Biomolecular information gained through *in vitro* evolution

Takuyo Aita · Yuzuru Husimi

Received: 26 April 2009 / Accepted: 22 October 2009 / Published online: 15 December 2009
© International Union for Pure and Applied Biophysics (IUPAB) and Springer 2009

Abstract An *in vitro* evolution is a simplified Darwinian evolution in well-controlled surroundings. This evolution process can be modeled as a hill-climbing or adaptive walk on a fitness landscape in sequence space. The evolving molecular system gains at least two kinds of information originating from the converged sequences and the fitness increment of the evolving biopolymer as the adaptive walker. These two represent two aspects of the biomolecular information, its extent and its content, respectively. Here, we review studies related to formulation of the “content” and “extent” of biomolecular information. The two aspects are interconnected through physicochemical properties of the biopolymer, contrary to the case of conventional information, which seems to be independent of matter. The interconnection was analyzed based on the analogy between the evolution process and thermodynamics. The linear combination of the two by a temperature-like fluctuation factor resulted in a free-energy-like monotonically increasing function during the evolution process.

Keywords Biological information · Fitness landscape · Free fitness · *In vitro* evolution · Pragmatic information · Quasi-species · Sequence space

Introduction

In vitro evolution is an artificial molecular evolution that is conducted in a laboratory and driven by the Darwinian

evolution mechanism. Study in this field began with Spiegelman’s experiment in 1967 (Mills et al. 1967) and has been further developed in both science and engineering, that is, the quest for the principle of emergence of functional biopolymers and practical applications to industry and medicine, e.g., evolutionary (or adaptive) drug design.

Here, we introduce several important terms to comprehend *in vitro* evolution theoretically. A quantitative measure of a molecular phenotype, that is, a certain physicochemical property (such as enzymatic activity or affinity to a ligand molecule or replication rate constant) of an evolving molecule is designated as the fitness. The conceptual space of all conceivable base or amino acid sequences (=genotypes) is designated as the sequence space (Maynard-Smith 1970). Each of the conceivable sequences is mapped onto their corresponding points in the sequence space. A distance between two arbitrary points is measured with the Hamming distance between the two corresponding sequences. The scalar field constructed by plotting the fitness value of each sequence into the corresponding point in the sequence space is designated as the fitness landscape, which is regarded as the evolutionary attribute of the biopolymer (Eigen and Schuster 1979; Kauffman 1993).

In an *in vitro* molecular evolution process, we have two parts: the *evolving system*, which consists of phenotype/genotype molecules as evolving entities and the *surroundings* (or environments) as an experimental setup. For example, we consider a typical case where we try to create specific peptides with high binding affinity to a given receptor protein through *in vitro* evolution (Fig. 1). The peptides correspond to the evolving system, and the shape of the receptor molecular surface including distribution of the electric charge on it and other experimental conditions of the solvent correspond to the surroundings. In this case, the fitness should be defined as the logarithm of the association constant between the peptide and the receptor, $\ln K_a$. Using

T. Aita (✉)
Graduate School of Science and Engineering, Saitama University,
Saitama 338-8570, Japan
e-mail: taita@mail.saitama-u.ac.jp

Y. Husimi
Innovative Research Organization, Saitama University,
Saitama 338-8570, Japan

Fig. 1 Emergence of a specific peptide with a high affinity to a receptor molecule, by absorbing the fitness information $\Delta\mathcal{W}/T_E$ from the surroundings, where T_E represents the evolutionary temperature and T_T represents the thermodynamical temperature. This is a case of $M=1$

the logarithm of the association constant means that the fitness is treated in the free-energy level. We must set and control the surroundings properly for correct evaluation of the fitness. Darwinian evolution is considered as an information gaining process from the surroundings. In order to simplify the information gaining process and treat it physically, we focused on the *in vitro* molecular evolution in a well-controlled environment as an extremely simplified process of biological evolution.

In this simplified system, an evolution process of a biopolymer is considered as an “adaptive walk or hill-climbing” on the corresponding fitness landscape in the sequence space (Fig. 2 right). Here, the evolving molecular system gains at least two kinds of information originating from the converged sequences and the fitness increment of the evolving biopolymer as the adaptive walker. These represent two aspects of the biomolecular information, extent and content, respectively. The two are interconnected through physicochemical properties of the biopolymer. The interconnection may be analyzed based on the analogy between the evolution process and thermodynamics. In fact, the right side of Fig. 2 is analogous to the conceptual view that protein folding is considered to be related to the downhill walk on the energy landscape in the conformation space (Wolynes and Luthey-Schulten 1997) (Fig. 2 left). Therefore, there have been many studies on the analogy between evolution and thermodynamics.

Extending the interpretation of evolution by thermodynamics-like concepts, we can clarify a view of the evolution process as an information gaining process from the surroundings. Eigen raised a question about the *extent* and *content* of

information in biological evolution (Eigen 2000). According to him, the extent of information is related to the constrained volume of the sequence space and can be treated within the classical information theory (Shannon and Weaver 1949). On the other hand, the content of information means the meaning or semantic value of information.

In this review, we summarize two typical studies on the evolutionary dynamics of *in vitro* evolution, which has been interpreted in terms of thermodynamics-like concepts leading to both aspects of the information concept, its extent and its content. One is a theory for a natural selection type model. That is the quasi-species theory developed by Eigen’s group (Eigen and Schuster 1979; Eigen 2000; Weinberger 2002). The other is a theory of an artificial selection type model, which was developed by the authors (Aita et al. 2004, 2005, 2007; Aita and Husimi 2006). These studies gave a formulation for the information gaining process from the surroundings. Other studies are also reviewed within the framework of the two typical studies.

