OPEN ACCESS

BIO-complexity

Research Article

Measuring Active Information in Biological Systems

Jonathan Bartlett*
The Blyth Institute, Tulsa, OK

Abstract

In computer search optimization theory, *active information* is a measurement of a search algorithm's internal information as it relates to its problem space. While it has been previously applied to evolutionary search algorithms on computers, it has not been applied yet to biological systems. Active information can be very useful in differentiating between mutational adaptations which are based on internally-coded information and those which are the results of happenstance. However, biological systems present many practical problems regarding measuring active information which are not present in digital systems. This paper describes active information, how it can be used in biology, and how some of these problems can be overcome in specific cases.

Cite as: Bartlett, J (2020) Measuring Active Information in Biological Systems. BIO-Complexity 2020 (2):1-11. **doi:**10.5048/BIO-C.2020.2.

Editor: Robert J. Marks II

Received: October 9, 2019; **Accepted:** March 14, 2020; **Published:** April 6, 2020

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2020.2.c.

*Email: jonathan.bartlett@blythinstitute.org

INTRODUCTION

Biological evolution operates in at least two well-known modes—either semi-directionalized, where the outputs of evolution are correlated with the selection pressures the organisms face [1], or as a non-directionalized drift, as neutral theory describes [2]. Historically, research into evolution has focused on the ability of natural selection to keep beneficial mutants in the population, and not how they originate to begin with.

Recent work in evolutionary theory, especially in evolutionary developmental biology, has led to the realization that the inputs to evolution (i.e. the evolutionary paths that organisms are endogenously predisposed to take and the existing developmental pathways that canalize these changes into useful phenotypes) are just as important as the processes of mutation and selection themselves. In the extended evolutionary synthesis, such predispositions are given first-class causal status [3].

Over the last several decades, several important experiments have shown that, not only do developmental pathways canalize evolution, but also the actual mechanics of mutation may be manipulated to some degree by cellular machinery. As an example, in the vertebrate adaptive immune system, when the system is confronted with an unknown antigen, invokes a process known as *somatic hypermutation* to generate mutations at a higher rate which have a higher than usual chance of success. While this process is partially stochastic, it is also tightly focused, so that it only creates mutations in the correct half of the correct gene where mutations are most likely to be beneficial, and mutations in other regions *do not occur* [4, 5].

There are a number of other systems in which the DNA and cellular machinery provide directionality for mutational processes, instead of the directionality being supplied by selection alone [6–8]. Caporale suggests that these might not be isolated incidents, but rather systemic features of the genome [9, 10]. Elsewhere, the present author groups these systems under the label "evolutionary teleonomy" to note that the directionality of the evolutionary process is partially programmed by the evolving organism's own genes and cellular machinery [11]. Therefore, a methodology is needed for detecting and measuring the degree to which mutations are based on endogenous cellular systems rather than on happenstance and/or copying failures in DNA replication processes.

Such measurements would allow for the characterization of genomes according to which evolutionary processes they are endogenously predisposed to accomplish. Such characterizations can then lead to improved understanding of the evolutionary potentials of different organisms. These mutational characterizations can also be used to match organisms for industrial processes such as waste management for which evolution is part of the solution. Finally, it could aid in drug development by

characterizing which types of selection pressures the organism is predisposed to evolve against.

In order to characterize the disposition of evolutionary machinery to selection pressures, it is best to model evolution under selection as an evolutionary search for a genetic program which can relieve selection pressures. This allows the application of analysis tools which are utilized in the study of search algorithms in computer science (including genetic algorithms) to evolution. The present paper will explore one such tool in-depth, active information.

AN ACTIVE INFORMATION TUTORIAL

In search optimization theory, one intriguing discovery known as the "No Free Lunch" theorem, states that all search algorithms are equally good (or, depending on how you look at it, equally bad) at finding valid solutions to search problems when averaged over all possible search environments [12]. In other words, there is no search algorithm that is universally better than any other in all circumstances. However, one can construct a search algorithm which is better than another one *if the person knows something about the search space.*

For instance, let's assume that we have 100 index cards, each with a distinct number written on each one, laying face down in a row on the table. Let's say we are looking for a card with the number 12 written on it. Without knowing anything about the order that the cards are in, any given method of searching is equivalent—including just picking cards up at random. If, however, we know something about the order of the cards, then we can pick a search algorithm that matches the way that the cards are ordered.

Let's say that the cards are in numerical order, but I don't know which numbers have been used. In that case, I want to pick the card in the middle, and then choose the first half of the cards if the number is greater than 12, and the second half of the cards if the number is less than 12. Then, I can repeat that procedure until I find the number 12. The maximum number of steps to find that card will be $\log_2(100)$, or about 7 attempts.

However, if the cards are separated out so that the odd-numbered cards are on the left and the even-numbered cards are on the right, then that method of searching won't work. Instead, I will simply have to search through the cards on the right-hand side of the line to find the card with the number 12. Here, the maximum number of steps to find the card will be 50.

So, as is evident, in order to have a search that is more efficient than a random search through the cards, the search algorithm needs to have some information about the order the cards are in. Active information quantifies the amount of information that a search algorithm has about the pattern of the areas to be searched [13].

Active information quantifies the amount of information that a search contains by comparing the effectiveness of the search on the search space to the effectiveness of a blind, random search. The blind, random search is chosen as a benchmark because (a) it has the same performance characteristics no matter what the search space, and (b) it performs equally well as all other search mechanisms when averaged over all possible search problems. Therefore, a purely randomized search is an effective benchmark to measure a search algorithm.

Active information is measured by comparing the probability of success of a single query of a given search to the probability of success of a single query of a pure random search after both probabilities have been converted into bits. Probabilities can be converted into bits by take the negative base 2 log of the probability. So, a probability of $\frac{1}{16}$ is equivalent to 4 bits, because $-\log_2(\frac{1}{16}) = 4$. Bits are often used in information calculations because it converts probabilities, which can be unwieldy to deal with, into values which can be more naturally added and subtracted from each other.

