# Functional information and the emergence of biocomplexity

**Robert M. Hazen*†, Patrick L. Griffin*, James M. Carothers‡, and Jack W. Szostak§**

*Geophysical Laboratory, Carnegie Institution, 5251 Broad Branch Road NW, Washington, DC 20015-1305; ‡California Institute for Quantitative Biomedical Research and Berkeley Center for Synthetic Biology, University of California, 717 Potter Street MC 3224, Berkeley, CA 94720-3224; and §Howard Hughes Medical Institute, Department of Molecular Biology and Center for Computational and Integrative Biology, 7215 Simches Research Center, Massachusetts General Hospital, Boston, MA 02114-2696

**Complex emergent systems of many interacting components, including complex biological systems, have the potential to perform quantifiable functions. Accordingly, we define "functional information," $I(E_x)$, as a measure of system complexity. For a given system and function, $x$ (e.g., a folded RNA sequence that binds to GTP), and degree of function, $E_x$ (e.g., the RNA–GTP binding energy), $I(E_x) = -\log_2[F(E_x)]$, where $F(E_x)$ is the fraction of all possible configurations of the system that possess a degree of function $\geq E_x$. Functional information, which we illustrate with letter sequences, artificial life, and biopolymers, thus represents the probability that an arbitrary configuration of a system will achieve a specific function to a specified degree. In each case we observe evidence for several distinct solutions with different maximum degrees of function, features that lead to steps in plots of information versus degree of function.**

origin of life | artificial life | evolution | aptamers | emergent systems

Complex emergent systems, in which interactions among numerous components or "agents" produce patterns or behaviors not obtainable by individual components, are ubiquitous at every scale of the physical universe, for example in neural networks (1), turbulent fluids (2), insect colonies (3), and spiral galaxies (4). Complex systems also appear in a range of artificial symbolic contexts, including genetic algorithms (5), cellular automata (6), artificial life (7), and models of market economies (8).

Life, with its novel collective behaviors at the scale of molecules, genes, cells, and organisms, is the quintessential emergent complex system. Furthermore, the ancient transition from a geochemical world to a living planet may be modeled as a sequence of emergent events, each of which increased the chemical complexity of the prebiotic world (9–11).

Given this ubiquity and diversity, it is desirable to understand the characteristics of emergent complex systems, as well as the factors that might promote complexity in evolving systems. However, complexity has proven difficult to define or measure with precision (12–14). A central objective of this study, therefore, is to examine "functional information" (15) as a quantitative measure of complexity that may be applicable to the analysis and prediction of attributes of a wide range of phenomena in physical and symbolic systems, including evolving biological systems.

An extensive literature explores historical developments and recent advances in the study of complexity and information (14, 16–18) as well as their application to understanding biological systems (3, 13, 19–24). Despite this rich literature, previous discussions of complexity have not generally focused on the relationship between information content and function (25). We propose to measure the complexity of a system in terms of functional information, the information required to encode a specific function.

## Systems and Their Functions

In this paper we consider the functional information of both symbolic systems (letter sequences and Avida artificial life genomes) and biopolymers (RNA aptamers). These systems share several characteristics: first, they consist of numerous individual components or "agents"; second, the agents can combine in a combinatorially large number of different configurations; and third, some configurations display functions that are not characteristic of the individual agents. Analyses of these systems address fundamental questions about the relationship between information content and function. For example, How much information does it take to encode a function? Are there multiple distinct solutions? How are solutions distributed in configuration space? How much more information does it take to encode a given improvement in function? What environmental factors might influence these relationships?

The function of some emergent systems is obvious: a sequence of letters communicates a specific idea, a computer algorithm performs a specific computation, and an enzyme catalyzes at least one specific reaction. Less obvious are the functions of systems of many interacting inanimate particles, such as molecules, sand grains, or stars, but these systems may also be described quantitatively in terms of function, for example, in terms of their ability to dissipate energy or to maximize entropy production through patterning (e.g., refs. 26–29). Living systems, by contrast, typically display multiple essential functions (21, 30, 31). This consideration of complexity in terms of the function of a system, as opposed to some intrinsic measure of its patterning or structural intricacy, distinguishes our treatment from many previous efforts.

**Quantifying Complexity.** Development of a quantitative measure of complexity has proven difficult for at least three reasons, each of which relates to the diversity of systems that may be labeled "complex."

1. Systems may be complex in terms of information content, physical structure, and/or behavior. Consider three stages in the life cycle of a multicellular organism: a fertilized egg, a live adult, and a postmortem adult. All three states are complex, but they are complex in different respects. All three states possess the sequence information (a genome) necessary to grow a living organism. Living and dead adult organisms also display complex anatomical structures, but only living organisms possess behavioral complexity. Any universal definition of complexity must thus have the potential to quantify complexity independently in terms of information, structure, or behavior.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "In the Light of Evolution I: Adaptation and Complex Design," held December 1–2, 2006, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program is available on the NAS web site at www.nasonline.org/adaptation_and_complex_design.

Author contributions: R.M.H. and P.L.G. designed research; R.M.H. and P.L.G. performed research; R.M.H. and P.L.G. analyzed data; R.M.H., P.L.G., J.M.C., and J.W.S. wrote the paper; and R.M.H., P.L.G., J.M.C., and J.W.S. developed concepts.

The authors declare no conflict of interest.