Natural selection type model of *in vitro* molecular evolution

We can classify models of *in vitro* molecular evolution into two types: natural selection type and artificial selection type. In the natural selection type, the fitness is the specific growth rate of evolving molecules. An example of this is self-replicating RNA molecules in a flow reactor. In the artificial selection type, the fitness is one of the physicochemical properties of evolving molecules, e.g., binding

Fig. 2 An analogy between protein folding and protein evolution. In protein folding, a folding polypeptide descends the energy landscape by emitting the thermal entropy $-\Delta H/T (= \Delta S_{\text{sur}})$ to the surroundings. In protein evolution, we interpret that the evolving sequences climb the fitness landscape by absorbing fitness information $\Delta\mathcal{W}/T (= \Delta I_{\text{fit}})$ as the negative entropy from the surroundings, which is defined in an experimental setup. Adapted from Aita and Husimi (2006)

free energy to a target receptor, and the selection process is conducted through a cycle of evaluation and amplification by the experimenter. Therefore, *in vitro* evolution of the artificial selection type is also called directed evolution.

Eigen and Schuster proposed the quasi-species model as the natural selection type model. This model describes the evolutionary dynamics of the ensemble of simple self-replicators in a flow reactor, such as self-replicating RNA molecules (Eigen and Schuster 1979; Eigen 2000). The genome sequence of the self-replicators consists of ν sites, and λ letters are available at each site, that is, there are $\lambda^\nu (\equiv n)$ possible sequences. Then, each genome sequence is mapped onto the corresponding point in the λ -valued ν -dimensional sequence space. In the replication processes, it is possible to replace each letter with one of the other $\lambda-1$ letters with a probability of the mutation rate μ . The mole fraction of a certain species s at time t , denoted by $x_s(t)$, obeys the following differential equation:

$$\frac{dx_s(t)}{dt} = (m_{ss}f_s - D(t))x_s(t) + \sum_{u \neq s} m_{su}f_u x_u(t) \quad (1)$$

($s, u = 1, 2, \dots, n$)

$$m_{su} \equiv (1 - \mu)^{\nu - d(s,u)} \left(\frac{\mu}{\lambda - 1} \right)^{d(s,u)}, \quad (2)$$

where f_s is the replication rate or fitness of s , and m_{su} is a probability of mutation from s to u . The $d(s, u)$ is the Hamming distance between s and u . $D(t)$ is the dilution rate and works to satisfy $\sum_{s=1}^n x_s(t) = 1$ and is given by the mean replication rate $D(t) = \sum_{s=1}^n f_s x_s(t)$. Equations 1 and 2 mean that any species propagates cooperatively with neighbor species in the sequence space through mutation. It should be noted that the quasi-species model assumes an infinite population and can be applied to any model of fitness landscapes.

Artificial selection type model of *in vitro* molecular evolution

We considered evolutionary dynamics of the following artificial selection type model (Aita et al. 2004, 2005). M parents (parent sequences) produce N offspring (mutant sequences), and the fitness value of each offspring is evaluated. Subsequently, the best M individuals among the N offspring become new parents in the next generation. N is the library size of mutants to be screened in a single generation. In the reproduction process, d -fold point mutations occur randomly, that is, d represents the Hamming distance between a parent and each of its offspring. This iterative process is called the adaptive walk or hill-climbing,

and the parents are regarded as the walkers or climbers with the step length of d on the landscape. The parameters M , N , and d are constant throughout the walking process.

The evolutionary dynamics of a finite population is dependent on local structures of fitness landscapes. We supposed the following NK landscape (Kauffman 1993) in the λ -valued ν -dimensional sequence space (Aita et al. 2005, 2007)¹. In this model, an arbitrary site in a sequence interacts with other k sites. The fitness W for a given sequence $A_1 A_2 \dots A_\nu$ is defined by

$$W = \sum_{j=1}^{\nu} w_j(A_j | A_{j_1}, A_{j_2}, \dots, A_{j_k}), \quad (3)$$

where w_j is the *site fitness*, i.e., a fitness contribution from the j th site, and A_j represents a particular letter at the j th site. The value of w_j is given as a function of $1+k$ letters at the j th site (A_j) and other k sites ($A_{j_1}, \dots, A_{j_k}$). The interacting k sites $\{j_1, \dots, j_k\}$ are randomly chosen from among all of the $\nu-1$ sites except the j th site. Once a set of letters $\{A_{j_1}, \dots, A_{j_k}\}$ at these k sites is given, the value of w_j for an arbitrary letter A_j , $w_j(A_j | A_{j_1}, \dots, A_{j_k})$ is assigned randomly from a given probability distribution. Here, we adopt a discrete uniform distribution in the range $[-\varepsilon, \varepsilon]$, where ε is a positive constant ($\varepsilon > 0$). Let σ^2 be the variance for the uniform distribution, then $\sigma^2 = \varepsilon^2/3$ for this case. On the whole, the ruggedness of the landscape is controlled by the parameter k . In the case of $k=0$, the resulting fitness landscape has a smooth surface and a single global peak. As the k value increases, the surface of the fitness landscape becomes more rugged and many local optima appear.

Analogy between evolution and thermodynamics

Thermodynamic concepts connect the evolution process to information. Here, we review studies related to the analogy between evolution and thermodynamics from various viewpoints.

Iwasa stressed the importance of genetic drift rather than of mutation in a finite population and introduced the *free fitness* $I \equiv \text{mean fitness} + (1/2N) \times \text{population's entropy}$, where N is population size, and demonstrated that free fitness I always increases with time in the evolution process. This scheme indicates the analogy of free fitness to free energy, the mean fitness to enthalpy, and fluctuation parameter $1/2N$ to temperature (Iwasa 1988). The special case of $N=\infty$ corresponds to Fisher's fundamental theorem

¹ The validity of using the NK landscape was demonstrated in Aita et al. (2007)

of natural selection (Fisher 1930). Blackburne and Hirst conducted a simulation of population dynamics using simple lattice model proteins (Blackburne and Hirst 2005). They also estimated the population using the analogy with Boltzmann distribution in thermodynamics, in which a temperature-like parameter was empirically derived as a function of the mutation rate and selection pressure.

