrm{M}_{k_1}\mathrm{E}$ or $\mathrm{D}_I\mathrm{M}_{k_1}\mathrm{E}$ and as U any of the emerging properties (F_A, L, A) distributions specified above. f is specified by our simulation system that implements and observes the circadian clock model. To quantify U, we plot it in the DME plots shown in Figures 8-9. This approach allows us to detect changes in the distribution of emerging features that are caused by differences in the underlying low-level $\mathrm{D}_D\mathrm{M}_{k_1}\mathrm{E}$ or $\mathrm{D}_I\mathrm{M}_{k_1}\mathrm{E}$.

Design of the DME plot. DMEs are notorious for being difficult to visualize due to conflicting requirements. A Biologist would typically want to get an overview of deleterious, neutral and advantageous mutations at the same time, which is simple on a linear scale. However recent results have shown that the DME^F is highly leptokurtic [1,2], implying that most mutations have very small effects and would thus be lost in something that looks like a bar around zero on a linear scale. Thus a logscale seems the most appropriate way to visually convey information about most DMEs. We decided to follow a pragmatic approach that combines the best of both worlds by neglecting parameter ranges that are biologically uninteresting and implemented a corresponding plotting function in R (http://www.r-project.org/). The code first constructs a histogram of bins for decreasing effects that are equally spaced on a logscale. Then it does the same for increasing effects. The focus of the plot is on values within user-defined upper and lower limits of interest, merely checking for the existence of other values. Then