The probability (in bits) of the random search is termed I_Ω and the probability (in bits) of the search under analysis is given as I_S. Therefore, active information (I_+) is given as:

$$I_+ = I_\Omega - I_S \tag{1}$$

Using the raw probabilities, the equation looks like this:

$$I_+ = -\log_2 P_\Omega + \log_2 P_S \tag{2}$$

This can be further simplified to:

$$I_+ = \log_2 \left(\frac{P_S}{P_\Omega} \right) \tag{3}$$

As an example, let's say that a random search of my card deck yielded success with an average success of $\frac{1}{100}$ probability, and a particular search algorithm yielded success with $\frac{1}{20}$ probability. To calculate the active information in this search we would convert each one into bits.

$$\begin{aligned} I_\Omega &= -\log_2 \left(\frac{1}{100} \right) && \approx 6.64 \\ I_S &= -\log_2 \left(\frac{1}{20} \right) && \approx 4.32 \\ I_+ &= I_\Omega - I_S \\ &= 6.64 - 4.32 \\ &= 2.32 \text{ bits} \end{aligned}$$

Therefore, we would say that the search algorithm contributed 2.32 bits to the search.

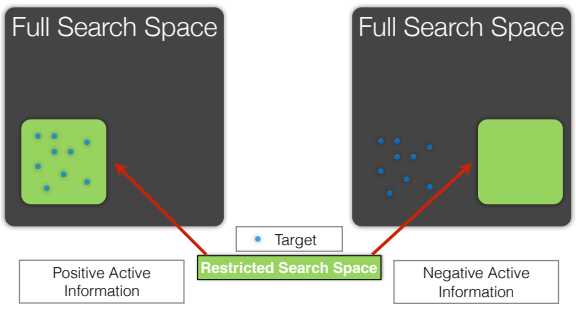

Figure 1: A Visual Representation of Active Information.
doi:10.5048/BIO-C.2020.2.f1

A search algorithm can also contribute negative active information. This means that the search is actually worse than a pure random search. Most search algorithms, if they are not tailored to the pattern of data in the search space, will yield an active information result nearing zero. Having a search algorithm that yields significant positive active information indicates that the search algorithm is tailored for the search space. In biological terms, we can say that if an organism exhibits a large amount of positive active information, then there is evidence that there is an internal mechanism geared towards generating solutions of this type.[1]

Figure 1 gives a visual representation for what positive and negative active information do with a search space.

For evolution, this means that we can measure the degree to which an organism is teleonomically aligned with an evolutionary problem by measuring the amount of active information the organism contributes to its own success responding to such problems. If an organism shows large amounts of active information for a particular problem, we can be confident that there is a mechanism of interest for generating that evolutionary pathway. In other words, the evolutionary result is likely directed by the organism's physiology, and not just a product of happenstance. Caporale calls such evolutionary directionality the *implicit genome* [10]. I have termed the internally coded programs that direct this process *evolutionary teleonomy* [11].

CLARIFICATIONS

A few items in the foregoing discussion need some clarification.

[1]More technically, active information indicates an alignment of internal mechanisms and external environment. While we are taking this as evidence that the organism contains information about the environment, this *could* be accomplished by the environment containing information about the organism. This is assumed to not be the case, since it is generally assumed that environments do not contain information about organisms.

Meanings of the Word *Endogenous Information*

One particular point of confusion with applying active information to biology is that both share a term, *endogenous information*, but have different meanings for it. In biology, endogenous information refers to useful information carried within the genome. Müller [14] defines endogenous information in biology as indicating primarily information-bearing DNA. He refers to "three sources of endogenous information, (1) the genetic information encoded in the nucleus and the mitochondria, (2) maternal cytoplasmic information which is directly (always?) derived from genetic information, (3) epigenetic information acquired by the interaction of the cells." This meaning can be seen in papers such as Quarton et al [15], which refers to the "operation of microRNAs as mediators of endogenous information and regulators of gene expression in synthetic biology."

Active information, however, has its origins in computer search theory. In information theory, the term "information" can actually refer to either (a) things that are explicitly coded for (i.e., this file contains a megabyte of information that tells you about X, Y, and Z), or the amount of entropy available in the system (i.e., this disk drive has the *capacity to hold* a gigabyte of storage). In active information, the term "information" actually refers to both. Endogenous and exogenous information refers to the entropy of the system before and after applying an algorithm to the search. Active information is the difference, and tells you how much information is contained in the algorithm itself.

However, since, in the case of organisms, the information *about the organism* would be stored in DNA or a similar organismal mechanism, the proper *biological* term for this is *endogenous information*. In order to bridge this problem, when the paper refers to biological endogenous information, this paper will expand the term to say endogenously-coded information. The bare term endogenous information will refer to the word from active information.

What Is Being Measured?

Measuring active information is measuring the information that the genome (as it presently stands) has about likely beneficial future configurations. Some have misinterpreted active information to mean that we are measuring information *being added* to the genome. One criticism of this measurement is that most organismal adaptation which helps with current selection pressures actually comes at a long-term cost to the organism [16], so therefore we would be incorrect to use selection and benefits to selection to measure active information.

This is an invalid view of active information. Active information does not claim that the resulting system contains more or less information than the system prior to mutation. Neither of those options are excluded using the active information calculation. What active information

measures is the alignment of the genome itself to the problem of finding viable genetic solutions to selection pressures.

This is wholly compatible with Behe's "First Rule of Adaptive Evolution," which states that evolution will "break or blunt any functional coded element whose loss would yield a net fitness gain." [16] The question that is posed by active information is a separate one. Does the genome contain information about what changes are likely to yield benefit? It may be that the most likely way to yield benefit is to blunt or break some particular system. If active information is present, then the blunting and breaking will be measurably tilted towards blunting and breaking systems that are likely to yield selection benefit by doing so.

