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 interactions1, 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 networks2–5, and complex dynamical systems exhibiting phenomena such as disease progression6–9 and neural computation10. Different directions of theoretical approaches include wide-ranging studies of driven systems showing self-organized criticality11,12 with implications to extinction dynamics13, dynamical systems views with analogies to equilibrium phase transition14, and concepts centered on autocatalysis, evolution, and selection15. Further studies in this interdisciplinary field include models of genetic regulatory networks and cell differentiation16, dynamic clustering in active media17, and the study of Boolean network dynamics18. 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 interactions19, for which well-established and quantitative theories now exist20. 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 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). Theoretical studies so far have focused mostly on the effects this conformational heterogeneity has on kinetic velocity. 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, 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, as well as highly successful recent developments in data-driven characterizations of biological genotype-phenotype spaces and reconstruction of the evolutionary history of extant metabolic pathways.

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, 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 and theoretical works based on it, 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 order in sequence space can occur purely from chemical driving forces without the mechanism of competition based on self-replication postulated in quasispecies theory. The system also equilibrates between different states and m:

where $S_n$ is the system in state n, $S_n$, R denotes a complex with a bound R, and the rate constants $k_{n}^R$ 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.

The system also equilibrates between different states and m:

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

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

where

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

In the absence of driving forces from the reservoir, the set of states $\{n\}$ satisfies the detailed balance condition $k_{n\rightarrow m} P_n^{(0)} = k_{m\rightarrow n} P_m^{(0)}$, 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 $S_n$, R could turn into states $m \neq n$, which may yield useful models for molecular motors. 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.

**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 $P_n^{(eq)} \propto e^{-E_n / k_B T}$ (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 $R \rightarrow P$ imposed by the reservoir:

\[ R + S_n \xrightarrow{k_n^{(1)}} S_n^R + S_n, \quad S_n^R \xrightarrow{k_n^{(-1)}} S_n + P, \quad (1) \]
We have \( \frac{p_n^{(0)}}{p_n^{(0)}} = k_{on}/k_{off} = \frac{p_n^{(eq)}}{p_n^{(eq)}} = e^{-E_n+E_v}, \) and may thus write

\[
p_n^{(0)} = \frac{e^{-E_n}}{Q},
\]

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

\[
p_n = p_n^{(0)} + p_n^{(r)} = \frac{1 + z_n}{Q} e^{-E_n} \equiv \frac{e^{-E_n}}{Q},
\]

where

\[
E_n^c = E_n - \ln(1 + z_n)
\]

is a generalized free energy. Since \( \Sigma_n p_n = 1 \), we have

\[
Q = \sum_n (1 + z_n) e^{-E_n} = \sum_n e^{-E_n},
\]

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 \( q_m \)

\[
\langle q \rangle = \sum_n q_n p_n = \frac{1}{Q} \sum_n q_n e^{-E_n},
\]

can be calculated from \( Q(f) = \sum_n e^{-E_n + f \delta_n} \) by \( \langle q \rangle = \frac{\partial}{\partial f} Q(f) \rightarrow f=0 \).

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

\[
v = \sum_n \left[ \left\{ \sum_{n'} \left( k_n^{(2)} p_n^{(r)} - c_p k_n^{(2)} p_n^{(0)} \right) \right\}
\]

\[
= \frac{1}{Q} \sum_n \left( c_r \frac{k_n^{(1)}}{k_n^{(2)}} \right) - \frac{c_p k_n^{(1)}}{k_n^{(2)}} - c_r e^{-E_n},
\]

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 \( S \) where \( S = S(n, \eta) \) 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 \( \dot{S} = -c_r \dot{\eta}_r - c_p \dot{\eta}_p = F v_i \), where \( \mu_i \) are the chemical potentials\(^{50} \) and \( F = \mu_r - \mu_p \) since \( \dot{\eta}_p = -n_r = v \). With the concentrations \( c_i = c_i e^{-E_i} \), where \( c_i \) is a constant, Eq. (11) gives a closed-form relation of the net flux \( v \) to \( \mu_i \).

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 \( k_{n}^{(2)}(\pm 1) = 0 \) 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 \( \mu M \) to \( m M \) range\(^{1} \). For an enzyme with \( K_m = 1 \mu M \), for instance, \( c_r^* \approx 22 \mu M \) 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 gradually revert to D: the organized structure in the system ‘dies’.

