Journal of Theoretical Biology 266 (2010) 584–594

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Evolutionary dynamics, epistatic interactions, and biological information

Christopher C. Strelioff [a,b,*], Richard E. Lenski [a], Charles Ofria [b]

[a] Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA
[b] Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

ARTICLE INFO

Article history:
Received 22 December 2009
Received in revised form
7 July 2010
Accepted 21 July 2010
Available online 24 July 2010

Keywords:
Information theory
Fitness landscapes
Quasispecies
Epistasis
Modularity

ABSTRACT

We investigate a definition of biological information that connects population genetics with the tools of information theory by focusing on the distribution of genotypes found in a population. Previous research has treated loci as non-interacting by making specific approximations in the calculation of information-theoretic quantities. We expand earlier mathematical forms to include epistasis, or interactions between mutations at all pairs of loci. Application of our improved measure of biological information to evolution on two-locus, two-allele fitness landscapes demonstrates that mutual information between loci reflects epistatic interaction of mutations. Finally, we consider four-locus, two-allele fitness landscapes with modular structure. As modular interactions are inherently epistatic, we demonstrate that our refined approximation provides insight into the underlying structure of these non-trivial fitness landscapes.

1. Introduction

Information and evolutionary theories describe seemingly unrelated methods for the acquisition, storage and transmission of information. On the one hand, information theory is focused on compression of data and the effective transmission of messages from sender to receiver over a noisy communication channel (Shannon and Weaver, 1949). On the other hand, evolutionary theory describes the acquisition, storage and transmission of genetic information governed by the biological processes of replication, heritable variation, and natural selection. A relation between these theories can be formed by considering the changing distribution of genotypes caused by evolutionary processes. Information-theoretic quantities such as entropy and mutual information that quantify properties of, and relations between, probability distributions provide the desired connection.

A natural measure of biological information exploits this association by considering the difference between an observed distribution of genotypes and a uniform distribution. The Kullback–Leibler (KL) divergence, also known as the relative entropy, quantifies the coding inefficiency created by assuming a particular distribution when the true distribution is different (Cover and Thomas, 1991). In the mathematical model of an infinite, asexual population to be considered here, a uniform distribution of genotypes is presumed to be representative of long-term evolution in an environment without selective pressure, heterogeneous mutation rates, or other sources of non-uniformity. The proposed measure of biological information quantifies the inefficiency of assuming a distribution of genotypes conforms to these assumptions of uniformity.

Although the application of information-theoretic methods to this problem is a recent development, it is interesting to note that this general viewpoint of evolutionary dynamics is not new. A well-known quote attributed to R.A. Fisher captures the spirit of the calculations in this paper: "natural selection is a mechanism for generating an exceedingly high degree of improbability" (Huxley, 1958). Kimura (1961) also noted the above quote in his paper on genetic information, where he employed the idea of a substitutional load rather than employing the tools of information theory. The KL divergence between an observed distribution of genotypes and a uniform distribution uses information-theoretic tools to quantify this "degree of improbability" in an intuitive way.

Previous research on biological information has employed a mathematical form that can be viewed as an approximation of the proposed KL divergence (Schneider, 1997, 2000; Adami et al., 2000; Adami and Cerf, 2000). Alternatively called biological information, biological complexity, or physical complexity, this research usually employed a simplifying assumption of no interaction between mutations. Notable exceptions considered correlations between sites in aligned DNA and RNA sequences using mutual information (Gutell et al., 1992; Adami, 2004; Bindewald et al., 2006). In the present work, we unify these

* Corresponding author at: Microbiology and Molecular Genetics, c/o Lenski Lab, Michigan State University, East Lansing, MI 48824, USA.
E-mail addresses: streliof@msu.edu (C.C. Strelioff), lenski@msu.edu (R.E. Lenski), ofria@msu.edu (C. Ofria).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.07.025

approaches by using the KL divergence as a foundation for all approximations of biological information. Application of the proposed measures on an evolving population demonstrates that mutual information is needed to capture the effect of epistatic interactions.

The need for an improved measure of biological information is motivated by a wide array of examples from experiment and theory that demonstrate (i) the existence of epistasis and (ii) the dramatic effects that mutational interactions can have on evolutionary dynamics. In fact, evolution is almost trivial if there are no epistatic interactions because each locus can be optimized independently. There is considerable evidence of epistasis provided by experiments in organisms including *Aspergillus niger*, *Escherichia coli* and *Drosophila melanogaster* (de Visser et al., 1997; Elena and Lenski, 1997; Whitlock and Bourguet, 2000). There is also evidence for epistasis in RNA viruses such as vesicular stomatitus virus and human immunodeficiency virus 1 (Michalakis and Roze, 2004; Bonhoeffer et al., 2004; Sanjuán et al., 2004). In addition to the important biological examples, mathematical theory and *in silico* evolution using self-replicating and evolving computer programs demonstrate that interacting mutations are fundamental to understanding evolution in rugged fitness landscapes (Jain and Krug, 2007) and the evolution of complex computational tasks (Lenski et al., 1999, 2003).

2. Models of evolution and biological information

Next, we consider the elements needed to define and explore an improved approximation of biological information in detail. First, in Section 2.1 we introduce simple fitness landscapes for two-locus, two-allele genotypes. In addition, we define epistasis in a quantitative way and provide an illustrative set of landscapes. In Section 2.2 we introduce the discrete-time quasispecies equation. This mathematical model describes mutation and selection in an infinite population. As a result, the quasispecies equation provides an ideal framework for developing a theory of biological information without concern for statistical fluctuations in finite populations. In Section 2.3 we define biological information in terms of the KL divergence and connect the genotype frequencies from the quasispecies equation to information-theoretic quantities of interest. Most important, we expand on the currently available approximations to include both single-locus information and mutual information between loci, corresponding to non-interacting and interacting mutations, respectively.

2.1. Fitness landscapes and epistasis

We first consider a population of asexual, haploid organisms evolving on a static two-locus, two-allele fitness landscape. We denote a genome as σ and the ith locus as σ_i. The first and second loci have possible alleles $\mathcal{A}_1 = \{a,A\}$ and $\mathcal{A}_2 = \{b,B\}$, respectively, resulting in the following set of possible genotypes:

$$\Lambda = \{ab,aB,Ab,AB\}. \tag{1}$$

A fitness landscape is specified by the fitness values, w, assigned to each of the possible genotypes. We use $\sigma = ab$ as the reference genome in our description of the fitness landscape and assign $w(ab)=1$, which corresponds to equal birth and death rates for the discrete-time quasispecies equation. Following Hartl and Clark (2007), the relative fitness of another genotype σ is given by a selective value s_σ:

$$\frac{w(\sigma)}{w(ab)} = 1+s_\sigma. \tag{2}$$

Assignment of $w(ab)=1$ for the reference genotype also means that $w(\sigma) = 1+s_\sigma$. Using these ideas, a complete fitness landscape for two-locus, two-allele genotypes can be described by three selective values: $\{s_{aB},s_{Ab},s_{AB}\}$. We note that these selective values can be positive, negative, or equal to zero and correspond to beneficial, deleterious, and neutral mutations, respectively.

Given a set of selective values, we must be able to quantify the nature of epistatic interactions in the resulting fitness landscape. Following Bonhoeffer et al. (2004), we define epistasis relative to the reference genotype:

$$E = \ln\frac{w(ab)w(AB)}{w(aB)w(Ab)} \tag{3a}$$

$$= \ln\frac{1+s_{AB}}{(1+s_{aB})(1+s_{Ab})}. \tag{3b}$$

We employ natural logarithms in this definition to maintain a connection between Malthusian fitness, m, and Darwinian fitness, $w : m = \ln w$ (Orr, 2009). Non-interacting mutations have $E=0$ whereas epistatic interactions between mutations can be positive or negative. These labels correspond to pairs of mutations that are more fit (positive epistasis) or less fit (negative epistasis) than expected from the individual effects of the contributing mutations.

