Devious Distortions: Durston or Myers?
or
Estimating the Probability of Functional Biological Proteins

Kirk Durston, Ph.D. Biophysics

A few days after his 2009 debate with me on the existence of God, well-known atheist, PZ Myers, posted a blog with the title, 'Durston's Devious Distortions', in which he attempted to address my claim that amino acid sequences coding for biological protein families have an astonishingly low probability … so low that they definitively falsify the hypothesis that biological proteins could have been assembled via an evolutionary process.

For his resource material, Myers referred to an online video of a lecture I gave at the University of Edinburgh about a year earlier. That video was for a general, interdisciplinary audience and did not actually discuss how the probabilities of biological proteins are calculated. Given this absence of information, Myers made some assumptions as to how I calculate protein probabilities. It appears that he did not check his own numbers by plugging them back into the relevant equation to see if they gave the results I had obtained. Those who tried to use Myers' numbers found that they were incorrect, not yielding anything close to the results for SecY or RecA that I had published in the literature and shown in the video. Unfortunately, some atheists who follow Myers, unquestioningly accepted his explanation without checking his numbers, thus spreading Myers' confusion.

In reality, the probabilities of biological proteins were calculated using a two-step procedure. The first step was to calculate the functional complexity required to code for a particular protein using a method I had published in 2007.[1] That paper was referenced in the video in a slide titled, 'Case 3: Average protein' so that inquiring minds could obtain further details. Once a value was obtained, the second step was to plug that value into the equation I showed in the video and solve for the probability. Here, I shall demonstrate how these probabilities can be calculated using real multiple sequence alignment data for protein families.

Note: For those with no science background, or those who do not wish to slug through the math and technical bits, skip sections 3 &4.

1. Functional Information

In 2003, Jack Szostak published a short article in Nature, pointing out that classical information theory did not consider the meaning or functionality of a message. For biology, however, whether or not a sequence is functional is very important. Szostak, therefore, introduced the need for a new measure of information, which he called *functional information*.[2] In 2007, Robert Hazen and three other colleagues, among whom was Szostak, published a definition of functional information $I(E_x)$ in the form of

$$I(E_x) = -\log_2[M(E_x)/N], \qquad (1)$$

where $M(E_x)$ is the number of different sequences that meet or exceed the required level of function within the cell and N is the total number of possible sequences, both functional and non-functional.[3] The ratio $M(E_x)/N$, therefore, represents the probability of a functional sequence within the larger set of all possible sequences.

To give a very simple example how Eqn. (1) can be applied, imagine a combination lock that requires a three-number combination, each number having 100 options (e.g., 0 to 99). The total number of possible combinations would be $N = 100^3$ or 1,000,000. Let us suppose that there are three combinations that will actually open the lock, so $M(E_x) = 3$. The probability $M(E_x)/N$ of dialing in three numbers and opening the lock is 3 chances in 1,000,000 or .000003. The amount of functional information $I(E_x)$ required to specify a functional combination, using Eqn. (1), would be 18 bits, where a 'bit' is a unit of information in classical information theory.

Unfortunately, for proteins $M(E_x)$ is an unknown number. That leaves us with two unknowns, $I(E_x)$ and $M(E_x)$, and only one equation, so we are unable to use Eqn. (1) by itself to calculate the probability $M(E_x)/N$ or the functional information $I(E_x)$ required to code for a functional protein sequence. Another equation is needed.

2. Measuring Functional Complexity

Functional Information has two subsets, prescriptive and descriptive information.[4] *Prescriptive Information* (PI) provides the instructions that determine the outcome. Computer software is an example of PI. *Descriptive Information* (DI) merely *describes* an outcome and is useful only to minds capable of appreciating a good description.

There is not yet general agreement as to how to measure functional information and there is still ongoing discussion as to what information actually is. Also, some prefer to use the term 'complexity' rather than 'information'. Furthermore, there is a difference between *defining* functional information and functional complexity and *measuring* it. Although Hazen *et al*. stated that they were providing a *definition* of functional information, it might be more cautious to say that they were providing a way to *measure* functional information. There may be various ways or methods to measure or estimate functional information, some of which may be more accurate than others. For the sake of this discussion, however, let us accept their definition and work from there.