Sato et al. referred to a mathematical relationship between fluctuation $V[X]_a$ and response $\langle X \rangle_{a+\Delta a} - \langle X \rangle_a$ in a biological system (where $\langle X \rangle_a$ and $V[X]_a$ are the average and variance of the variable X at the initial parameter value a), and demonstrated that the relationship they found, $\langle X \rangle_{a+\Delta a} - \langle X \rangle_a \propto V[X]_a \times \Delta a$, is similar to Einstein's relation in the fluctuation-dissipation theorem in Brownian motion (Sato et al. 2003). They confirmed the relationship through an experimental observation in which X represents the logarithm of fluorescence intensity per *E. coli* cell including mutant GFP proteins and Δa represents the synonymous mutation rate of their genes. Ao presented the relationship between Darwinian evolution and thermodynamics from the viewpoint of Langevin dynamics (Ao 2008). His theory describes the dynamics on a potential surface in genetic frequency space, where each coordinate axis represents the frequency of a species. From the viewpoint of molecular imprinting, Pande et al. developed statistical mechanics of protein folding and design (Pande et al. 1997), introducing the design temperature T_{des} , which controls the probability of the occurrence of amino acid sequences with low energy (designed sequences) in a given canonical ensemble. They obtained a phase diagram for model heteropolymers in a two dimensional $T_{\text{des}}-T$ (T is thermodynamic temperature) space.

Quasi-species

We summarized the result of the mathematical analysis of the natural selection type model (Eq. 1). In a special case of $\mu=0$, the solution of Eq. 1 is easily obtained by introducing the variable $z_s(t) \equiv x_s(t) \exp(-\int_0^t D(t)dt)$. Let s^* be the fittest species that has a maximum replication rate among all the n species. In the stationary state, we can observe

$$\lim_{t \rightarrow \infty} x_s(t) = \begin{cases} 1, & \text{if } s = s^* \\ 0, & \text{if } s \neq s^*. \end{cases}$$

That is, only the fittest species s^* exists in the reactor.

In general cases of $\mu > 0$, Eq. 1 must be transformed in the following manner. Considering the n -dimensional matrix $[m_{su}f_u]$, we denote the q th eigenvalue and eigenvector of this matrix by Λ_q and $\mathbf{r}_q = (r_{q1}, r_{q2}, \dots, r_{qn})$ (for $q = 1, 2, \dots, n$), respectively. By diagonalizing the matrix $[m_{su}f_u]$ by the unitary matrix $U \equiv (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ and introducing

${}^t(y_1(t), y_2(t), \dots, y_n(t)) \equiv U^{-1} {}^t(x_1(t), x_2(t), \dots, x_n(t))$, Eq. 1 is transformed to

$$\frac{dy_q(t)}{dt} = (\Lambda_q - D(t))y_q(t) \quad (q = 1, 2, \dots, n), \quad (4)$$

where q is designated as the quasi-species. The dilution rate is rewritten by $D(t) = \sum_{q=1}^n \Lambda_q y_q(t)$. The solution of Eq. 4 is easily obtained in the same manner as mentioned above. Let q^* be the quasi-species that has a maximum eigenvalue (according to Frobenius theorem, $\Lambda_{q^*} > 0$ and $r_{q^*i} > 0$ ($q = 1, 2, \dots, n$)). In the stationary state, we can observe

$$\lim_{t \rightarrow \infty} y_q(t) = \begin{cases} 1, & \text{if } q = q^* \\ 0, & \text{if } q \neq q^* \end{cases}$$

That is, only the quasi-species q^* is realized in the reactor. The $x_s(t)$ in the stationary state is given by

$$\lim_{t \rightarrow \infty} x_s(t) = \lim_{t \rightarrow \infty} \sum_{q=1}^n r_{qs} y_q(t) = r_{q^*s}.$$

The realized eigenvector $\mathbf{r}_{q^*}$ is designated as the quasi-species distribution. The quasi-species distribution is caused by the mutation-selection balance.

It is important to note that the quasi-species distribution is strongly dependent on the shape of the fitness landscape ($\{f_s\}$) and mutation rate (μ). For proper landscapes, the quasi-species distribution shows a phase transition at several critical points of the mutation rate. When the mutation rate μ exceeds a certain critical point called the error threshold,

$$\mu_{cr} = \frac{\ln f_m - \ln \langle f_s \rangle_{s \neq m}}{\nu} \quad (\text{for } \lambda = 2), \quad (5)$$

an error catastrophe or a localization-delocalization transition occurs. In the delocalization state, all species have the identical mole fraction. In Eq. 5, the species m is the master sequence and $\langle f_s \rangle_{s \neq m}$ is the mean fitness over all the species except m .

An example of a fitness landscape demonstrating a sharp localization-localization transition is that of asymmetric twin peaks consisting of a sharp high peak and a broad low peak. When μ is very small, the quasi-species members localize at the high peak. When μ becomes greater than some critical value (and less than the error threshold), the quasi-species members become localized at the broad low peak because the population at the broad peak is mutationally robust based on mutational interconnectedness. The transition is very narrow for μ and shows a critical slowing down phenomenon (Husimi 1988; Schuster and Swetina 1988). Wilke et al. called this situation the survival of the flattest (Wilke et al. 2001).

These results can be interpreted by thermodynamics-like concepts. Mutation causes the species to diffuse in the sequence space, while selection causes them to converge on

the local area. Therefore, the mutation rate μ corresponds to a temperature-like parameter T . When $\mu=0$, the fittest s^* that has a maximum fitness f_{s^*} is realized. This is analogous to the case of $T=0$ in thermodynamics because the thermodynamic system realizes the minimum energy state. On the other hand, when $\mu>0$, the quasi-species q^* that has a maximum eigenvalue Λ_{q^*} is realized. This is analogous to the case of the thermodynamic system realizing the minimum free energy state. In the above mentioned asymmetric twin peak case, the localization at the broad low peak is an analogous state with an intermediate conformation X of a protein in the unfolding process (native $\leftrightarrow X \leftrightarrow$ denatured). Therefore, the eigenvalue Λ_q could be called the *free fitness* (Husimi 1988).

In thermodynamics, the phase transition temperature between state A and state B is given by $T_{cr} = \Delta H / \Delta S$, where ΔH and ΔS are the enthalpy change and entropy change between A and B, respectively. In Eq. 5, the numerator represents energy- or enthalpy-like quantity, and the denominator represents entropy-like quantity ($\nu = \log_2 2^\nu - \log_2 1$). Therefore, Eq. 5 is analogous to $T_{cr} = \Delta H / \Delta S$.