Given the definitions of fitness landscapes and epistasis, we can now consider some examples. In Table 1 we present a set of seven landscapes that have a variety of interesting properties. Our focus is the dynamics of the genotype frequencies and biological information, starting from a population consisting of only the reference genotype ab, and ending at the asymptotic distribution of genotypes for that landscape.

To start, we have chosen three fitness landscapes (*FL*) without epistatic interactions to provide a baseline understanding of the dynamics. In *FL* 1, all selection coefficients are zero, resulting in neutral evolution. For *FL* 2, a mutation to the second locus is beneficial. Finally, *FL* 3 has beneficial mutations at both loci, where the combination of these mutations is exactly what is expected when there is no interaction between them.

Next, we constructed four fitness landscapes with epistatic interactions. The first two examples, *FL* 4 and *FL* 5, are modifications of the single peak landscape *FL* 3. For *FL* 4, the combined value of the two mutations is less than expected from their separate effects. In *FL* 5, the selection coefficient of the double mutant is larger than the corresponding value without epistasis. These fitness landscapes provide examples of negative and positive epistasis, respectively.

Table 1
Two-locus, two-allele fitness landscapes.

	Selection coefficients			E
	s_{Ab}	s_{aB}	s_{AB}	
Simple fitness landscapes				
FL 1 Neutral	0.0	0.0	0.0	0.0
FL 2 Beneficial locus	0.0	0.03	0.03	0.0
FL 3 Single peak	0.1	0.03	0.133	0.0
Epistatic fitness landscapes				
FL 4 Negative epistasis	0.1	0.03	0.1	−0.03
FL 5 Positive epistasis	0.1	0.03	0.2	+0.06
FL 6 Fitness valley	−0.1	−0.03	0.1	+0.23
FL 7 Polymorphic	0.03	0.03	0.0	−0.06

The first three fitness landscapes *FL* 1–*FL* 3 are simple and have no interacting mutations. *FL* 4–*FL* 7 are more complex landscapes that have epistatic interactions between mutations. Selection coefficients and epistasis are defined relative to the reference genome $\sigma = ab$, as described in Eqs. (2) and (3), respectively.

Finally, we include two examples that might seem redundant because they are further examples of positive and negative epistasis. However, fitness landscapes such as *FL* 6 are often discussed in the context of sign epistasis, where individually deleterious mutations combine to produce an overall benefit. This example can also be seen as a fitness valley, where a population must pass through a deleterious intermediate state to reach a higher peak in the fitness landscape. Our final example, *FL* 7, can be called polymorphic because genotypes *Ab* and *aB* coexist at equal frequency in the long-term, at least given an infinite population size. This behavior is of particular interest because we might also expect to observe this type of sustained coexistence in changing fitness landscapes. Although we will not consider dynamic fitness landscapes here, it is instructive to discuss an example with a sustained genetic polymorphism.

2.2. Discrete-time quasispecies equation

Our model of evolutionary dynamics is the discrete-time formulation of the quasispecies equation. Originally developed by Eigen (1971) and Eigen and Schuster (1977) to describe molecular evolution, this mathematical model is now used as a general tool for the description of mutation–selection dynamics (Nowak, 1992; Wilke, 2005).

The quasispecies equation describes the frequency of each genotype σ at generation t, $X(\sigma,t)$. Changes in these frequencies occur through selection and mutation processes. Selection is reflected in the fitness values assigned to each genotype. The rate of mutation from one genotype to another is given by

$$M(\sigma,\hat{\sigma}) = \left(\frac{\mu}{|\mathcal{A}|-1}\right)^{d(\sigma,\hat{\sigma})}(1-\mu)^{L-d(\sigma,\hat{\sigma})}. \tag{4}$$

In the above equation μ is the per-locus mutation rate, $|\mathcal{A}|$ is the number of alleles per locus, L is the number of loci, and $d(\sigma,\hat{\sigma})$ is the Hamming distance between genotypes σ and $\hat{\sigma}$. For the two-locus, two-allele examples discussed above we have $|\mathcal{A}|=2$ and $L=2$. Combining these elements, the discrete-time version of the quasispecies equation is given by

$$X(\sigma,t+1) = \sum_{\hat{\sigma}\in\Lambda}\frac{M(\sigma,\hat{\sigma})w(\hat{\sigma})}{W(t)}X(\hat{\sigma},t) \quad \forall\sigma\in\Lambda, \tag{5a}$$

$$W(t) = \sum_{\hat{\sigma}\in\Lambda} w(\hat{\sigma})X(\hat{\sigma},t), \tag{5b}$$

where $W(t)$ is the average fitness of the population at time t.

In many papers that employ the quasispecies equation, the focus is on the asymptotic distribution of genotypes. In that case, the form of Eq. (5) can be linearized and the asymptotic behavior is given by the eigenvector corresponding to the largest eigenvalue for the combined mutation–selection matrix (Wilke, 2005; Jain and Krug, 2007). However, our focus includes the transient dynamics that occur during evolution from the reference genotype to the asymptotic distribution. As a result, we iterate Eq. (5) to obtain the desired genotype distribution at each generation.

2.3. Biological information

Biological information is calculated using the KL divergence between a uniform distribution, $\Pr(\sigma)=|\mathcal{A}|^{-L}$, and an observed distribution of genotypes, $X(\sigma,t)$. In an effort to keep the discussion compact, technical details of the derivation and definitions of basic information-theoretic quantities are shown in the Appendix. In this section we provide only a brief overview of the derivation and present the resulting form.

Using the definition of the KL divergence provided in Eq. (C.1) and the assumptions introduced above, the most general form for the biological information is

$$I(\sigma,t) = L\log_2|\mathcal{A}| - H(\sigma,t), \tag{6}$$

where $H(\sigma,t)$ is the entropy of the observed genotype distribution, defined in Eq. (B.1).

In practice it is difficult to estimate accurately the entropy of the observed genotype distribution if the length of the genome is large or the population size of interest is small. As a result, previous work has approximated the observed entropy by considering each locus individually and adding the contributions (Schneider, 2000; Adami et al., 2000; Huang et al., 2004). This choice results in the first approximation of the biological information:

$$I_1(\sigma,t) = \sum_{i=1}^{L} I(\sigma_i,t), \tag{7}$$

which ignores all epistatic interactions between loci. The single-locus information employed in this approximation is defined as

$$I(\sigma_i,t) = \log_2|\mathcal{A}| - H(\sigma_i,t), \tag{8}$$

where $H(\sigma_i,t)$ is the Shannon entropy for the observed marginal distribution at locus σ_i, defined in Eq. (B.2a).

We extend the approximation of biological information provided in Eq. (7) to include the effect of epistatic interaction between pairs of mutations. This refinement produces

$$I_2(\sigma,t) = I_1(\sigma,t) + \sum_{i=1}^{L}\sum_{j>i}^{L} I(\sigma_i:\sigma_j,t), \tag{9}$$

where $I(\sigma_i:\sigma_j,t)$ is the mutual information between locus σ_i and locus σ_j, as defined in Eq. (B.3). To be clear, the double sum in Eq. (9) is over all unique pairs of loci. The new mutual information terms measure correlations between the distributions of alleles at different loci. As a result, we expect these terms to be nonzero when epistatic interactions are present.

3. Results

In this section we apply Eqs. (7) and (9) as approximations of the biological information. First, we consider fitness landscapes *FL* 1–*FL* 3, which are non-epistatic. $I_1(\sigma,t)$ accurately captures the properties of these simple landscapes because there is no mutual information between loci. Next, we consider epistatic fitness landscapes *FL* 4–*FL* 7. Interaction between mutations results in mutual information between loci and requires the new approximation $I_2(\sigma,t)$ to reflect accurately the biological information. Finally, we consider four-locus, two-allele fitness landscapes where both approximations are inexact. In all cases, we plot the true biological information given by Eq. (6), the approximations in Eqs. (7) and (9), and the single-locus information provided in Eq. (8). Mutual information terms, given in Eq. (B.3), are only shown for examples with epistasis. Consideration of the elements that make up these approximations provides additional insight into the features of the fitness landscape that create different types of biological information. For all examples, we employ a per-locus mutation rate of $\mu=10^{-3}$.