The goal of active information is not to be a universal quantification of all aspects of information in biology, but rather to assess the narrow question of the information that cells contain that assist in their own evolution.

Additionally, it should be pointed out that the measurement is a valid measurement whether or not any active information is found. That is, if cells do not contain information that assist their evolution, then active information measurements will yield near-zero or negative values. If cells only contain information that assist their evolution in specific ways, active information measurements will be able to determine which ways the cell contains information for.

PROBLEMS MEASURING ACTIVE INFORMATION IN BIOLOGICAL SYSTEMS

While measuring active information is fairly straightforward in digital systems, it is decidedly much harder in living systems. There are several problems that make measuring active information more difficult in living systems.

First, in living systems, determining a valid set of target sequences that would correspond to a biologically-meaningful response to a selection pressure is difficult. This is because life optimizes for a number of variables, both short-term and long-term. Therefore, it is difficult to know what a solution necessarily looks like. Is a short-term benefit with long-term detriment a solution? It is not always possible to know, so there will inevitably be some ambiguity to the question of the fitness of a solution. However, this is not any more ambiguous than other questions of measuring fitness of organisms generally.

Second, and more significantly, there is the question of finding I_Ω. This is the biggest technical challenge, because living systems cannot be turned off without unforeseen repercussions. Therefore, it is difficult to find out how effective an organism would have been without the specificity of the mutation system. Ideally, we would be able to experimentally create a set of mutations using a computerized random algorithm and test the organisms in order to determine a solid value for I_Ω.

With digital algorithms and digital organisms, each aspect of the evolutionary search can be taken apart and evaluated. Each digital organism can be carefully modified and tracked (see Ewert et al [17] for an example of this). With living organisms, this is a practical impossibility.

Therefore, the remainder of this paper will focus on several different techniques to overcome this problem in living systems.

MEASURING ACTIVE INFORMATION IN SOMATIC HYPERMUTATION

The most straightforward mutational system for which to measure active information in living systems is the somatic hypermutation (SMH) process by which the vertebrate immune system adapts to new antigens to generate new immunoglobulins.[2] This system increases the mutation rate in order to better fit the immunoglobulin to the antigen. The question is then, "does the cellular system contribute information to the search for the solution, and, if so, how much?"

While the somatic hypermutation system has been correctly characterized as a "shotgun approach" due to its stochastic nature, it is only a shotgun approach to a very restricted range of base pairs, and in a limited phase of B-cell development. As mentioned previously, the actual mutations are limited to a single half of a single gene where mutations are likely to be beneficial.

Immunoglobulins have a "constant region" (C) and a "complementary-determining region" (CDR). The constant region is what signals to the immune system that an antigen is present. Thus, mutations in the constant region are unlikely to be helpful, as they would merely degrade the immune response. The complementary-determining region is the region that attaches to the antigen. Thus, for an antigen that the immune system has not seen before, changes to the complementary-determining region are needed in order to get a correct fit to the new antigen. And, as has been shown, this is indeed where nearly all of the mutations take place [4, 5].

The question is, how do we quantify how much *information* the cell is providing to its mutational process through this range limitation?

Several facts will help us with the computation:

- The somatic hypermutation process is usually successful. Therefore, we can focus on successes as the norm without worrying about cherry-picking.
- The somatic hypermutation process usually succeeds by only mutating a very small number of base pairs (usually about three). This means that this process probably yields the smallest mutation

[2]Note that the biological literature abbreviates somatic hypermutation as both SMH and SHM, depending on the source.

which will perform the task. Because it is the smallest size of mutation, this means that we can use the probability of this particular mutation to stand in as an estimate for the probability of achieving success generally.[3] Therefore, we don't need to know the specifics about all possible mutations that would yield a beneficial result—we just need a reasonable estimate of the smallest workable mutation.

- The somatic hypermutation process operates by restricting the physical range where possible mutations can occur. This makes it easy to calculate I_S.

The main detail we are concerned with here is the *localization* of the mutations. During somatic hypermutation, the mutation rate increases dramatically, but only within a small region of the genome. The mutation rate increases for the immunoglobulin gene. Not only that, however, the mutations are focused on the CDR (which attaches to the antigen) and are almost non-existent in the C region (which is involved in determining the immune system response).

In other words, only the gene most likely to have a beneficial effect is targeted, and then, only the part of the gene which would provide benefit is targeted. The hypermutation targeting is limited to 2 kilobases starting at the transcription start site [5]. The effective range is much shorter than that, as the mutations follow an exponential decay curve, with mutations occurring near the 2kb end of the range only happening less than 1% of the time.

Note that the mistargeting of this process can end in lymphomas, indicating that its precision is necessary to the health of the organism [19].

Therefore, since the mutation space is reduced to the same area which has likely beneficial targets, we can estimate the amount of active information this target space reduction brings. To start with, let's look at the case where only one mutation is needed to be successful.

Since we are only looking the target space (*which* base pairs to mutate but not *what* they should mutate into), we need only look at the relative sizes of the whole genome and the target space. The whole genome is roughly 3,000,000,000 base pairs, and the somatic hypermutation target is roughly 2,000 base pairs. Since we know biologically that the somatic hypermutation target is roughly the same location as the actual necessary mutations needed, this entire space reduction can be considered to be active information.

Because we are aware of the biology, we can assume that the reduction of mutation space is correct. Therefore, if the probability of finding the right base to mutate within the somatic hypermutation target (P_S) is $\frac{1}{2{,}000}$ and the probability of finding the right base to mutate within the entire genome (P_Ω) is $\frac{1}{3{,}000{,}000{,}000}$ then the active information (I_+) for this reduction is $\log_2\left(\frac{3{,}000{,}000{,}000}{2{,}000}\right)$, or approximately 20.5 bits.