†To whom correspondence should be addressed. E-mail: rhazen@gl.ciw.edu.





www.pnas.org/cgi/doi/10.1073/pnas.0701744104

2. It has been difficult to define complexity in terms of a metric that applies to all complex systems. No obvious common thread exists in comparing the complexity of symbolic systems, such as language, with those of physical agents, such as cells. Parameters useful in characterizing symbolic systems (e.g., algorithm- or information-based complexity metrics) generally differ from those used to analyze systems of interacting particles (e.g., Newtonian dynamics or maximum entropy models). Gell-Mann (12) concludes, "A variety of different measures would be required to capture all our intuitive ideas about what is meant by complexity."
3. Complex emergent systems are diverse in terms of their dimensionality. Sequences of letters, computer code, or bipolymers can be treated as one-dimensional strings of symbolic information (or as points in a high-dimensional sequence space). On the other hand, many physical emergent systems, including those composed of many interacting sand grains, cells, organisms, or stars, exhibit time-dependent behaviors in two or three spatial dimensions. It is desirable for a complexity formalism to apply to this range of dimensionalities.

Despite this diversity, a common thread is present: All complex systems alter their environments in one or more ways, which we refer to as functions (32). In the words of von Baeyer (18), "Information gathering by itself, without observable effects on the gatherer's behavior, is a pointless pursuit." Function is thus the essence of complex systems. Accordingly, we focus on function in our operational definition of complexity. Therefore, although many previous investigators have explored aspects of biological systems in terms of information (e.g., ref. 33), we adopt a different approach and explore information in terms of the function of a system (including biological systems).

Szostak and coworkers (15, 34) introduced "functional information" as a measure of complexity. They proposed that the complexity of an information-rich system, such as RNA aptamers (RNA structures that bind a target molecule), can be quantified in the context of specific functions of the system, in contrast to prior formalisms based on genomic, sequence, or algorithmic information (e.g., refs. 13 and 35). Here we examine applications of this formalism to letter sequences, the artificial life platform Avida (36), and RNA apatmers.

**Functional Information as a Measure of System Complexity.** Many emergent systems of interacting agents can be described in terms of their potential to accomplish one or more quantifiable tasks. Consider a system that can exist in a combinatorially large number of different configurations (i.e., a 100-nt RNA strand comprised of four different nucleotides, A, U, G, and C, with $4^{100}$ different possible sequences). Assume that a small fraction of these configurations accomplishes a specified function $x$ to a high degree (corresponding to a high information content). Typically, a significantly greater number of configurations will prove somewhat less efficient in accomplishing function $x$ (corresponding to lower information content), whereas the majority of configurations will display little or no function (34, 35).

Accordingly, "degree of function $x$" ($E_x$) is a measure of a configuration's ability to perform the function $x$. For example, in an enzymatic system $E_x$ might be defined as the increase in a specific reaction rate that is achieved by the enzyme. In the case of a sequence of letters, $E_x$ might represent the probability that the sequence conveys a desired message to a particular recipient. And in a system with water flowing over sand ripples, $E_x$ might be defined as the rate of energy dissipation by turbulence, compared with flow over a smooth, unpatterned surface. The units or scale of $E_x$ may be somewhat arbitrary and will depend on the nature of function $x$. Thus, for example, catalytic efficiency might be recorded in terms of rate enhancement or in terms of decreased activation energy (proportional to the log of the rate enhancement).

In the formalism of Szostak (ref. 15; see also ref. 19, p. 252), functional information [$I(E_x)$] is calculated with reference to a specific degree of function $x$, designated $E_x$. Typically, a small fraction, $F(E_x)$, of all possible configurations of a system achieves at least the specified degree of function, $\geq E_x$. Accordingly, we define functional information in terms of $F(E_x)$:

$$I(E_x) = -\log_2[F(E_x)].$$

Thus, in a system with $N$ possible configurations (e.g., a sequence of $n$ RNA nucleotides, which has $N = 4^n$ discrete possible sequences):

$$I(E_x) = -\log_2[M(E_x)/N],$$

where $M(E_x)$ is the number of different configurations that achieves or exceeds the specified degree of function $x$, $\geq E_x$.

In every system, the fraction of configurations, $F(E_x)$, capable of achieving a specified degree of function will generally decrease with increasing $E_x$ (15). The largest possible functional information of a system is exhibited in the case of a single configuration that displays the highest possible degree of function, $E_{\max}$:

$$I(E_{\max}) = -\log_2[1/N] = \log_2 N,$$

where $I$ is measured in bits. This maximum functional information is thus equivalent to the maximum number of bits necessary and sufficient to specify any particular configuration of the system.

Alternatively, the minimum functional information of a system is zero for configurations with the lowest degree of function, $E_{\min}$, because all possible states have $E_x \geq E_{\min}$:

$$I(E_{\min}) = -\log_2(N/N) = -\log_2(1) = 0 \text{ bits}.$$

In this formulation, functional information increases with degree of function, from zero for no function (or minimum function) to a maximum value corresponding to the number of bits necessary and sufficient to specify completely any configuration of that system.

Functional information is defined only in the context of a specific function $x$. For example, the functional information of a ribozyme may be greater than zero with respect to its ability to catalyze one specific reaction but will be zero with respect to many other reactions. Functional information therefore depends on both the system and on the specific function under consideration. Furthermore, if no configuration of a system is able to accomplish a specific function $x$ [i.e., $M(E_x) = 0$], then the functional information corresponding to that function is undefined, no matter how structurally intricate or information-rich the arrangement of its agents.

It is important to emphasize that functional information, unlike previous complexity measures, is based on a statistical property of an entire system of numerous agent configurations (e.g., sequences of letters, RNA oligonucleotides, or a collection of sand grains) with respect to a specific function. To quantify the functional information of any given configuration, we need to know both the degree of function of that specific configuration and the distribution of function for all possible configurations in the system. This distribution must be derived from the statistical properties of the system as a whole [as opposed, for example, to the statistical properties of populations evolving in a fitness landscape (37)]. Any analysis of the functional information of a specific functional sequence or object, therefore, requires a deep understanding of the system's agents and their various interactions.

Three examples (letter sequences, the artificial life platform Avida, and RNA aptamers) serve to illustrate the concept of functional information.