3.1. Simple fitness landscapes

We first consider *FL* 1, a fitness landscape with no selective advantage for any genotype. The panels in the left-most column of Fig. 1 show the frequencies of all genotypes, single-locus information dynamics and the biological information and its

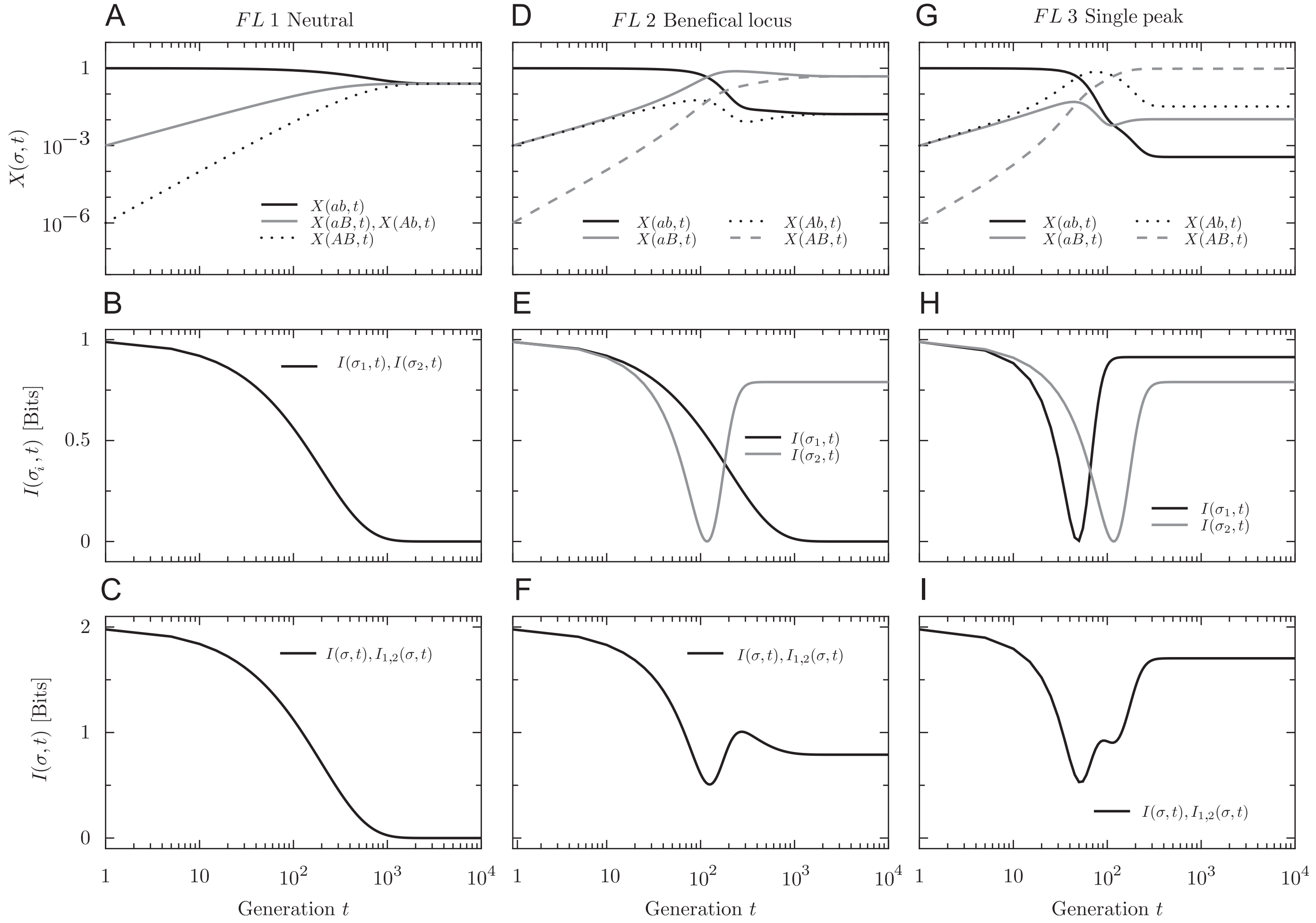

Fig. 1. Population dynamics and biological information on simple fitness landscapes. Dynamics of genotype frequencies and information measures are provided for fitness landscapes *FL*1–*FL* 3. The top row plots genotype frequencies, the middle row single-locus information, and the bottom row biological information and its approximations over time. The ordinate scales are provided on the far-left and are exactly the same for all panels in a row. The scale for the abscissa in all panels is provided at the bottom of each column. Labels at the top of each column provide the fitness landscape under consideration. Mutual information between loci for these examples is zero at all times (not shown in the figure).

approximations. In panel (A) the population, which initially consists of only $\sigma = ab$, equilibrates to equal probability for all genotypes. Single-locus information for both loci, shown in panel (B), starts at one bit, reflecting the initial population state, and decays to zero bits. The true biological information and both approximations start at two bits and end at zero bits (Fig. 1C). This illustration provides a baseline understanding of a fitness landscape without biological information.

Next we consider *FL* 2, a fitness landscape where mutation from allele *b* to *B* at the second locus provides a selective benefit. However, like our first example, mutation at the first locus is neutral and there are no interactions between mutations. The panels in the middle column of Fig. 1 show all quantities of interest. In panel (E), a new dynamic appears in the behavior of the single-locus information at locus two. Information decreases during the selective sweep of the new *B* allele because there are comparable frequencies of the competing alleles. However, panels (E) and (F) show that the long-term information value for the second locus recovers and reflects the selective value of allele *B*.

Our final non-epistatic example, *FL* 3, is a smooth fitness landscape where mutation at either locus provides a constant selective benefit, regardless of the genetic background. The dynamics of interest are shown in the right-most column of Fig. 1. For this landscape, the benefit of the *A* allele is greater than the *B* allele. As a result of this difference in selective value, $I(\sigma_1,t)$ dips first in panel (H) and rises to higher long-term value than $I(\sigma_2,t)$. The biological information, shown in panel (I), displays intricate dynamics that result from the sum of the single-locus information dynamics in panel (H).

3.2. Epistatic fitness landscapes

Our first example of an epistatic fitness landscape is *FL* 4, where the fitness of the double mutant is less than expected from the individual effects of the contributing mutations. The genotype frequencies, single-locus information, mutual information between loci, and total biological information for *FL* 4 are provided in the left column of Fig. 2. In panel (B), the single-locus information shows that the first locus has information in the long-term, whereas the second locus has none. This difference from the dynamics found in *FL* 3 reflects the equal fitness of genotypes *Ab* and *AB*. Panel (C) illustrates a typical pattern in the dynamics of mutual information between loci: the value peaks during a selective sweep involving epistatic interactions and settles to a more moderate value in the long-term. Although the contribution of the mutual information between loci to the total biological information is small in this example, the fact that this value is nonzero means that the new approximation, $I_2(\sigma,t)$, should be employed (Fig. 2D).

