Hidden long evolutionary memory in a model biochemical network

Md. Zulfikar Ali, Ned S. Wingreen, and Ranjan Mukhopadhyay

Phys. Rev. E **97**, 040401 — Published 20 April 2018

DOI: 10.1103/PhysRevE.97.040401

Hidden long evolutionary memory in a model biochemical network

Md. Zulfikar Ali,[1] Ned S. Wingreen,[2,*] and Ranjan Mukhopadhyay[1,†]

[1]*Department of Physics, Clark University, Worcester, MA 01610*

[2]*Department of Molecular Biology, Princeton University, Princeton, NJ 08540*

We introduce a minimal model for the evolution of functional protein-interaction networks using a sequence-based mutational algorithm, and apply the model to study neutral drift in networks that yield oscillatory dynamics. Starting with a functional core module, random evolutionary drift increases network complexity even in the absence of specific selective pressures. Surprisingly, we uncover a hidden order in sequence space that gives rise to long-term evolutionary memory, implying strong constraints on network evolution due to the topology of accessible sequence space.

Within even the simplest living cells there is a highly complex web of interacting molecules, with biological function typically emerging from the actions of a large number of different factors [1, 2]. What is the relationship between the architecture of such interaction networks and the underlying processes of evolution? Much of the theory related to evolution focuses on the evolution of individual phenotypic traits or on population dynamics (see, for example, [3]); however, in general, individual genes do not determine individual traits. Rather, many traits arise from the dynamics of interacting components. With this in mind, we formulated and analyzed a minimal physically-based protein-protein interaction model that allows us to map from sequence space to interactions and, consequently, to network dynamics and fitness. Surprisingly, the model reveals a long-term memory of network origins hidden in the space of sequences.

Recently, bottom-up approaches to molecular evolution, typically in the context of the folding properties/thermodynamics of individual proteins or RNAs [4–8] have led to new insights into evolutionary outcomes, for example regarding a power-law distribution of protein family sizes. Here we generalize such bottom-up studies to functional networks. We focus on oscillatory networks of interacting enzymes, both due to the relevance of biological oscillators (e.g. cell cycle, circadian rhythms) [9–11] and due to the simplicity of defining function and fitness. As such a network evolves, are the original nodes still both necessary and sufficient or does the network redistribute function over new nodes? If new nodes do become essential, is there still memory of the original network?

In order to address these questions, we develop a model of protein-protein interaction networks consisting of two classes of enzymes, activators (e.g. kinases) and deactivators (e.g. phosphatases). Each of these can be in either an active state or an inactive state and only function when in the active state. To model cooperativity, we assume that activation or deactivation of a target (either an activator or a deactivator) requires h independent binding/modification events, with partially modified intermediates being short lived. The resulting chemical kinetic processes are

$$h\mathrm{A}_i^* + \mathrm{T}_l \xrightarrow{k_{il}} h\mathrm{A}_i^* + \mathrm{T}_l^*, \quad h\mathrm{D}_j^* + \mathrm{T}_l^* \xrightarrow{\tilde{k}_{jl}} h\mathrm{D}_j^* + \mathrm{T}_l, \quad (1)$$

where A/A*, D/D*, and T/T* denote activator, deactivator, and target in inactive/active states respectively. We note here that, in our model, the same protein species act both as enzymes (represented as A* or D* in the equations) as well as targets (represented as T/T*). The corresponding chemical kinetic equation can be approximated as (see Supplementary Material (SM) [12], section I for details)

$$\frac{d[\mathrm{T}_l^*]}{dt} = \sum_{i=1}^{m} k_{il}[\mathrm{A}_i^*]^h[\mathrm{T}_l] - \sum_{j=1}^{n} \tilde{k}_{jl}[\mathrm{D}_j^*]^h[\mathrm{T}_l^*] + \alpha[\mathrm{T}_l] - \alpha'[\mathrm{T}_l^*], \quad (2)$$

where m and n are the number of distinct types of activators and deactivators respectively. In Eq. 2, α and α' are background activation and deactivation rates. We further assume that the total concentration of each species is constant, such that $\mathrm{T}_l = c_0 - \mathrm{T}_l^*$.