**The Functional Information of Letter Sequences.** Systems of many interacting components can occur in a combinatorially large number of different configurations. Functional information depends on the fraction of all possible configurations that achieve at least a specified degree of function. Sequences of letters provide a conceptually familiar example.

Consider various sequences of $n$ letters that convey the message: "A fire has just started in a house at the corner of Main Street and Maple Street." Many different sequences of letters are capable of conveying that information. To determine the functional information of any particular sequence we must specify three parameters:

1. $n$, the number of letters in the sequence.
2. $E_x$, the degree of function $x$ of that sequence. In the case of the fire example cited above, $E_x$ might represent the probability that a local fire department will understand and respond to the message (a value that might, in principle, be measured through statistical studies of the responses of many fire departments). Therefore, $E_x$ is a measure (in this case from 0 to 1) of the effectiveness of the message in invoking a response.
3. $M(E_x)$, the total number of different letter sequences that will achieve the desired function, in this case, the threshold degree of response, $\geq E_x$.

The functional information, $I(E_x)$, for a system that achieves a degree of function, $\geq E_x$, for sequences of exactly $n$ letters is therefore

$$I(E_x) = -\log_2[M(E_x)/26^n].$$

Note that $26^n$ is the total number of possible arrangements of 26 letters in a sequence of $n$ letters, and in this treatment we assign equal probability to all possible sequences. The important more general case of configurations of unequal probabilities is a straightforward extension of the treatment of Shannon (38, 39), as discussed by Carothers *et al.* (34). Greater clarity of expression can be added through additional characters such as "space," "capital," and "period"; however, in this example we use only 26 letters. As in all combinatorially large emergent systems, most sequences convey no information (i.e., have no discernable function). Functional information is determined by identifying the fraction of all sequences that achieve a specified outcome.

Consider, for example, sequences of 10 letters that have a high probability ($E_x \cong 1$) of evoking a positive response from the fire department. Such sequences might include "FIREONMAIN," "MAINSTFIRE," or "MAPLENMAIN." Additionally, some messages containing phonetic misspellings (FYRE or MANE), mistakes in grammar or usage (FIREOFMAIN), or typing errors (MAZLE or NAPLE) may also yield a significant but lower probability of response ($0 \ll E_x < 1$). Given these variants, on the order of 1,000 combinations of 10 letters might initiate a rapid response to the approximate location of the fire. Thus,

$$I(1) \approx -\log_2[1000/26^{10}] \approx 36 \text{ bits}.$$

Numerous additional 10-letter sequences convey some relevant information but would result in a lower probability of response ($0 < E_x < 1$): "FIREHELPME," "DANGERFIRE," or "BURNINGNOW." A lower degree of function, $E_x$, will generally correspond to a larger number of effective letter sequences, $M(E_x)$.

The formulation of functional information also applies to systems in which sequences of varied lengths are combined. For letter sequences of any length from 1 to $n$ letters,

$$I(E_x) = -\log_2\left\{ M(E_x) \Big/ \left[ \sum_1^n (26^n) \right] \right\}.$$

Varying the maximum length, $n$, of the letter sequence has a significant effect on the maximum possible degree of function, $E_x$, as well as the number of states, $M(E_x)$, that achieve that degree of function. Sequences of 1, 2, or 3 letters are unlikely to convey sufficient information to achieve any response. With 4 letters, however, a few suggestive configurations exist (FIRE, MAIN, or MAPL), although all such sequences possess a high degree of ambiguity (i.e., $E_x \ll 1$).

On the other hand, with longer letter sequences ($n \gg 10$), the number of messages of a given degree of function increases dramatically, with new opportunities for explicit instructions (and hence maximum degree of function, $E_x = 1$). With a sufficient number of letters, any arbitrary degree of accuracy and precision in a message can be communicated. Note, however, that arbitrarily long sequences are not necessarily more effective at conveying information and thus may not increase the functional information of a system. For example, consider sequences of letters that begin with the following 22 letters:

FIREATMAINSTANDMAPLEST . . .

Such a sequence should invariably summon the fire department, no matter what or how many additional letters are placed at the end of the sequence. Thus, for this admittedly contrived fire department scenario, the fraction of sequences that achieve the desired outcome attains a maximum value at ≈20 letters. In competitive systems, notably genetic information constrained by length-selective pressure in living systems (e.g., refs. 40–43), longer sequences may prove inefficient and do not necessarily confer an advantage. (Indeed, in the case of reporting a fire, an overly long and detailed message might delay response time.)

Note that in this formulation of functional information the maximum possible value, $I(E_{\max})$, arises when a message is so specific that only a single letter sequence out of all possible letter sequences achieves a desired outcome. In the case of a sequence of $n$ letters, that maximum functional information occurs when $M(E_x) = 1$:

$$I(E_{\max}) = -\log_2[1/26^n] = \log_2 26^n \approx 4.7n \text{ bits}.$$

Although this conceptual example is qualitative, it introduces key concepts that are required to quantify functional information in any emergent system with numerous configurations. Of special interest is the relationship between information and degree of function. Letter sequences point to the existence of discrete "classes" of functional configurations, based in this case on the appearance of familiar words ("FIRE" and "MAIN") as well as their mutations ("FYRE" and "MANE"). We explore the role of such multiple classes of solutions in the subsequent sections on Avida and RNA aptamers.

We conclude that rigorous analysis of the functional information of a system with respect to a specified function $x$ requires knowledge of two attributes: (*i*) all possible configurations of the system (e.g., all possible sequences of a given length in the case of letters or RNA nucleotides) and (*ii*) the degree of function $x$ for every configuration.

These two requirements are difficult to meet in many systems. In the case of letter sequences, for example, the sequence is obvious, but it is difficult to determine quantitatively the degree of function of many sequences. By contrast, it is relatively straightforward to determine the degree of function (for exam-

ple, the ligand affinity) of any given RNA sequence, but impossible with present technology to measure all sequences in a large population, e.g., $\approx 10^{14}$ randomly generated 100-mers as used in some aptamer evolution studies (although single-molecule methods may ultimately provide a technical solution to this challenge). However, these concepts may be placed on a firmer footing in the case of computational systems, such as the artificial life platform Avida.