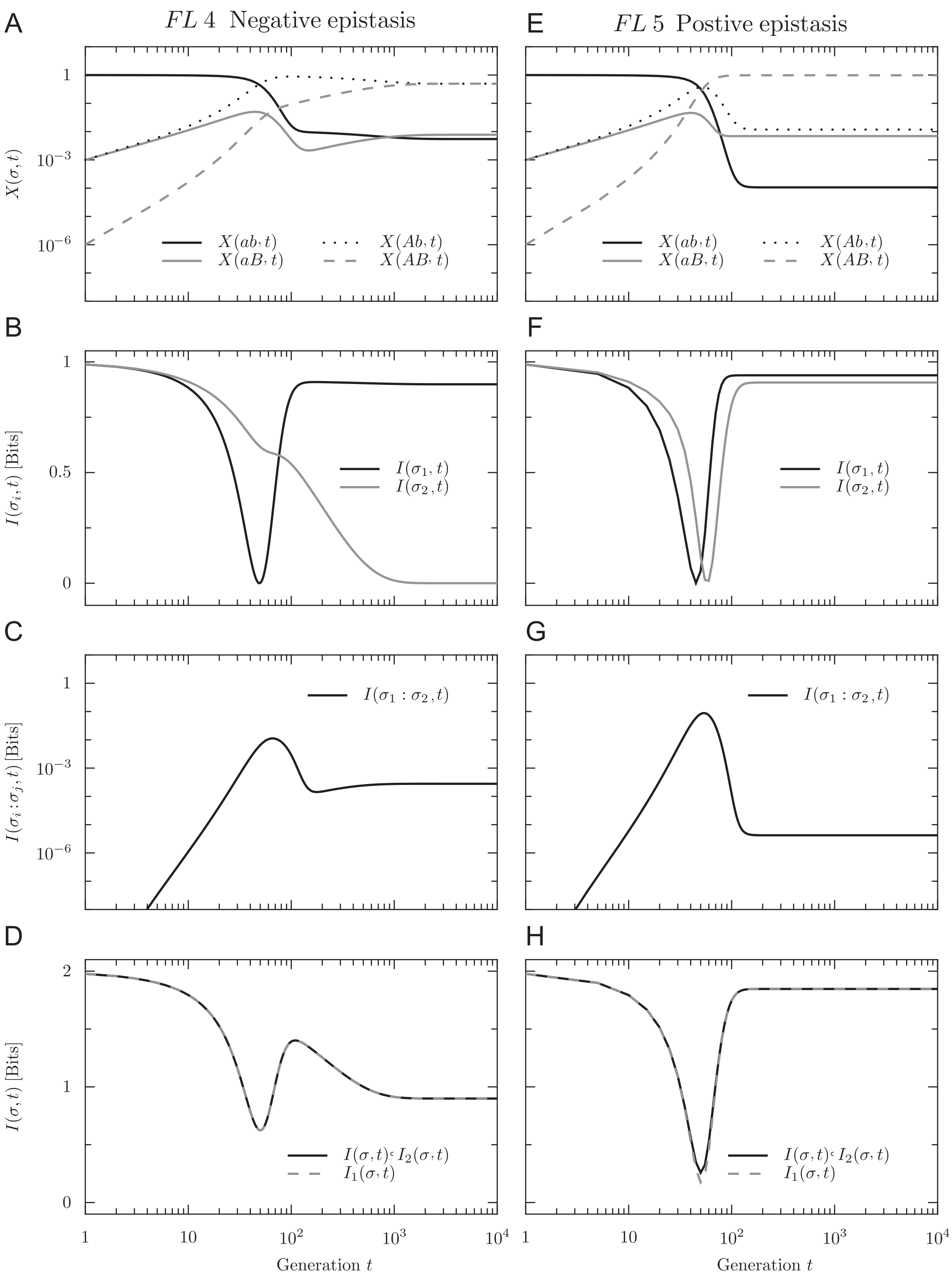

Fig. 2. Population dynamics and biological information on epistatic fitness landscapes I. Dynamics of genotype frequencies and information measures are provided for fitness landscapes *FL* 4 and *FL* 5. The top row plots genotype frequencies, the second row single-locus information, the third row mutual information, and the bottom row biological information and its approximations over time. The ordinate scales are provided on the far-left and are exactly the same for all panels in a row. The scale for the abscissa in all panels is provided at the bottom of each column. Labels at the top of each column provide the fitness landscape under consideration.

In fitness landscape *FL* 5, the double mutant is more fit than expected from the individual effects of substituting the *A* or *B* allele. Plots of the genotype frequencies, single-locus information, mutual information between loci, and complete biological information are provided in right column of Fig. 2. The single-locus information in panel (F) shows that the fitness effect of a mutation at the first locus is greater than for the second locus. This behavior is similar to the dynamics found for non-epistatic fitness landscape *FL* 3. However, unlike *FL* 3, panel (G) reveals the nonzero mutual information between loci that results from positive epistasis.

The next example, *FL* 6, provides an illustration of sign epistasis. Genotype *ab* is a local peak in the fitness landscape whereas the double mutant, *AB*, is the global maximum. Genotype frequencies, single-locus information, mutual information between loci, and complete biological information are plotted in the left column of Fig. 3. The genotype frequencies in panel (A) show a transition from the local peak to the global peak with little change in the "valley" genotype frequencies. The single-locus information in panel (B) shows that $I(\sigma_2,t)$ dips slightly before $I(\sigma_1,t)$ and rises to higher long-term value. This dynamic is a result of the fact that mutation at the second locus is less deleterious than at the first locus. Unlike previous examples, the magnitude of

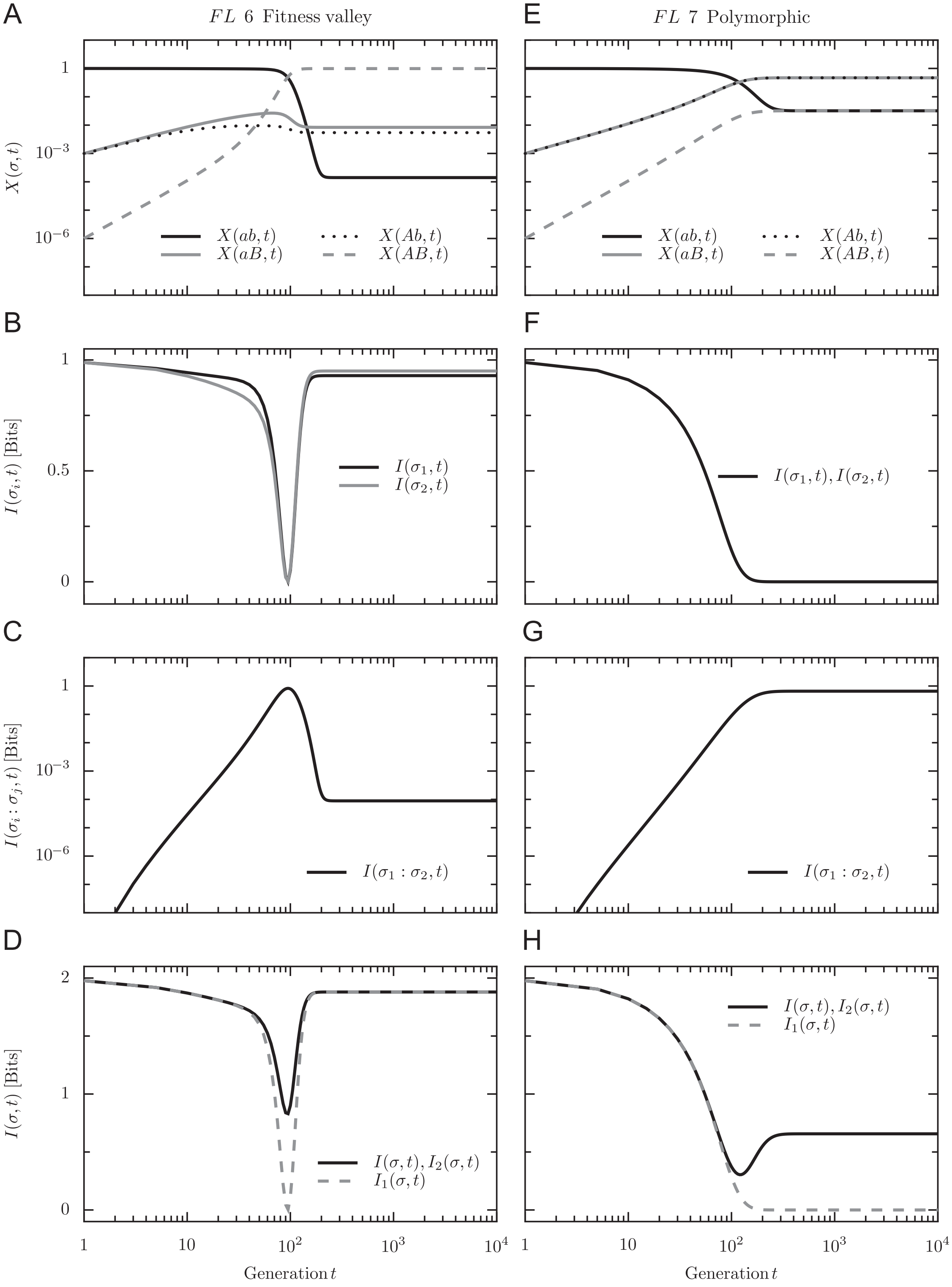

Fig. 3. Population dynamics and biological information on epistatic fitness landscapes II. Dynamics of genotype frequencies and information measures are provided for fitness landscapes *FL* 6 and *FL* 7. The top row plots genotype frequencies, the second row single-locus information, the third row mutual information, and the bottom row biological information and its approximations over time. The ordinate scales are provided on the far-left and are exactly the same for all panels in a row. The scale for the abscissa in all panels is provided at the bottom of each column. Labels at the top of each column provide the fitness landscape under consideration.

the mutual information, plotted in panel (C), approaches one bit. As a result of strong correlations, approximation $I_1(\sigma,t)$ dramatically underestimates the true biological information, shown in panel (D). Based on this example, we might expect a large burst of mutual information between loci involved in a move between fitness peaks.