Protein-protein interaction strengths are generally determined by amino-acid-residue interactions at specific molecular interfaces. Moreover, it has been estimated that $> 90\%$ of protein interaction interfaces are planar with the dominant contribution coming from hydrophobic interactions [13, 14]. For simplicity, we therefore assume each protein possesses a pair of interaction interfaces, an in-face and an out-face, and we associate a binary sequence, $\vec{\sigma}_{\mathrm{in/out}}$, of hydrophobic residues (1s) and hydrophilic residues (0s) to each interface (our approach builds on previous studies [15, 16]). The interaction strength between an enzyme (denoted by index i) and its target (denoted by index l) is determined by the interaction energy $E_{il} = \epsilon \vec{\sigma}_{\mathrm{out}}^{(i)} \cdot \vec{\sigma}_{\mathrm{in}}^{(l)}$ between the out-face of the enzyme and in-face of its target. (All energies are expressed in units of the thermal energy $k_{\mathrm{B}}T$.) The effective reaction rate is then given by

$$k_{il} = k_0(1 + \exp[-(E_{il} - E_0)])^{-h}, \quad (3)$$

where E_0 plays the role of a threshold energy, e.g. accounting for the loss of entropy due to binding. The background activation and deactivation rates are set equal

and define the unit of time via $\alpha = \alpha' = 1$. In our simulations we set $k_0 = 10^4$, $\epsilon = 0.2$, cooperativity $h = 2$, $E_0 = 5$, $c_0 = 1$, and we take the length of each sequence representing an interface to be $N = 25$. These interaction parameters were chosen to provide a large range for the rate constants k_{il} as a function of sequence and to keep the background rates small compared to the highest enzymatic rates; cooperativity was introduced to allow oscillations in relatively simple biomolecular networks.

For our evolutionary scheme, we assume a population sufficiently small that each new mutation is either fixed or entirely lost [17, 18]. We consider only point mutations – namely replacing a randomly chosen hydrophobic residue (1) in the in- or out-face of one enzyme by a hydrophilic residue (0), or vice versa. In this study, mutations are accepted if and only if they satisfy the selection criterion that the network remains oscillatory and moreover that the network exhibits oscillatory dynamics independent of the choice of initial concentrations of the active fractions (global oscillators). For this purpose we identified the fixed points of the chemical dynamics and carried out linear stability analysis (SM [12], section II).

In order to address the question of network drift – how function could redistribute over new nodes in an evolving network – we construct a 3-component oscillator (see Fig. 1A for a schematic) by starting with a 2-component oscillator, with one activator and one deactivator, and adding a second activator with all 0s for the sequences representing in- and out-interfaces (so that initially Activator 2, representing a new node, has minimal interactions with the other two components). We then let the system evolve, accepting only mutations corresponding to global oscillators. To characterize network drift, we studied the time evolution of the essentiality of each activator for a random sample of starting sequences that corresponded to oscillators, as depicted in Fig. 2A, where we characterize a component as being "essential" if the system stops oscillating when the component is removed, or equivalently, in our model, if we set the total concentration c_0 of that component to zero [19]. Initially Activator 2 is inessential (since Deactivator and Activator 1 generate oscillations), and in Fig. 2B we exhibit the distribution of the number of accepted mutational steps before it become essential for two distinct starting sequences. While the two distributions peak at very different values for the number of mutational steps, the interaction strengths for the two initial states do not differ appreciably (Fig. 2B, inset), highlighting the importance of the underlying sequence in governing evolutionary dynamics. Returning to Fig. 2A, we find relatively rapid flips between states where both activators are essential to states where only one of the activators is essential.

