SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
BIOLOGICAL PHYSICS
EVOLUTIONARY THEORY
STATISTICAL PHYSICS

Received
4 September 2013

Accepted
8 November 2013

Published
25 November 2013

Correspondence and requests for materials should be addressed to H.J.W. (hgjun.woo@gmail.com)

Stationary distribution of self-organized states and biological information generation

Hyung Jun Woo

Henry M. Jackson Foundation for the Advancement of Military Medicine, 2405 Whittier Drive, Frederick, Maryland, 21702, USA.

Self-organization, where spontaneous orderings occur under driven conditions, is one of the hallmarks of biological systems. We consider a statistical mechanical treatment of the biased distribution of such organized states, which become favored as a result of their catalytic activity under chemical driving forces. A generalization of the equilibrium canonical distribution describes the stationary state, which can be used to model shifts in conformational ensembles sampled by an enzyme in working conditions. The basic idea is applied to the process of biological information generation from random sequences of heteropolymers, where unfavorable Shannon entropy is overcome by the catalytic activities of selected genes. The ordering process is demonstrated with the genetic distance to a genotype with high catalytic activity as an order parameter. The resulting free energy can have multiple minima, corresponding to disordered and organized phases with first-order transitions between them.

Despite enormous progress in understanding the characteristics of life's building blocks and their interactions[1], many aspects of processes occurring in living organisms continue to pose challenges to physics-based explanations. A major difficulty is in characterizing their organization and function, which tend to appear spontaneously under suitable conditions, in stark contrast to common experience with nonliving matter ruled by increasing disorder. The term *self-organization* has been used widely to describe these spontaneous appearances of highly ordered structures, not only within biological systems but also in higher-level organizations including networks[2–6], and complex dynamical systems exhibiting phenomena such as disease progression[7–9] and neural computation[10]. Different directions of theoretical approaches include wide-ranging studies of driven systems showing self-organized criticality[11,12] with implications to extinction dynamics[13], dynamical systems views with analogies to equilibrium phase transition[14], and concepts centered on autocatalysis, evolution, and selection[15]. Further studies in this interdisciplinary field include models of genetic regulatory networks and cell differentiation[16], dynamic clustering in active media[17], and the study of Boolean network dynamics[18]. However, one characteristic common to current approaches of studying self-organization is the lack of concrete connections to equilibrium statistical mechanics.

We focus in this paper on chemically driven systems and describe an approach extending the equilibrium statistical mechanical concepts to cover the stationary distribution of self-organized states. Our approach, which is based on a combination of equilibrium theory and enzyme kinetics, will allow us to distinguish self-organization from self-assembly, a related but distinct class of phenomena in which ordered structures are favored in equilibrium because of certain structural features present within the constituents. A familiar example is the formation of micelles and lipid bilayers stabilized by hydrophobic interactions[19], for which well-established and quantitative theories now exist[20].

Self-organization, in contrast, is a sustainable nonequilibrium process in which a system spontaneously increases and maintains its degree of ordering as a result of interactions and exchanges of matter with its surrounding. Typically the system is in thermal and barometric equilibrium with its surrounding, but is driven by supply and extraction of chemical species; the 'food' and 'waste' molecules. The point of view adopted in this paper is that the ordering occurs when the system has the potential to catalyze reactions involving these externally controlled species in the favored direction. We use a simple description of this effect to derive a biased distribution of system configurations away from equilibrium. The driving force increasingly favors organized states that would have negligible probabilities of occupation in equilibrium.

One immediate consequence of such effects amenable to current experimental investigations is the shift in conformational distributions of protein enzymes while under stationary working conditions compared to

equilibrium. Recent developments in single-molecule experimental techniques[21,22] have uncovered many surprises challenging the traditional views on how enzymes operate, including the classical 'induced fit' mechanism, which assumed that an enzyme mostly inhabits a single conformation, only changing it as a result of substrate binding. Increasing body of experimental evidence suggests that many enzymes instead sample a wide range of conformational states without ligands, while a substrate binding event shifts the equilibrium into states optimal for catalytic activities ('conformational selection' mechanism)[23–25]. Theoretical studies so far have focused mostly on the effects this conformational heterogeneity has on kinetic velocity[25,26]. The theoretical consideration presented in this paper provides a simple statistical mechanical description of the nonequilibrium enzyme conformational distribution that can be used to interpret single-molecule experiments.

As a second, more fundamental application, we address the issue of how modern proteins capable of exhibiting such ordering that defies entropic costs – e.g., via conformational changes leading to the activation of its enzymatic activity – or rather the biological information encoding their structure and function, could have arisen spontaneously. Viewed from the information theoretic perspective[27], the generation of such biological information is a self-organization process where random heteropolymers with maximum entropy were replaced by highly conserved genes. Considerations of this process can serve as a bridge between equilibrium statistical mechanics and current well-developed statistical approaches to modeling evolution[28], as well as highly successful recent developments in data-driven characterizations of biological genotype-phenotype spaces[29,30] and reconstruction of the evolutionary history of extant metabolic pathways[31].

There are two broad classes of approaches put forward to describe such an ordering of biochemical sequences. One highly influential perspective is to center on the properties of the minimal networks of autocatalytic species[15,32–36], where self-organization becomes possible when the degree of diversity of the autocatalytic set exceeds a critical value. A different approach, represented by the quasispecies theory[37,38] and theoretical works based on it[39–45], takes the self-replication population dynamics of sequences as a starting point. The order-disorder transition of genetic information occurs as the mutation rate of the replication process crosses a threshold value.

Based on the general consideration of the stationary distribution under driven conditions, we consider in this paper the question of whether the ordering in sequence space can occur purely from chemical driving forces without the mechanism of competition based on self-replication postulated in quasispecies theory. It is shown that there exists a first-order transition from the phase characterized by random sequences to those dominated by the few active sequences irrespective of the specific population dynamics of heteropolymers. Our thermodynamic consideration thus connects the biochemical self-organization more directly to chemical driving forces and reveals its close correspondence to equilibrium phase transitions, complementing the autocatalytic kinetics-based and population-based approaches.