The final two-locus, two-allele fitness landscape that we will consider is *FL* 7. In this example, single mutants with respect to *ab* have the same fitness, resulting in equal frequency of genotypes *Ab* and *aB* in the long-term. The genotypes frequencies, single-locus information, mutual information between loci, and complete biological information are plotted in the right column of Fig. 3. In panel (F) the single-locus information dynamics show a decay close to zero bits at each locus. This apparent information reduction results from the equal frequency of *aB* and *Ab*, creating a maximum entropy distribution at each locus. Panel (G) shows the mutual information between loci, which increases to approximately 0.7 bits in the long-term. All of the long-term information, shown in panel (H), comes in the form of correlations. As a result, the new approximation $I_2(\sigma,t)$ must be used to reflect the presence of information. The lack of single-locus information and the high level of mutual information between loci is a sign of a sustained genetic polymorphism.

3.3. Modular landscapes

Next, we consider a pair of four-locus, two-allele fitness landscapes. The motivation for these final illustrations is two-fold. First, these examples demonstrate that the behaviors observed in the two-locus examples translate to a longer genome in consistent ways. Second, the four-locus landscapes provide a more demanding application of the new approximation of the biological information. In the two-locus examples, the true biological information was exactly equal to one or both of the approximations, $I_1(\sigma,t)$ and $I_2(\sigma,t)$. However, in these four-locus fitness landscapes, the true biological information and the two-locus approximation are not generally equal.

We present the four-locus, two-allele fitness landscapes in Table 2. As with the two-locus examples, we employ upper and lower case letters for the possible alleles. At the first locus, possible alleles are $\mathcal{A}_1=\{a,A\}$, with the rest of the genome following a similar pattern: $\mathcal{A}_2=\{b,B\}$, $\mathcal{A}_3=\{c,C\}$ and $\mathcal{A}_4=\{d,D\}$. We assign $\sigma = abcd$ to be the reference genotype and will use it as the sole member of the starting population. As in Eq. (2), we assign $w(abcd)=1$ and describe the fitness landscape using selective values relative to the reference genome: $w(\sigma)=1+s_\sigma$.

Table 2
Modular four-locus, two-allele fitness landscapes.

σ	s_σ
Two-module landscape	
****	0.0
*AB***	0.2
ABCD	0.44
One-module landscape	
****	0.0
ABCD	0.44

Genotypes that match the allele pattern in the σ column have the selective advantage s_σ given in the next column (* is treated as a wild card). Similar to the two-locus landscapes, the selective values are relative to genotype $\sigma = abcd$, which has fitness one. The two-module fitness landscape has a hierarchical structure with two genetic units providing selective benefits of $s_{AB^{**}}=0.2$ and $s_{ABCD}=0.44$. The one-module landscape eliminates the benefits for AB^{**} genotypes, resulting in a single genetic unit consisting of all four loci.

The two-module fitness landscape requires that alleles *A* and *B* both be present before any fitness benefit is obtained. These loci make up the first genetic module. A further fitness benefit can be obtained by substituting alleles *C* and *D* at the third and fourth loci, resulting in a second module. However, the first module must still be in place to receive the fitness benefit for the second module. The one-module fitness landscape requires that all alleles be changed for any fitness benefit to accrue. By way of contrast, these examples demonstrate the advantages and limitations of our new approximation.

Fig. 4 provides an overview of the biological information and the various approximations for the four-locus fitness landscapes. Panels (A)–(C) correspond to the two-module landscape, and panels (D)–(F) correspond to the single-module landscape. In panel (A) of Fig. 4 we plot the true biological information as well as the two approximations of this quantity for the two-module example. A single dip in the biological information is apparent as the reference genotype is replaced by the superior genotypes AB^{**}, and then *ABCD*. Although both approximations of the biological information are quite accurate away from the selective sweep, the new approximation is more effective during this period of greater genetic diversity because mutual information becomes important. The fact that the new approximation is not equal to the true biological information demonstrates that there are non-trivial correlations involving three or four loci in this two-module example.

The details of the two-module example can be further analyzed by considering the factors that make up the $I_2(\sigma,t)$ approximation. In panel (B) we plot the single-locus information for all four loci, and in panel (C) we show the mutual information between pairs of loci. The values of $I(\sigma_1,t)$ and $I(\sigma_2,t)$ dip first, reflecting their importance in the first genetic module. At the same time, the mutual information between these loci peaks. As we saw in our two-locus examples, this pattern is typical of a sweep involving an epistatic interaction. A similar pattern in the single-locus and mutual information plays out for the third and fourth loci after the first module starts to become dominant in the population. This ordering of events reflects the need for the appropriate genetic background to be in place before the second module is beneficial.

In addition to the interaction of mutations at loci within the same genetic module, there are also effects that cross this boundary. Mutual information across the module barrier peaks between sweeps of the first and second modules. For example, $I(\sigma_1:\sigma_3,t)$ is high, but not as great as the mutual information between loci in the same module, reflecting interaction between modules. Also, single-locus information at the first and second loci is greater than at the third and fourth, further reflecting the genetic hierarchy. Combined, the elements that make up the new approximation of the biological information provide an improved if not quite complete picture of this non-trivial fitness landscape.

Finally, we turn to the one-module fitness landscape. This example is of particular interest because we can contrast the results with the two-module landscape discussed above. First, we consider biological information and its approximations in panel (D) of Fig. 4. As with the two-module example, both of the approximations are very accurate in the short- and long-term. However, the new approximation does not perform as well in this example as in the previous one. The reason for this is simple: strong epistatic interactions exist among more than two loci due to the single genetic module.

Differences between the one- and two-module fitness landscapes become more apparent when the components of the biological information approximations are considered. The single-locus information values for all loci in the single-module example dip at the same time and increase to the same long-term level