**The Functional Information of Avida Populations.** We have adapted the artificial life platform Avida (35–36) to explore the distribution of function in an emergent system. The digital organisms that populate the virtual world of Avida are "computer programs that self-replicate, mutate, and adapt by natural selection" (44) and as such share many (although not all) of the attributes ascribed to biological life. Accordingly, artificial life models have been used as a means of exploring ideas about organic biology that are not readily amenable to experimentation. Here we explore the functional information of randomly generated populations of Avida organisms. Understanding the origin and evolution of complex biological systems motivates this work; however, the first task is to demonstrate an approach for quantifying the relationship between information and functional behavior in a well characterized emergent system, whether or not unambiguous biological insight is immediately revealed.

Avida organisms consist of multiple lines of machine instructions, termed its "genome." Each organism operates as a formal computer similar to that outlined by Turing (45), and the computational properties of each organism are determined by the sequence of machine instructions stored in its memory. A population of Avida organisms can be thought of as a multitude of identical computers running many different simple programs, where differences between any two members of the population arise solely from the differences in the programs being run.

This research focuses on the ability of a small fraction of all randomly generated Avida organisms to perform computational tasks that arise through the coordinated execution of multiple machine instructions (35). None of these computational tasks can be performed by the execution of a single instruction; indeed, the shortest functional program requires five instructions. The computational ability (function) of Avida organisms thus emerges from the interaction of instructions (the agents), making Avida an ideal model for characterizing complex emergent systems.

In a typical Avida experiment, we generate $10^7$ random instruction sequences (i.e., $10^7$ different individual genomes), each sequence 100–500 instructions in length, from the default set of 26 different machine instructions. Although most sequences display no function, a small subset of sequences code for the ability to compute logic operations (such as "not" or "and") or arithmetic functions (addition and subtraction).

The set of computational tasks Avida organisms can perform allows for varied solutions, analogous to variations seen in nature. This characteristic is underscored by the fact that in its evolution apparatus Avida does not consider how a task is accomplished but only the resulting function, i.e., whether or not it is executed. The Avida platform does not specify preferred approaches to problem solving, which allows novel solutions to appear through evolution. There may be great variety among these solutions, and they may be very different from those that might have been arrived at by design (44).

**Measures of Avida Function.** Just as there is no unique measure of function in natural systems, there is no unique measure of the degree of function in an Avida sequence population. We chose to consider three distinct measures of function: (*i*) the number of times a sequence is able to compute a specific task, for example, addition or not/and; (*ii*) the total number of all tasks the sequence is able to compute, because many sequences can perform multiple distinct operations; and (*iii*) the total number of different tasks the sequence is capable of computing.

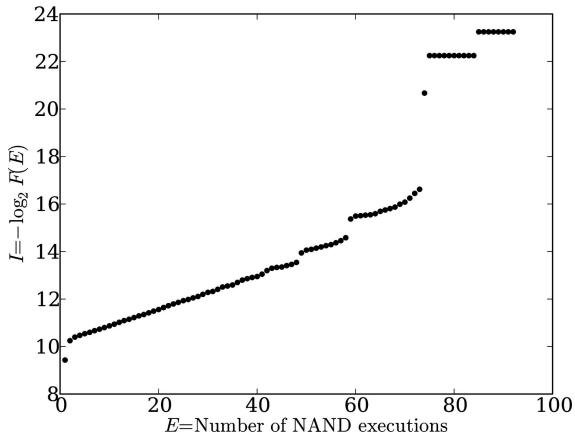


**Fig. 1.** Distribution of the not/and (NAND) function in 300-line Avida genomes in a randomly generated sample of $10^7$ genomes. The degree of function, *E*, is the number of times NAND is executed by the genome, whereas functional information, *I* (in bits), is $-\log_2$ of the fraction of all sequences that achieves at least that degree of function, *F*(*E*). Note the discontinuities, which are a recurrent feature in these experiments.

Each of these measures of function correlates to strategies that biological organisms employ to increase their fitness. Some organisms rely on the ability to perform one action very well, others rely on the ability to perform multiple actions moderately well, and still others take advantage of flexibility, the ability to do many different tasks (46). However, unlike with living organisms, quantifying the extent of these traits in Avida is straightforward and unambiguous. Most of the discussion that follows, however, focuses on execution of a single task.

Functional sequences constitute a tiny minority of the Avida genome space. Therefore, to explore fully the distribution of function within a sequence space, a large number of randomly generated sequences (i.e., equal probability) must be surveyed (see *Methods*). Such random explorations of genome space are similar to the strategies used in the directed evolution of RNA structures (e.g., refs. 47–48). Note, however, that this type of random sampling is not possible with living organisms because the portion of genome space explored in an evolution experiment will be constrained by the topology of the underlying fitness landscape and the particular configuration of the environment maxima (25, 49–51).

**Avida Results.** Random sampling of genome space has yielded several interesting results related to the frequency and distribution of functional configurations. By using Avida's default set of 26 machine instructions, a randomly generated sequence with length of a magnitude of $\approx 10^2$ lines was found to be functional (i.e., was able to perform at least one logic or arithmetic operation at least once) with probability $P \approx 10^{-3}$. The functional fraction of a population decreases with decreasing sequence length until it reaches zero for populations with sequences of a length of four machine instructions or less.