Surprisingly, we also note the prevalence of much longer time periods where Activator 1 is always essential or where Activator 2 is always essential. This is true independent of initial conditions. These long evolutionary periods presumably reflect the division of sequence space into two regions or "phases": Phase 1 where Activator 1 is always essential and Phase 2 where Activator 2 is always essential. The system starts in Phase 1 (Activator 2 is inessential), then when Activator 1 first become inessential we infer that the system has entered Phase 2,

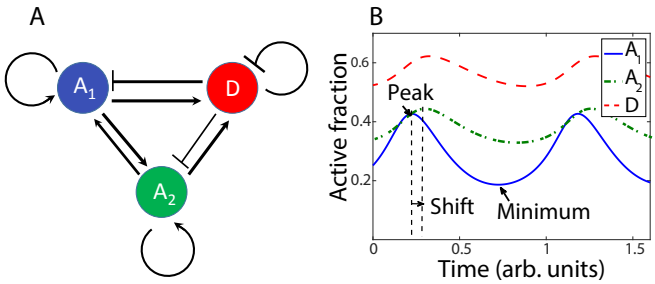

FIG. 1: Oscillatory protein-protein interaction network. (A) Schematic of a 3-component network with two activators (A_1, A_2) and one deactivator (D). The symbols,$\rightarrow$ and $\dashv$, indicate the chemical process of activation and deactivation, respectively. (B) Steady-state oscillations of the active fractions of the components of the network in (A). The dashed vertical lines indicate peaks of the activator oscillations, and the horizontal arrow indicates the time shift between these peaks.

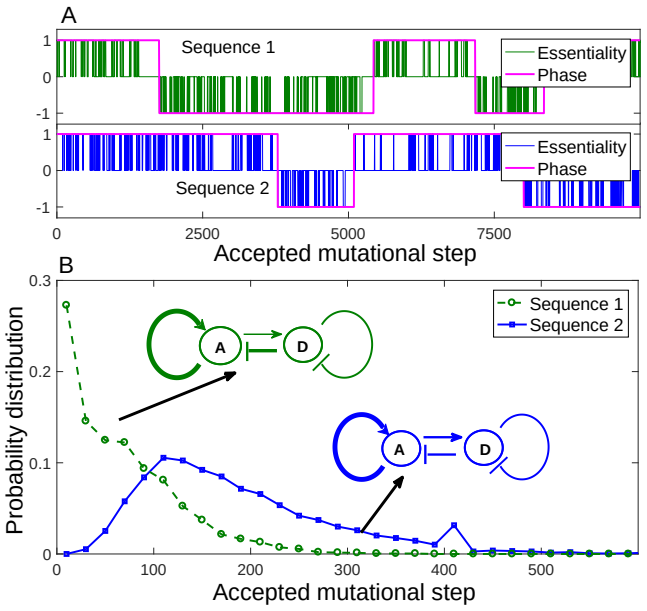

FIG. 2: Temporal evolution of essentiality of activators in 3-component systems. (A) Temporal evolution for two different initial sequences (the two sequences are specified in Supplementary Material (SM) [12]). On the y-axis, +1 indicates only Activator 1 is essential, -1 indicates only Activator 2 is essential, and 0 indicates both activators are essential [19]. (B) Histograms of the number of accepted mutational steps before Activator 2 first becomes essential, for the two distinct initial sequences. Inset: interaction strengths of the two initial states.

and so on. These results imply that while, naively, one might have expected that the starting state of the network (e.g., the identity of the solely essential activator) would be effectively forgotten as soon as both activators became essential, the system retains a hidden memory of the starting conditions in terms of persistence in the starting phase (Phase 1, in this case). Thus the long duration in each phase (in comparison to the duration between successive flips in essentiality) constitutes a long-term memory in our evolving network.