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

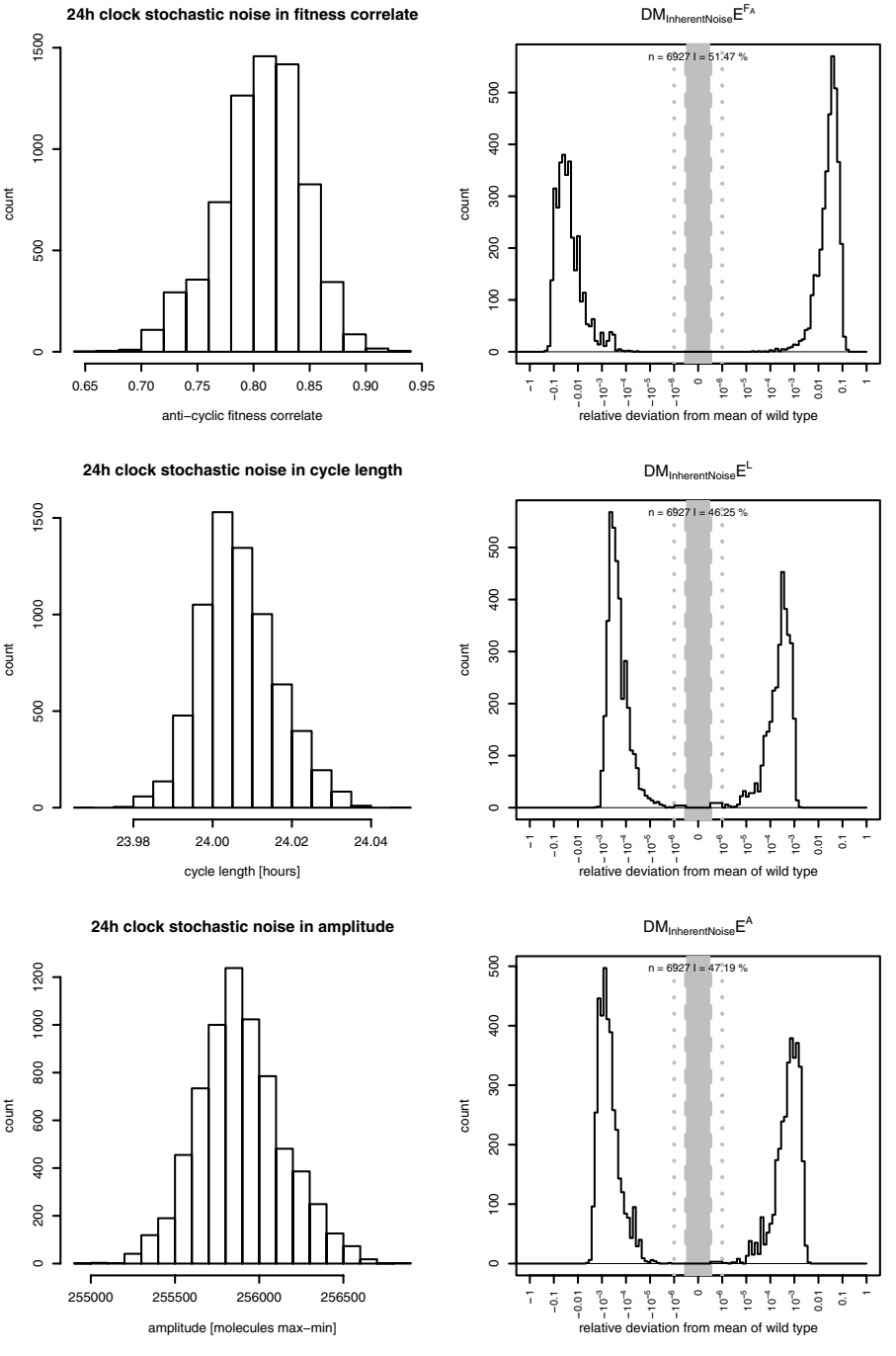

Fig. 6. The stochastic variability of the emerging features of the 24h-clock parameter combination. See Section 5.5 for an explanation of the right part of the figure.