Results

Self-organization. We consider a system in thermal equilibrium with a reservoir (Fig. 1), characterized by a set of (coarse-grained) states n and corresponding (free) energy E_n. The equilibrium canonical distribution of finding the system in state n is $P_n^{(\mathrm{eq})} \propto e^{-E_n}$ (we measure energy and entropy in units of temperature and Boltzmann constant, respectively). An example of n is a variable indicating whether an enzyme is in one conformational state or another, defined for instance in terms of an angle or distance between certain subdomains within the protein. When driven by the reservoir, each state n has some degree of catalytic activity toward a reaction $\mathrm{R} \rightarrow \mathrm{P}$ imposed by the reservoir:

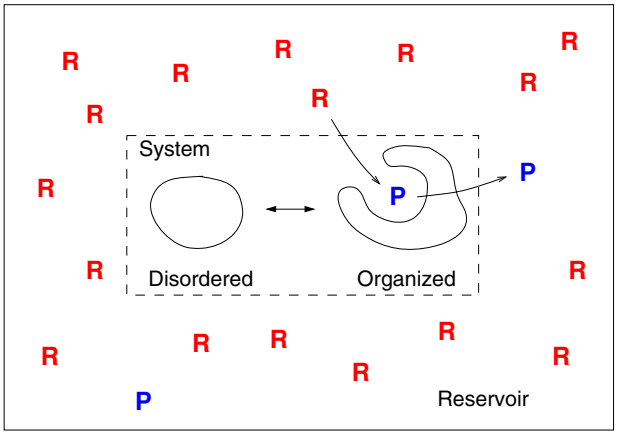

Figure 1 | Self-organization under chemical driving force. A nonequilibrium driving force of reservoir biases the equilibrium inside the system, where ingredients for an enzyme switches between disordered (D) and organized (O) states. The enzyme becomes active only in the latter state, catalyzing the reaction from substrate R into product P. The reservoir supplies R in excess quantities versus P.

$$\mathrm{R}+\mathrm{S}_n \underset{k_n^{(-1)}}{\overset{k_n^{(1)}}{\rightleftharpoons}} \mathrm{S}_n\cdot\mathrm{R} \underset{k_n^{(-2)}}{\overset{k_n^{(2)}}{\rightleftharpoons}} \mathrm{S}_n+\mathrm{P}, \tag{1}$$

where S_n is the system in state n, $\mathrm{S}_n\cdot\mathrm{R}$ denotes a complex with a bound R, and the rate constants $k_n^{(\pm 1,2)}$ vary depending on n. This scheme of enzyme catalysis under different conformational states corresponds to the conformational selection mechanism (as opposed to the induced fit), now supported by a growing body of experimental evidence[24,46]. The system also equilibrates between different states n and m:

$$\mathrm{S}_n \underset{k_{nm}}{\overset{k_{mn}}{\rightleftharpoons}} \mathrm{S}_m, \tag{2}$$

where k_{mn} is the rate of conversion from n to m. The rate equation for the probability $P_n^{(r)}$ of being in state n with a bound R is

$$\dot{P}_n^{(r)} = c_r k_n^{(1)} P_n^{(0)} - k_n^{(-1)} P_n^{(r)} - k_n^{(2)} P_n^{(r)} + c_p k_n^{(-2)} P_n^{(0)}, \tag{3}$$

where $P_n^{(0)}$ is the probability for n without R, and c_r, c_p are the concentrations of R and P, respectively. At steady state,

$$P_n^{(r)} = z_n P_n^{(0)}, \tag{4}$$

where

$$z_n = \frac{c_r k_n^{(1)} + c_p k_n^{(-2)}}{k_n^{(-1)} + k_n^{(2)}}, \tag{5}$$

or if $c_p = 0$, $z_n = c_r\kappa_n = c_r/K_m$ where $\kappa_n^{-1} = K_m = \left(k_n^{(-1)} + k_n^{(2)}\right) \Big/ k_n^{(1)}$ is the Michaelis constant.

In the absence of driving forces from the reservoir, the set of states $\{n\}$ satisfies the detailed balance condition, $k_{mn}P_n^{(0)} = k_{nm}P_m^{(0)}$, which remains valid even when $P_n^{(r)} > 0$, because Eq. (3) does not couple n and m. This assumption could be violated if for instance the reaction scheme Eq. (1) is generalized such that $\mathrm{S}_n \cdot \mathrm{R}$ could turn into states $m \neq n$, which may yield useful models for molecular motors[47–49]. We will not consider such cases in this paper. In contrast to Eq. (2), the two main reaction steps in Eq. (1) do not satisfy detailed balance except in equilibrium.

We have $P_m^{(0)}/P_n^{(0)} = k_{mn}/k_{nm} = P_m^{(\mathrm{eq})}/P_n^{(\mathrm{eq})} = e^{-E_m+E_n}$, and may thus write

$$P_n^{(0)} = \frac{e^{-E_n}}{Q}, \tag{6}$$

where Q is a normalization constant. From Eqs. (4) and (6), the total probability P_n to be in state n is

$$P_n = P_n^{(0)} + P_n^{(r)} = \frac{1+z_n}{Q} e^{-E_n} \equiv \frac{e^{-E_n^*}}{Q}, \tag{7}$$

where

$$E_n^* = E_n - \ln(1+z_n) \tag{8}$$

is a generalized free energy. Since $\Sigma_n P_n = 1$, we have

$$Q = \sum_n (1+z_n) e^{-E_n} = \sum_n e^{-E_n^*}, \tag{9}$$

which is therefore a generalized partition function. The significance of partition function in equilibrium statistical mechanics carries over to this nonequilibrium extension: the expectation value of a state-dependent quantity q_n,

$$\langle q \rangle = \sum_n q_n P_n = \frac{1}{Q} \sum_n q_n e^{-E_n^*}, \tag{10}$$

can be calculated from $Q(f) = \Sigma_n e^{-E_n^* + f q_n}$ by $\langle q \rangle = \partial \ln Q/\partial f|_{f=0}$.