Can these two phases be distinguished in terms of measurable dynamical quantities or rate constants? Since the two phases presumably relate to an asymmetry in the roles of the two activators, we quantify this asymmetry via the relative peak-to-valley ratio (PVR) of the oscillations of their active fractions, where relative PVR is ((PVR A_1 - PVR A_2)/(PVR A_1 + PVR A_2)). The peak-to-valley ratio (PVR) of a component is obtained by determining the peak value and the valley (minimum) of the active concentration for steady-state oscillations (see Fig. 1B) and taking the ratio of the two. From Fig. 3A (top panel) and Fig. 3B, we see that relative PVR correlates with the phase, and we display the distribution quantifying this correlation. A corollary is that the probability that an activator is essential also correlates with the relative PVR (Fig. 3C), so that if an activator has a relatively larger PVR it is also more likely to be essential. Moreover, we find that the phase-shift between peaks in the active fractions of the two activators also correlates with the phase (Fig. 3D), so that Activator 1 typically leads in Phase 1 and Activator 2 in Phase 2. Finally in order to determine how these observations relate to the underlying rate constants, we constructed the covariance matrix for the covariation of the nine rate constants k_{ij} and carried out a principal component analysis (SM [12], section IV). We find that the projected component of the rates onto the eigenvector with the largest eigenvalue (PC1 = 94.93%) strongly correlates with the phase (Fig. 3A, lowest panel, and Fig. 3E); we find no such correlation for projections onto any of the remaining eigenvectors. On examining the top eigenvector, we find that it primarily consists of a linear superposition of the difference in auto-activation rates of the two activators and the difference in their deactivation rates. This suggests that strong auto-activation coupled with strong deactivation produces an activator that peaks first during each oscillation cycle and also has a large PVR (see SM [12], section VIII, for a physical explanation of the correlation). However, the co-occurrence of these features does not by itself explain the observed long intervals of the two distinct phases.

The question remains: what is the origin of the observed long-term memory? We first quantify the duration of long-term network memory by constructing a histogram of the number of mutational steps that the system spends in each phase before flipping. As shown

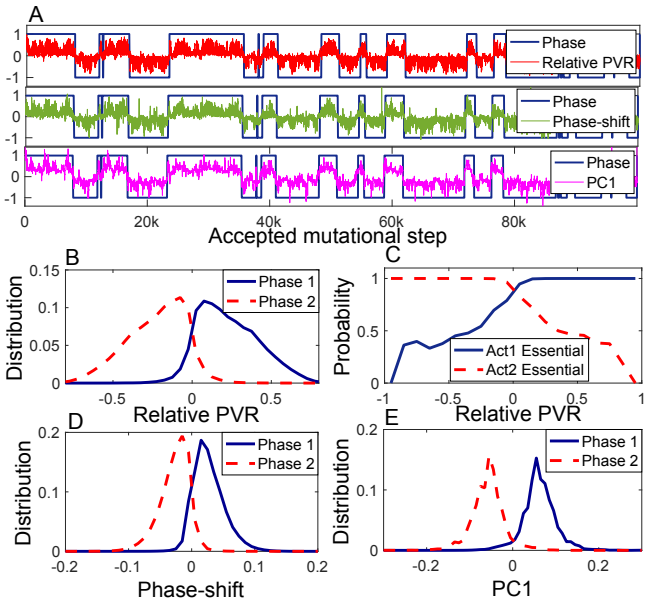

FIG. 3: Temporal evolution of phases in 3-component system. (A) Depiction of the temporal evolution where a value of +1 indicates Phase 1 and -1 indicates Phase 2. Along with the phase, the three panels show (i) normalized relative PVR of the two activators (red, top panel), (ii) phase-shift between their oscillatory peaks (green, middle panel), and (iii) projected component of the chemical rates on the principal eigenvector from PCA analysis (magenta, bottom panel). (B) Distributions of relative PVR of the two activators in Phase 1 and in Phase 2. (C) Probability that each activator is essential as a function of its relative PVR. (D) Distribution of phase-shifts between active fraction peaks of the two activators in Phase 1 and Phase 2. (E) Distribution of projected rate constants on the principal eigenvector, obtained from PCA analysis, in Phase 1 and Phase 2.