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

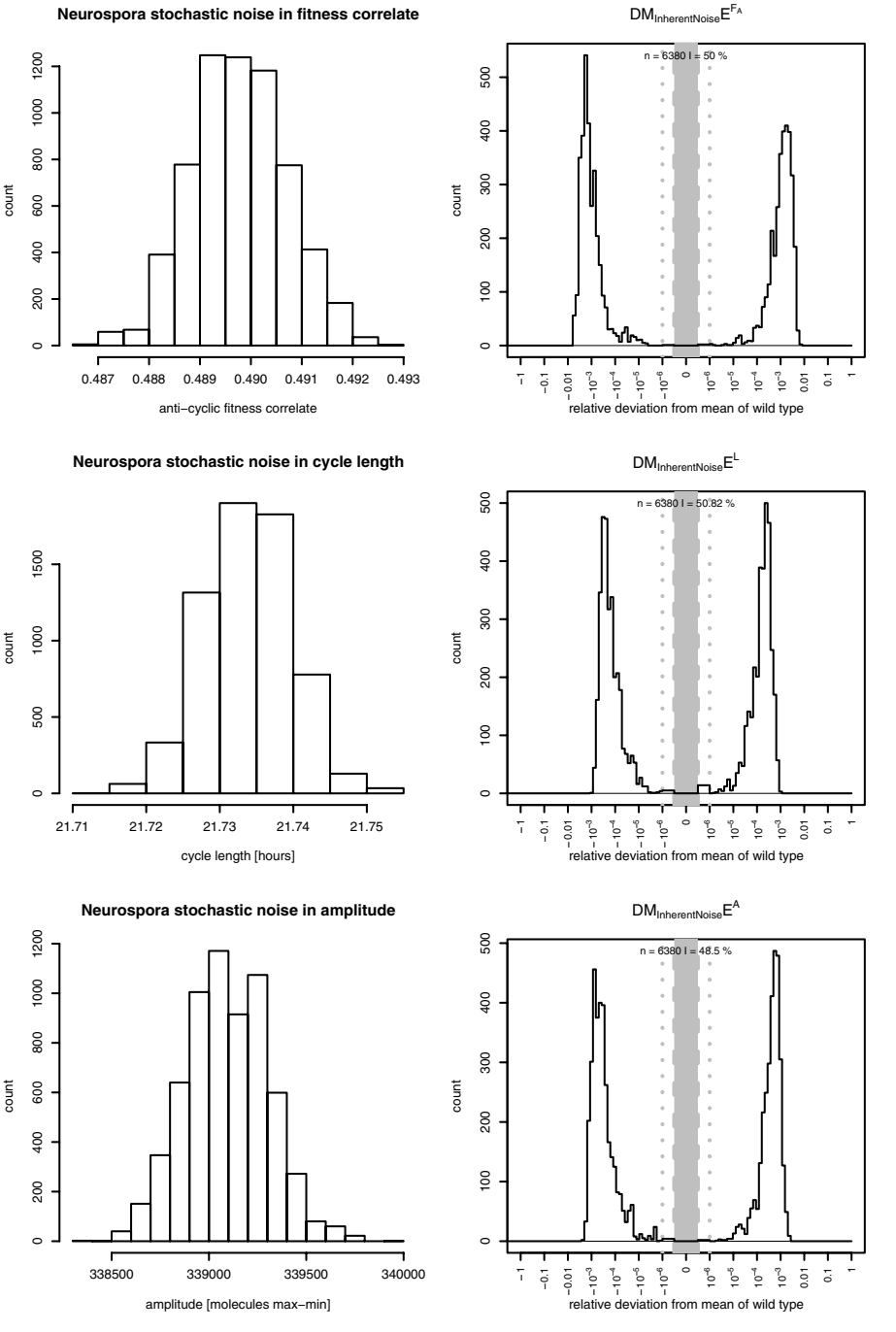

Fig. 7. The stochastic variability of the emerging features of the Neurospora clock parameter combination. See Section 5.5 for an explanation of the right part of the figure.