GOING BEYOND SMH

The methodology described for the somatic hypermutation system can be generalized to any mutational system for which the following are reasonable parameters:

- The cell reduces the mutation space to an area that still fully contains (or almost fully contains) the solution space.
- The number of mutations that are required are small enough so that they can be reasonably thought of as the smallest mutation to accomplish the effect.

When this occurs, we can deduce a number of basic formulas. Within these formulas, g will be the genome size of the organism, z will be the reduced mutational space, and m will be the number of mutations required for success. For the calculations, we are only concerned with which positions are mutated, not what they are mutated to, because we are looking at the effect of the reduction in search space, and we are assuming that the range of mutations at these positions are not affected by the reduction.

[3] To see why this one probability can stand in for the others, imagine that I have a combination locks with five dials (each dial has the digits $0 - -9$ as options). Two codes will open the safe—one of them is a a three-digit code (but you don't know which of the digits it contains) and the other uses all five digits. The chance of hitting the right combination on the smaller set is a sufficiently accurate approximation for getting either one right. The relevant probability for both combinations is $\frac{1}{1000} + \frac{1}{100000} \approx \frac{1}{990}$ as opposed to $\frac{1}{1000}$ for just the smaller combination. Even if there were 20 working 5-digit combinations, that would still be a probability of $\frac{1}{833}$. Maintaining the same probability as the number of d-sized digits increase will require multiplying the number of successful trials by d^n, where n is the number of digits to add. So, for lock with ten positions, if there is one combination that works for a 3-digit combination, there would need to be one hundred combinations to have the same probability on a 5-digit lock. In short, expanding the number of digits in an n-position lock by d digits would only be compensated by multiplying the number of solutions by n^d. Therefore, larger-sized combinations would have to be absolutely filled with working combinations to make a significant impact on the probability. Biologically, this does not appear to be the case [18]. It is also possible that there could be multiple combinations of the same length. However, again, unless the space were replete with such combinations, it would make little impact on the probability—especially since we are, in the end, measuring it with an order-of-magnitude measuring system (bits). As the search space increases, the impact of these possibilities becomes exponentially smaller. With these considerations in mind, we can say that, from a practical perspective, the smallest valid combination can serve as an approximation for the probability of all possible combinations.

For a single required mutation for success,

$$I_\Omega = \log_2(g) \tag{4}$$
$$I_S = \log_2(z) \tag{5}$$
$$I_+ = I_\Omega - I_S \tag{6}$$
$$I_+ = \log_2(g) - \log_2(z) \tag{7}$$

If success requires multiple mutations, we can use combination laws to determine that:

$$I_\Omega = \log_2\left(\frac{g!}{(g-m)!m!}\right) \tag{8}$$
$$I_S = \log_2\left(\frac{z!}{(z-m)!m!}\right) \tag{9}$$
$$I_+ = I_\Omega - I_S \tag{10}$$
$$I_+ = \log_2\left(\frac{g!}{(g-m)!m!}\right) - \log_2\left(\frac{z!}{(z-m)!m!}\right) \tag{11}$$

Additionally, where g is large and m is small, $\frac{g!}{(g-m)!}$ can be approximated as g^m (and likewise for z). This yields:

$$I_\Omega \approx \frac{\log_2(g^m)}{\log_2(m!)} = \log_2(g^m) - \log_2(m!) \tag{12}$$
$$I_S \approx \frac{\log_2(z^m)}{\log_2(m!)} = \log_2(z^m) - \log_2(m!) \tag{13}$$
$$I_+ = I_\Omega - I_S \tag{14}$$
$$I_+ \approx (\log_2(g^m) - \log_2(m!)) - (\log_2(z^m) - \log_2(m!)) \tag{15}$$
$$I_+ \approx \log_2(g^m) - \log_2(m!) - \log_2(z^m) + \log_2(m!) \tag{16}$$
$$I_+ \approx \log_2(g^m) - \log_2(z^m) \tag{17}$$
$$I_+ \approx m(\log_2(g) - \log_2(z)) \tag{18}$$
$$I_+ \approx m\left(\log_2\left(\frac{g}{z}\right)\right) \tag{19}$$

In the case of somatic hypermutation, the average number of mutations needed is three. Therefore, the process as a whole confers approximately $20.5 \cdot 3 \approx 61.5$ bits of active information to the evolutionary search process.[4]

As is apparent, there is a significant amount of active information in the somatic hypermutation mechanism. Likewise, as is apparent from the biochemical data, the mechanism to implement this biological search is also significant. This lends credence to our original claim that significant amounts of active information indicate a system which guides mutations to beneficial ends.

[4]As an example of how close the approximation of Equation 19 is to Equation 11, using *Mathematica* with Equation 11 we calculated the number of bits with the exact formula to be within 0.0022 bits of the simplified approximation formula.

A GENERAL METHOD

In addition to specific models for evaluating the active information of specific systems, it would be helpful to establish a generalized method for performing active information experiments on living systems. Here we will attempt to establish the general parameters needed for such experiments, at least for single-celled organisms. The point of this methodology is not to strictly bind experimenters to this method, but rather to give a general outline of the problems faced and how they can be overcome mathematically.

What we would like is to let organisms produce their own mutations, measure the success rate of the organism's experiments, and then let the experimenter produce random mutations, and measure the success rate of those experiments.