in Fig. 4A, we find an approximately exponential distribution, $P(\tau) \propto e^{-\tau/\tau_0}$, where $\tau_0 \simeq 3200 \pm 48$ mutational steps. An exponential distribution implies a fixed, history-independent rate of flipping between the two phases, which in turn suggests that flipping corresponds to barrier crossing. Since our model treats all oscillatory states as equally fit, the only barriers are entropic, i.e., there must be relatively speaking very few boundary points connecting phases (SM [12], section V). To check this hypothesis, we studied the neighborhood of states in Phase 1 and Phase 2. In Phase 1, for example, we distinguished between states where only Activator 1 is essential and states where both are essential. For states where only Activator 1 is essential we found no examples of sequences that were one Hamming distance away (that is, separated by a single point mutation) for which Activator 1 stops being essential. Of the states in Phase 1 where both activators are essential, for only 3% of states the Hamming distance 1 neighborhood contained one or more states where Activator 1 was inessential. The relative rarity of such states (which can be considered as

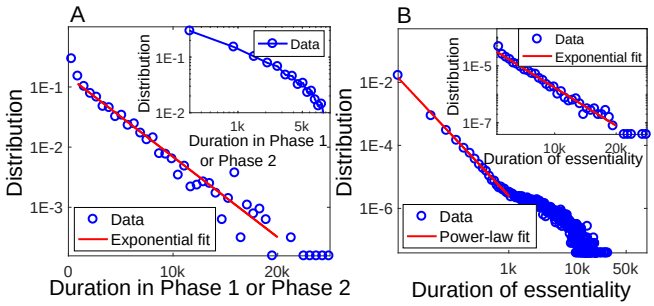

FIG. 4: Distribution of accepted mutational steps between flips. (A) Distribution of the number of accepted mutational steps between flips from one phase to the other, on a semi-log scale to highlight the exponential distribution (data is binned with bin size 600). Inset: same data on log-log scale. (B) Distribution of the number of accepted mutational steps where an activator is essential for the whole duration, on a log-log scale showing a power-law fit $f(x) \sim x^{-2.3\pm0.05}$ for short times (bin size 50). Inset: same distribution over longer times on semi-log scale (bin size of 600).

boundary states) is consistent with our hypothesis that in sequence space the two phases touch at a relatively small number of boundary points.

Interestingly, in contrast to flipping between phases, the distribution of the number of mutational steps that an activator remains essential exhibits a power-law distribution for short times, as depicted in Fig. 4B. For Activator 1, for example, this power-law part of the distribution is dominated by cases where the system is in Phase 2, with Activator 1 switching between being essential and inessential. Thus the power-law distribution is related to the presence of domains within Phase 2 where Activator 1 is also essential (and likewise for Activator 2 in Phase 1). For longer times, the periods of essentiality correspond to the duration of phases, and thus the distribution decays exponentially (Fig. 4B, inset). In contrast to exponential decay, a power-law distribution implies a history-dependent switching rate, with the escape rate from a domain proportional (on average) to the inverse of the time elapsed since the system entered the domain (SM [12], section IX; see also section VI for a toy model exhibiting mixed power-law and exponential distributions).

It is not *a priori* obvious how the above observations of two phases generalize to more complex networks. We therefore extended our study by starting with a 3-component oscillator and adding a fourth component (Activator 3) with all its sequences initially set to 0s. Once again we find that Activator 3 becomes essential relatively rapidly (typically in ∼100 mutational steps). If we continue to follow the evolution of essentiality for the activators, we find for each activator long periods (∼1000+ mutational steps) where that activator remains essential, separated by similarly long periods where that activator is intermittently essential/inessential (Fig. 5A). This suggests that for each activator, the sequence space of oscillators divides into two regions: one region where that activator is essential at every point and a second region consisting of smaller domains where the activator is essential interspersed with domains where it is inessential. Note that time periods where one activator remains essential sometimes overlap with periods where one of the other activators remains essential, implying that the region where one activator is essential at every point has some overlap with the regions where other activators are essential at every point. This contrasts somewhat with the 3-component system where Phase 1, the region in which Activator 1 is essential at every point, is complementary to Phase 2. By contrast, as shown in Fig. 5B, the distribution of mutational steps over which any one of the activators is essential for the 4-component system is quite similar to that of the 3-component system, being power-law at short times with a similar exponent, and exponential for longer times, albeit with a shorter decay time $\tau_0 \simeq 1750 \pm 54$ mutational steps. As for 3-component systems, we also find strong correlation between normalized/relative PVR of oscillation, phase-shift, and essentiality for pairs of activators. We find that when the normalized PVR of an activator is higher, the probability that it is essential is also higher (Figs. 5C and 5D); these results generalize to much larger systems of activators and deactivators (SM [12], section X).