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

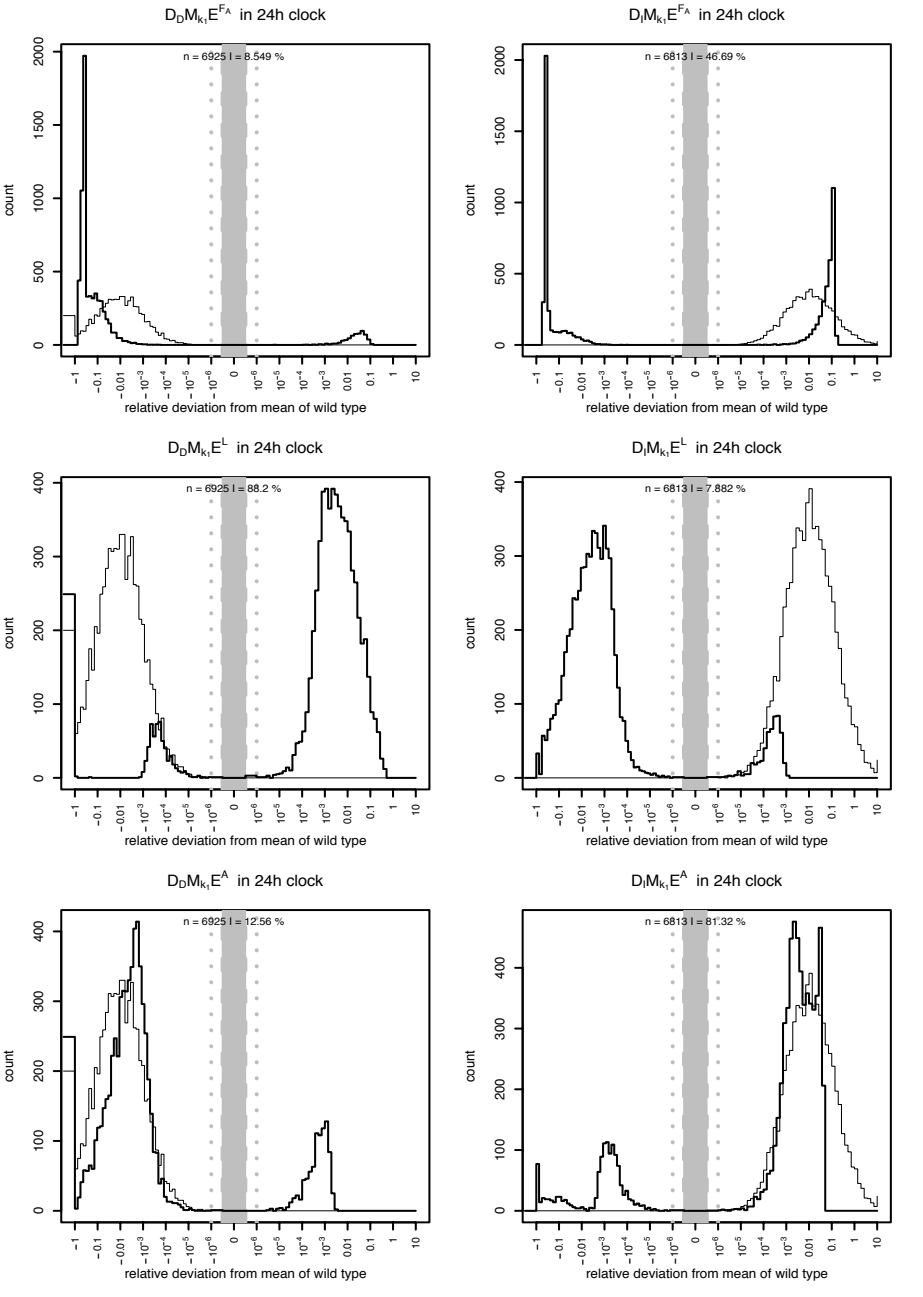

Fig. 8. These DMEs show the effects of assuming a low level lognormal $D_D M_{k_1} E$ and $D_I M_{k_1} E$ as distribution of generated genotypes on the emerging phenotypic features anti-cyclic fitness F_A, cycle length L and amplitude A for the 24h-clock parameter combination. The thick line gives the high level DME, the thin line the low level DME, n the sample size and I the fraction of increasing effects on a high level.