From Eq. (1), the stationary velocity v can be written as

$$\begin{aligned} v &= \sum_n \left[k_n^{(2)} P_n^{(r)} - c_p k_n^{(-2)} P_n^{(0)} \right] \\ &= \frac{1}{Q} \sum_n \frac{c_r k_n^{(1)} k_n^{(2)} - c_p k_n^{(-1)} k_n^{(-2)}}{k_n^{(-1)} + k_n^{(2)}} e^{-E_n}, \end{aligned} \tag{11}$$

which is a constitutive relation connecting the nonequilibrium flux v to thermodynamic forces: at steady state, the system entropy is constant, and the total entropy production rate is $\dot{S}$ where $S = S(n_r, n_p)$ is the entropy of the reservoir, a function of the number of species n_i (i = R, P). Its rate of change due to the catalyzed reaction is $\dot{S} = -\mu_r \dot{n}_r - \mu_p \dot{n}_p = Fv$, where μ_i are the chemical potentials[50] and $F = \mu_r - \mu_p$ since $\dot{n}_p = -\dot{n}_r = v$. With the concentrations $c_i = c_i^\circ e^{\mu_i}$ where c_i° is a constant, Eq. (11) gives a closed-form relation of the net flux v to μ_i.

A simplest special case is where there are only two states, n = D, O, each corresponding to disordered and organized states (Fig. 1), and $\Delta E = E_O - E_D > 0$. From Eq. (8), if $k_D^{(\pm 1)} = 0$ and $c_p = 0$, the relative stability of D over O is reversed when $c_r > c_r^* = K_m(e^{\Delta E} - 1)$. Typical K_m values of enzymes range from μM to mM ranges[1]. For an enzyme with $K_m = 1$ μM, for instance, $c_r^* \simeq 22$ mM for $\Delta E = 10$. It is worth noting that in contrast to self-assembly, the stabilization of states with $\Delta E > 0$ arises strictly from the chemical driving force of the reservoir. If the matter flow is cut off, the system would quickly revert to D: the organized structure in the system 'dies'.

For the two state model, Eq. (11) becomes

$$v = \frac{k^{(2)} c_r - c_p k^{(-1)} k^{(-2)}/k^{(1)}}{(1+e^{\Delta E}) K_m + c_r + c_p k^{(-2)}/k^{(1)}}, \tag{12}$$

where the rate constants are those for state O. When $c_p = 0$, Eq. (12) reduces to the Michaelis-Menten expression with the substrate concentration for half-maximum velocity increased by a factor of $1 + e^{\Delta E}$ due to the presence of the noncatalytic state D. Equilibrium is reached when $c_r/c_p = k^{(-1)}k^{(-2)}/k^{(1)}k^{(2)}$ and $v = 0$.

More general cases can be illustrated further by a simple toy model of an enzyme with a continuous angular degree of freedom θ. This angle represents the degree of closing of the binding pocket at the center. As θ becomes smaller, the catalytic activity increases linearly, while the thermal stability decreases:

$$E(\theta) = g_0(1-\theta/\pi), \quad K_m^{-1} = \kappa(1-\theta/\pi), \tag{13}$$

and $c_p = 0$ such that $z = c_r/K_m$. Figure 2 shows the changes in angular distribution and mean angle from the equilibrium to driven cases as the driving force $\zeta = c_r\kappa$ increases.

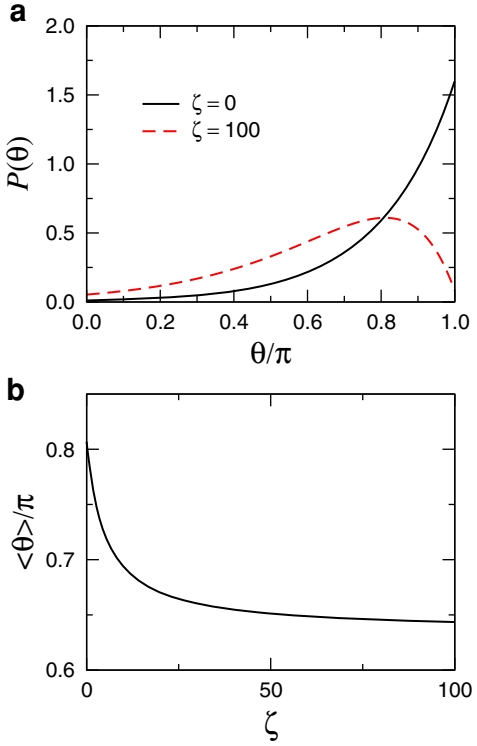

Figure 2 | Organization induced under driven conditions. The model defined by Eq. (13) was used with $g_0 = 5$. (a) Angular distribution $P(\theta)$. (b) Mean angle $\langle\theta\rangle$ in units of π as a function of the reduced reactant concentration $\zeta = c_r\kappa$.

Information generation. Virtually all self-organized structures in biological systems are based on biopolymers, including proteins and nucleic acids (RNAs and DNAs), which carry biological information that are copied over generations with mutations. A satisfactory theory of self-organization therefore must address how such biological information could have been generated. We show below that the chemical driving force imposed by the reservoir induces an order-disorder transition in sequence space, where the stationary distribution of genotypes becomes dominated by sequences with catalytic activity.

We consider a chemically driven environment where nucleotide sequences of a fixed length l are continually synthesized and degraded, e.g., against a solid support. Each nucleotide at different sites on the chain can contain one of four possible bases, $b = \mathbf{a}, \mathbf{g}, \mathbf{c}, \mathbf{u}$, with the total number of all possible sequences $\mathbf{s}_n = \{b_1, ..., b_l\}$, $\Omega = 4^l$. Without additional driving forces other than the chain synthesis, the resulting sequences would be mostly random. This pool of random sequences corresponds to the system in the disordered D state in Fig. 1. The stationary probability P_n for an RNA chain randomly picked from the population to have sequence $\mathbf{s}_n$ is $P_n \propto e^{-E_n}$, where E_n describes the intrinsic relative stability of each sequence.