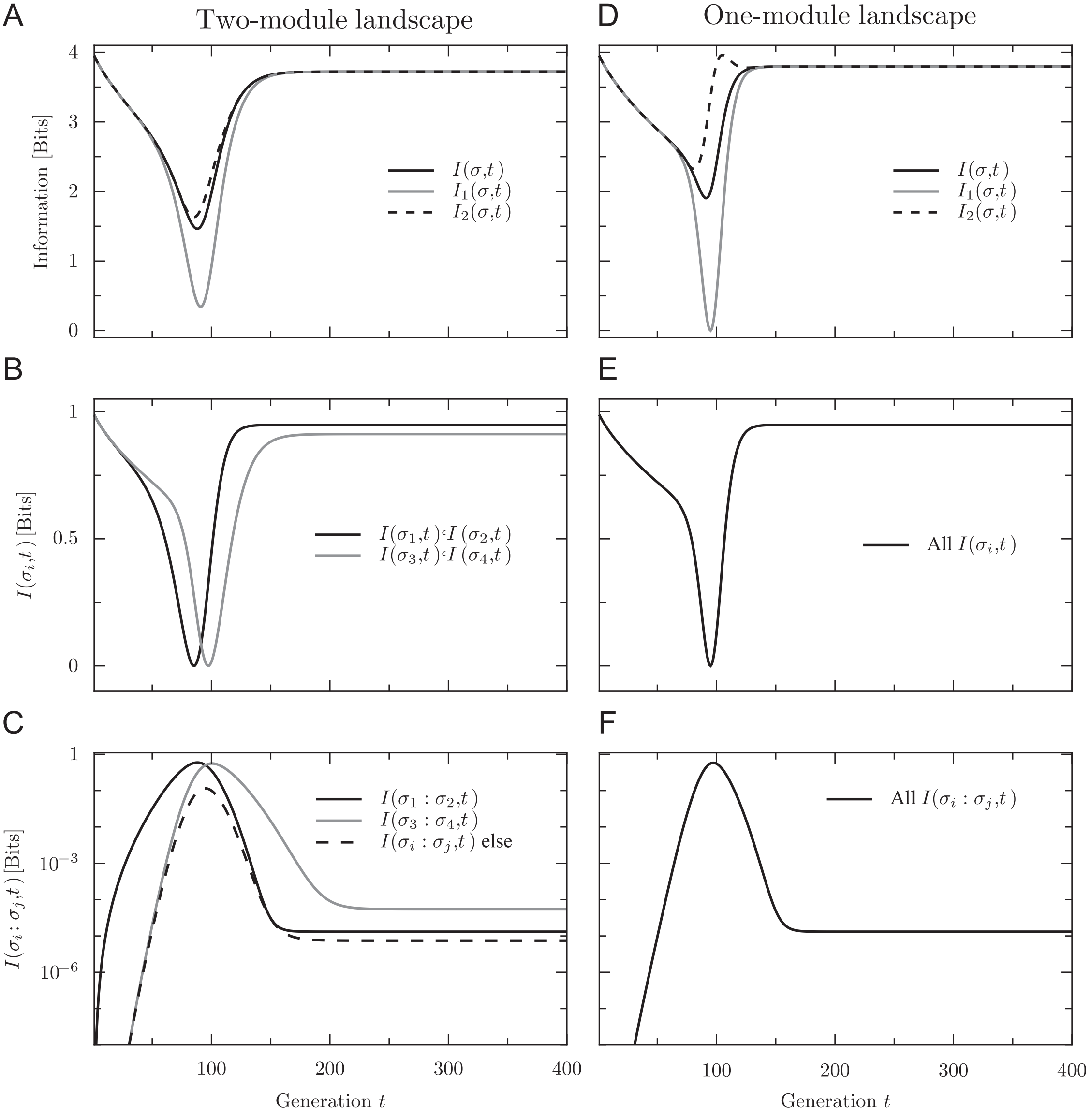

Fig. 4. Biological information on modular four-locus fitness landscapes. Dynamics of the biological information, single-locus information, and mutual information between loci is provided for the modular four-locus, two-allele fitness landscapes. The top row plots the biological information and its approximations, the middle row single-locus information, and the bottom row mutual information between loci over time. The ordinate scales for biological information, single-locus information, and mutual information are provided on the far-left and are exactly the same for all panels in the row. The scale for the abscissa in all panels is provided at the bottom of each column. Labels at the top of each column provide the fitness landscape under consideration.

(Fig. 4E). Mutual information between all pairs of loci follows the same pattern, peaking when the single-locus information dips (Fig. 4F). All information moves in concert because there is only the single module. The lack of subtleties in the timing and magnitude of the biological information dynamics when compared with the two-module fitness landscape in our previous example provides an informative contrast. This difference reflects the greater strength of the higher order correlations in the single-module fitness landscape when compared with the hierarchical nature of the two-module example.

4. Discussion

It is clear from the examples presented above that the temporal dynamics of biological information, computed solely from the distribution of genotypes, can provide insights into features of the underlying fitness landscape. Elements of the proposed approximations have natural interpretations as epistatic and non-epistatic contributions to biological information, given by mutual information and single-locus information terms, respectively. As a result, the addition of mutual information reveals epistatic interactions and modular structure that were not captured previously. As discussed in the Introduction, there are many examples of epistasis in biological and computational systems, and therefore this addition will help identify and quantify interactions between loci that contribute to evolutionary dynamics in many populations.

It is important to consider what biological information, as we have discussed it here, actually means. In principle, a single genome with properly encoded genes has information about its environment. However, information theory was constructed to consider properties of probability distributions. This requirement results in a choice between at least two competing ideas. The first approach is to take a single genome of interest and generate an artificial distribution of genotypes that reflects the local fitness landscape as closely as possible. An alternative is to use the observed population distribution that reflects the processes of mutation and selection as they happen. The former method attempts to assign information to a single genome, whereas the latter method attributes information to the population.

The first approach has been applied to self-replicating and evolving computer programs in the Avida platform, where single-locus information was considered by applying mutation–selection balance to a group of genotypes with all possible mutations at a single locus (Huang et al., 2004; Ofria et al., 2008). From the resulting distribution, single-locus information was estimated and an approximation of the biological information was obtained using Eq. (7). In principle, this approach for estimating information in a single genome should be fairly accurate when the chosen genome sits on a peak in the fitness landscape. As we saw in the examples with the four-locus genomes, approximations of the biological information were quite accurate far away from selective sweeps (i.e., at a local fitness peak). However, in fitness landscapes where there is a sustained genetic polymorphism, such as during a selective sweep, this approach provides misleading results. In general, the presence of epistatic interactions would be difficult to capture accurately without generating huge numbers of double mutants (and even then higher-order epistasis would be missed). In particular, creating properly normalized probability distributions with sensible marginal distributions (properties described in Appendix A) would be difficult when more than one locus is considered at a time. Even in computational environments like Avida, this becomes a restrictive requirement.

The second approach to biological information is the one presented in this paper. As a result of our focus on populations, rather than individuals, the meaning of biological information is different. We do not assign an information quantity to any single genome, even though it may contain information about the environment. Instead, the fitness effects of alternative genes are reflected by their numerical representation in the population of interest. This basis also means that population size, mutation rate, and the nature of the fitness landscape also all play a role in the amount and nature of the biological information. For example, a mutation rate that exceeds the error threshold (Eigen, 1971; Wilke, 2005) creates a diffuse distribution of genotypes even though some might be more fit than others and limits biological information as we define it.

In addition to the approaches discussed above, certain other research also connects evolutionary dynamics with methods from information theory and statistical inference. We will briefly contrast this work with the results presented here. Weinberger (2002) considers the "semantic content", or meaning, of information in a theory of pragmatic information. Although there are formal similarities between this work and our own, specifically the use of a KL divergence, the goal is quite different. Our intent is to quantify the difference from randomness, reflecting the effects of evolutionary processes on the distribution of genotypes. This objective results in a different choice for the distributions used in the KL divergence. In Weinberger's pragmatic information, the initial and current genotype distributions are employed. In our work, a uniform distribution is used instead of the initial distribution of genotypes, reflecting the expected distribution if a population experiences no selection.

Fisher information, an idea from statistical inference, has also been applied to evolutionary dynamics (Frieden et al., 2001; Frank, 2009). This measure quantifies uncertainty in the estimation of a parameter from a statistical model, given some set of data. Applying this idea to evolutionary biology, the population of genotypes can be thought of as performing statistical inference on the underlying fitness landscape. Although we will not pursue this idea here, the fundamental connection of statistical inference and information theory suggests that future attempts to reconcile Fisher information with the formulation of biological information discussed here would be fruitful.

There are many other directions that the present research in biological information can be taken. A first step toward application to real data is to consider evolutionary models with finite populations. In this context, the size of the population becomes a potentially limiting factor in the ability to acquire biological information. Development of statistical inference methods for this step will require an investigation into the connections between biological information and inference. Also of interest is the application of these ideas to temporally changing fitness landscapes. Specifically, mechanistic models that include effects such as resource competition and predator–prey interactions would result in dynamic fitness values.