Since Eqn. (1) has two unknowns when applied to proteins, it cannot be used by itself. I and my colleagues published a different method that enabled us to measure the functional complexity required to code for a functional protein.[1] Whether one prefers to call it functional information or functional complexity, both Hazen's Eqn. (1) and our method are based upon the same concept of Shannon uncertainty with the joint variable of function. Ours focuses on the probability distributions of the amino acids at each site in the multiple aligned sequences to calculate the Shannon uncertainties, and Hazen's approach focuses on the probability distribution of the entire set of sequences. In their Eqn. (1), all sequences are assumed to be equally probable. Thus, we have two equations to solve two unknowns. Using the method in our paper, the functional complexity required for a functional sequence is measured by the change in functional uncertainty $H(X_f(t))$ of some data set $(X_f(t))$ between the ground state (*g*) and the functional state (*f*) at some time *t* and is described as

$$\zeta = \Delta H\,(X_g(t_i), X_f(t_j)), \qquad (2)$$

where ζ is equivalent to Hazen's $I(E_x)$ used in Eqn. (1) and

$$\Delta H\,(X_g(t_i), X_f(t_j)) = H(X_g(t_j)) - H(X_f(t_i)) \qquad (3)$$

and

$$H(X_f(t)) = -\sum P(X_f(t)) \log P(X_f(t)). \qquad (4)$$

If all options are equally probable, then

$$H(X_g(t_i)) = -\sum (1/W) \log (1/W) = \log W \qquad (5)$$

where *W* represents the total number of possible options. To apply this method of measuring functional information to proteins of length *L* amino acids, *W* = 20 amino acids, Eqns. (2-5) can be assembled to give a measure of functional complexity as follows,

$$\zeta = I\left(E_x\right) = \sum_{j=1,L}\left[\log 20 - \sum_{i=1,20} P\left(X_f\left(t\right)\right)\log P\left(X_f\left(t\right)\right)\right]. \qquad (6)$$

The summation within the brackets is done for the probabilities of each of the 20 amino acids at a particular site *j* in the amino acid sequence. The results are then summed for all the sites from 1 to *L*. For more details, please see our paper.[1] Hazen's Eqn. (1) and our Eqn. (6) now put us in a position of being able to calculate $I(E_x)$, $M(E_x)$ and the probability $M(E_x)/N$. First, we can use Eqn. (6) to solve for $I(E_x)$. Once a value for $I(E_x)$ has been obtained, we can insert it into Eqn. (1) to solve for the probability $M(E_x)/N$. Thus, the two equations together can give us the probability $M(E_x)/N$ of a functional protein.

3. An Example

This method was applied to 35 protein families, including the universal protein RecA mentioned in the video, which is an average length protein found in all life forms.[1] A total of 1,553 sequences (not to be confused with $M(E_x)$ which is an unknown and likely much greater than the known functional sequences) were downloaded from the online protein family database Pfam. The functional information (ζ or $I(E_x)$) required to code for a functional member of the RecA protein family was found to be 832 Functional Bits (Fits). This value can be inserted into Hazen's Eqn. (1) to get

$$832 = -\log_2[M(E_x)/N]$$

allowing us to solve for the probability $M(E_x)/N$, which turns out to be 10^{-250}, the number shown in the video mentioned earlier. To clarify, the probability of obtaining a functional sequence for RecA in a single sampling is approximately 1 chance in 1 with 250 zeros after it. Evolutionary biology, contends that there were a vast number of evolutionary samplings. Using published estimates for fast mutation rates, total number of individual life forms on earth, length of genomes, and so forth, we can estimate that the total number of different sequences possible in four billion years was not more than 10^{42}. This makes the very generous assumption that no sequences were ever repeated. A published, 'extreme upper limit' puts the maximum number of samplings at 10^{43}.[5] To put this in perspective, we have only 10^{43} opportunities to find something that, on average, would require closer to 10^{250} trials. In other words, the entire sum of mutations, insertions and deletions, operating over four billion years, would fall short by more than 200 orders of magnitude of producing a functional RecA sequence. Perhaps we got unbelievably lucky, for it is the nature of probabilities that one might not have to use 10^{250} tries to obtain a functional RecA sequence; one could be fortunate enough to obtain it on the very first attempt. Life requires, however, tens of thousands of proteins. Getting massively lucky tens of thousands of times does not qualify as a scientific explanation for the origin of life.