If the reservoir exerts a driving force $F = \mu_r - \mu_p > 0$ for a reaction R → P, Eq. (1) describes the catalysis with S_n referring to a chain with sequence n: RNA chains with suitable sequences can fold and

catalyze reactions, as does the catalytic core of modern ribosomes carrying out the synthesis of proteins during translation[51,52]. Equation (8) gives the biased energy E_n^* under the driven condition, where we assume $E_n = 0$ for simplicity. We note that z_n given by Eq. (5) in this case represents the enzymatic activity of the sequence n, or its *fitness*: the generalized free energy E_n^* is more negative for sequences with higher fitness. Equation (7) therefore implies that as the system reaches the stationary state under the influence of reservoir driving force, the distribution would become peaked around sequences with low energy, or high fitness. The analogy of the evolution of populations toward genotypes with higher fitness – the climbing of peaks on the fitness landscape – to the minimization of energy in equilibrium systems has been observed before, notably by Kauffman[15] (but also note the proposal that fitness landscapes are intrinsically dynamic quantities[53]). Here, we see this correspondence more directly from thermodynamic considerations.

This identification of the fitness of genotypes as their catalytic activity contains its common definition in evolutionary theories – the replication rate – as a special case where R → P is the self-replication reaction. The order-disorder transition we describe below, on the other hand, does not rely on the assumption of the existence of self-replication machinery, and may provide an understanding of how the large variety of genes found in modern genomes coding for enzymes with many different functions could have originated.

Emergence of a gene. To describe the biased population of sequences under driven conditions in more detail, the fitness (or energy) landscape E_n^* needs to be specified. Physically, we only need to be convinced of the existence of sequences with high catalytic activities toward the reaction imposed by the reservoir. The fitness landscape describes the dependence of fitness values as we depart from these genotypes in sequence space. The theory of self-organization allows for a quantitative description of the dominance of highly fit genotypes with a discontinuous first-order transition.

For concreteness, we adopt the class of landscapes widely studied in applications of the quasispecies theory[28,37,39,41,42,54], where the fitness is given as a function of Hamming distance to the master sequence, which becomes the natural order parameter. It is worth emphasizing, however, that in contrast to the error catastrophe transition in the quasispecies theory, the order-disorder transition we derive below is thermodynamic in origin independent of specific population dynamics, and is applicable to any reaction for which a sufficiently strong enzymatic genotype exists in the sequence space.

The distance h_{nm} between two sequences n and m is the total number of nucleotide positions at which the base identities differ. For l-mers, $0 \le h \le l$. The probability P_h of sequences with distance h is

$$P_h = \Omega_h \frac{e^{-E_h^*}}{Q} = \frac{e^{-G_h}}{Q}, \tag{14}$$

where the free energy

$$G_h = E_h^* - S_h \tag{15}$$

is given in terms of the entropy $S_h = \ln \Omega_h$ and the number of genotypes Ω_h of distance h to the master sequence. This number can be written as

$$\Omega_h = C_h^l \cdot 3^h, \tag{16}$$

which gives the number of different ways of choosing h sites within l nucleotide positions, each site having 3 possible bases that differ from that of master sequence. The prefactor is the binomial coefficient, $C_h^l = l!/h!(l-h)!$.

In Fig. 3(a), a model fitness landscape

$$\kappa(h) = \kappa e^{-h^2/2\xi^2}, \tag{17}$$

where the fitness values are distributed by a Gaussian function centered at the master sequence ($h = 0$) and $c_p = 0$ such that $z_h = c_r\kappa(h)$, was used to calculate the free energy profile as a function of h for three different values of the driving force $\zeta = c_r\kappa$. Under $\zeta = 0$, the minimum occurs near $h = 3l/4$, which is the average distance of random sequences of length l from the master sequence, because random sequences have the maximum entropy. As ζ increases, a new minimum develops near $h = 0$, whose location is determined by the balance of energy E_h^* and entropy S_h. The stationary distribution P_h peaked at the minimum of G_h can be interpreted physically as follows: the probability of finding genotypes away from the master sequence is affected by the fitness cost (the energy increases with increasing h, reducing P_h) and the number of possible sequences (the entropy increases with increasing h, making P_h larger).

The qualitative features of G_h for different driving forces resembles the order-disorder transitions observed in equilibrium fluids, where a condensed phase can coexist with the disordered phase. In the corresponding phase diagram shown in Fig. 3(b) as a function of distance h and driving force ζ, the coexistence region between the disordered D phase and the organized O phase shrinks as the fitness peak width ξ increases and ζ decreases, vanishing at a critical point where the D-O transition becomes continuous. The D phase is characterized by random sequences, while in the O phase, the sequence

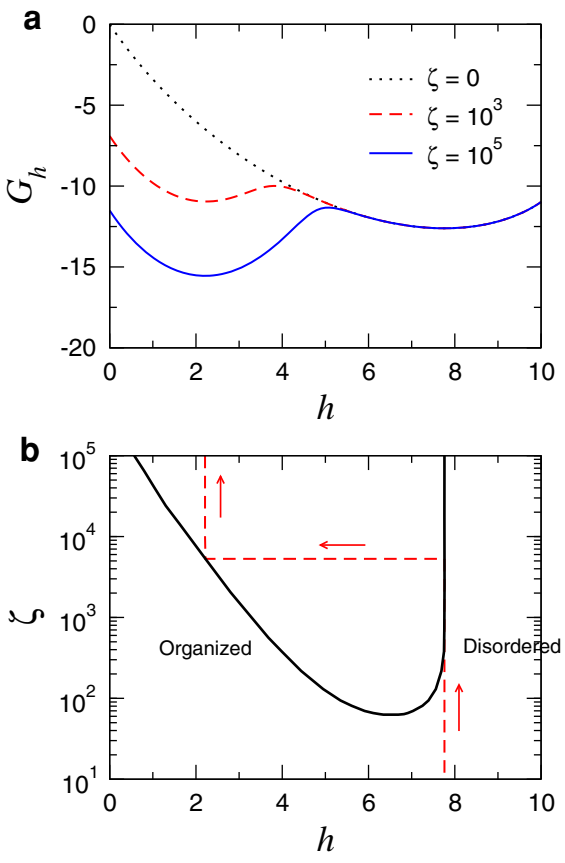

Figure 3 | Stability of catalytic sequences under driven conditions. (a) Free energy profile G_h as a function of distance h to the master sequence given by Eq. (15) for $l = 10$. The fitness landscape (17) was used with $\xi = 1$. (b) Phase diagram showing the coexistence curve of the disordered and organized phases. The dashed line in (b) shows an 'isotherm' (order parameter h as a function of driving force ζ) for $\xi = 1$ with the arrows indicating the direction of increasing ζ.

distribution is peaked at a short distance from the catalytically active master sequence. The position of the minimum in G_h in the O phase in Fig. 3(a) corresponds to the most likely distance value h expected within this population.