Finally, we consider the long-term goal of this research: to apply the methods developed here to real sequence data. A natural place to start is by considering experimental evolution in computational (Lenski et al., 1999, 2003) and bacterial (Lenski and Travisano, 1994) settings. In the computational arena, Avida (Ofria and Wilke, 2004; Adami, 2006) provides a non-trivial genotype-to-phenotype map, one that allows the evolution of complex computational tasks. Application of the refined approximation developed here promises to elucidate details in the evolutionary emergence of these complex features. Consideration of biological data may seem like a distant possibility, but is increasingly probable. With the ability to sequence full populations of bacteria and phage with increasing coverage (Wichman et al., 2005; Barrick and Lenski, 2009), the ability to apply the techniques developed here to biological populations might soon be a reality.

Acknowledgements

This research was funded by DARPA Grant HR0011-09-1-0055 and NSF Grants CNS-0751155, CCF-0643952. The authors would like to acknowledge helpful discussions with Jeffrey Barrick, Art Covert and Matthew Rupp.

Appendix A. Marginal distributions

In this appendix we describe the calculation of marginal distributions from the genotype frequencies $X(\sigma,t)$. It is important to note that the quasispecies equation, described in Eq. (5), is constructed to ensure the distribution of genotypes is properly normalized at all times (generations) t:

$$\sum_{\sigma \in \Lambda} X(\sigma,t) = 1. \tag{A.1}$$

A marginal distribution describes the distribution of alleles at some subset of loci. In this endeavor, it is useful to think of a genotype as being made up of its loci: $\sigma = \sigma_1\sigma_2$ in the two-locus landscape and $\sigma = \sigma_1\sigma_2\sigma_3\sigma_4$ in the four-locus landscape. We can then express the genotype distribution as $X(\sigma_1\sigma_2,t)$ for a two-locus fitness landscape. Similar expressions hold for longer genomes.

Given the more explicit notation introduced above, we can now consider the distribution of alleles at a subset of loci. As in the main text, we use the notation $\mathcal{A}_i$ to indicate the set of allowed alleles at locus σ_i. A marginal distribution for the first locus in a two-locus genotype is constructed using

$$X(\sigma_1,t) = \sum_{\sigma_2 \in \mathcal{A}_2} X(\sigma_1\sigma_2,t) \quad \forall \sigma_1 \in \mathcal{A}_1. \tag{A.2}$$

As a concrete example of Eq. (A.2), we consider a two-locus, two-allele fitness landscape:

$$X(\sigma_1 = a,t) = X(ab,t) + X(aB,t), \tag{A.3a}$$

$$X(\sigma_1 = A,t) = X(Ab,t) + X(AB,t), \tag{A.3b}$$

where $\mathcal{A}_1 = \{a,A\}$ and $\mathcal{A}_2 = \{b,B\}$ as introduced in Section 2.1. It is important to note that the marginal distribution is appropriately normalized. This property can be seen by summing both sides of Eq. (A.2) over $\sigma_1 \in \mathcal{A}_1$. The result of this calculation is consistent with Eq. (A.1), as required.

A set of similar calculations can be performed to construct the marginal distribution for any subset of loci, irrespective of the length of the genome. In general, we notate the resulting distribution as $X(\sigma_i,t)$ or $X(\sigma_i\sigma_j,t)$ to indicate all loci except σ_i or $\sigma_i\sigma_j$, respectively, have been summed (integrated) out.

Appendix B. Entropy and mutual information

Here we define the set of information-theoretic quantities that are used throughout this paper. This appendix is not meant to be an exhaustive discussion of information theory. The interested reader should consult a reference such as Cover and Thomas (1991) for a more thorough discussion of these and related topics. In all of the formulae to follow, we assume the distribution of genotypes, $X(\sigma,t)$, as well as the marginal distributions, $X(\sigma_i,t)$ and $X(\sigma_i\sigma_j,t)$, are properly normalized. We choose a base-two logarithm for these definitions, resulting in units of bits.

We start with entropy, a quantity that reflects the average surprisal (Tribus, 1961). Entropy is maximum when all elements of the distribution are equally likely and is equal to zero when one element has probability one. More formally, the entropy for the distribution over genotypes can be written as

$$H(\sigma,t) = -\sum_{\sigma \in \Lambda} X(\sigma,t)\log_2 X(\sigma,t), \tag{B.1}$$

where $0\log_2 0 = 0$ based on continuity arguments (Cover and Thomas, 1991). In this example, the maximum entropy is $\log_2|\Lambda|$ bits, where $|\Lambda|$ is the number of possible genotypes.

One- and two-locus entropy can also be found in a straightforward manner. These calculations employ marginal distributions created using the methods detailed in Appendix A. The resulting forms for the entropy in these cases are

$$H(\sigma_i,t) = -\sum_{\sigma_i \in \mathcal{A}_i} X(\sigma_i,t)\log_2 X(\sigma_i,t), \tag{B.2a}$$

$$H(\sigma_i\sigma_j,t) = -\sum_{\sigma_i \in \mathcal{A}_i}\sum_{\sigma_j \in \mathcal{A}_j} X(\sigma_i\sigma_j,t)\log_2 X(\sigma_i\sigma_j,t). \tag{B.2b}$$

The maximum entropy is $\log_2|\mathcal{A}_i|$ bits for the single-locus entropy and $\log_2|\mathcal{A}_i\|\mathcal{A}_j|$ bits for the two-locus entropy, where $|\mathcal{A}_i|$ is the number of possible alleles at locus σ_i.

Finally, we consider the mutual information between two loci. This quantity describes the interaction between distributions of alleles at two loci. The definition of mutual information between locus σ_i and σ_j, and its relation to the entropy defined above, are given by

$$I(\sigma_i : \sigma_j,t) = \sum_{\sigma_i \in \mathcal{A}_i}\sum_{\sigma_j \in \mathcal{A}_j} X(\sigma_i\sigma_j,t)\log_2 X(\sigma_i\sigma_j,t) \tag{B.3a}$$

$$-\sum_{\sigma_i \in \mathcal{A}_i}\sum_{\sigma_j \in \mathcal{A}_j} X(\sigma_i\sigma_j,t)\log_2 X(\sigma_i,t)X(\sigma_j,t),$$

$$= H(\sigma_i,t) + H(\sigma_j,t) - H(\sigma_i\sigma_j,t). \tag{B.3b}$$

The form in Eq. (B.3b) is obtained by applying the entropy definitions provided in Eq. (B.2).

Appendix C. Derivation of the biological information

In this appendix we provide a derivation of biological information and its approximations. As we discussed in the main text, the starting point is the KL divergence, which provides a measure of the difference between two distributions. For example, the KL divergence between arbitrary distributions $P(\sigma)$ and $Q(\sigma)$, over the genotypes σ, is given by

$$D[P\|Q] = \sum_{\sigma \in \Lambda} P(\sigma)\log_2\frac{P(\sigma)}{Q(\sigma)}. \tag{C.1}$$

It is important to note that this quantity is not a distance in the usual sense. It is not symmetric under exchange of P and Q and does not obey a triangle inequality. However, the KL divergence is equal to zero when the distributions under consideration are identical and greater than zero when they are different.

Our definition of biological information employs Eq. (C.1), using observed and uniform distributions of genotypes. These are $X(\sigma,t)$, from the quasispecies equation, and a uniform distribution for all genotypes, $\Pr(\sigma) = |\Lambda|^{-1}$, respectively. We can write the most general form of biological information as

$$I(\sigma,t) = \sum_{\sigma \in \Lambda} X(\sigma,t)\log_2\frac{X(\sigma,t)}{|\Lambda|^{-1}} \tag{C.2a}$$

$$= \log_2|\Lambda| - H(\sigma,t), \tag{C.2b}$$

where $H(\sigma,t)$ is the entropy defined in Eq. (B.1). We change the notation for this KL divergence to $I(\sigma,t)$, indicating that the biological information we are considering is for the population of genotypes. Although this notation is a little misleading, we feel it allows the series of approximations described in this appendix to be more clearly understood and emphasizes the biological meaning of the quantity.

