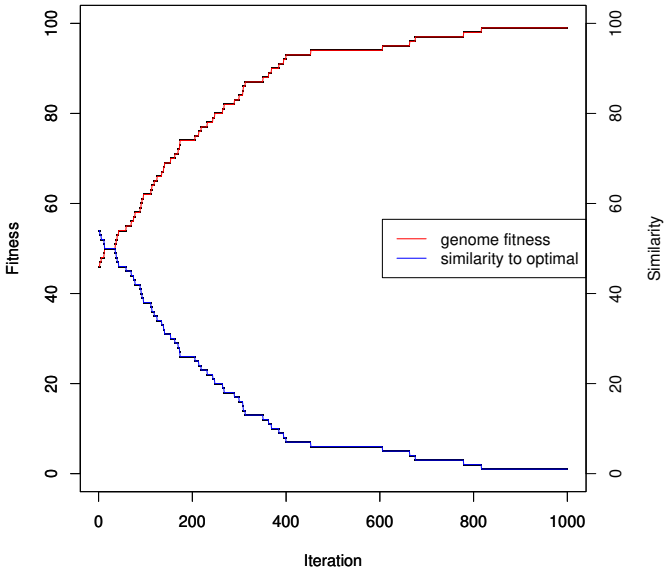
Figure 2: A Deceitful Landscape



Is Active Information Applicable to Biology?

Jonathan Bartlett

DOI: 10.33014/issn.2640-5652.2.2.bartlett.2

Active information was originally introduced in 2009 by William Dembski and Robert Marks II (Dembski and Marks II, 2009). Active information identifies how much information that a search has compared to a "random search." Introduced in the context of information science, it was originally utilized towards identifying information sources in various computerized forms and simulations of evolution.

The reason why active information works is because there is no general "best" algorithm for searching. A search that is good in one context will be terrible in another. There might be a best search *for a particular situation*, but not one that serves all situations equally. In fact, it turns out that, for any particular search situation, a random search has *average* performance characteristics compared to any other search algorithm.

Therefore, for any search situation, we have the capability of determining what the average success rate for a search should be (note that the success rate is how many times the search algorith has to "look" before finding a successful hit). If we simply perform a random search and measure successes, we can determine the average value.

In terms of statistics, this average value is the expected value for the success rate for a search strategy chosen arbitrarily. That is, if the search strategy is chosen arbitrarily, we would expect that the success rate should be roughly equivalent to that of a random search.

Active information measures the distance between the success rate that we actually observe and the success rate that we would expect from an arbitrarily chosen search strategy. If this is measured prior to selection affecting the success rate, we can then measure the distance between the success rate that the cell's own mutational machinery is having and the success rate that we would expect from arbitrary mutation strategies. This will tell us the amount of information that the cell's mutational machinery has for finding a solution in a given selective process.

Recently, I demonstrated how this could be measured in biological systems, giving examples for how different types of systems might be measured (Bartlett, 2020). Since this is a fairly new approach for thinking about mutations in the genome, there are many confusions about what is actually being claimed and proposed. This note intends to clarify, explain, and defend the notions presented in the paper.

Addressing Misconceptions

I want to start by clarifying that active information does not (a) hold that mutations form a uniform random distribution, (b) hold that mutations *should* form a uniform random distribution, or (c) hold that standard evolutionary theory holds that mutations should form a uniform random distribution. Instead, active information attempts to simulate a uniform random distribution of mutations *in order to get an expected value* for the success rate of other mutational strategies. This follows not from evolutionary theory but rather from information theory, which states that such a search *will give you the expected value* for the success rate of *other* searches. This distinction is critical and forms the basis of the logic of applying active information to biology.

Another important clarification is that, as stated in the paper, it does not matter if evolution is *ontologically* a search.

Many incorrectly reject the application of the mathematics of search to evolution on the basis that evolution isn't truly a search for anything. Whether or not that is true is irrelevant. Evolution (or at least certain situations in evolution) matches the mathematical preconditions of a search, and, therefore, search mathematics applies whether or not it is a search ontologically. If an organism is undergoing selective pressure, we can define a "successful search" as an organismal configuration that relieves that selective pressure beyond a certain threshold. This is easiest to understand and measure when the selection is lethal. The search space (the genome), the search activity (mutation), and the search target (any genome configuration that relieves the selection pressure) are clearly defined. Also note that some people incorrectly believe that the mathematics of search imply that we are looking for a specific target (i.e., DNA sequence), or that we know what the target(s) (DNA sequences) are ahead-of-time. This is not the case either. We merely have to have a well-defined definition of the target. In this case, it is a genome configuration that relieves the selective pressure. We identify it not by sequence (since we don't know what sequence(s) that will be) but by result (relieving the selective pressure).

Methodological Concerns

One potential concern is that we are excluding the effects of the active information supplied by natural selection. The general method presented does not fall prey to that criticism, as it focuses on single-generation results (thus not allowing for natural selection to work). However, it is true that, if trying to apply active information to biology in some other way, this could be an issue. Pachón and Marks II (2020) presents a way of calculating the active information of selection, which may point towards a way of measuring the active information in the biological system in experiments where selection also supplies active information as well.

The active information supplied by selection may actually be a contributing factor to the success of *E. coli* developing the *Cit*$^+$ mutation described in "Relative Active Information" section of Bartlett (2020). Further research will be required to determine how much of an impact this has on the calculation.

Another potential issue with the measurement technique presented in Bartlett (2020) is that, to replicate to a population size adequate to perform the study, variation in the genome will already be introduced prior to the study in question. This might already introduce variety in the population that needs to be accounted for either experimentally or mathematically. A simple way to adapt for this is to begin with a replica plating technique to filter out colonies that already have a successful hit.