In this paper, we focused on oscillatory networks and introduced a sequence-based evolutionary scheme, in contrast to schemes where mutations are directly implemented by changes in rate constants (see, for example, [20]). We studied how function can become distributed over new nodes due to random network drift. For a 3-node network, the typical timescale for the new node to become essential for oscillation is ∼100 point accepted mutations, which, given the total of 150 sites, corresponds to around 66% accepted mutations [21]. Surprisingly, our model also revealed a much longer term memory (around 2000 point accepted mutations per 150 amino acids for a 3-node system) with exponential decay, indicative of a barrier crossing process in the space of sequences.

We expect our model to be broadly useful for exploring principles of protein network evolution. While simple and easy to implement, the model is biologically grounded in sequence-based evolution, and also physically grounded insofar as all proteins interact via binding with other proteins. In this approach, any component is allowed to interact with all other components and no specialized topology is introduced by hand. Moreover, there is no fine tuning and the degree of cooperativity utilized for the studies in this paper is modest and easily achievable in practice by biochemical networks [22]. The model provides a natural framework to study the interplay between selection pressure and sequence-based

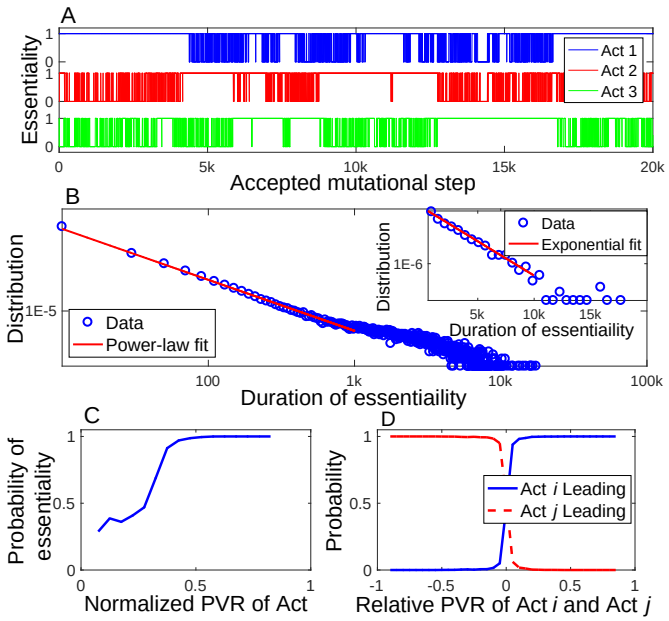

FIG. 5: Temporal evolution of essentiality of each activator in 4-component systems. (A) Depiction of temporal evolution where, on the y-axis, +1 indicates that the activator is essential and 0 indicates that it is not essential. (B) Distribution of the number of accepted mutational steps where the activator is essential for the whole period, on a log-log scale showing the power-law distribution $f(x) \sim x^{-2.15\pm0.02}$ for short times (bin size 20). Inset: same distribution on a semi-log scale. (C) Probability of Activator i being essential as a function of its normalized PVR defined as PVR A_i/(PVR A_1 + PVR A_2 + PVR A_3). (D) For any pair of activators, the probability that Activator i leads Activator j as a function of their relative PVR.

designability/accessibility. It can moreover be readily extended to larger networks, networks with other functions, and also to other mutation-selection regimes (for example, the concurrent mutations regime expected for larger populations [23]).