We observe regular, reproducible structure in the distribution of task execution frequency, for example, in the number of not/and or addition operations executed ($E_x$) versus functional information (Fig. 1). This plot, which illustrates the distribution of function for $10^7$ randomly generated 300-instruction genomes, is continuous over most values of $E_x$, for example, between 2 and 48. However, at several values of $E_x$, discontinuities appear. At $E_x > 73$ these discontinuities point to isolated individual ge-

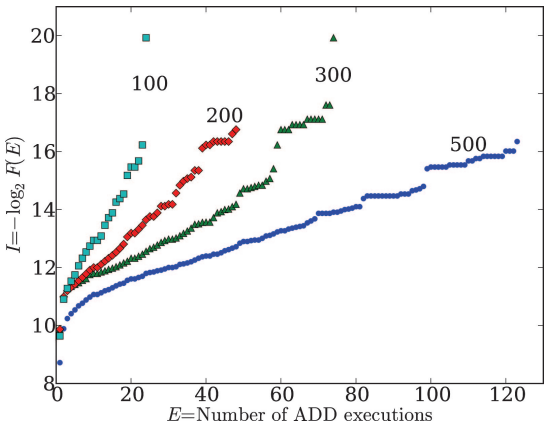


**Fig. 2.** The frequency of the ADD function in 100-, 200-, 300-, and 500-line linear Avida genomes in randomly generated samples of $10^6$ genomes. Degree of function, *E*, is the number of times the ADD function is executed by the genome, whereas functional information, *I* (in bits), is $-\log_2$ of the fraction of all sequences that achieves at least that degree of function, *F*(*E*). Note that maximum *E* increases with genome length.

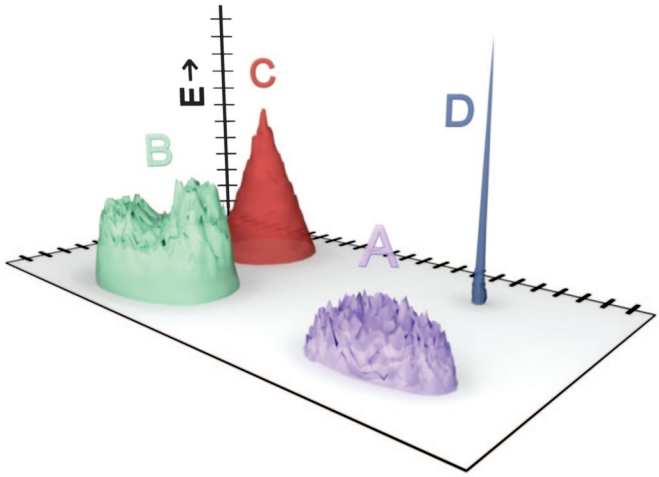


**Fig. 3.** Schematic representation of four discrete functional classes, or ''islands,'' of solutions that display function. The vertical axis is degree of function, *E*, whereas the horizontal plane represents a two-dimensional projection in sequence space. The number of sequences with degree of function $\geq E$ corresponds to the area intersected by the horizontal plane at that height along the *E* axis. Increasing *E* above the heights of the flat-topped islands A and B will result in discontinuities in the function *E* versus *I*, as illustrated in Figs. 1 and 2. Island C is a cone-shaped distribution, and island D represents a discrete solution of the type that might not be discovered in random sampling experiments.

nomes of high functionality; such outliers always appear, but they may occur at different values of $E_x$ for repetitions of this experiment. However, other discontinuities (notably those between 48/49 and 58/59) are robust, always appearing in experiments on 300-instruction genomes. Thus these gap-like features reflect an intrinsic behavior of Avida genomes.

We also find that the number and specific location of these gaps, as well as the maximum values of $I(E_x)$ and $E_x$, depend on the length of the sequences being studied (Fig. 2). For example, we examined the number of executions of the addition function for $10^6$ randomly generated genomes of 100, 200, 300, and 500 instructions. We find that the maximum number of addition executions, $E_x$, increases with genome length. We often observe discrete highly functional genomes, representing outlier solutions, as well as reproducible gaps. For randomly generated genomes of 100, 200, 300, and 500 instructions, the first significant gap in addition execution frequency occurs at 19, 39, 59, and 69 executions, respectively.

**Islands of Function.** What is the source of the reproducible discontinuities in Figs. 1 and 2? We suggest that the population of random Avida sequences contains multiple distinct classes of solutions, perhaps with conserved sequences of machine instructions similar to those of words in letter sequences or active RNA motifs (52). Each class has a maximum possible degree of function; therefore, the discontinuities occur at degrees of function below which a major class of sequences is represented and above which it is not represented.

Fig. 3 demonstrates one possible model for this stepped behavior, based on discrete ''islands'' of solutions. In Fig. 3, the islands, each of which represents a specific distinct set of solutions to the function [i.e., fitness (*z* axis)], are conceptually represented as being close to each other in sequence space (projected on the *x*–*y* plane). Note, however, that these islands are a visual simplification. For example, in the case of RNA sequences, any given ''island'' of closely related functional solutions may be more realistically represented by a densely interconnected network that spans all of sequence space (25, 53, 54). Similar consideration of function topologies has been applied to neural network connections (55) and viroid solutions to infecting the same plant host (56). Avida may be similar, because the commands relevant to a given solution do not necessarily need to appear sequentially at a specific location in the string but can occur in different registers and can be spread apart by neutral commands.

Consider a case where four classes of solutions for the same function, labeled A–D, occur in the population (Fig. 3). Each class may contain a normal distribution of degrees of function, but each has a different topology in sequence space and a different maximum degree of function, $E_x$. For relatively low values of $E_x$, all four islands contribute functional sequences. As the value of $E_x$ increases to just above the heights of flat-topped islands A and then B, discontinuities in the plot of $E_x$ versus $I(E_x)$ will occur (i.e., in Fig. 1 the height corresponding to island A would be $E_x = 48$ and the height of island B would be $E_x = 58$). This model also matches the observation that the continuous stretches of $E_x$ versus $I(E_x)$ are longest for populations of long sequences: Longer sequences allow for a greater number of distinct solutions whose superposition would serve to drown out individual discontinuities.