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

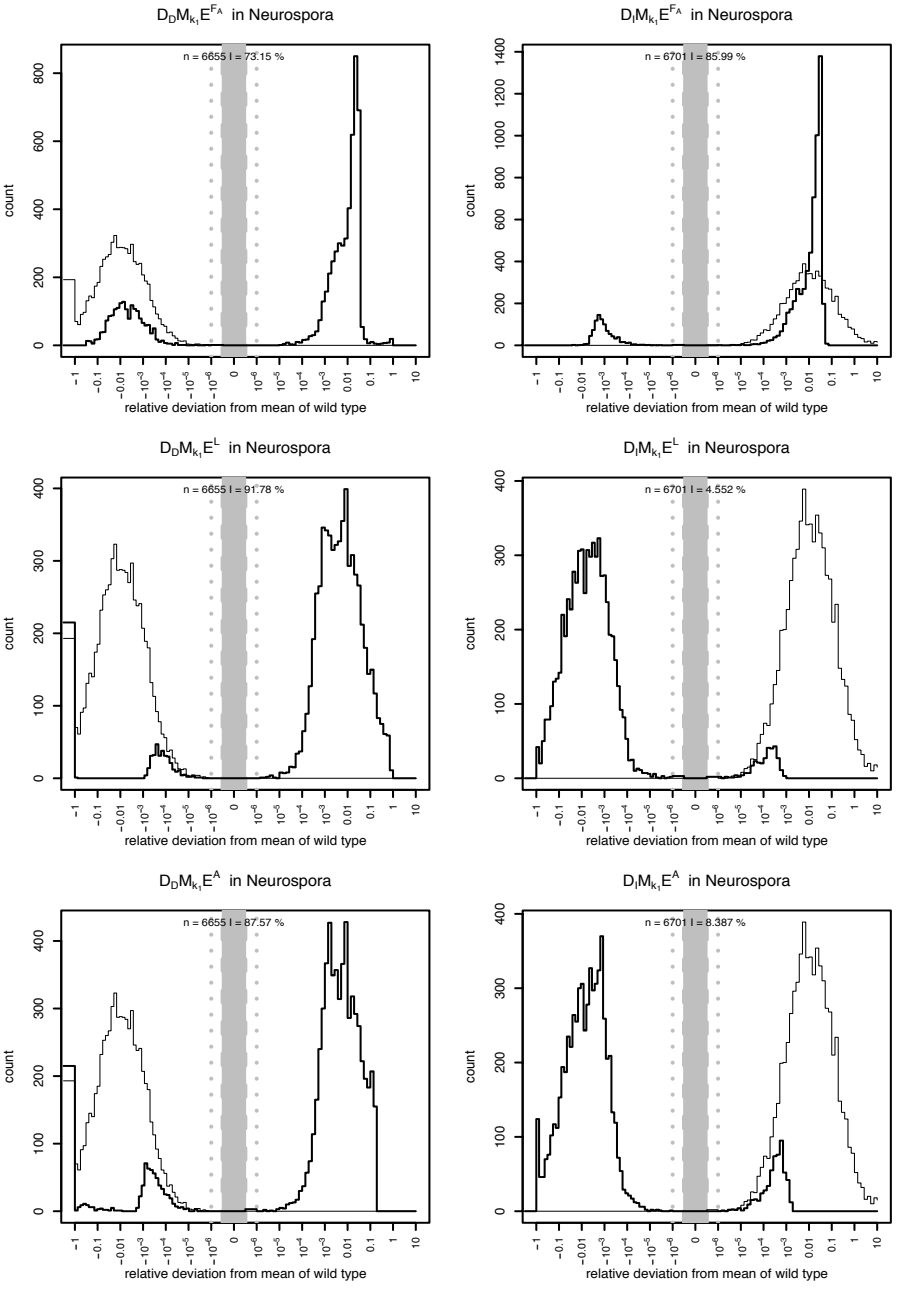

Fig. 9. These DMEs show the effects of assuming a low level lognormal $D_D M_{k_1} E$ and $D_I M_{k_1} E$ as distribution of generated genotypes on the emerging phenotypic features anti-cyclic fitness F_A, cycle length L and amplitude A for the Neurospora parameter combination. The thick line gives the high level DME, the thin line the low level DME, n the sample size and I the fraction of increasing effects on a high level.

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

a combined array of bin boundaries is built that is then used to construct a histogram with unequal bin width on a linear scale, but equal bin width in the ranges of interest on the positive and negative logscale. Finally, a special linear scaling is constructed that allows the final plot to be produced in a standard linear plotting environment. To read these plots, the following features have to be understood:

- The smallest borders of the smallest bins of interest are marked by the axis labels closest to zero. This is indicated by the grey dotted vertical lines.
- All values that are closer to 0 than the specified range are sorted into the 3 bins defined by the limits of user interest and $\pm 10^{-15}$. Thus it is easy to see what data might have been missed. The bin borders of$\pm 10^{-15}$ are plotted at values that allow for easy visual distinction from zero.
- The break in the scales is indicated by the massive greying around zero.

In the production of these plots it is of paramount importance to have a precise reference point, which in our case is taken to be the parameter combination that we used to start our explorations. We obtained these reference points by computing large numbers of single simulations for the 'wild type' *Neurospora* clock and the 'wild type' 24h-clock and combining their elementary statistics to obtain the aggregated estimates reported in Table 3.

Table 3. High precision estimates of the emerging features of the two 'wild type' clock parameter combinations that are used as a starting points for exploring DMEs here. N denotes the *Neurospora* clock, 24h, the 24h-clock with the respective parameters specified in Table 2.

	Mean	StDev	StErr	CV	n
N: F_A	0.4897883	0.0009251	1.45×10^{-7}	0.001888	6380
N: L	21.7338705	0.05796	1.07×10^{-7}	0.002667	542300
N: A	339093.583	1836	0.00339	0.005415	542300
24h: F_A	0.807005	0.03768	5.44×10^{-6}	0.04669	6927
24h: L	24.0066962	0.06960	1.30×10^{-7}	0.002899	533379
24h: A	255891.547	1620	0.00304	0.006331	533379

Since the reference points used for constructing DMEs are infinitesimally small and our model system exibits a significant amount of stochasticity, any repeated observation of an identical parameter combination will lead to what looks like many small increasing and decreasing changes. The amount of such stochastic noise determines how close the corresponding peaks will be to zero on the logscale. It is important to obtain a null-observation for the DME that determines its natural stochasticity, to avoid reporting spurious mutational effects that supposedly increase or decrease fitness. We report such an observation in Figures 6 and 7 for our two clock parameter combinations that we use as starting points for estimating real DMEs.