The number of amino acid sites in the multiple sequence alignment for RecA was 240. For proteins, there are commonly 20 different amino acid options at each site, so the total number of possible sequences is $N = 20^{240}$ (not 10^{43} as Myers incorrectly assumed). We can then solve for the total number of functional sequences that might code for RecA as follows,

$$M(E_x) = 20^{240}(10^{-250})$$

or $M(E_x) = 10^{62}$. This is an enormous number of possible functional sequences for RecA (not merely '1' as Myers incorrectly assumed). The sheer number of possible non-functional sequences, however, makes the probability of assembling a functional sequence virtually zero for all practical purposes.

4. The Probabilities Get Worse

This measure of functional information is a good first-pass estimate, but the situation is actually far worse for an evolutionary search. In the method described above and as noted in our paper, each site in an amino acid protein sequence is assumed to be independent of all other sites in the sequence. In reality, we know that this is not the case. There are numerous sites in the sequence that are mutually interdependent with other sites somewhere else in the sequence. A more recent paper shows how these interdependencies can be located within multiple sequence alignments.[6] These interdependencies greatly reduce the number of possible functional protein sequences by many orders of magnitude, reducing the probabilities by many orders of magnitude. In other words, the numbers we obtained for RecA above are exceedingly generous; the actual situation is far worse for an evolutionary process. Nevertheless, for the purpose of this example, let us use the very generous numbers calculated above.

5. Implications for Evolution

RecA is a universal protein, found in all life forms. For that reason, evolutionary biologists believe it has descended from the Last Universal Common Ancestor (LUCA). Thus, it would have been a component of the earliest life forms with very little time to appear. We have just seen that the probability of obtaining Rec A is one chance in 10^{250} (or one chance in 1 with 250 zeros after it). The common response to this is that there has been more than just one chance to generate a sequence for RecA. In fact, it could be as many as 10^{43} chances. Of course, that is more than 200 orders of magnitude too few chances to reasonably expect to obtain RecA, but surely if we took the entire universe as our physical system, there would be enough chances to get life going somewhere, and maybe our planet is the winner of the lottery.

The number of particles in the observable universe is often estimated to be around 10^{80}. Let us suppose that each one of these particles was actually a sequence of 300 amino acids forming an average length protein. How many possible combinations could be sampled if all 10^{80} proteins in this imaginary protein universe recombined once per second for 13 billions years? The answer turns out to be approximately 10^{97} different combinations. Doubtless, many of them would be repeats of earlier combinations, but let us ignore that. The entire evolutionary capacity of the universe, if it were nothing but proteins recombining every second, is still more than 150 orders of magnitude too inadequate to expect to produce any of the 10^{62} possible functional sequences for RecA.

Of course, the problem is much greater than merely finding one protein family in an evolutionary search. The simplest life form is thought to contain approximately 151 to 250 protein-coding genes.[7, 8] But the problem is not merely limited to getting the first life form. To get the full diversity of life requires thousands of additional protein families. The usual response is to appeal to evolutionary biology's god-of-the-gaps, and say that natural selection did it. However, novel biological protein families represent needle-in-the-haystack problems for an evolutionary algorithmic search, not hill-climbing problems as many assume. Thus, natural selection is of no use in guiding an evolutionary trajectory across non-folding sequence space. The genesis of novel protein families must proceed via a blind, unguided random walk across non-folding sequence space.[9] Natural selection actually hinders this process, tending to drive an evolving amino acid sequence back into functional sequence space.[10] Novel proteins are relatively easy to assemble; any amino acid sequence will do, but those that produce stable-folding, functional sequences seem to be extremely rare.

A more formal way to evaluate the hypothesis that biological proteins were obtained in a blind evolutionary process is to apply the Universal Plausibility Metric (UPM), where an hypothesis is definitively operationally falsified if the UPM for that hypothesis is less than 1.[11] For RecA occurring somewhere within the universe during its history to date, the UPM = 10^{-142}, which means that the hypothesis that it could be located by an evolutionary process is definitively falsified. On the other hand, RecA requires only 832 Fits (Functional Bits) of information to sequence, a quantity of functional information that intelligence can easily generate.

6. Fingerprints of Intelligence

It is increasingly acknowledged in the literature that the functional information encoded in the genomes of life is essentially software designed for molecular computers.[12-14] It has also been pointed out in the literature that functional sequence complexity is only observed in human languages, computer code and in DNA, and requires 'rational agency' to encode.[15] It follows from this that statistically significant levels of functional complexity or functional information is the fingerprint of a rational agent, or an intelligent mind.