Free fitness

We summarized the result of the statistical analysis of the artificial selection type model. Denoting the mean fitness over the M parents by $\mathcal{W}$ for every generation, we focus on the statistical properties of a time course of $\mathcal{W}$ through the hill-climbing process. Consider that the hill-climbing starts from the foot of the landscape ($\mathcal{W} = 0$). The mean fitness $\mathcal{W}$ increases exponentially and tends toward a stationary value denoted by $\mathcal{W}^*$. In the stationary state, the value of $\mathcal{W}$ fluctuates around the attractor $\mathcal{W}^*$. As a result, under extreme conditions where λ , ν , d , and N have large values², the attractor $\mathcal{W}^*$ is given by:

$$\mathcal{W}^* = 2\sigma\nu\sqrt{\frac{\ln(N/M)}{d(1+k)}}, \quad (6)$$

where $\ln(N/M)$ is the selection pressure because we select the best M individuals from among N offspring, and $d(1+k)$ is the expected number of the affected sites by random d -fold point mutations because a single point mutation changes the site-fitness values of on average $1+k$ sites.

This stationary state is caused by the mutation-selection-random drift balance.

In order to interpret the evolutionary dynamics mentioned above, we introduce the following thermodynamics-like quantities. The analogy between the concepts in evolution and those in thermodynamics is compiled in Table 1. We denote the fitness coordinate by W . The frequency distribution of fitness over all conceivable sequences (of λ^ν) is given approximately by the following normal distribution:

$$\Omega(W) = \frac{\lambda^\nu}{\sqrt{2\pi\mathcal{V}}} \exp\left(-\frac{W^2}{2\mathcal{V}}\right) \quad \text{for } -H \leq W \leq H, \quad (7)$$

where $H \equiv \varepsilon\nu$ and $\mathcal{V} = \sigma^2\nu$. The average of fitness over the whole sequence space corresponds to the foot of the landscape, while regions where $W<0$ correspond to below sea level and are negligible for the adaptive walks that start from random points, which are likely to be located at the foot of the landscape. Since the fitness at the global peak takes about $H (= \varepsilon\nu)$, H corresponds to the height of the landscape from the foot to the global peak. In this paper, we focus on the regions from the foot of the landscape to the global peak: $0 \leq W \leq H$.

We define the *landscape constant* k_L and the *entropy* S as a function of W as follows:

$$k_L \equiv \sigma\sqrt{1+k} \quad (8)$$

$$S(W) \equiv k_L \times \ln \Omega(W), \quad (9)$$

$$= -\frac{k_L W^2}{2\mathcal{V}} + \text{const.} \quad (10)$$

Following the definition of thermodynamic temperature, we define the *evolutionary temperature* T as follows:

$$T \equiv -\left(\frac{dS}{dW}\bigg|_{W=\mathcal{W}^*}\right)^{-1}, \quad (11)$$

where $\mathcal{W}^*$ is a stationary value of mean fitness $\mathcal{W}$ for the M walkers. Using Eqs. 6 and 10, T is given by

$$T = \sqrt{\frac{d}{4\ln(N/M)}}. \quad (12)$$

In Eq. 8, the $\sigma\sqrt{1+k}$ means the standard deviation of a change in fitness for a unit Hamming distance. That is, the landscape constant k_L indicates the degree of ruggedness of the landscape. In Eq. 12, the d indicates the degree of diffusion in the sequence space by random mutation, while $\ln(N/M)$ indicates the degree of convergence of sequence

² The conditions are $\ln(N/M) \ll d(1+k) \ll \nu$ and $M \ll N \ll \binom{\nu}{d}(\lambda-1)^d$, where $\binom{\nu}{d}(\lambda-1)^d$ is the size of the d -boundary of any sequence. The derivation of Eq. 6 is described in Aita et al. (2007).

Table 1 An analogy between thermodynamics and *in vitro* evolution

Thermodynamics	Evolution
Number of moles: M	Number of walkers: M
Molar energy or enthalpy: E_m, H_m	Mean fitness of walkers: $\mathcal{W}$
Boltzmann constant: k_B	Landscape constant: $k_L \equiv \sigma\sqrt{1+k}$
Molar entropy: $S_m \equiv k_B \ln \Omega$	Walker's entropy: $S \equiv k_L \ln \Omega$
Temperature: $T \equiv \partial E_m / \partial S_m$	Evolutionary temperature: $T \equiv -dW/dS$ $T = \sqrt{d/4 \ln(N/M)}$
Chemical potential: $G_m \equiv H_m - T S_m$	Evolutionary potential: $\phi \equiv \mathcal{W} + T S$
Free energy: $G \equiv G_m \times M$	Free fitness: $G \equiv \phi \times M$
Generalized force: $X \equiv d(G_m/T)/dx$	Evolutionary force: $X \equiv d(\phi/T)/d\mathcal{W}$
Generalized flux: J	Expectation of fitness change: $J \equiv \mathbf{E}[\Delta\mathcal{W}]$
Linear transport equation: $J \approx L X$	Linear relationship between J and X : $J \approx L X$
Diffusion coefficient: D	Diffusion coefficient: D
Einstein relation: $L = MD/k_B$	$L = MD/k_L$
Energy-fluctuation formula: $\mathbf{SD}[E_m] \propto k_B T / \sqrt{M}$	Fitness fluctuation: $\mathbf{SD}[\Delta\mathcal{W}] = k_L T \times 4/(\sqrt{M} + 1)$
$\Delta S_{\text{system}}/M = \Delta S_m$	Shannon information gain : $\Delta I_{\text{Sha}} = -\Delta S$
$\Delta S_{\text{surroundings}}/M = -\Delta H_m/T$	Fitness information gain: $\Delta I_{\text{fit}} = \Delta\mathcal{W}/T$
$\Delta S_{\text{universe}}/M = -\Delta G_m/T$	Biomolecular information gain: $\Delta I_{\text{bio}} = \Delta\phi/T$

diversity by selection. T is the ratio of these conflicting effects. By using these parameters, the stationary value $\mathcal{W}^*$ given by Eq. 6 is rewritten by:

$$\mathcal{W}^* = \frac{\mathcal{V}}{k_L T}. \quad (13)$$

According to thermodynamics of protein folding, the most probable energy of a canonical ensemble is given by an equation similar to Eq. 13 (e.g., Eq. 11 in Wolynes and Luthey-Schulten 1997). Therefore, we can say that k_L and T

are analogous to the Boltzmann constant and thermodynamic temperature of the thermal bath, respectively.