Equipped with such a framework, we can now present an example analysis of a simple system using DMEs. Figure 8 and 9 show the DMEs that result from computing many samples varying the rate k_1 with which the repressor accumulates in the nucleus. We ran four sets of simulations. In one set the rate was decreased by an amount that

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

was sampled from a lognormal distribution, while keeping the smallest resulting rates at zero. The lognormal distribution had location $\mu = 0.1$ and shape $\sigma = 2$ on the log scale. In the other set the rate was increased by an amount that was sampled from the same lognormal distribution. We chose a lognorm distribution, because it had performed well in previous population genetical tests [2]. This analysis was performed for the *Neurospora* and 24h-clock parameter sets to obtain an impression for how different DMEs are when sampled from different points in parameter space.

The results in Figure 8 and 9 show that it depends on the starting point and other parameters, whether a particular parameter change will be advantageous or deleterious. For example decreases in k_1 led to frequent increases in F_A and A for Neurospora, but mostly decreased these emerging properties for the 24h-clock. In other cases the high level effects appear to follow the low level effects rather closely. For example, A in Figure 8 follows the distribution of k_1 so closely that one would expect that at this point the intra-molecular structural effects of the corresponding enzymes have a larger influence on clock amplitude than the larger biochemical reaction network. However, this depends on other properties of the system, as the same is not true in Neurospora. It is not the purpose of this paper to discuss all corresponding DMEs in this simple clock model, but rather to demonstrate that the approach which has been presented is capable of producing the raw data that is needed for more comprehensive analyses.

6 Discussion

We have introduced a new framework for quantifying distributions of mutational effects using molecular systems biological models and presented a compact notation for navigating the complex multi-layered world of DMEs. We have demonstrated how this new approach works in principle and addressed fundamental challenges by estimating $\mathrm{DM}_{k_1}\mathrm{E}^{F_A}$, $\mathrm{DM}_{k_1}\mathrm{E}^{L}$ and $\mathrm{DM}_{k_1}\mathrm{E}^{A}$ in a simple model of a circadian clock. We were able to observe significant changes in the DME that exceeded the noise present in our system. The changes in this simple model show that it is important which parameter combination is used as a starting point for estimating a DME. While some parameter combinations lead to a majority of decreases, others mostly increase fitness. A more comprehensive analysis of this and other models is needed to determine how frequent fitness increasing effects are on a larger scale.

We deliberately did not include entrainment here to focus the reader's attention on our new framework. Our analysis showed that circadian clocks without entrainment will in most cases have a fitness in the wild that will approximate the absence of a circadian clock. Other, non-circadian clocks might still be important, but realistic models of robust circadian clocks in the wild need to include entrainment. This can be done by allowing for time-dependent changes in the reaction rates used in propensity functions.

Future work can improve the accuracy of our estimates by using more computing power. This will help building more realistic models, which is important, as the quality of our DME^{F} estimates depends on the quality of the molecular systems biological models used. Recent advances in molecular systems biology provide hope for the construction of quality models in an increasing number of model systems. Any such models are likely to be closer to biological reality than most of the extremely abstract and simple

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

models of mutational effects that have been used in population genetics so far. As any DME^F that has been observed by this new approach is specific to a very specific model, one can start comparing many different DME^Fs from many different systems. Such work will show how specific such DMEs are, and how often general features emerge that are robust to much of the underlying complexity.

7 Conclusions

We presented the first comprehensive application of our new framework for estimating distributions of mutational effects. Using the example of a simple circadian clock we demonstrated several fundamental features of this approach. Circadian clocks have been analysed before with systems biology methods [14,16,22,23], but the distribution of mutational effects has not been quantified in these systems before. Many more models need to be analysed in order to determine the general features that DMEs may exhibit.

Acknowledgements. We thank Ozgur Akman for extensive discussions of molecular clocks, Martha Loewe for help with LaTeX, John Welch and three anonymous reviewers for helpful comments on this manuscript and the BBSRC and EPSRC for funding. The Centre for Systems Biology at Edinburgh is a Centre for Integrative Systems Biology (CISB) funded by BBSRC and EPSRC, reference BB/D019621/1.