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

\[
v = \frac{k_r^{(2)} c_r - c_r k_r^{(-1)} k_r^{(-2)}}{1 + e^{E_r} K_m + c_r + c_r k_r^{(-2)} / k_r^{(1)}},
\]

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 \( 1 + e^{E_r} \) due to the presence of the noncatalytic state D. Equilibrium is reached when \( c_r \mu_r = k_r^{(2)} k_r^{(-1)} k_r^{(-2)} / k_r^{(1)} \) 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 \( \theta \). This angle represents the degree of closing of the binding pocket at the center. As \( \theta \) becomes smaller, the catalytic activity increases linearly, while the thermal stability decreases:

\[
E(\theta) = g_0(1 - \theta / \pi), \quad K_m^{-1} = \kappa(1 - \theta / \pi),
\]

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

**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, \( b = a, g, c, u \), with the total number of all possible sequences \( s_o = \{b_1, \ldots, b_l\}, \Omega = 4^l \).

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 \( s_o = P_n \propto e^{-E_n} \), 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 \( R \to P \), 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.\textsuperscript{51,52} Equation (8) gives the biased energy $E_n^b$ 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 $E_n^b$ 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\textsuperscript{17} (but also note the proposal that fitness landscapes are intrinsically dynamic quantities\textsuperscript{20}). 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 $R \rightarrow P$ 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 $E_n^b$ 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,\textsuperscript{28,29,30,41,42,54} 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, $0 \leq h \leq l$. The probability $P_h$ of sequences with distance $h$ is

$$P_h = \Omega_h e^{-E_h^b} / Q,$$

where the free energy

$$G_h = E_h^b - S_h$$

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

$$\Omega_h = C_h^l z_0^h,$$

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

$$\kappa(h) = \kappa e^{-h^2/2\sigma^2},$$

where the fitness values are distributed by a Gaussian function centered at the master sequence ($h = 0$) and $c_p = 0$ such that $z_0 = c_\kappa(k(h))$, was used to calculate the free energy profile as a function of $h$ for three different values of the driving force $\zeta = c_\kappa(h)$. 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 $\zeta$ increases, a new minimum develops near $h = 0$, whose location is determined by the balance of energy $E_h^b$ 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 $h$, and $P_h$), 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 $\zeta$, the coexistence region between the disordered D phase and the organized O phase shrinks as the fitness peak width $\zeta$ increases and $\zeta$ 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...
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 $\xi$ 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 enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes. The amount of this information generated is quantified in distribution $P_n$ by the reduction in Gibbs entropy associated with enzymes.

\[
I_c = S_{eq}^P + \frac{1}{l} \sum \ln P_n = -\ln Q + \left(\ln P_n \right)/l, \tag{18}
\]

where $S_{eq}^P$ is the entropy of $P_{eq}$, which is $4$ (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 $\xi$ for different values of fitness peak width $\xi$. The information content makes a jump as the D-to-O transition occurs, with the asymptotic value increasing with decreasing $\xi$.

The information content makes a jump as the D-to-O transition occurs, with the asymptotic value increasing with decreasing $\xi$.

**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 $h = (h_1, h_2)$ with two components specifying distances to the master sequences $s_1$ and $s_2$. It can then be shown that Eq. (8) becomes

\[
E_h^* = -\ln [1 + c_1 K_1(h_1) + c_2 K_2(h_2)],
\]

where $c_1$ and $c_2$ are the concentrations of two reactants. We adopt $K_1(h) = K(h)$ with Eq. (17) and let $\xi_1 = c_1 K$ and $\xi_2 = c_2 K$ be the set of reduced driving forces.

The calculation of $\Omega_h$ involves counting the number of sequences with given distances $h$ to the master sequences. In Fig. 5, the two master sequences $s_1$ and $s_2$ 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 $s_n$ with distances $h_1$ to $s_1$ and $h_2$ to $s_2$, we start with $s_1$ and first mutate a subset of length $m$ from section II, such that $h = (m, d + m)$. The number of ways of doing this is $C_{m}^{5-2d}$ 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 $s_2$, which results in $h = (m + k, d + m - k)$. The number of ways of doing this step is $C_k^{5-d}$ without any nucleotide multiplicity because the target sequence is fixed. By choosing $k = d + m - h_2$, the distance $h_2$ to $s_2$ is achieved. Finally, we choose $p$ additional sites from section I and mutate into nucleotides distinct from both $s_1$ and $s_2$, after which $h = (m + k + p, d + m - k)$. The number of ways for this third step is $C_p^{5-k-2d}$ 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 $s_2$. The total number of sequences is then given by

\[
\Omega_h = \sum_{m=m_0}^{m_1} C_{m}^{5-d} C_{d}^{m} C_{h_1-h_2-d-2m}^{5} - 2 h_1 + h_2 - d - 2m, \tag{20}
\]

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

\[
m_0 = \max \{0, h_1 - d, h_2 - d\}, \tag{21a}
\]

\[
m_1 = \min \left\{h_1, h_2, l, d, \left[\frac{h_1 + h_2 - d}{2}\right]\right\}, \tag{21b}
\]