Here, we define the *evolutionary potential* ϕ and *free fitness* G for the M walkers with mean fitness $\mathcal{W}$ as follows:

$$\phi(\mathcal{W}) \equiv \mathcal{W} + T \times S(\mathcal{W}) \quad (14)$$

$$G(\mathcal{W}) \equiv \phi(\mathcal{W}) \times M. \quad (15)$$

Using Eq. 10, ϕ is rewritten as the following convex function of $\mathcal{W}$:

$$\phi(\mathcal{W}) = -\frac{k_L T}{2\mathcal{V}} \left(\mathcal{W} - \frac{\mathcal{V}}{k_L T} \right)^2 + \text{const.} \quad (16)$$

$$= -\frac{k_L T}{2\mathcal{V}} (\mathcal{W} - \mathcal{W}^*)^2 + \text{const.} \quad (17)$$

Equation 17 is derived by substituting Eq. 13 into Eq. 16. We can see that the evolutionary potential ϕ and free fitness G take the maximum values at $\mathcal{W} = \mathcal{W}^*$. Furthermore, ϕ and G are Lyapunov functions of the evolution process (proof is shown in the next section).

In addition to these quantities, we define the *evolutionary force* X by

$$X(\mathcal{W}) \equiv \frac{d}{d\mathcal{W}} \frac{\phi(\mathcal{W})}{T} \quad (18)$$

$$= -\frac{k_L}{\mathcal{V}} (\mathcal{W} - \mathcal{W}^*). \quad (19)$$

Substituting Eq. 14 into Eq. 18, the evolutionary force X is decomposed into the *fitness force* X_{fit} and the *entropy force* X_{ent} : $X = X_{\text{fit}} + X_{\text{ent}}$, where

$$X_{\text{fit}} \equiv \frac{1}{T} \quad \text{and} \quad X_{\text{ent}} \equiv \frac{dS(\mathcal{W})}{d\mathcal{W}} = -\frac{k_L}{\mathcal{V}} \mathcal{W}. \quad (20)$$

X_{fit} is caused by a selection event and pushes the walkers upward, while X_{ent} is caused by a mutation event and pushes the walkers downward. The mutation-selection-random drift balance occurs when X_{fit} and X_{ent} cancel each other out.

Based on the definition we mentioned above, we can describe the dynamics of the adaptive walk as follows (see Fig. 3). Driven by the evolutionary force X , the walkers tend to achieve the maximum free fitness state (or maximum evolutionary potential state) in which the fitness force X_{fit} and the entropy force X_{ent} cancel each other out. The evolutionary force X depends strongly on the evolutionary temperature T . Here, consider the walkers are

Fig. 3 Interpretation of evolutionary dynamics by thermodynamics-like concepts. The *solid* and *dotted* lines represent the entropy S and free fitness G , respectively, as a function of fitness. The *right* and *left dotted* lines are for the cases of low and high evolutionary temperature T , respectively. $\mathcal{W}^*$ represents the stationary point, in which G takes the maximum under a given T . The vectors (arrows) represent the evolutionary force X that acts on the walkers with the mean fitness $\mathcal{W}$. Details are described in the text. Adapted from Aita and Husimi (2006)

located at the middle point on the landscape. If $T=\infty$ (which is the case where $N/M \approx 1$), then $\mathcal{W}^*$ lies at the foot of the landscape and a negative force ($X < 0$) acts on the walkers so the walkers are pushed downward. In this case, the maximum entropy state is realized. As T becomes lower (which is the case where $N/M \gg 1$), the $\mathcal{W}^*$ becomes higher up near the top of the landscape and a positive force ($X > 0$) acts on the walkers so the walkers are pushed upward. In this case, the (nearly) maximum fitness state is realized. If $T=0$ (which is the case of $d=0$), the walkers cannot move in the sequence space, and evolution does not occur.

Fitness fluctuation

We denote the change in mean fitness $\mathcal{W}$ after a single generation by $\Delta\mathcal{W}$ (top of Fig. 4). Since the exploration is done by random sampling of N mutants from among the underlying mutant population, the fitness change $\Delta\mathcal{W}$ is a stochastic quantity and is described by the theory of extreme value statistics. Let J and Σ be the expectation and standard deviation of $\Delta\mathcal{W}$, respectively. These are described in the following based on Aita et al. (2004).

The expectation J ($\equiv \mathbf{E}[\Delta\mathcal{W}]$) is described as follows:

$$J \approx LX, \quad L = \frac{MD}{k_L}, \quad (21)$$

where L is the linear transport coefficient or mobility. D ($\equiv k_L^2 d/M$) represents the diffusion coefficient for the mean fitness $\mathcal{W}$ along the fitness coordinate when walkers

Fig. 4 Upper A probability density of the mean-fitness change after a single step walk (=generation) from the mean fitness $\mathcal{W}$. J and Σ represent the expectation and standard deviation of the change. Middle The fitness-information change, ΔI_{fit} , is the fitness change digitized by the fluctuation size, $k_L T = \Sigma (\times (\sqrt{M} + 1)/4)$. Bottom The concept of ΔI_{fit} is analogous to the thermal entropy, $\Delta S_{\text{sur}} = -\Delta H/T$, which is the enthalpy change digitized by the fluctuation size of thermal energy per degree of freedom, $k_B T$. Adapted from Aita and Husimi (2006)

perform a random walk in the sequence space. The random walk occurs when $N = M$ because there is no selection pressure ($X_{\text{fit}} = 0$). Equation 21 is analogous to Einstein's relation in Brownian motion (Einstein 1905). Sato et al. found a similar scheme to Eq. 21 for the relationship between fluctuation and response in a biological system (Sato et al. 2003). Iguchi has tried to extend Eq. 21 to a case of coevolving biopolymers in a framework similar to Onsager's reciprocal relations (Iguchi 2008).