Ideally, for producing random mutations, the experimenter would create a library of random edits and apply them to organisms. However, this is not necessarily practical. As an alternative, we are supposing that the experimenter is introducing a mild but broad-spectrum mutagen (or set of mutagens) which will cause mutations haphazardly within the genome. The mutations do not need to follow an exact uniform random distribution, only simply to be sufficiently broad-spectrum and uncorrelated with fitness so as to produce a significant sampling of mutations. Additionally, if possible, the mutagen should act as directly on the DNA as possible. If it instead merely activated mutagenic potential, then it is difficult to distinguish between the actions of the mutation system and actual randomized mutations (however, see the section "Relative Active Information" for measuring relative amounts of active information in situations where finding an absolute I_Ω is a practical impossibility). Finally, the mutagen should minimize non-genetic stress as much as possible to prevent interference with the solution should it be found.

Supposing an edit library or a proper mutagenic substance can be found, the next problem is that it is difficult to disable an organism's internal processes which create mutations. In other words, while we can measure an organism's success rate for producing its own mutations, it is very difficult to find out what the success rate for random mutations (i.e., I_Ω) is alone. Since an organism is also mutating on its own, the successful mutations in the population affected by the mutagen are not just those in I_Ω, but a combination of both I_S and I_Ω. Therefore, we will have to develop a mathematical model to be able to estimate what these parameters are, even though they are mixed within our experiment.

Figure 2 provides a conceptual model of what the issues are. In the conceptual model, Plate 1 consists of *only* the mutations that the organism itself undergoes. Therefore, using Plate 1, an accurate measurement of I_S (the probability of success using the organism's internal

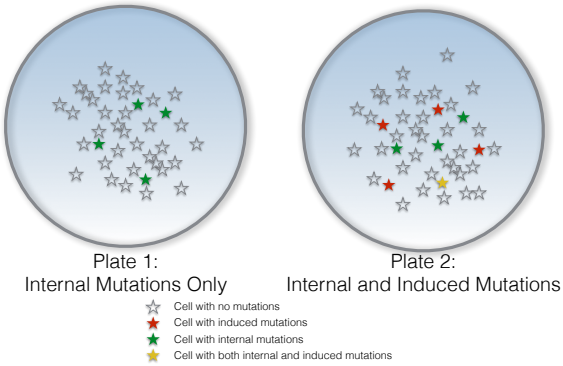

Figure 2: A Conceptual Model of the Experimental Method.
doi:10.5048/BIO-C.2020.2.f2

program) can be obtained. Plate 2 is where we attempt to model and measure I_Ω by introducing a mutagen to produce random mutations. However, in Plate 2, *both* random and induced mutations are occurring, sometimes within the same individual.

Therefore, we need a way of mathematically separating out the mutations that are happening due to an organism's internally-generated mutations and those that are happening due to the experimenter's attempt to induce random mutations (which we will call externally-generated mutations).

We can conceptualize the following measurements and calculations which will allow for the measurement of active information in such an experiment.[5]

N_{C_1} **and** N_{C_2} This is the total count of organisms under study in Plate 1 or Plate 2 of the Conceptual Model. This is determined experimentally.

U_{C_1} **and** U_{C_2} This is the total count of organisms that were successful in Plate 1 or Plate 2. This is determined experimentally.

U_{S_1} This is the total count of organisms on Plate 1 that were successful due to internally-generated mutations. Since internally-generated mutations are the only ones available in Plate 1,

$$U_{S_1} = U_{C_1}. \tag{20}$$

G This is the average genome size for the organism.

M_S This is the per-base-pair mutation rate for internally-generated mutations for the given organism. This is determined experimentally beforehand.

M_Ω This is the per-base-pair mutation rate for externally-generated mutations for the given organism. This is determined experimentally beforehand. This is modeled as being independent of M_S.

O_S This is the per-organism mutation rate for internally-generated mutations. This can be either given, determined experimentally, or calculated from M_S. The calculation for O_S is

$$O_S = 1 - (1 - M_S)^G. \tag{21}$$

Technically, this only applies if M_S is constant throughout the genome (an assumption not made in this paper). However, if a calculation of M_S is itself based on this assumption already, working backwards back to O_S using the same assumptions should be unproblematic.

O_Ω This is the per-organism mutation rate for externally-generated mutations. This is given by

$$O_\Omega = 1 - (1 - M_\Omega)^G. \tag{22}$$

Since these are indeed supposed to be random mutations, the per-base pair mutation rate can be extrapolated in this way.

P_C This is the probably that, for any organism mutated on Plate 2, that mutation was successful. In other words, this is the probability of the success of the combined action of both internal and external mutations. Using the law of large numbers, this can be calculated simply as

$$P_C \approx \frac{U_{C_2}}{N_{C_2}}. \tag{23}$$

P_S This is the probability that, for an organism mutated by internally-generated mutations, the process yielded a successful organism. If this is not already known, it can be determined in terms of U_{S_1}. Indeed, since $U_{S_1} \sim \mathrm{B}\left(N_{C_1}, P_S\, O_S\right)$ where $P_S\, O_S$ is the probability that any particular organism will get an internally-generated mutation, then, taking also into account that N_{C_1} is very large and making use of the law of large numbers, we obtain[6]

$$U_{S_1} \approx \mathbb{E}\left[U_{S_1}\right] = P_S \cdot O_S \cdot N_{C_1}. \tag{24}$$

[5] As a bit of an explanation of the nomenclature: M refers to per-base-pair mutation rates, O refers to per-organism mutation rates, U refers to the successful mutation counts (or expected values), and P refers to the probability of success of an *organism* that was mutated. For the subscripts, S refers only to the internally-generated mutations, Ω refers only to the externally-generated mutations, and C refers to their combination. Plate-specific values can further be subscripted by a 1 or 2. Ω results do not have a plate number as they are only applicable to Plate 2.

[6] This is a slight abuse of notation, as I am using the same name for both the random variable as the individual outcome. However, given the multiplied number of variables so far, introducing another level of nomenclature seems more confusing than having the same name stand in for both the random variable and the specific outcome.