References

1. Eyre-Walker, A., Keightley, P.D.: The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8, 610–618 (2007)
2. Loewe, L., Charlesworth, B.: Inferring the distribution of mutational effects on fitness in Drosophila. Biology Letters 2, 426–430 (2006)
3. Keightley, P.D., Eyre-Walker, A.: Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177, 2251–2261 (2007)
4. Martin, G., Lenormand, T.: A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60, 893–907 (2006)
5. Kitano, H.: Towards a theory of biological robustness. Mol. Syst. Biol. 3, 137 (2007)
6. Kitano, H.: A robustness-based approach to systems-oriented drug design. Nat. Rev. Drug Disc. 6, 202–210 (2007)
7. Brommer, J.E.: The evolution of fitness in life-history theory. Biol. Rev. Camb. Philos. Soc. 75, 377–404 (2000)
8. Stearns, S.C.: The evolution of life histories. Oxford University Press, Oxford (1992)
9. Rust, M.J., Markson, J.S., Lane, W.S., Fisher, D.S., O'Shea, E.K.: Ordered phosphorylation governs oscillation of a three-protein circadian clock. Science 318, 809–812 (2007)
10. Panda, S., Hogenesch, J.B., Kay, S.A.: Circadian rhythms from flies to human. Nature 417, 329–335 (2002)
11. Brunner, M., Káldi, K.: Interlocked feedback loops of the circadian clock of Neurospora crassa. Mol. Microbiol. 68(2), 255–262 (2008)
12. Gjuvsland, A.B., Plahte, E., Omholt, S.W.: Threshold-dominated regulation hides genetic variation in gene expression networks. BMC Syst. Biol. 1, 57 (2007)
13. Efron, B., Tibshirani, R.D.: An introduction to the bootstrap. Chapman & Hall, New York (1993)

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock

14. Leloup, J.C., Gonze, D., Goldbeter, A.: Limit cycle models for circadian rhythms based on transcriptional regulation in Drosophila and Neurospora. J. Biol. Rhythms 14(6), 433–448 (1999)
15. Goodwin, B.C.: Oscillatory behavior in enzymatic control processes. Adv. Enzyme Regul. 3, 425–438 (1965)
16. Gonze, D., Halloy, J., Goldbeter, A.: Deterministic versus stochastic models for circadian rhythms. J. Biol. Phys. 28, 637–653 (2002)
17. Bundschuh, R., Hayot, F., Jayaprakash, C.: Fluctuations and Slow Variables in Genetic Networks. Biophys. J. 84, 1606–1615 (2003)
18. Arkin, A.P., Rao, C.V.: Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. J. Chem. Phys. 11, 4999–5010 (2003)
19. Cao, Y., Gillespie, D.T., Petzold, L.: Accelerated Stochastic Simulation of the Stiff Enzyme-Substrate Reaction. J. Chem. Phys. 123(14), 144917–144929 (2005)
20. Cao, Y., Gillespie, D.T., Petzold, L.: Adaptive explicit-implicit tau-leaping method with automatic tau selection. J. Chem. Phys. 126, 224101 (2007)
21. Gillespie, D.T.: Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. 58, 35–55 (2007)
22. Bradley, J.T., Thorne, T.: Stochastic Process Algebra models of a Circadian Clock. In: Nicol, D.M., Priami, C., Nielson, H.R., Uhrmacher, A.M. (eds.) Simulation and Verification of Dynamic Systems, Dagstuhl Seminar Proceedings, Dagstuhl, Germany (2006), http://drops.dagstuhl.de/opus/volltexte/2006/705
23. Stenico, M.: Modelling molecular systems with discrete concentration levels in the context of process algebra PEPA: Stochastic and deterministic interpretations. MSc.Thesis, University of Trento (2006)

Lecture Notes in Bioinformatics (2008) 5307:156-175. http://dx.doi.org/10.1007/978-3-540-88562-7_14
Proceedings of the 6th Conference on Computational Methods in Systems Biology 11-15 Oct 2008 Rostock