A 'god-of-the-gaps' argument follows the idea that if we do not know what did it, then God did. At best, it is a weak argument based upon ignorance. In the case of functional information or functional complexity, we do know what can produce such an effect; rational agency. We do it all the time whenever we send an email, type out a text, write some software. An intelligent mind is an empirically verified candidate for the sequencing of statistically significant levels of functional sequence complexity. Chance and necessity can produce statistically trivial levels of functional complexity or functional information but nothing of any statistical significance. Thus, we have only one empirically verified candidate for the origin of functional information or functional complexity ….. rational agency. Statistically significant levels of functional information or functional complexity are a positive marker for intelligence.

The argument can be summarized as follows:

1. A unique attribute of intelligence is the ability to produce statistically significant levels of functional information.
2. The gene that codes for RecA encodes a statistically significant level of functional

information.
3. Therefore, RecA has the positive fingerprints of intelligence all over it.

It should be pointed out that premise (1) is verifiable and falsifiable. If one is *a priori* committed to the belief that rational agency was not involved in the origin of life, then one must falsify premise (1) either in the lab or through computer simulations that utilize a fitness function that accurately models a random walk through non-folding sequence space to solve a needle-in-the-haystack problem.

One response has been to suggest that nylonase[16] is an example of nature producing functional information. This response illustrates a chronic lack of rigor that is often evident in such assertions. A method has been published for measuring the change in functional complexity due to an evolutionary process.[1] The change in functional complexity required to produce the novel function of breaking down nylon is actually trivial, of no statistical significance. It is a mistake to assume that a novel function is equivalent to a statistically significant increase in functional information or functional complexity. Novel functions can be achieved with little or no change in functional information. Any claims that nature can produce a non-trivial, statistically significant gain in functional information needs to be supported by some actual numbers. A method is available in the literature for doing that.

Functional complexity, or functional information as defined by Hazen *et al*., provides a marker or fingerprint of intelligence. Thus, the software of life, encoded in the genomes of life, has the fingerprints of intelligence all over them. *To clarify, the presence of functional information encoded in a genome provides positive evidence for an intelligent origin.* The functional information may be deteriorating, as it always does in all information storage media, but its source has to be a rational agent. It is the only empirically verified option on the table. Everything science is doing right now in the rational design of artificial genomes and proteins supports the hypothesis that rational design is an essential component of DNA and protein design.

7. Discussion

Axe found that only about one in 10^{64} sequences forms any working domain.[17] The extreme improbability of finding a single, functional protein was noted in the literature as early as 2001 by Taylor when he pointed out that even a protein library with the mass of the entire earth would only compose a miniscule portion of sequence space. He stated that ‘intelligent design’ in the lab will be required to create novel protein scaffolds.[18] This prediction has certainly been borne out in recent advances in designing artificial proteins where ‘rational design’ has been shown to be necessary.[19-21] For example, a recent paper on designed proteins uses the word ‘designed’ 29 times in the text, title, captions and abstract, the word ‘design’ 11 times, and ‘strategy’ 3 times.[20] The point to note here is that the ongoing synthesis of artificial proteins is an example of *intelligent design in action*.

The origin of RNA replication reduces to the problem of how to obtain the properly coded functional information to produce the relevant components needed for RNA replication. Evolutionary biologist Eugene Koonin has determined that the probability of obtaining RNA replication is so low that the best explanation for the origin of life is that there must be an infinite number of universes.[22, 23] Invoking an infinite number of universes is purely metaphysical and would constitute another of science's own 'god-of-the-gaps'.

Barring the existence of an untestable, infinite number of universes, there is an alternate, testable hypothesis for the origin of the functional information encoded in the genomes of life; premise (1) given in the previous section. We can empirically verify intelligence can produce huge amounts of functional information; every way we can test the theory that mutations can do it is soundly, and consistently falsified. There is only one empirically verifiable option on the table; rational agency and we are demonstrating that in the lab every time we construct an artificial folding protein using information we have reverse engineered from biological proteins.

8. Conclusions

Here, I have shown how to estimate the probability of a functional protein using real, multiple sequence alignment data. In explaining how this is done, I have hopefully added some clarity to this process so that others will not make the kind of incorrect assumptions Myers did.