Therefore,

$$P_S \approx \frac{U_{S_1}}{O_S \cdot N_{C_1}}. \tag{25}$$

To calculate P_Ω, which, as defined for active information, is the probability of an organism hitting the target by a random search, we will use the law of total probability.

Essentially, we start by noting that there are two sources of successful mutations in this experiment—Ω (random mutations) and S (internally-generated mutations). U_{C_2} is the count of all of the successful mutations. We will also use this term below to represent membership in the group of successful mutations.

The probability of success of any organism in the combined experiment (P_C) can be given by

$$\begin{aligned} P_C = \Pr[U_{C_2}|S\Omega]\Pr[S\Omega] + \Pr[U_{C_2}|\overline{S}\Omega]\Pr[\overline{S}\Omega] + \\ \Pr[U_{C_2}|S\overline{\Omega}]\Pr[S\overline{\Omega}] + \Pr[U_{C_2}|\overline{S}\,\overline{\Omega}]\Pr[\overline{S}\,\overline{\Omega}]. \end{aligned} \tag{26}$$

Since we are assuming that the probability of being in S and the probability of being in Ω are independent, we can modify this equation to read:

$$\begin{aligned} P_C = \Pr[U_{C_2}|S\Omega]\Pr[S]\Pr[\Omega] + \\ \Pr[U_{C_2}|\overline{S}\Omega]\Pr[\overline{S}]\Pr[\Omega] + \\ \Pr[U_{C_2}|S\overline{\Omega}]\Pr[S]\Pr[\overline{\Omega}] + \\ \Pr[U_{C_2}|\overline{S}\,\overline{\Omega}]\Pr[\overline{S}]\Pr[\overline{\Omega}]. \end{aligned} \tag{27}$$

Note that we have the following identities:

$$\Pr[U_{C_2}|S\overline{\Omega}] = P_S \tag{28}$$

$$\Pr[U_{C_2}|\overline{S}\Omega] = P_\Omega \tag{29}$$

$$\Pr[S] = O_S \tag{30}$$

$$\Pr[\overline{S}] = 1 - O_S \tag{31}$$

$$\Pr[\Omega] = O_\Omega \tag{32}$$

$$\Pr[\overline{\Omega}] = 1 - O_\Omega \tag{33}$$

Also note that all successes will come from mutations, and all mutations will come from either S, Ω, or both. Therefore,

$$\Pr[\overline{S}\,\overline{\Omega}] = 0. \tag{34}$$

Using these facts, the equation becomes

$$P_C = \Pr[U_{C_2}|S\Omega]\,O_S\,O_\Omega + P_\Omega\,(1-O_S)O_\Omega + P_S\,O_S(1-O_\Omega) \tag{35}$$

The probability we are wanting to know is P_Ω, so we can rearrange to solve for this.

$$P_\Omega = \frac{P_C - P_S\,O_S(1-O_\Omega) - \Pr[U_{C_2}|S\Omega]\,O_S\,O_\Omega}{O_\Omega(1-O_S)} \tag{36}$$

The probability that we don't know is $\Pr[U_{C_2}|S\Omega]$. This is essentially an error term relating to the organisms that were affected by both internal and external processes. We don't have a clean way of separating out the effects of internal and external processes in this term. However, since it is a probability, it has to be between 0 and 1, so therefore we can deduce a minimum and maximum P_Ω based on thos two values. This yields

$$P_{\Omega_{\max}} = \frac{P_C - P_S\,O_S(1-O_\Omega)}{O_\Omega(1-O_S)} \tag{37}$$

$$P_{\Omega_{\min}} = \frac{P_C - P_S\,O_S(1-O_\Omega) - O_S\,O_\Omega}{O_\Omega(1-O_S)} \tag{38}$$

We can then use our Active Information equation (Equation 2) to deduce that

$$\log_2 P_S - \log_2 P_{\Omega_{\min}} \leq I_+ \leq \log_2 P_S - \log_2 P_{\Omega_{\max}} \tag{39}$$

While that is sufficient for the standard case, there is also the possibility that P_Ω is too small to actually hit by experimental random mutation. This would lead to an I_+ of infinity.

In order to counteract that case, let us define a "configurational probability." This is the probability, based on what we know about the mutation rate and the number of mutations needed to achieve the smallest known successful mutation, that the mutation will be found by chance. Essentially, if the chance rate winds up being below our experimental detection threshold, this allows us to simulate it mathematically. With G as the genome size, and L being the smallest known successful mutation (and taking into account Footnote 3), we can calculate the configurational probability as

$$P_{\text{config}} = \frac{\binom{G}{L}^{-1}\left(\frac{1}{3}\right)^L\binom{G}{L}(M_S)^L(1-M_S)^{G-L}}{O_S}. \tag{40}$$

This equation combines (a) the chances of mutations hitting the correct mutable locations for the success, (b) the chance of getting the correct new base pair for each of those locations, (c) the chances of having that many mutations within a single cell, and (d) the chance of a particular organism having any mutations (since the P set of values is specifically only for organisms for which some mutation occurred). We used here the probability of getting exactly the right number of mutations. One could argue for using the probability of at least the right number of mutations instead.[7] This also assumes that the experiment is setup so that so that the time frame for cellular reproduction is limited to a single generation.

As you can see, even in the face of difficult experimental factors, it can sometimes be possible to develop calculations for measuring the degree to which mutational

[7]Note that even though $\binom{G}{L}^{-1}$ and $\binom{G}{L}$ cancel out, they were both left in the equation in order to better identify how this was being calculated, especially since (c) could be changed from an exact count to an "at least" count, and this facilitates modifying the equation as such.

activity is aligned with an organism's benefit. Unfortunately, the parameters of these calculations indicate that in order for this particular setup to work well (i.e., generate an I_+ within a sufficiently tight range in Inequality 39 to be meaningful) are fairly narrow.