This model for generating discontinuities is plausible because multiple distinct solutions may exist in sequence space for a given task. For example, the shortest possible sequence (''gene'') for accomplishing subtraction is five lines long (35). However, an alternative unrelated subtraction gene 10 lines long can be constructed within the Avida language using two's-complement arithmetic (57). This second class of solutions reinforces the concept of ''islands'' of function in sequence space, where two or more types of solutions exist that achieve the same task but do so in an unrelated fashion.

We note, by contrast, that purely random statistical functions do not display steps. For example, if the degree of function is defined as the frequency of the appearance of the number ''1'' in randomly generated sequences of 100 digits, then functional information follows a well behaved smooth curve (Fig. 4). Maximum functional information arises for the solitary state with 100 consecutive 1s, whereas an obvious uniform distribution follows for lesser degrees of function. This statistically random case is not stepped. By comparison, the structures depicted in Figs. 1 and 2 suggest that the tasks being considered as functions are neither trivial, nor are they achieved by essentially arbitrary

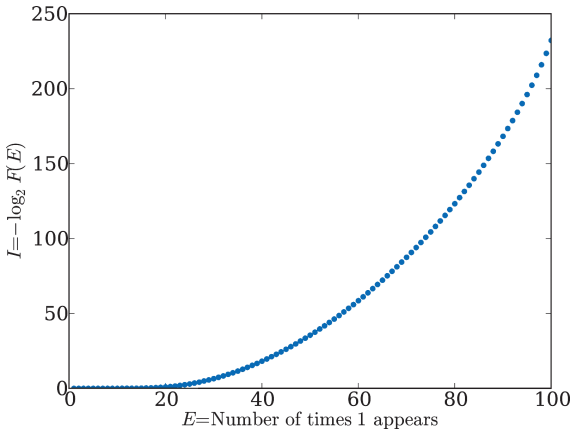


**Fig. 4.** $I(E)$ versus $E$ for the statistically random system, where $E$ is the number of times the digit 1 appears at least that many times in a sequence of 100 digits. This statistically random case is not stepped, in contrast to the topology of Avida genomes.

or random, albeit rare, configurations of the system. The interactions in the Avida system, and perhaps many other complex systems, lead to distribution of function that prove far richer than in systems possessing statistically trivial function. It remains to be seen, however, whether the observed stepped relationship between $I(E_x)$ and $E_x$ is a general feature of functional information or an idiosyncratic characteristic of Avida genomes.

**Functional Information and RNA Polymers.** The previous two examples, sequences of letters and Avida machine commands, illustrate the utility of the functional information formalism in characterizing the properties of symbolic systems that can occur in combinatorially large numbers of configurations. Functional information also has applicability to complex biological and biochemical systems; indeed, it was originally developed (15, 34) to analyze aptamers (RNA structures that bind target ligands) and ribozymes (RNA structures that catalyze specific reactions). Thus, the degree of function, $E_x$, of these linear sequences of RNA letters (A, C, G, and U) can be defined quantitatively as the binding energy to a particular molecule or the catalytic increase in a specific reaction rate. We can easily specify every possible RNA sequence of length $n$, and we can (at least in principle) synthesize RNA strands and measure the degree of function of any given sequence. The behavior of aptamers and ribozymes thus lends itself to the type of quantitative analysis that we applied previously to letter sequences and Avida populations (34).

In general, a single RNA nucleotide will display minimal catalytic or binding function, $x_{\min}$. It follows that a minimum sequence length ($n_{\min}$ nucleotides) will be required to achieve any significant degree of ribozyme or aptamer function, $E_x > E_{\min}$. Increasing the number of nucleotides ($n > n_{\min}$) will generally lead to many more functional sequences, some of which will have a greater degree of function. Furthermore, for any given catalytic or binding function there exists an optimal RNA sequence of length $n_{\text{opt}}$ that attains the maximum possible degree of function, $E_{\max}$. That sequence thus possesses the maximum possible functional information:

$$I_{\max}(E_{\max}) = -\log_2\left\{ 1 \Big/ \left[ \sum\nolimits_{1-n_{\text{opt}}} (4^n) \right] \right\}.$$

For degrees of function less than the maximum ($E_x < E_{\max}$), an intermediate functional information obtains [$I(E_x) < I_{\max}(E_{\max})$].

The *in vitro* evolution of RNA aptamers (e.g., refs. 47 and 48) provides a dramatic illustration of the evolution and selection of systems with high functional complexity. Aptamer evolution experiments begin with large populations (up to $10^{16}$ randomly generated RNA sequences), which are subjected to a selective environment, a test tube coated with a target molecule, for example. A small fraction of the random RNA population will selectively bind to the target molecules. Those RNA strands are recovered, amplified with mutations (through reverse transcription, PCR, and transcription), and the process is repeated several times. Each cycle yields a more restricted RNA population with improved binding specificity (i.e., a higher degree of function, $E_x$).

Carothers *et al.* (34), who analyzed the distribution of functional RNA aptamers in a random population, provide data on a specific example. They identify 11 distinct classes of GTP-binding RNAs, which are distinguished from each other both by nucleotide sequences (RNA motifs) (52) and secondary stem–loop structures. The degree of function of these aptamers can be defined by a solution dissociation constant, a measure of the binding strength between GTP and the folded aptamer. Carothers and coworkers find that a 10-fold increase in GTP binding strength requires ≈10 additional bits of information (i.e., a 1,000-fold decrease in abundance in a population of random sequences). Such a finding is in accord with studies of biopolymers (58, 59) that show functionally similar peptides with dissimilar primary structures, as well as reports of many distinct classes of protease enzymes (60, 61).