The transition from the disordered to organized phases is accompanied by the dominance of catalytically active states leading to a bias in distribution P_n, under which sequences carry information encoding enzymes. The amount of this information generated is quantified by the reduction in Gibbs entropy associated with P_n, known as the information content (per site)[27,55,56],

$$I_c = S_c^{(\mathrm{eq})} + \frac{1}{l}\sum_n P_n \ln P_n = \ln 4 - \left(\ln Q + \left\langle E_h^* \right\rangle\right)/l, \qquad (18)$$

where $S_c^{(\mathrm{eq})}$ is the entropy of $P_n^{(\mathrm{eq})}$, which is ln 4 (or 2 bits) here with $E_n = 0$. The Gibbs entropy in Eq. (18) is that of the system, which remains constant over time in stationary states, while entropy is produced steadily in the reservoir. Figure 4 shows the growth of I_c with increasing driving force ζ for different values of fitness peak width ξ. The information content makes a jump as the D-to-O transition occurs, with the asymptotic value increasing with decreasing ξ. The information content of 2 bits per site is the upper limit reached when the final biased distribution is very sharply peaked.

Multiple genes. Equation (1) can be easily generalized to cases where there are more than one externally imposed reactions, which would lead to the coexistence of multiple genes. We consider the case of two reactions (e.g., RNA elongation and another reaction such as ATP hydrolysis). The order parameter is a vector $\mathbf{h} = (h_1, h_2)$ with two components specifying distances to the master sequences $\mathbf{s}_1$ and $\mathbf{s}_2$. It can then be shown that Eq. (8) becomes

$$E_{\mathbf{h}}^* = -\ln[1 + c_1\kappa_1(h_1) + c_2\kappa_2(h_2)], \qquad (19)$$

where c_1 and c_2 are the concentrations of two reactants. We adopt $\kappa_{1,2}(h) = \kappa(h)$ with Eq. (17) and let $\zeta_1 = c_1\kappa$ and $\zeta_2 = c_2\kappa$ be the set of reduced driving forces.

The calculation of $\Omega_{\mathbf{h}}$ involves counting the number of sequences with given distances $\mathbf{h}$ to the master sequences. In Fig. 5, the two master sequences $\mathbf{s}_1$ and $\mathbf{s}_2$ are depicted with their sites grouped into two sections, I and II, where nucleotides are different and identical between the sequences, respectively. The length of section I is $h_{12} = d$. To count the number of sequences $\mathbf{s}_n$ with distances h_1 and h_2 to $\mathbf{s}_1$ and $\mathbf{s}_2$, we start with $\mathbf{s}_1$ and first mutate a subset of length m from section II, such that $\mathbf{h} = (m, d + m)$. The number of ways of doing this is $C_m^{l-d}\cdot 3^m$ because there are three nucleotides different from each site in section II. We then mutate k sites from section I into the corresponding nucleotides of $\mathbf{s}_2$, which results in $\mathbf{h} = (m + k, d + m - k)$. The number of ways of doing this step is C_k^d without any nucleotide multiplicity because the target sequence is fixed. By choosing $k = d + m - h_2$, the distance h_2 to $\mathbf{s}_2$ is achieved. Finally, we choose p additional sites from section I and mutate into nucleotides distinct from both $\mathbf{s}_1$ and $\mathbf{s}_2$, after which $\mathbf{h} = (m + k + p, d + m - k)$. The number of ways for this third step is $C_p^{d-k}\cdot 2^p$ because there are two nucleotides that can be chosen for each site. Taking $p = h_1 - m - k = h_1 + h_2 - d - 2m$, we achieve the distance h_1 to $\mathbf{s}_1$. The total number of sequences is then given by

$$\Omega_{\mathbf{h}} = \sum_{m=m_0}^{m_1} C_m^{l-d}\cdot 3^m \cdot C_{d+m-h_2}^{d} C_{h_1+h_2-d-2m}^{h_2-m}\cdot 2^{h_1+h_2-d-2m}, \qquad (20)$$

where the lower and upper limits of the summation can be deduced by requiring the binomial coefficients to be well-defined:

$$m_0 = \max\{0,\, h_1 - d,\, h_2 - d\}, \qquad (21a)$$

$$m_1 = \min\left\{h_1,\, h_2,\, l - d,\, \left\lfloor \frac{h_1 + h_2 - d}{2} \right\rfloor\right\}, \qquad (21b)$$

where $\lfloor x \rfloor$ is the largest integer not exceeding x.

Figure 6 shows the free energy $G_{\mathbf{h}} = E_{\mathbf{h}}^* - S_{\mathbf{h}}$ as a function of $\mathbf{h}$ for equilibrium ($\zeta_i = 0$) and strongly driven ($\zeta_1 = \zeta_2 = 10^4$) cases. The landscape is bounded by values of $\mathbf{h}$ for which $\Omega_{\mathbf{h}} = 0$. The allowed region can be deduced by requiring $m_0 \le m_1$ in Eqs. (21): $0 \le h_i \le l$, $h_1 \le h_2 + d$, $h_2 \le h_1 + d$, and $h_1 + h_2 \ge d$. The single minimum in Fig. 6(a) at $\mathbf{h} = (8, 8)$ corresponds to the disordered D phase, which is the only phase in equilibrium. Under strong driving forces, in contrast, up to two additional minima develop near the h_2 and h_1-axes [Fig. 6(b)], which correspond to organized O_1 and O_2 phases dominated by sequences close to $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively. The global minimum switches from D to O_i (or both when $\zeta_1 = \zeta_2$) with increasing ζ_i.

The phase diagram in Fig. 7 shows such stability changes within the ζ_i-space, where the three phases are separated by boundaries on which neighboring phases can coexist. There is a triple point where D, O_1 and O_2 phases can all coexist. As suggested by the single-gene case in Fig. 3(b), increasing the fitness peak width parameter ξ shifts the D-O_i boundaries toward smaller ζ_i values. The boundaries disappear at a critical point. The symmetry $1 \leftrightarrow 2$ in Figs. 6 and 7 is a result of our choice of the same fitness landscape, Eq. (17), for the two genes, and would be broken in more general cases.