There are several important considerations to keep in mind when using this procedure in measuring active information. First of all, since the goal is measuring the active information of the mutational mechanisms, it is important to note that if an adaptation is a multistep process, each *selectable* step must be independently evaluated. Without this consideration, the active information being measured may be that of the selection process rather than the mutation process.

Second, the organism should have a mutation rate and genome size which are low enough to allow the introduction of mutations without having too many organisms containing mutations from both processes, since, as is evident above, they cause the window of P_Ω values to be too wide to be worthwhile.

Finally, some organisms have extremely high mutation rates under selection. Under normal conditions, the mutation rate per base pair of single-celled DNA-based organisms is on the order of 10^{-11} [20]. Under certain selective conditions, however, the mutation rate may skyrocket by several orders of magnitude [21]. Rates remotely approaching (or above) G^{-1} are too high to perform this particular experiment directly, since, as O_S approaches 1, the ability to determine whether any success was from internal or external processes approaches zero.

RELATIVE ACTIVE INFORMATION

Active information is, in actuality, a relative measurement, similar to decibels. In engineering, a decibel refers to the relative loudness of a noise compared to a reference point. The reference point itself can vary—decibels simply measure the loudness *relative to* this point.

However, sometimes decibels are used alone without specifying a reference point. In these cases, the decibel is measured against a standard reference point. For audio, a decibel, when the reference point is unspecified, is measured against a sound pressure level of 20 micropascals.

So far, we have been measure active information against the standard of the genome undergoing happenstance (i.e., random) mutations. However, we can use other reference points as well to measure a relative active information against those reference points.

Richard Lenski has been performing a long-term evolution experiment that has stretched over 60,000 generations of *E. coli* evolution [22]. In this experiment, Lenski's team grows about 6.7 generations per day, and takes a 1% sample each day to transfer to a new flask.

During this experiment, Lenski's team detected a rare mutation of *E. coli* which caused it to be able to utilize citrate as a carbon source after approximately 31,500 generations [23]. The important potentiating mutation, however, occurred at approximately generation 20,000. Prior to this generation, attempts to re-evolve the gene by similar means have failed, but, after this generation, it evolves quite easily. Therefore, we will focus on the potentiating mutations that occur at generation 20,000.

In later experiments, a different lab showed that the same series of mutations can occur in as few as 12 generations if the organism is under strong selection [24]. Hofwegen et al [24] suggested that their results were due to an increase in mutation rate because their colony was under selection. However, in order for that to be the case, the number of mutations that each individual would have to undergo would be phenomenal. Thus, it is likely that, as well as an increase in mutation rate, a targeting process also occurred that made the citrate mutation more likely.

Now, we cannot know if the base citrate mutation obtained by Lenski is due to random mutation, or if there is an deeper underlying logic to the mutation which focuses it on mutations which, even if not selective, at least are more likely to be sensical. Such a hedge-betting strategy would likely be stochastic to some extent, but would not necessarily be random. In other words, even if the specific site selected is not based on function, the specific list of mutable sites may be skewed towards function [25, 26].

However, because active information is a relative measurement, we don't have to know whether or not this is true. What we can do is measure the active information of the mutation under selection compared to base rate of the mutation not under selection. This will give us the amount of active information that the organism has towards solving this problem over and above what is present in the basal rate.

Lenski [27] notes that, after 20,000 generations, each of his twelve populations have encountered between $3 \cdot 10^8$ and $1.5 \cdot 10^9$ mutations. Therefore, we will use 10^{10} as the estimate of the number of mutational events that happened across all twelve populations combined in order to get the potentiating mutation.

Hofwegen et al [24] does not provide a number of mutational events or even a mutation rate, but we can estimate from the number of generations tested. It is unclear what the mutation rate was, but the population per generation was probably equivalent due to their attempt to replicate the Lenski experiment in other aspects of their experimentation. Since the mutation rate is not given, it is possible that there was an increased mutation rate due to hypermutating cells. Foster [21] says that hypermutable cells can have a 200-fold mutation rate, but that hypermutators represent only about 1% of cells

under selection. Therefore, we can estimate that being under selection will triple the number of mutational events.

If the average per-generation mutational events in Lenski is $5{\cdot}10^4$, then the per-generation mutational events under selection will be approximately $1.5{\cdot}10^5$. Therefore, across 12 generations, we will have approximately $1.8{\cdot}10^6$ mutational events, out of which we will get a mutation.

Therefore, we can calculate the active information that the cell has about the selection in such cases as:

$$\begin{aligned} I_+ &= I_\Omega - I_S \\ &= -\log_2\left(10^{10}\right) + \log_2\left(1.8 \cdot 10^6\right) \\ &= 33.2 - 20.8 \\ &= 12.4 \end{aligned}$$

Therefore, *E. coli* contributes approximately 12.4 additional bits of information towards the search for the Cit^+ mutation when under selection. This number is *relative* to the ordinary predisposition of *E. coli* to produce this mutation when not under selection, which has not been determined.

CONCLUSION

At its core, active information is a quantitative tool for understanding evolution and its potentials. In general, the purpose of calculating active information is to see if the cell has mutational mechanisms geared to solving the given evolutionary problem presented to it. If the active information in a process is significantly above zero, then it is likely that the cell has significant resources devoted to solving that evolutionary problem or class of evolutionary problems. If the active information is near zero or even negative, then that indicates that the cell does not have resources dedicated to solving that sort of biological problem.

Knowing whether or not an organism has active information targeting a particular problem or class of problems can help in fundamental research by identifying whether or not we should be searching for a teleonomic mutational system for generating such mutations. It can take considerable lab work to detect and analyze the workings of mutational machinery. Active information can therefore be used to establish the likelihood that there is a mechanism worth finding prior to investigation. If an interesting mutation is found but active information is near or below zero, then it was likely to be merely a fortuitous occurrence. On the other hand, if active information is significantly positive, this provides the justification for expending the cost needed to search for a corresponding mechanism. Additionally, establishing patterns of problems for which organisms have active information can be identified, which would be good first steps to finding the mechanisms responsible.