The standard deviation Σ ($\equiv \text{SD}[\Delta \mathcal{W}]$) is described as follows:

$$\Sigma \approx k_L T \times \frac{4}{\sqrt{M} + 1}. \quad (22)$$

Σ is the degree of fluctuation of the change in mean fitness $\mathcal{W}$ after a single generation, while $k_L T$ is analogous to the thermal fluctuation energy per degree of freedom $k_B T$ (k_B is the Boltzmann constant and T is temperature). Therefore, Eq. 22 is analogous to the energy-fluctuation formula in thermo-statistical mechanics.

Information gained through *in vitro* evolution

Biomolecular information

Suppose an adaptive walk from the foot of the landscape ($\mathcal{W} = 0$) up to the stationary state ($\mathcal{W} = \mathcal{W}^*$). Here, using entropy S in Eq. 9 and evolutionary potential ϕ in

Eq. 14, we introduce the following three quantities as functions of mean fitness $\mathcal{W}$:

$$I_{\text{fit}}(\mathcal{W}) \equiv \frac{\mathcal{W}}{T} \quad (23)$$

$$I_{\text{Sha}}(\mathcal{W}) \equiv S(0) - S(\mathcal{W}) \quad (24)$$

$$I_{\text{bio}}(\mathcal{W}) \equiv \frac{\phi(\mathcal{W})}{T} - \frac{\phi(0)}{T} \quad (25)$$

$$= I_{\text{fit}}(\mathcal{W}) - I_{\text{Sha}}(\mathcal{W}), \quad (26)$$

and let ΔI_{fit} , ΔI_{Sha} , and ΔI_{bio} be the changes in I_{fit} , I_{Sha} , and I_{bio} , respectively, after a single generation. Since Eq. 24 describes the change in entropy S between the initial state and a certain state (Fig. 5), I_{Sha} is interpreted as the Shannon information (Shannon and Weaver 1949).

On the other hand, from Eq. 22, the meaning of $\Delta I_{\text{fit}} (= \Delta \mathcal{W}/T)$ is as follows:

$$\frac{\Delta I_{\text{fit}}}{k_L} = \frac{\Delta \mathcal{W}}{k_L T} \approx \frac{\Delta \mathcal{W}}{\Sigma} \times \frac{4}{\sqrt{M} + 1}. \quad (27)$$

Fig. 5 Schematic representation of the fitness information I_{fit} and Shannon information I_{Sha} . The cone represents a fitness landscape schematically. The height represents the fitness W divided by T , while the area of the cross section represents the entropy $S(W) \equiv k_L \times \ln \Omega(W)$, where $\Omega(W)$ is the number of all sequences with a given fitness value W . $I_{\text{fit}}(W)$ is defined as the change in W/T from the foot of the landscape, while $I_{\text{Sha}}(W)$ is defined as the negative change in S from the foot. ΔI_{fit} and ΔI_{Sha} are the changes in these quantities after a single step walk (=generation)

We can see that ΔI_{fit} is the fitness change digitized by the fitness fluctuation size Σ (in the middle of Fig. 4). Here, the analog to digital conversion is realized as the significant figures of the fitness change observed by the adaptive walkers with the observation error Σ . This is analogous to the thermal entropy change $-\Delta H/T$, that is the enthalpy change digitized by the fluctuation size of thermal energy $k_B T$, when a system emits the heat $-\Delta H$ to the surroundings (bottom of Fig. 4) (Atkins 1978). According to the analogy with thermodynamics, we can interpret $\Delta I_{\text{fit}} (= \Delta \mathcal{W}/T)$ as the negative entropy that the evolving system absorbs from the surroundings (Fig. 2 right). Here, the surroundings mean the experimental setup (e.g., a column of affinity-chromatography) around the biopolymer as an evolving entity (Fig. 1). We designate I_{fit} as the *fitness information*. We can say that the evolving entity gains the fitness information from the surroundings (Figs. 1 and 2).

The expectation of ΔI_{bio} is given by

$$\mathbf{E}[\Delta I_{\text{bio}}] = JX \approx LX^2 \geq 0. \quad (28)$$

Equation 28 proves the theorem that I_{bio} (or $\phi(\mathcal{W})$ or $G(\mathcal{W})$) is a Lyapunov function of the evolution process. Therefore, we conclude that the evolution is driven in the direction in which I_{bio} increases, and then we designate I_{bio} as the *biomolecular information* in *in vitro* evolution.

In the stationary state ($\mathcal{W} = \mathcal{W}^*$), I_{fit} , I_{Sha} , and I_{bio} satisfy the following simple relation:

$$I_{\text{bio}}(\mathcal{W}^*) = I_{\text{Sha}}(\mathcal{W}^*) = \frac{I_{\text{fit}}(\mathcal{W}^*)}{2} \quad (29)$$

$$= \frac{\mathcal{V}}{2k_L T^2} \approx t^* \times k_L \ln(N/M), \quad (30)$$

where t^* represents the mean generation (or mean step number) up to the stationary state and is defined as follows: $\mathbf{E}[\mathcal{W}_{t^*}] \equiv (1 - e^{-2})\mathcal{W}^*$, where $\mathbf{E}[\mathcal{W}_t]$ is the expectation of the mean fitness $\mathcal{W}_t$ at the t th generation. The t^* in this model is given approximately by $t^* \approx 2\nu/d(1+k)$. Since in each selection step, the best M mutants are selected from among the N mutants, the term $k_L \ln(N/M)$ on the right-hand side corresponds to the *operational* information gain by selection. Therefore, the result shown in Eq. 30 is reasonable in that I_{bio} in a stationary state is approximately equivalent to the sum of the operational information gain over generations up to the stationary state: $t^* \times k_L \ln(N/M)$.