Actual $M(E_x)/N$ probabilities estimated from real world data available on Pfam exposes a colossal problem for the kind of evolution that would be required to produce novel life forms. That is why, after a century and a half of science smashing its head against a solid rock, the 'main forces directing long-term molecular evolution remain obscure.'[10]

The idea that a Darwinian process can produce the kind of functional information required to code for the average functional protein family, not to mention all of biological life, is a popular one, but when actually tested against real data, is definitively falsified. The probability of coding the functional information into a genome to specify a functional, biological protein is so small, we cannot expect it to happen even once in the history of the universe. Of course, we can predict that mutations, insertions and deletions are easily capable of producing novel proteins, if we simply define a protein as an amino acid sequence, but the kind of biological proteins that are essential to life are highly sophisticated components that are extremely rare in sequence space, according to the multiple sequence alignment data. It is increasingly being recognized that the functional information encoded in the genomes of life is highly complex and sophisticated computer code or 'DNA software' as Craig Venter describes it.[13] Intelligence can easily produce the levels of functional information we observe in biological life and computer software. There is only one observable, testable and scientifically verifiable option on the table;

rational agency. The genomes of life have the fingerprints of intelligence all over them.

References

1. Durston KK, Chiu DK, Abel DL, Trevors JT: **Measuring the functional sequence complexity of proteins**. *Theor Biol Med Model* 2007, **4**:47.
2. Szostak JW: **Functional information: Molecular messages**. *Nature* 2003, **423**(6941):689.
3. Hazen RM, Griffin PL, Carothers JM, Szostak JW: **Functional information and the emergence of biocomplexity**. *Proc Natl Acad Sci U S A* 2007, **104 Suppl 1**:8574-8581.
4. Abel DL: **The first gene**, 1st edn. New York, NY: LongView Press--Academic Biological Research Division; 2011.
5. Dryden DT, Thomson AR, White JH: **How much of protein sequence space has been explored by life on Earth?** *J R Soc Interface* 2008, **5**(25):953-956.
6. Durston KK, Chiu DD, Wong AK, Li GC: **Statistical discovery of site inter-dependencies in sub-molecular hierarchical protein structuring**. *EURASIP J Bioinform Syst Biol* 2012, **2012**(1):8.
7. Huang CH, Hsiang T, Trevors JT: **Comparative bacterial genomics: defining the minimal core genome**. *Antonie van Leeuwenhoek* 2013, **103**(2):385-398.
8. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome**. *Trends Genet* 2009, **25**(3):107-110.
9. Blanco FJ, Angrand I, Serrano L: **Exploring the conformational properties of the sequence space between two proteins with different folds: an experimental study**. *J Mol Biol* 1999, **285**(2):741-753.
10. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA: **Epistasis as the primary factor in molecular evolution**. *Nature* 2012, **advance online publication**.
11. Abel DL: **The Universal Plausibility Metric (UPM) & Principle (UPP)**. *Theor Biol Med Model* 2009, **6**:27.
12. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M: **What is a gene, post-ENCODE? History and updated definition**. *Genome Res* 2007, **17**(6):669-681.
13. O'Connell C: **Passing the baton of life-from Schrodinger to Venter**. *New Scientist* 2012:3.
14. Liberman EA, Minina SV: **Cell molecular computers and biological information as the foundation of nature's laws.** *Bio Systems* 1996, **38**(2-3):173-177.
15. Abel DL, Trevors JT: **Three subsets of sequence complexity and their relevance to biopolymeric information**. *Theor Biol Med Model* 2005, **2**:29.
16. Ohno S: **Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence**. *Proc Natl Acad Sci U S A* 1984, **81**(8):2421-2425.

17. Axe DD: **Estimating the prevalence of protein sequences adopting functional enzyme folds**. *J Mol Biol* 2004, **341**(5):1295-1315.
18. Taylor SV, Walter KU, Kast P, Hilvert D: **Searching sequence space for protein catalysts**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10596-10601.
19. Baker M: **Protein engineering: navigating between chance and reason**. *Nature methods* 2011, **8**(8):623-626.
20. Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH: **De novo designed proteins from a library of artificial sequences function in Escherichia coli and enable cell growth**. *PLoS ONE* 2011, **6**(1):e15364.
21. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D: **Principles for designing ideal protein structures**. *Nature* 2012, **491**(7423):222-227.
22. Koonin EV: **The cosmological model of eternal inflation and the transition from chance to biological evolution in the history of life**. *Biology direct* 2007, **2**:15.
23. Koonin EV: **The logic of chance : the nature and origin of biological evolution**. Upper Saddle River, N.J.: Pearson Education; 2012.