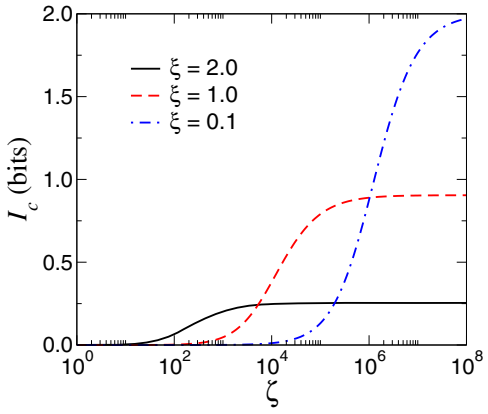

Figure 4 | Information content versus driving force. The value given by Eq. (18) (divided by ln 2 to convert I_c into bits) is shown as a function of driving force ζ for three different landscape width (ξ) values. The maximum for nucleotide sequences is 2 bits per site.

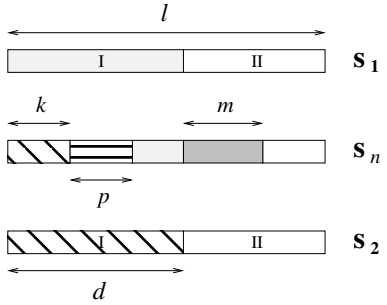

Figure 5 | Calculation of entropy for two genes. The number of sequences $\Omega(h_1, h_2)$ is counted, where h_1 and h_2 are the distances to the two master sequences $\mathbf{s}_1$ and $\mathbf{s}_2$. Sections I and II are sites where the two sequences are different and identical, respectively. New sequences $\mathbf{s}_n$ are generated by first choosing m sites from section II of $\mathbf{s}_1$ and mutating them, k sites from section I into the corresponding sequences of $\mathbf{s}_2$, and p sites away from both $\mathbf{s}_1$ and $\mathbf{s}_2$, such that $\mathbf{h} = (m + k + p, d + m - k)$. The number of sites k and p are chosen such that $\mathbf{h} = (h_1, h_2)$. Segments with the same filled patterns represent the same sequences.

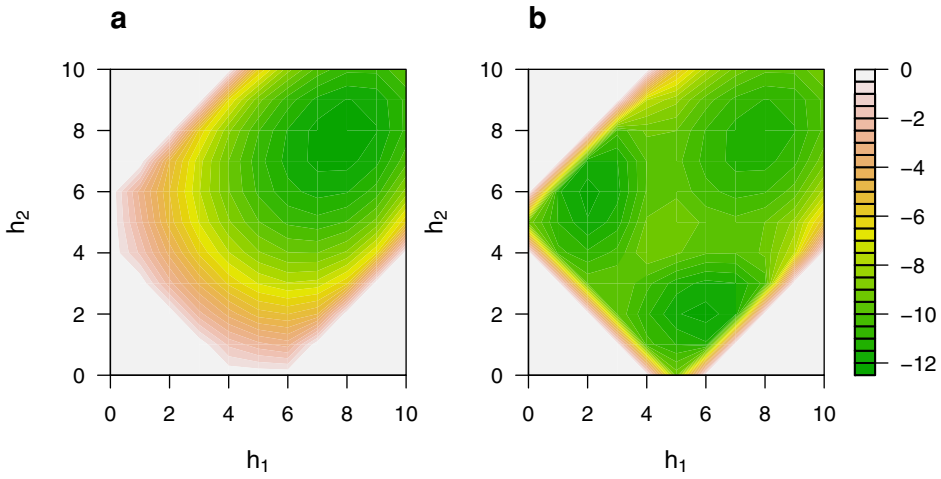

Figure 6 | Free energy landscape for two genes. The free energy $G_{\mathbf{h}}$ as a function of $\mathbf{h} = (h_1, h_2)$ is shown for $l = 10$, $d = 5$, and $\xi = 1$. (a) Equilibrium condition where $\zeta_1 = \zeta_2 = 0$. There is a single minimum at $\mathbf{h} = (8, 8)$, corresponding to the D phase. (b) Strongly driven case where $\zeta_1 = \zeta_2 = 10^4$. A pair of minima at $\mathbf{h} = (2, 6)$, $(6, 2)$, corresponding to the O_1 and O_2 phases, becomes more stable than the D phase.

Discussion

In this paper, we presented a statistical mechanical description of biological self-organization under chemically driven conditions: the system possesses a small subset of coarse-grained states which can catalyze the reactions imposed externally by the reservoir. The entropic cost of observing these organized states is overcome when the driving forces are sufficiently strong, leading to biased stationary distributions dominated by such organized states, or *phases*.

As a major application, we focused on the appearance of biological information encoded into sequences of nucleotide strands. Chemical driving forces lead to one or multiple peaks in sequence space, centered on sequences that can catalyze the externally imposed reactions. In this viewpoint, genes encoding enzymes spontaneously appear and dominate the nucleotide sequence populations if the driving forces are sufficiently strong to overcome the Shannon entropy costs. Our analyses demonstrate that this transition into self-organization in sequence space has much in common with equilibrium phase transitions. Although the global stability of phases is dictated by the phase diagram (Fig. 7) as in equilibrium, metastable phases (and genes) can still be present together with the dominant phase outside the coexistence region, with the relative population of different sequences given by Eq. (7).

The phase transition observed here can occur irrespective of how interconversion between sequences actually takes place (random synthesis on solid support or replication of existing chains). The nature of sequence evolution, however, would affect the *dynamics* of ordering, which was not considered here. The establishment of stationary distribution, Eq. (7), requires sufficient exploration of all sequences via Eq. (2), fastest with random synthesis but still taking relaxation times that grow exponentially with l. High-fidelity replications would slow down this relaxation, while allowing for the preservation of information already discovered.