We also believe our results for network drift will apply beyond the context of oscillators studied here. It has been suggested that protein networks evolve primarily by two biological mechanisms: (i) gene duplication, and (ii) random mutations in proteins leading to neo-functionalization, that is, the *de novo* creation of new relationships with other proteins [24]. Our studies illustrate the significance of neo-functionalization in the context of functional networks where protein-protein interactions are physically grounded, i.e., described via quantitative interaction strengths rather than Boolean variables. Our discovery of hidden order in sequence space leading to evolutionary long-term memory could also be quite general, highlighting the strong constraints to network evolution that emerge from the topology of accessible sequence space. It will be interesting to see if the presence of "phases" generalizes to other network types. Future studies may profitably include the evolutionary dynamics of nodes, address other network functions (e.g. signal integration), and explore the role of graded selection in the *de novo* evolution of new functions.

We acknowledge helpful discussions with Yigal Meir and Ammar Tareen. The research was supported in part by DARPA Biochronicity program, Grant D12AP00025, National Science Foundation Grant PHY-1305525, and National Institutes of Health Grant R01 GM082938.

* Electronic address: wingreen@princeton.edu
† Electronic address: ranjan@clarku.edu

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, Molecular Biology of the Cell. Taylor and Francis; 2002.
[2] U. Alon, An Introduction to Systems Biology: Design Principles of Biological Systems. *Chapmann and Hall*; 2009.
[3] S. J. Maynard, The Theory of Evolution. *Cambridge:Cambridge Univ. Press*; 1993.
[4] M. Eigen, Self-organization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften* **58**, 465-523 (1971).
[5] M. Eigen and P. Schuster, J. Mol. Evol. **19**, 47-61(1982).
[6] J.D. Bloom, A. Raval, O. Wilke, *Genetics* **175**, 255-266 (2007).
[7] A. W. R. Serohijos and E. I. Shakhnovich, Curr. Op. in Str. Bio. **26**, 84-91 (2014).
[8] K.B. Zeldovich and E.I. Shakhnovich, Annu Rev Phys Chem. **59**,105-27 (2008).
[9] Goldbeter, Biochemical Oscillations and Cellular Rhythms: The molecular bases of periodic and chaotic behaviour. *Cambridge University Press*, Cambridge; 1996.
[10] J. L. Ditty, S. R. Mackey and C. H. Johnson, Bacterial circadian programs. *Springer, New York*; 2009.
[11] M. Nakajima, K. Imai, H. Ito, T. Nishiwaki, Y. Murayama, H. Iwasaki, T. Oyama and T. Kondo, Science **308**, 414-415 (2005).
[12] See Supplementary Material at [URL will be inserted by publisher] for detailed chemical kinetics, further analysis, toy model, and robustness of the model.
[13] M. Heo, S. Maslov and E. I. Shakhnovich, Proc. Nat. Acad. of Sci. USA **108**, 4258-4263 (2011).
[14] Z. Keskin, A. Gursoy, B. Ma and R. Nussinov, Chem. Rev. **108**, 1225-1244 (2008).
[15] M.E. Johnson and G. Hummer, J. Phys. Chem. B. 117, 13098-13106 (2013).
[16] I.M. Nooren and J.M. Thornton, EMBO J. 22, 3486-92 (2003).
[17] P. A. P. Moran, Math. Proc. of the Cambridge Philosophical Society **54**, 60-71 (1958).
[18] M. A. Nowak, Evolutionary Dynamics: Exploring the Equations of Life. *Belknap Press* 2006.
[19] Since we find that states where both activators are individually inessential are very rare, approximately 0.001% of the total number of oscillatory states, we ignore such states for the purposes of the figure.

[20] P. Francois, N. Despierre and E. D. Siggia, PLOS Comp. Bio., DOI: 10.1371/journal.pcbi.1002585 (2012).
[21] J. Pevsner, Bioinformatics and Functional Genomics (2nd ed.). *Wiley-Blackwell*; (2009).
[22] J. E. Ferrell, Trends in Biochem. Sci. **21**, 460-466 (1996).
[23] M. M. Desai and D. S. Fisher, Genetics **176**, 1759-98 (2007).
[24] G. J. Peterson, S. Presse, K. S. Peterson and K. A. Dill, PLOS One **7**, e39052 (2012).