Extent and content of information

According to Eigen (2000), the extent of information is related to the constrained volume of the sequence space. Therefore, we define explicitly the extent of information in *in vitro* evolution in the following. Let p_s be the probability (or frequency) of occurrence of sequence s in a population. Entropy for this state is given by

$$S = - \sum_{s=1}^{\lambda^\nu} p_s \log p_s, \quad (31)$$

while the maximum entropy is given by $S_{\text{max}} = \nu \log \lambda$, which is for the case where every sequence in the sequence space occurs with the same probability of $p_s = 1/\lambda^\nu$. Particularly, S_{max} is called the *potential (or source) entropy*. The extent of information is defined as

$$I_{\text{extent}} \equiv S_{\text{max}} - S. \quad (32)$$

Several concepts of the *genomic complexity* (Adami et al. 2000), R_{sequence} (Schneider 2000; Kim et al. 2003), *functional information* (Szostak 2003), *functional sequence complexity* (Durston et al. 2007), and I_{Sha} (Eq. 24) can be classified to I_{extent} . For example, Szostak introduced the functional information, which is defined as $-\log_2$ of the fraction of functional sequences that have fitness values (activity of a biopolymer) greater than a specified value (Szostak 2003). For example, suppose all possible RNA sequences of 4^{70} and the fraction of functional sequences among them is 10^{-11} , the functional information in this case is 37 bit compared with 140 bit to specify a unique 70-mer sequence. Szostak suggested the importance of the activity-functional information relationship, which affects the evolvability of the biopolymer.

On the other hand, the content of information means the meaning or semantic value of information. One may say that the concept of fitness corresponds to the content of information in biological evolution. However, the fitness cannot be treated within the same level of Shannon information. Therefore, it is desirable to introduce a novel measure for the content of information.

Pragmatic information

As an answer to the issue of the content of information, Weinberger proposed *pragmatic information*, which quantifies the impact of a message on the receiver's subsequent actions (Weinberger 2002). His theory is based on a communication system consisting of a *decision maker* and an *effector* (Fig. 6) and the assumption that the semantic value of information stems from its usefulness in making an informed decision.

Fig. 6 Conceptual framework for pragmatic information. Adapted from Weinberger (2002)

Suppose that the decision maker, in some current state $\mathcal{S}$, receives a set of $\mathcal{M}$ messages ($m=1,2,3, \dots, \mathcal{M}$) and chooses a message m among the set with the probability of φ_m . Subsequently, based on the practical meaning of the chosen message m , the effector selects an outcome o from among a set of $\mathcal{N}$ outcomes ($o=1,2,3, \dots, \mathcal{N}$). The probability that an outcome o is selected without any message is given by q_o , and the conditional probability that an outcome o is selected when message m is given is $P_{o|m}$. The pragmatic information of the message ensemble is defined by

$$I_{\text{pra}} \equiv \sum_{m=1}^{\mathcal{M}} \varphi_m \sum_{o=1}^{\mathcal{N}} P_{o|m} \log_2 \left(\frac{P_{o|m}}{q_o} \right). \quad (33)$$

Then, the pragmatic information is the average of the relative entropy between $\{P_{o|m}\}$ and $\{q_o\}$ over the message ensemble.

Weinberger demonstrated that the pragmatic information is a global Lyapunov function for the quasi-species model. Here, the setup of the flow reactor effectively decides on the fitness of phenotype corresponding to each given genotype, where a phenotype's fitness is defined to be its reproduction rate. At each time t , the flow reactor receives *messages* about the fitness of a particular replicator via the number of copies of that replicator's genome. Prior to receipt of the messages, the initial probability of selecting a species s at random from the system is $q_s = x_s(0)$. The probability of selecting a species s at subsequent time t is $P_{s|m} = x_s(t)$, in which the process's state at various times is the only *message* ($\mathcal{M}=1$) received, that is, $\varphi_1=1$. Then, the pragmatic information for the quasi-species model is given by

$$I_{\text{pra}} = \sum_{s=1}^n x_s(t) \log_2 \left(\frac{x_s(t)}{x_s(0)} \right). \quad (34)$$

Regardless of arbitrary initial distribution, $\{x_s(0)\}$, $dI_{\text{pra}}/dt > 0$ holds for all finite times. We can say that the pragmatic

information is generated through the process of evolution for the quasi-species and answer Eigen's call for a value parameter for the level of evolution.

For the natural selection type model (quasi-species model) and artificial selection type model, the Lyapunov functions are the pragmatic information I_{pra} (Eq. 34) and the biomolecular information I_{bio} (Eq. 25), respectively. We have not yet demonstrated the relationship between them. In addition, these concepts should be related with thermodynamic entropy production in the real evolving system and the surroundings (Smith 2008a, 2008b, 2008c), in a similar way as the problem of Maxwell's demon (Szilard 1929; Brillouin 1956).

Fitness information

We provide further discussion about the concept of the fitness information I_{fit} defined in Eq. 23. The evolving entity (=biopolymer) gains from the surroundings a content of information for adaptation and existence under given conditions. We interpret that the fitness information I_{fit} corresponds to the content of information. For example, concerning the emergence of a specific peptide with a high affinity to a receptor molecule (see Fig. 1), as mentioned in the "Introduction" section, the emerging peptides gain I_{fit} from the receptor under experimental conditions. In this case, fitness should be defined as a natural logarithm of the association constant between the peptide and the receptor: $W \equiv \ln K_a$. The value of the content of information is quantified by dividing a fitness change, $\Delta \ln K_a$, by evolutionary temperature T , or in other words, by scaling $\Delta \ln K_a$ by the accuracy of observation by the adaptive walkers, that is, the fluctuation of the fitness change after a single generation, $\text{SD}[\Delta \ln K_a]$. On the other hand, I_{Sha} represents the extent of information. The biomolecular information in *in vitro* evolution, I_{bio} , consists of the content and the extent of information (see Fig. 5).

Acknowledgments We thank Dr. Kazumoto Iguchi for helpful discussions and providing information.