How Targeted is Somatic Hypermutation?

Bartlett (2020) also shows how, using certain assumptions, the active information calculation can be simplified. One particular simplification was given for the somatic hypermutation process. Essentially, if it can be shown that a particular mutational system occurs by restricting the targets of mutation, and that the shortest mutational targets are contained within this restricted space, then a simplified calculation can be used based on the size of the genome, the size of the restricted mutational space and the number of mutations required to hit a target.

Some have called into question whether or not the somatic hypermutation process actually fits the given criteria. For instance, there is evidence that sometimes Activation-induced Deaminase (AID), the mechanism behind somatic hypermutation, sometimes hits targets outside of the space suggested by this characterization (see, for instance, Álvarez-Prado et al. (2018)). What is at issue is not the relevance of the simplified formulas to situations matching the criteria, but of whether or not the specific case of somatic mutation matches the criteria. Additionally, the goal of the formula (and, in fact, any formula) is to generalize, so whether or not this criticism successfully prevents applying the simplified formula will depend on the quantity of exceptions.

Álvarez-Prado et al. (2018) itself does not specifically address these issues, as it is itself working with a modified mutational process intended to identify potential AID targets from a biochemical perspective. In fact, the paper itself shows that the mutational process with all components intact actually removes the vast majority of "misses." The biochemistry of AID *acting alone* targets a number of regions (275 identified by the paper), but the combination of AID with the other components of the mutational process limits the actual mutated targets (i.e., targets with an actual final sequence change) to only a handful (Liu et al., 2008).

Since the mathematics of the process are based on order-of-magnitude reductions in search space, it is unlikely that having a handful of additional targets would actually significantly change the results, especially if they occurred at a lower frequency than those in the primary targeted area.

Thus, while it is certainly possible to be more precise in the measuring of active information of somatic hypermutation,

it seems that being used as a simplified measurement is still well-justified. In fact, such papers as Álvarez-Prado et al. (2018) show how important the targeting is (justifying the criteria for using the formula), by showing the prevalence of cancerous effects of mistargeted mutations.

Isn't This Already Well-Known?

One criticism is that we already knew that there are targeted mutations without active information. This is at least partially true. While there are groups who recognize this reality, many evolutionary biologists do not. In fact, I've talked with several practicing biologists (evolutionary, molecular, and otherwise) who were shocked to find out even that such phenomena existed. Some were familiar with somatic hypermutation as a general idea, but had not mentally linked it to the question of directed mutation. Many biologists still believe (and most textbooks still teach) that mutations are uncorrelated with their fitness effects. This could wind up being true or false in the general case. Active information provides a mechanism for measuring this question from the data.

However, the more important goal is not to determine the existence of such phenomena, but rather to be able to *measure* the phenomena. Currently, directed mutations are only known *after* we know the mechanism in detail. The goal of active information is to provide a measurement *prior* to knowing the mechanism (in fact, specifically to see if there is a mechanism worth finding).

There are some who agree that mutations are not uncorrelated with fitness, but don't believe that comparing against a random background is a correct way to quantify the phenomena. However, I have not heard any such critic present an alternative means of quantifying directedness. I think the mathematics of active information (and the biological application of it) is sufficiently sound for experimental use. However, if there is a better means of quantification, I would be interested in comparing the two.

Additional Notes

The mathematics of Bartlett (2020) are a little hard to follow, so I wanted to present a combined formula here. The meanings of the components of the formula are given in Bartlett (2020).

$$I_{+_{\max}} = \log_2\left(\frac{U_{C_1}}{N_{C_1}}\right) - \log_2\left(\frac{\frac{U_{C_2}}{N_{C_2}} - \frac{U_{C_1}}{N_{C_1}}(1-O_\Omega) - O_S\,O_\Omega}{O_\Omega(1-O_S)}\right) \tag{1}$$

$$I_{+_{\min}} = \log_2\left(\frac{U_{C_1}}{N_{C_1}}\right) - \log_2\left(\frac{\frac{U_{C_2}}{N_{C_2}} - \frac{U_{C_1}}{N_{C_1}}(1-O_\Omega)}{O_\Omega(1-O_S)}\right) \tag{2}$$

Additionally, a supplementary spreadsheet to assist calculating active information using the techniques found in the paper is available (Supplement 1), with example possibilities provided to give a feel for how different outcomes affect active information.[1]

Álvarez-Prado, Á F et al. (2018). "A Broad Atlas of Somatic Hypermutation Allows Prediction of Activation-induced Deaminase Targets". In: *Journal of Experimental Medicine* 215.3, pp. 761–771. DOI: 10.1084/jem.20171738.

Bartlett, J (2020). "Measuring Active Information in Biological Systems". In: *Communications of the Blyth Institute* 2020.2, pp. 1–11. DOI: doi:10.5048/BIO-C.2020.2.

Dembski, W A and R J Marks II (2009). "Conservation of Information in Search: Measuring the Cost of Success". In: *IEEE Transactions on Systems, Man and Cybernetics A, Systems & Humans* 5.5, pp. 1051–1061. DOI: 10.1109/TSMCA.2009.2025027.

Liu, M et al. (2008). "Two levels of protection for the B cell genome during somatic hypermutation". In: *Nature* 451, pp. 841–846. DOI: 10.1038/nature06547.

Pachón, D A Díaz and R J Marks II (2020). "Active Information Requirements for Fixation on the Wright-Fisher Model of Population Genetics". In: *BIO-Complexity* 4, pp. 1–6. DOI: doi:10.5048/BIO-C.2020.4.

[1] Supplement 1 is available online at https://journals.blythinstitute.org/ojs/index.php/cbi/article/view/68/66.