Furthermore, although the data of Carothers *et al.* (34) are too few to draw definitive conclusions, there is a suggestion of a stepped relationship between binding strength ($E_x$) and functional information ($I$), a relationship analogous to that displayed by populations of Avida organisms (e.g., Fig. 1). These steps, if real, are likely caused by the existence of separate classes of GTP-binding solutions. Functional classes with greater numbers of stems represent a significantly smaller fraction of all RNA sequences, but they have the potential to display greater GTP-binding affinities.

**Functional Information in Higher-Dimensional Systems.** Functional information provides a measure of complexity by quantifying the probability that an arbitrary configuration of a system of numerous interacting agents (and hence a combinatorially large number of different configurations) will achieve a specified degree of function. This concept was originally discussed in the context of biopolymer sequences that perform specific binding or catalytic functions (15, 34). In the preceding sections we demonstrated that the extension of functional information analysis to one-dimensional systems of letters or Avida computer code is conceptually straightforward, requiring only specification of the degree of function of each possible sequence.

We suggest that the functional information formalism may also be applicable to complex physical structures in higher-dimensional systems. Of special interest in this regard are biological systems that display complex emergent behavior, for example, through long-range chemical signaling among a collection of cells in social amoebas (62–64), cooperation among consortia of host organisms and symbionts (65), or colonies of social insects (3, 22, 66). We propose that functional information can be applied, at least in principle, to any such emergent system that has the ability to perform a function.

Many emergent systems can be analyzed in terms of their ability to dissipate energy or maximize entropy production (27, 29, 67, 68). For example, consider the functional information of an assemblage of sand grains subjected to a steady flow of wind or water (e.g., refs. 69 and 70). The formation of periodic sand dunes or ripples serves to initiate turbulent flow and thus increase energy dissipation. Functional information of the system can thus be measured as the fraction of all possible sand

configurations, $F(E_x)$, that achieve at least the corresponding energy dissipation, $E_x$. Such a problem might be analyzed with Monte Carlo simulations of numerous gravitationally stable sand configurations. The analytical challenge remains to determine the degree of function of a statistically significant random fraction of all possible configurations of the system so that the relationship between $I(E_x)$ and $E_x$ can be deduced.

## Conclusions

A complexity metric is of little utility unless its conceptual framework and predictive power result in a deeper understanding of the behavior of complex systems. Analysis of complex systems in terms of functional information reveals several characteristics that are important in understanding the behavior of systems composed of many interacting agents. Letter sequences, Avida genomes and biopolymers all display degrees of functions that are not attainable with individual agents (a single letter, machine instruction, or RNA nucleotide, respectively). In all three cases, highly functional configurations comprise only a small fraction of all possible sequences. Furthermore, these three examples reveal that several discrete classes of functional configurations exist, a situation that can lead to distinctive step features in plots of information versus function.

The functional information formalism may also point to key factors in the origin and emergence of biocomplexity. In particular, functional information quantifies the probability that, for a particular system, a configuration with a specified degree of function will emerge. Furthermore, analysis of the relationship between information and function may reveal how much more information is required to encode a given improvement in function. The formalism also points to strategies, such as increasing the concentration and/or diversity of molecular agents, that might maximize the effectiveness of chemical experiments that attempt to replicate steps in the origin of life.

## Methods

Determination of the computational properties of a randomly generated instruction sequence is accomplished within Avida's analyze mode. The trace feature in analyze mode generates detailed information on the state of the virtual computer at each step in the processing of a genome, including a notation of when a recognized function has been executed. An automated script parsed these logs to collect all of the data necessary to determine the functional properties of each sequence and cataloged the genomes found to be functional to permit later study. Detailed documentation of the Avida software, including descriptions of the trace function and analyze mode, can be found online at the Digital Evolution Laboratory at Michigan State University web site (http://devolab.cse.msu.edu/software/avida/doc).

We thank John Avise and Francisco Ayala for organizing this Sackler-Colloquium; H. J. Cleaves, K. Esler, R. Lenski, H. Morowitz, C. Ofria and D. Sverjensky for valuable comments and suggestions. This work was supported in part by the National Aeronautics and Space Administration Astrobiology Institute, the National Science Foundation, and the Carnegie Institution. J.W.S. is an Investigator of the Howard Hughes Medical Institute.