References

- *Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. *Proc Natl Acad Sci USA* 97:4463–4468
- Aita T, Husimi Y (2006) An interpretation of evolutionary dynamics in in vitro molecular evolution from thermodynamical and informational viewpoint (in Japanese with English abstract). *Seibutsu Butsuri (Biophysics)* 46:137–143
- **Aita T, Morinaga S, Husimi Y (2004) Thermodynamical interpretation of evolutionary dynamics on a fitness landscape in an evolution reactor, I. *Bull Math Biol* 66:1371–1403
- *Aita T, Morinaga S, Husimi Y (2005) Thermodynamical interpretation of evolutionary dynamics on a fitness landscape in an evolution reactor, II. *Bull Math Biol* 66:1371–1403
- Aita T, Hayashi Y, Toyota H, Husimi Y, Urabe I, Yomo T (2007) Extracting characteristic properties of fitness landscape from in vitro molecular evolution: a case study on infectivity of fd phage to *E.coli*. *J Theor Biol* 246:538–550
- Ao P (2008) Emerging of stochastic dynamical equalities and steady state thermodynamics from Darwinian dynamics. *Commun Theor Phys* 49:1073–1090
- Atkins PW (1978) *Physical chemistry*. Oxford University, New York
- Brillouin L (1956) *Science and information theory*. Academic Press, New York
- *Blackburne BP, Hirst JD (2005) Population dynamics simulations of functional model proteins. *J Chem Phys* 123:154907
- Durston KK, Chiu DK, Abel DL, Trevors JT (2007) Measuring the functional sequence complexity of proteins. *Theor Biol Med Model* 4:47
- *Eigen M (2000) Natural selection: a phase transition? *Biophys Chem* 85:101–123
- Eigen M, Schuster P (1979) *The hypercycle*. Springer, Berlin
- Einstein A (1905) On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat (in German). *Ann D Phys* 17:549–560
- Fisher RA (1930) *The genetical theory of natural selection*. Clarendon, Oxford
- Husimi Y (1988) Selective value landscape on the base sequence space and concept of free selective value: a model of quasi-species (in Japanese with English abstract). *Viva Origino* 16:136–141
- Iwasa Y (1988) Free fitness that always increases in evolution. *J Theor Biol* 135:265–281
- Iguchi K (2008) Reciprocal relations in evolutionary processes. *Prog Theor Phys Suppl* 173:235–242
- Kauffman SA (1993) *The origin of order*. Oxford University Press, Oxford
- Kim JT, Martinetz T, Polani D (2003) Bioinformatic principles underlying the information content of transcription factor binding sites. *J Theor Biol* 220:529–544
- Maynard-Smith J (1970) Natural selection and the concept of a protein space. *Nature* 225:563–564
- Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc Natl Acad Sci USA* 58:217–224
- Pande VS, Grosberg AY, Tanaka T (1997) Statistical mechanics of simple models of protein folding and design. *Biophys J* 73:3192–3210
- **Sato K, Ito Y, Yomo T, Kaneko K (2003) On the relation between fluctuation and response in biological systems. *Proc Natl Acad Sci USA* 100:14086–14090
- Schneider TD (2000) Evolution of biological information. *Nucleic Acid Res* 28:2794–2799
- Schuster P, Swetina J (1988) Stationary mutant distributions and evolutionary optimization. *Bull Math Biol* 50:635–660
- Shannon CE, Weaver W (1949) *The mathematical theory of communication*. University of Illinois Press, Champaign
- Smith E (2008a) Thermodynamics of natural selection I: energy flow and the limits on organization. *J Theor Biol* 252:185–197
- Smith E (2008b) Thermodynamics of natural selection II: chemical Carnot cycles. *J Theor Biol* 252:198–212
- Smith E (2008c) Thermodynamics of natural selection III: Landauer's principle in computation and chemistry. *J Theor Biol* 252:213–220
- Szilard L (1929) Über die Entropieverminderung in einem thermodynamischen System bei eingriffen intelligenter Wesen. *Z Physik* 53:840–856
- *Szostak JW (2003) Functional information: molecular messages. *Nature* 423:689
- **Weinberger ED (2002) A theory of pragmatic information and its application to the quasi-species model of biological evolution. *Biosystems* 66:105–119
- Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412:331–333
- Wolynes PG, Luthey-Schulten Z (1997) The energy landscape theory of protein folding. In: Flyvbjerg H, Hertz J, Jensen MH, Mouritsen OG, Sneppen K (eds) *Physics of biological systems: from molecules to species*. Springer, New York, pp 61–79

Note that we limited the selected references described here to those published after the year 2000, although there are many important references published before that year.

*Aita et al. 2004. Based on additive fitness landscapes, the evolutionary dynamics of the artificial selection type model is analyzed and interpreted by showing an analogy to thermodynamics. This interpretation is extended to quantify information gain from the surroundings (environments) in an evolution process.

**Eigen 2000. A question about the extent and content of information in biological evolution is raised. The quasi-species theory is summarized, and the evolutionary dynamics including phase transitions of the quasi-species is discussed by showing an analogy to thermodynamics and from the viewpoint of the extent and content of information.

*Sato et al. 2003. A mathematical relationship between fluctuation and response in a biological system is described by showing an analogy to the fluctuation-dissipation theorem in physics. An application to an experimental observation is demonstrated.

**Weinberger 2002. As an answer to Eigen's call for a value parameter, that is a content of information, for biological evolution, a theory of pragmatic information is described. The proof of the pragmatic information being a global Lyapunov function for the quasi-species model is given.

*Adami et al. 2000. A simulation study of the evolution of digital organisms. For analysis of the evolutionary dynamics, the concept of genomic complexity, which is a quantity increasing through the evolution process, is proposed.

*Aita et al. 2005. The thermodynamics-like concepts and information-theoretical concepts proposed in Aita et al. 2004 are extended to Kauffman's NK fitness landscapes.

*Blackburne and Hirst 2005. A simulation study of population dynamics of lattice model proteins. The population is estimated using the analogy to thermodynamics.

*Szostak 2003. Based on a question as to how we can define and quantify the information content of biopolymer sequences, Szostak introduced the functional information by a concise description.