Additional simplification of the biological information can be made when the number of alleles per locus is constant. This assumption is often true in theoretical models and certainly applies when these methods are used on DNA, RNA and protein sequences. In this case, we assume there are $|\mathcal{A}|$ alleles per locus and the genotype has L loci. This property means that the number of possible genotypes is given by $|\Lambda| = |\mathcal{A}|^L$ and the general form for the biological information can be simplified to

$$I(\sigma,t) = L\log_2|\mathcal{A}| - H(\sigma,t). \tag{C.3}$$

Next, we consider approximations of the complete biological information, which are important for two reasons. First, it is often difficult to accurately infer genotype frequencies in cases where finite data samples are considered. This issue motivates the division of the task into smaller parts, namely considering distributions of alleles at a single locus or pairs of loci. Second, the division of the biological information into parts nicely mirrors biological discussion of mutations and their effects. A single-locus approximation tells us about non-interacting mutations whereas the multiple-locus approximation provides details of epistatic interactions and modular structure.

The first approximation ignores correlations between loci (Schneider, 1997, 2000; Adami et al., 2000; Adami, 2004) by summing the entropy of the marginal distribution at each locus:

$$H_1(\sigma,t) = \sum_{i=1}^{L} H(\sigma_i,t), \tag{C.4}$$

where $H(\sigma_i,t)$ are calculated as described in Appendix B. In general, Eq. (C.4) overestimates the true entropy for the population by ignoring correlations: $H(\sigma,t) \le H_1(\sigma,t)$. Using the fact that $\log_2|\mathcal{A}|$ is the maximum entropy for one locus, the single-locus information can be written as

$$I(\sigma_i,t) = \log_2|\mathcal{A}| - H(\sigma_i,t). \tag{C.5}$$

Summing Eq. (C.5) over all loci provides our first approximation of the biological information:

$$I_1(\sigma,t) = \sum_{i=1}^{L} I(\sigma_i,t). \tag{C.6}$$

This expression is a lower bound on the true biological information because of the approximation made in writing Eq. (C.4).

The refinement of Eq. (C.6) includes correlations between all pairs of loci by expanding the approximation of the entropy given in Eq. (C.4) to include mutual information:

$$H_2(\sigma,t) = \sum_{i=1}^{L} H(\sigma_i,t) - \sum_{i=1}^{L}\sum_{j>i}^{L} I(\sigma_i : \sigma_j,t), \tag{C.7}$$

where the double sum is over all unique pairs of loci. We substitute Eq. (C.7) into Eq. (C.3) and employ Eq. (C.5) to produce

$$I_2(\sigma,t) = \sum_{i=1}^{L} I(\sigma_i,t) + \sum_{i=1}^{L}\sum_{j>i}^{L} I(\sigma_i : \sigma_j,t). \tag{C.8}$$

For the two-locus, two-allele fitness landscapes considered in much of this paper, $I_2(\sigma,t)$ is not an approximation, the result is exact.

References

Adami, C., 2004. Information theory in molecular biology. Phys. Life Rev. 1, 3–22.

Adami, C., 2006. Digital genetics: unravelling the genetic basis of evolution. Nat. Rev. Genet. 7, 109–118.

Adami, C., Cerf, N.J., 2000. Physical complexity of symbolic sequences. Physica D 137, 62–69.

Adami, C., Ofria, C., Collier, T.C., 2000. Evolution of biological complexity. Proc. Natl. Acad. Sci. USA 97, 4463–4468.

Barrick, J.E., Lenski, R.E., 2009. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. Cold Spring Harb. Symp. Quant. Biol. 74, 119–129.

Bindewald, E., Schneider, T.D., Shapiro, B.A., 2006. CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. Nucl. Acids Res. 34, W405–W411.

Bonhoeffer, S., Chappey, C., Parkin, N.T., Whitcomb, J.M., Petropoulos, C.J., 2004. Evidence for positive epistasis in HIV-1. Science 306, 1547–1550.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley-Interscience, New York, NY.

de Visser, J.A.G.M., Hoekstra, R.F., van den Ende, H., 1997. Test of interaction between genetic markers that affect fitness in *Aspergillus niger*. Evolution 51, 1499–1505.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58, 465–523.

Eigen, M., Schuster, P., 1977. A principle of natural self-organization. Naturwissenschaften 64, 541–565.

Elena, S.F., Lenski, R.E., 1997. Test of synergistic interactions among deleterious mutations in bacteria. Nature 390, 395–398.

Frank, S.A., 2009. Natural selection maximizes Fisher information. J. Evolution Biol. 22, 231–244.

Frieden, B.R., Plastino, A., Soffer, B.H., 2001. Population genetics from an information perspective. J. Theor. Biol. 208, 49–64. doi:10.1006/jtbi.2000.2199.

Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J., Stormo, G.D., 1992. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis. Nucl. Acids Res. 20, 5785–5795.

Hartl, D.L., Clark, A.G., 2007. Principles of Population Genetics, fourth ed. Sinauer Associates Inc., Sunderland, MA.

Huang, W., Ofria, C., Torng, E., 2004. Measuring biological complexity in digital organisms. In: Proceedings of the 9th International Conference on Artificial Life, pp. 315–321.

Huxley, J., 1958. The evolutionary process. In: Huxley, J., Hardy, A.C., Ford, E.B. (Eds.), Evolution as a Process second ed. George Allen & Unwin Ltd., London, pp. 5.

Jain, K., Krug, J., 2007. Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. Genetics 175, 1275–1288.

Kimura, M., 1961. Natural selection as the process of accumulating information in adaptive evolution. Genet. Res. 2, 127–140.

Lenski, R.E., Ofria, C., Collier, T.C., Adami, C., 1999. Genome complexity, robustness and genetic interactions in digital organisms. Nature 400, 661–664.

Lenski, R.E., Ofria, C., Pennock, R.T., Adami, C., 2003. The evolutionary origin of complex features. Nature 423, 139–144.

Lenski, R.E., Travisano, M., 1994. Dynamics of adaption and diversification: a 10,000-generation experiment with bacterial populations. Proc. Natl. Acad. Sci. USA 91, 6808–6814.

Michalakis, Y., Roze, D., 2004. Epistasis in RNA viruses. Science 306, 1492–1493.

Nowak, M.A., 1992. What is a quasispecies? Trends Ecol. Evol. 7 118–121.

Ofria, C., Huang, W., Torng, E., 2008. On the gradual evolution of complexity and the sudden emergence of complex features. Artif. Life 14, 255–263.

Ofria, C., Wilke, C.O., 2004. Avida: a software platform for research in computational evolutionary biology. J. Artif. Life 10, 191–229.

Orr, H.A., 2009. Fitness and its role in evolutionary genetics. Nat. Rev. Genet. 10, 531–539.

Sanjuán, R., Moya, A., Elena, S.F., 2004. The contribution of epistasis to the architecture of fitness in an RNA virus. Proc. Natl. Acad. Sci. USA 101, 15376–15379.

Schneider, T.D., 1997. Information content of individual genetic sequences. J. Theor. Biol. 189, 427–441 doi:10.1006/jtbi.1997.0540.

Schneider, T.D., 2000. Evolution of biological information. Nucl. Acids Res. 28, 2794–2799.

Shannon, C.E., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.

Tribus, M., 1961. Thermostatics and Thermodynamics. Van Nostrand, Princeton, NJ.

Weinberger, E.D., 2002. A theory of pragmatic information and its application to the quasi-species model of biological evolution. Biosystems 66, 105–119.

Whitlock, M.C., Bourguet, D., 2000. Factors affecting the genetic load in *Drosophila*: synergistic epistasis and correlations among fitness components. Evolution 54, 1654–1660.

Wichman, H.A., Millstein, J., Bull, J.J., 2005. Adaptive molecular evolution for 13,000 phage generations: a possible arms race. Genetics 170, 19–31.

Wilke, C.O., 2005. Quasispecies theory in the context of population genetics. BMC Evol. Biol. 5, 44.