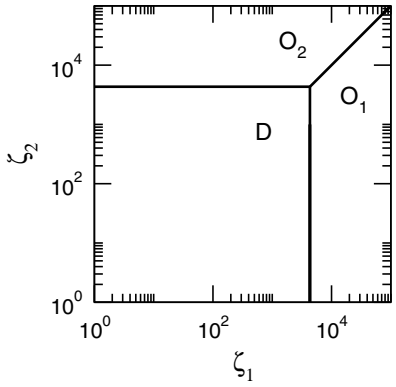

Figure 7 | Phase diagram for two genes. The parameter values are $l = 10$, $d = 5$, $\xi = 1$. Lines represent phase boundaries on which neighboring phases coexist.

The organized phases in Figs. 6 and 7 consist of groups of heteropolymers independently acting as enzymes. One of these groups can be polymerases catalyzing the synthesis of polymers. An aspect of evolutionary transitions we may presume to have occurred, in particular, is that of the emergence of 'selfishness', or the ability of the polymerase gene to limit its action to its own replication only, excluding others. This assumption is one of the starting points of the quasispecies theory[37]. The selfishness is also likely to be closely related to the conjoining of genes into *genomes*. It will be of interest to see how we may understand the evolution of selfishness within the statistical mechanical perspective.

The viewpoint we adopted for biological self-organization – the stabilization of structures capable of catalyzing reactions imposed by chemical driving forces – may also have relevance to the question of how one may usefully define living organisms. Ruiz-Mirazo et al.[57] emphasized two main elements in such a definition: autonomy and open-ended evolution. The former is a subset of self-organization capable of auto-regulation, while the latter requires the establishment of a division of labor between record-keeping (DNA) and expression (proteins) of biological information. The perspective adopted and elaborated in this paper shows how thermodynamic driving forces both constrain and enable self-organization, which may prove useful in understanding higher-level structures.

The statistical mechanical expression for stationary states and its partition function, Eq. (9), can form a basis for calculating properties of systems with more complex features than assumed here. In particular, one may have multiple reactions coupled to each other, a common situation in biochemical systems, which would lead to an extension of Eq. (19) to a coupled 'hamiltonian'. The calculation of partition function would be akin to that for interacting systems in equilibrium such as the Ising model. Such an extension would also allow one to consider fairly large systems in which the system components catalyze the formation of one another, forming an autocatalytic network[15,32,33]. In such systems, the thermodynamic phase transitions studied here may therefore precede and combine with the transition to self-sustained autocatalytic organizations.

Methods

For the toy model defined by Eq. (13), after replacing discrete sums with integrals and adding a field f, Eq. (9) reads

$$Q(f) = \int_0^\pi d\theta (1 + c_r/K_m) e^{-E(\theta) + f\theta} = \frac{1}{g_0/\pi + f}\left[e^{\pi f} - (1+\zeta)e^{-g_0} + \zeta \frac{e^{\pi f} - e^{-g_0}}{g_0 + \pi f}\right] \quad (22)$$

and the mean angle (Fig. 2) can be calculated by

$$\langle\theta\rangle = \frac{\partial \ln Q}{\partial f}\Big|_{f=0} = \frac{\pi}{g_0}\left\{\frac{g_0(g_0+\zeta) - \zeta(1-e^{-g_0})}{g_0[1-(1+\zeta)e^{-g_0}] + \zeta(1-e^{-g_0})} - 1\right\}. \quad (23)$$

In obtaining Fig. 3, Eq. (15) was used with Eq. (16) and $C_h^l = \Gamma(l+1)/\Gamma(h+1)\Gamma(l-h+1)$, where $\Gamma(z)$ is the gamma function, such that nonintegral values of distances could be included. For the two-dimensional landscapes in Fig. 6, $h_{1,2}$ were restricted to integers because of the summation in Eq. (20).

The phase diagram in Fig. 3(b) was obtained by varying the width parameter ξ of the landscape (17), and locating the values of the driving force ζ for which the two phases – organized (O) and disordered (D) – have the same free energy. The coexistence values of ζ decreases with increasing ξ from top to bottom.

The validity of Eq. (20) for the total number of sequences with given distances to two master sequences was verified by enumerating all genotypes for small l and counting the number of sequences for each set of possible distance values.

1. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry* (Freeman, New York, 2002).
2. Newman, M. E. J. *Networks: An Introduction* (Oxford, New York, 2010).
3. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
4. Ahn, Y.-Y., Ahnert, S. E., Bagrow, J. P. & Barabási, A.-L. Flavor network and the principles of food pairing. *Sci. Rep.* **1**, 196 (2011).
5. Bagrow, J. P. Communities and bottlenecks: trees and treelike networks have high modularity. *Phys. Rev. E* **85**, 066118 (2012).
6. Colomer-de-Simón, P., Serrano, M. Á., Beiró, M. G., Alvarez-Hamelin, J. I. & Boguñá, M. Deciphering the global organization of clustering in real complex networks. *Sci. Rep.* **3**, 2517 (2013).
7. Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nat. Phys.* **8**, 32–39 (2012).
8. Meloni, S. *et al.* Modeling human mobility responses to the large-scale spreading of infectious diseases. *Sci. Rep.* **1**, 62 (2011).
9. Boguñá, M., Castellano, C. & Pastor-Satorras, R. Nature of the epidemic threshold for the susceptible-infected-susceptible dynamics in networks. *Phys. Rev. Lett.* **111**, 068701 (2013).
10. Hopfield, J. J. Physics, computation, and why biology looks so different. *J. Theor. Biol.* **171**, 53–60 (1994).
11. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality: an explanation of the 1/*f* noise. *Phys. Rev. Lett.* **59**, 381–384 (1987).
12. Drossel, B. Complex scaling behavior of nonconserved self-organized critical systems. *Phys. Rev. Lett.* **89**, 238701 (2002).
13. Newman, M. E. J. & Palmer, R. G. *Modeling Extinction* (Oxford, New York, 2003).
14. Haken, H. *Synergetics: An Introduction* (Springer, Berlin, 1983).
15. Kauffman, S. A. *The Origins of Order: Self-Organization and Selction in Evolution* (Oxford, New York, 1993).
16. Hanel, R., Pöchacker, M., Schölling, M. & Thurner, S. A self-organized model for cell-differentiation based on variations of molecular decay rates. *PLOS ONE* **7**, e36679 (2012).
17. Theurkauff, I., Cottin-Bizonne, C., Palacci, J., Ybert, C. & Bocquet, L. Dynamic clustering in active colloidal suspensions with chemical signaling. *Phys. Rev. Lett.* **108**, 268303 (2012).
18. Gehrmann, E. & Drossel, B. Boolean versus continuous dynamics on simple two-gene modules. *Phys. Rev. E* **82**, 046120 (2010).
19. Safran, S. A. *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes* (Addison-Wesley, New York, 1994).
20. Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* **437**, 640–647 (2005).
21. Min, W. *et al.* Flucuating enzymes: lessons from single-molecule studies. *Acc. Chem. Res.* **38**, 923–931 (2005).
22. Kou, S. C., Cherayil, B. J., Min, W., English, B. P. & Xie, X. S. Single-molecule Michaelis-Menten equations. *J. Phys. Chem. B* **109**, 19068–19081 (2005).
23. Bohr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
24. Ma, B. & Nussinov, R. Enzyme dynamics point to stepwise conformational selection in catalysis. *Curr. Opin. Chem. Biol.* **14**, 652–659 (2010).
25. Weikl, T. R. & von Deuster, C. Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. *Proteins* **75**, 104–110 (2009).
26. Min, W., Xie, X. S. & Bagchi, B. Role of conformational dynamics in kinetics of an enzymatic cycle in a nonequilibrium steady state. *J. Chem. Phys.* **131**, 065104 (2009).
27. Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (Univ. of Illinois Press, Urbana, 1949).
28. Drossel, B. Biological evolution and statistical physics. *Adv. Phys.* **50**, 209–295 (2001).
29. Ferrada, E. & Wagner, A. Evolutionary innovations and the organization of protein functions in genotype space. *PLOS ONE* **5**, e14172 (2010).
30. Ferrada, A. & Wagner, A. Protein robustness promotes evolutionary innovations on large evolutionary time-scales. *Proc. Roy. Soc. B-Biol. Sci.* **275**, 1595–1602 (2008).
31. Braakman, R. & Smith, E. The emergence and early evolution of biological carbon-fixation. *PLOS Comput. Biol.* **8**, e1002455 (2012).
32. Farmer, J. D., Kauffman, S. A. & Packard, N. H. Autocatalytic replication of polymers. *Physica D* **22**, 50–67 (1986).
33. Kauffman, S. A. Autocatalytic sets of proteins. *J. Theor. Biol.* **119**, 1–24 (1986).
34. Hanel, R., Kauffman, S. A. & Thurner, S. Phase transition in random catalytic networks. *Phys. Rev. E* **72**, 036117 (2005).
35. Fontana, W. & Buss, L. W. "The arrival of the fittest": toward a theory of biological organization. *B. Math. Biol.* **56**, 1–64 (1994).
36. Dittrich, P. & di Fenizio, P. S. Chemical organization theory. *B. Math. Biol.* **69**, 1199–1231 (2007).
37. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Natur-wissenschaften* **58**, 465–523 (1971).
38. Swetina, J. & Schuster, P. Self-replication with errors. A model for polynucleotide replication. *Biophys. Chem.* **16**, 329–345 (1982).
39. Tannenbaum, E. & Shakhnovich, E. I. Semiconservative replication, genetic repair, and many-gened genomes: extending the quasispecies paradigm to living systems. *Phys. Life Rev.* **2**, 290–317 (2005).
40. Altmeyer, S. & McCaskill, J. S. Error threshold for spatially resolved evolution in the quasispecies model. *Phys. Rev. Lett.* **86**, 5819–5822 (2001).
41. Saakian, D. B. & Hu, C.-K. Exact solution of the Eigen model with general fitness functions and degradation rates. *Proc. Natl. Acad. Sci. USA* **103**, 4935–4939 (2006).
42. Saakian, D. B., Muñoz, E., Hu, C.-K. & Deem, M. W. Quasispecies theory for multiple-peak fitness landscapes. *Phys. Rev. E* **73**, 041913 (2006).
43. Muñoz, E., Park, J.-M. & Deem, M. W. Quasispecies theory for horizontal gene transfer and recombination. *Phys. Rev. E* **78**, 061921 (2008).
44. Park, J.-M., Muñoz, E. & Deem, M. W. Quasispecies theory for finite populations. *Phys. Rev. E* **81**, 011902 (2010).
45. Woo, H.-J. & Wallqvist, A. Nonequilibrium phase transitions associated with DNA replication. *Phys. Rev. Lett.* **106**, 060601 (2011).
46. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends. Bichem. Sci.* **35**, 539–546 (2010).
47. Reimann, P. Brownian motors: noisy transport far from equilibrium. *Phys. Rep.* **361**, 57–265 (2002).
48. Woo, H.-J. Analytical theory of the nonequilibrium spatial distribution of RNA polymerase translocations. *Phys. Rev. E* **74**, 011907 (2006).
49. Woo, H.-J. Relaxation dynamics near nonequilibrium stationary states in Brownian ratchets. *Phys. Rev. E* **79**, 021101 (2009).
50. Callen, H. B. *Thermodynamics and an Introduction to Thermostatistics* (Wiley, New York, 1985).
51. Cech, T. R. Structural biology. The ribosome is a ribozyme. *Science* **289**, 878–879 (2000).
52. Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418**, 214–221 (2002).
53. Klimek, P., Thurner, S. & Hanel, R. Evolutionary dynamics from a variational principle. *Phys. Rev. E* **82**, 011901 (2010).
54. Itan, E. & Tannenbaum, E. Semiconservative quasispecies equations for polysomic genomes: the general case. *Phys. Rev. E* **81**, 061915 (2010).
55. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431 (1986).
56. Adami, C. Information theory in molecular biology. *Phys. Life Rev.* **1**, 3–22 (2004).
57. Ruiz-Mirazo, K., Peretó, J. & Moreno, A. A universal definition of life: autonomy and open-ended evolution. *Origins Life Evol. B.* **34**, 323–346 (2004).

Author contributions

H.J.W. performed the research and wrote the manuscript.

Additional information

Competing financial interests: The author declares no competing financial interests.

How to cite this article: Woo, H.J. Stationary distribution of self-organized states and biological information generation. *Sci. Rep.* **3**, 3329; DOI:10.1038/srep03329 (2013).