Active information can also be important for bioengineering. Knowing the types of problems different organisms are geared to solve will help in determining the likely future evolutionary paths of organisms. This can aid in selecting organisms for industrial problems (such as medicine and waste management) which rely on an organism's mutational abilities.

ACKNOWLEDGMENTS

I want to thank all of the people who gave comments on early versions of this manuscript, especially Matthew Mcintosh and Winston Ewert. I especially want to thank Asatur Khurshudyan, who helped walk me through various issues in the statistical calculations. I also want to thank the many reviewers of this paper for their feedback—it is a much-improved paper due to their efforts. In particular, one reviewer suggested a change to the statistical method being used which tightened the bounds in the "General Method."

1. Darwin C (1859) On the Origin of Species By Means of Natural Selection. John Murray. doi:10.5962/bhl.title.28875
2. Kimura M (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press. doi:10.1017/CBO9780511623486
3. Pigliucci M, Müller GB (2010) Elements of an extended evolutionary synthesis. In: Pigliucci M, Müller GB, eds. Evolution - The Extended Synthesis. The MIT Press pp 3–17. doi:10.7551/mitpress/9780262513678.003.0001
4. Papavasiliou FN, Schatz DG (2002) Somatic hypermutation of immunoglobulin genes: Merging mechanisms for genetic diversity. Cell 109:S35–S44. doi:10.1016/s0092-8674(02)00706-7
5. Rada C, Milstein C (2001) The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. The EMBO Journal 20:4570–4576. doi:10.1093/emboj/20.16.4570
6. Hall BG (1999) Transposable elements as activators of cryptic genes in *E. coli*. Genetica 107:181—187. doi:10.1023/A:1003936706129
7. Zhang Z, Saier MH (2009) A mechanism of transposon-mediated directed mutation. Mol Microbiol 74:29–43. doi:10.1111/j.1365-2958.2009.06831.x
8. Henderson IR, Owen P, Nataro JP (2002) Molecular switches—the on and off of bacterial phase variation. Mol Microbiol 33:919–932. doi:10.1046/j.1365-2958.1999.01555.x
9. Caporale LH (1999) Chance favors the prepared genome. Ann NY Acad Sci 870:1–21. doi:10.1111/j.1749-6632.1999.tb08860.x
10. Caporale LH (2006) An overview of the implicit genome. In: Caporale LH, ed. The Implicit Genome. Oxford University Press pp 3–25.
11. Bartlett J (2018) Evolutionary teleonomy as a unifying principle for the extended evolutionary synthesis. BIO-Complexity 2018 (2):1–7. doi:10.5048/BIO-C.2017.2
12. Wolpert DH, Macready WG (1997) No free lunch theo-

rems for optimization. IEEE Trans Evol Comput 1:67–82. doi:10.1109/4235.585893

13. Dembski WA, Marks II RJ (2009) Conservation of information in successful search: Measuring the cost of success. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 5:1051–1061. doi:10.1109/TSMCA.2009.2025027
14. Müller WA (1996) From the Aristotelian soul to genetic and epigenetic information: The evolution of the modern concepts in developmental biology at the turn of the century. The International Journal of Developmental Biology 40:21–26.
15. Quarton T, Ehrhardt K, Lee J, Kannan S, Li Y, Ma L, Bleris L (2018) Mapping the operational landscape of microRNAs in synthetic gene circuits. npj Systems Biology and Applications 4. doi:10.1038/s41540-017-0043-y
16. Behe M (2010) Experimental evolution, loss-of-function mutations, and "the first rule of adaptive evolution". The Quarterly Review of Biology 85:419–445. doi:10.1086/656902
17. Ewert W, Dembski WA, Marks RJ (2009) Evolutionary synthesis of nand logic: Dissecting a digital organism. In: Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, 3047–3053. doi:10.1109/ICSMC.2009.5345941
18. Axe DD (2004) Estimating the prevalence of protein sequences adopting functional enzyme folds. Journal of Molecular Biology 341:1295–1315. doi:10.1016/j.jmb.2004.06.058
19. Teng G, Papavasiliou FN (2007) Immunoglobulin somatic hypermutation. Annual Review of Genetics 41:107–120. doi:10.1146/annurev.genet.41.110306.130340
20. Roth JR, Kugelberg E, Reams AB, Kofoid E, Andersson DI (2006) Origin of mutations under selection: The adaptive mutation controversy. Annual Review of Microbiology 60:477–501. doi:10.1146/annurev.micro.60.080805.142045
21. Foster PL (2004) Adaptive mutation in *Escherichia coli*. Journal of Bacteriology 186:4846–4852. doi:10.1128/JB.186.15.4846-4852.2004
22. Lenski RE (2017) Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. The ISME Journal 11:2818–2194. doi:10.1038/ismej.2017.69
23. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proceedings of the National Academy of Sciences 105:7899–7906. doi:10.1073/pnas.0803151105
24. van Hofwegen DJ, Hovde CJ, Minnich SA (2016) Rapid evolution of citrate utilization by *Escherichia coli* by direct selection requires citT and dctA. Journal of Bacteriology 198:1022–1034. doi:10.1128/JB.00831-15
25. Bartlett J (2008) Statistical and philosophical notions of randomness in creation biology. Creation Research Society Quarterly 45:91–99.
26. Bartlett J (2009) Towards a creationary classification of mutations. Answers Research Journal 2:169–174.
27. Lenski RE (2010) Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*. In: Janick J, ed. Plant Breeding Reviews. John Wiley & Sons, Ltd pp 225–265. doi:10.1002/9780470650288.ch8