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

Figure 6 shows the free energy $G_h = E_h^* - \Omega_h$ as a function of $h$ for equilibrium ($\xi_1 = 0$) and strongly driven ($\xi_1 = \xi_2 = 10^4$) cases. The landscape is bounded by values of $h$ for which $\Omega_h = 0$. The allowed region can be deduced by requiring $m_0 \leq m_1$ in Eqs. (21): $0 \leq h_1 \leq l$, $h_2 \leq h_1 + d$, $h_2 \leq l + h_2 + d$, and $h_2 \leq h_1 + d$. The single minimum in Fig. 6(a) at $h = (8, 8)$ 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 $h_1$ and $h_2$-axes (Fig. 6(b)), which correspond to organized O1 and O2 phases dominated by sequences close to $s_1$ and $s_2$, respectively. The global minimum switches from D to O1 (or both when $\xi_1 = \xi_2$) with increasing $\xi$.

The phase diagram in Fig. 7 shows such stability changes within the $\xi$-$\zeta$ space, where the three phases are separated by boundaries on which neighboring phases can coexist. There is a triple point where D, O1 and O2 phases can all coexist. As suggested by the single-gene case in Fig. 3(b), increasing the fitness peak width parameter $\zeta$ shifts the D-O1 boundaries toward smaller $\xi$ 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](https://example.com/figure4.png)  
**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 $\zeta$ for three different landscape width ($\xi$) values. The maximum for nucleotide sequences is 2 bits per site.
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 phasé.

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.

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. 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. 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. 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

$$Q(f) = \int \frac{d\theta}{\theta} \left[ e^{\theta f} - (1 + \xi e^{-\theta}) \right] \frac{e^{-\theta f}}{C_1}
$$

and the mean field (Eq. 2) can be calculated by

$$\langle \theta \rangle = \frac{\ln Q(f)}{f} - 0
$$

In obtaining Fig. 3, Eq. (15) was used with Eq. (16) and $C_{1,2} = G(1+1/G(1+G(1-1/G(1-h-1)))$, where $G(\tau)$ is the gamma function, such that noninteger 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 $\xi$ of the landscape (17), and locating the values of the driving force $f$ for which the two phases – organized (O) and disorganized (D) – have the same free energy. The coexistence values of $\xi$ decreases with increasing $\xi$ 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 of possible distace 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. & Baraba, A.-L. Flavor network and the kinetic differences and mutational analysis. Proc. Natl. Acad. Sci. USA 103, 4935–4939 (2006).
5. Saakian, D. B., Muñoz, E., Hsu, C.-K. & Deem, M. W. Quasispecies theory for multiple-peak fitness landscapes. Phys. Rev. E 73, 041913 (2006).
6. Muñoz, E., Park, J.-M. & Deem, M. W. Quasispecies theory for horizontal gene transfer and recombination. Phys. Rev. E 78, 061921 (2008).
7. Park, J.-M., Muñoz, E. & Deem, M. W. Quasispecies theory for finite populations. Phys. Rev. E 81, 011902 (2010).
8. Hsu, H.-J. & Wallqvist, A. Nonquasispecies phase transitions associated with DNA replication. Phys. Rev. Lett. 106, 060601 (2011).
9. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends. Biochem. Sci. 35, 539–546 (2010).
10. Reimann, P. Brownian motors: noisy transport far from equilibrium. Phys. Rep. 361, 57–265 (2002).
11. Saakian, D. B., Muñoz, E., Hsu, C.-K. & Deem, M. W. Quasispecies theory for multiple-peak fitness landscapes. Phys. Rev. E 73, 041913 (2006).
12. Hsu, H.-J., Wallqvist, A. Nonquasispecies phase transitions associated with DNA replication. Phys. Rev. Lett. 106, 060601 (2011).
13. Joyce, G. F. The antiquity of RNA-based evolution. Nature 418, 214–222 (2002).
14. Klimek, P., Thurner, S. & Hanel, R. Evolutionary dynamics from a variational principle. Phys. Rev. E 82, 011901 (2010).
15. Itan, E. & Tannenbaum, E. Semiconservative quasispecies equations for polygenic genomes: the general case. Phys. Rev. E 81, 061915 (2010).
16. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188, 415–437 (1986).
17. Adami, C. Information theory in molecular biology. Phys. Life Rev. 1, 3–22 (2004).
18. Ruiz-Mirazo, K., Pereto, J. & Moreno, A. A universal definition of life: autonomy and open-ended evolution. Origins Life Evol. B 34, 323–346 (2004).

Author contributions
HJW. 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 carbonization. PLOS ONE 11, e014729 (2015).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/