1. Deamer DW, Evans J (2006) in *Life As We Know It: Cellular Origin, Life in Extreme Habitats and Astrobiology*, ed Seckbach J (Springer, New York), Vol 10.
2. Frisch M (1995) *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge Univ Press, Cambridge, UK).
3. Camazine S, Deneubourg JL, Franks NR, Sneyd H, Theraulaz G, Bonabeau E (2001) *Self-Organization in Biological Systems* (Princeton Univ Press, Princeton).
4. Carlberg R (1992) in *The Astronomy and Astrophysics Encyclopedia*, ed Maran SP (Van Nostrand Reinhold, New York), pp 268–270.
5. Mitchell M (1996) *An Introduction to Genetic Algorithms* (MIT Press, Cambridge, MA).
6. Wolfram S (2002) *A New Kind of Science* (Wolfram, Champaign, IL).
7. Adami C (1995) *Physica D* 80:154–170.
8. Holland JH (1995) *Hidden Order* (Helix Books, Reading, MA).
9. De Duve C (1995) *Vital Dust: Life as a Cosmic Imperative* (Basic Books, New York).
10. Morowitz HJ (2002) *The Emergence of Everything* (Oxford Univ Press, New York).
11. Hazen RM (2005) *Genesis: The Scientific Quest for Life's Origins* (Joseph Henry, Washington, DC).
12. Gell-Mann M (1995) *Complexity* 1:16–19.
13. Adami C (2003) *Complexity* 8:49–56.
14. Shalizi CR (2006) ArXiv: nlin.AO/0307015.
15. Szostak JW (2003) *Nature* 423:689.
16. Kåhre J (2002) *The Mathematical Theory of Information* (Kluwer, Boston).
17. Gell-Mann M, Lloyd S (2003) in *Nonextensive Entropy: Interdisciplinary Applications*, eds Gell-Mann M, Tsallis C (Oxford Univ Press, New York), pp 387–389.
18. Von Baeyer HC (2003) *Information: The New Language of Science* (Weidenfeld & Nicolson, London).
19. Morowitz HJ (1978) *Foundations of Bioenergetics* (Academic, New York).
20. Bell G (1997) *The Basics of Selection* (Chapman & Hall, New York).
21. Allen C, Bekoff M, Lauder G, eds (1998) *Nature's Purposes: Analyses of Function and Design in Biology* (MIT Press, Cambridge, MA).
22. Solé R, Goodwin B (2000) *Signs of Life: How Complexity Pervades Biology.* (Basic Books, New York).
23. Avery J (2003) *Information Theory and Evolution* (World Scientific, Singapore).
24. Ricard J (2003) *C R Soc Biol* 326:133–140.
25. Lehman N, Donne MD, West M, Dewey TG (2000) *J Mol Evol* 50:481–490.
26. Bertalanffy L (1968) *General System Theory: Foundations, Applications, Development* (Braziller, New York).
27. Nicolis G, Prigogine I (1977) *Self-Organization in Non-Equilibrium Systems: From Dissipative Structures to Order through Fluctuations* (Wiley, New York).
28. Swenson R, Turvey MT (1991) *Ecol Psychol* 3:317–348.
29. Emanuel K (2006) *Phys Today* August:74–75.
30. Ayala FJ (1999) *Hist Philos Life Sci* 21:3–33.
31. McShea DW (2000) *Biol Philos* 15:641–668.
32. Bigelow J, Pargetter RP (1998) in *Nature's Purposes: Analyses of Function and Design in Biology*, eds Allen C, Bekoff M, Lauder G (MIT Press, Cambridge, MA).
33. Schneider TD, Stormo GD, Gold L, Ehrenfeuch A (1986) *J Mol Biol* 188:415–431.
34. Carothers JM, Oestreich SO, Davis JH, Szostak JW (2004) *J Am Chem Soc* 126:5130–5137.
35. Lenski RE, Ofria C, Pennock RT, Adami C (2003) *Nature* 423:139–144.
36. Adami C (1998) *Introduction to Artificial Life* (Springer, New York).
37. Wright S (1942) *Bull Am Math Soc* 48:223–246.
38. Shannon C (1948) *Bell System Technical J* 27:379–423, 623–656.
39. Klir G (2006) *Uncertainty and Information: Foundations of Generalized Information Theory* (Wiley, Hoboken, NJ).
40. Mills DR, Peterson RL, Spiegelman S (1967) *Proc Natl Acad Sci USA* 58:217–224.
41. Andersson JO, Andersson SG (1999) *Curr Opin Genet Dev* 9:664–671.
42. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) *Nature* 407:81–86.
43. Nakabuchi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M (2006) *Science* 314:267.
44. Lenski RE, Ofria C, Collier TC, Adami C (1999) *Nature* 400:661–664.
45. Turing A (1936) *Proc London Math Soc* 42:230–265.
46. Wilson EO (1992) *The Diversity of Life* (MIT Press, Cambridge, MA).
47. Ellington AE, Szostak JW (1990) *Nature* 346:818–822.
48. Wilson DS, Szostak JW (1999) *Annu Rev Biochem* 68:611–647.
49. Van Nimwegen E, Crutchfield JP, Huymen M (1999) *Proc Natl Acad Sci USA* 96:9716–9720.
50. Taverna DM, Goldstein RA (2000) *Biopolymers* 53:1–8.
51. Sasaki A, Nowak MA (2003) *J Theor Biol* 224:241–247.
52. Knight R, Yarus M (2003) *RNA* 9:218–230.
53. Huynen MA, Stadler PF, Fontana W (1996) *Proc Natl Acad Sci USA* 93:397–401.
54. Reidys C, Forst CV, Schuster P (2001) *Bull Math Biol* 63:57–94.
55. Ebner M, Shackleton M, Shipman R (2002) *Complexity* 7:19–33.
56. Codoñer FM, Daròs J-A, Solé RV, Elena SF (2006) *PLoS Pathogens* 2:1187–1193.
57. Zarowski JC (2004) *An Introduction to Numerical Analysis for Electrical and Computer Engineers* (Wiley, New York).
58. Aronson HG, Royer WE, Hendrickson WA (1994) *Protein Sci* 3:1706–1711.
59. Wang QS, Unrau PJ (2005) *RNA* 11:404–411.
60. Rawlings ND, Barrett AJ (1993) *Biochem J* 290:205–218.

61. Rawlings ND, Morton FR, Barrett AJ (2006) *Nucleic Acids Res* 34:D270–D272.
62. Goldbeter A (1996) *Biochemical Oscillations and Cellular Rhythms* (Cambridge Univ Press, New York).
63. Brännström Å, Dieckmann U (2005) *Proc R Soc London Ser B* 272:1609–1616.
64. Schaap P, Winckler T, Nelson M, Alvarez-Curto E, Elgie B, Hagiwara H, Cavender J, Milano-Curto A, Rozen DE, Dingermann T, *et al*. (2006) *Science* 314:661–663.
65. Moran NA (2007) *Proc Natl Acad Sci USA* 104(Suppl):8627–8633.
66. Strassman JE, Queller DC (2007) *Proc Natl Acad Sci USA* 104(Suppl):8619–8626.
67. Lorenz RD (2003) *Science* 299:837.
68. Whitfield J (2005) *Nature* 436:905–907.
69. Bagnold RA (1988) *The Physics of Sediment Transport by Wind and Water* (Am Soc of Civil Eng, New York).
70. Hansen JL, van Hecke M, Hanning A, Ellegaard C, Andersen KH, Bohr T, Sams T (2001) *Nature* 410:324.