On The Origin of Protein Folds

A common objection to the theory of intelligent design (ID) is that it has no power to make testable predictions, and thus there is no basis for calling it science at all. While recognising that testability may not be a sufficient or necessary resolution of the "Demarcation Problem", this article will consider one prediction made by ID and discuss how this prediction has been confirmed.

What is Intelligent Design?

To understand the basis for a prediction of ID, we first have to understand what the theory asserts. ID has been defined as the claim that certain patterns in nature bear the hallmarks of having been designed by an intelligent cause, rather than an unguided natural process. The theory of ID maintains that there are two criteria, both of which need to be met, in order to justify an inference to design. These are:

(1) Complexity.
(2) Specificity.

To satisfy criterion (1), the feature under investigation must be highly improbable with respect to the available probabilistic resources, thereby rendering appeals to chance unreasonable. To satisfy criterion (2), the feature under investigation must conform to some independently given pattern. This may include semantic meaning, such as spelling out words or sentences, or functional specificity. We use this method routinely in our everyday lives, and it has application in several scientific disciplines, including forensic science, SETI research, and archaeology. Mathematician William Dembski has offered a very conservative estimate of the maximum number of physical events that have transpired in the visible universe, which he gives as 10^{150} (Dembski, 1998), thereby setting a "universal probability bound" of 1 in 10^{150} below which the probabilistic resources of the Universe are exhausted. For a more thorough discussion and mathematical treatment of this method of design detection, I refer readers to William Dembski's books "The Design Inference" (Dembski, 1998) and "No Free Lunch" (Dembski, 2001).

There are thus two major questions that we can ask. Firstly, is the design inference based on a reliable method of detecting design? The design inference is predicated upon our uniform and repeated experience of cause-and-effect. In all of our experience, we know of only one type of cause – one category of explanation – that is able to produce specified complexity. That cause is deliberative activity. Thus, based on the historical abductive method – the major basis of scientific reasoning about the past – it is concluded that the most causally adequate candidate explanation for this phenomenon is intelligence (Meyer, 2009).

The second question that we can ask is this: Based on this methodology, do features of interest exhibit evidence of design? This essay will evaluate the hypothesis that protein folds exhibit evidence of design.

The Nature of Protein Folds

The design inference with respect to the origin of protein folds makes a testable prediction which will be the subject of our present discussion. Firstly, however, we need to discuss a little protein biochemistry. Proteins are natural polymers, made up of smaller subunits (monomers) called amino acids, each of which differs from the others in its chemically unique appendages called side chains. These amino acids are linked together by peptide bonds to form chains, and the side-chains protrude from the main chain.

The sequence of amino acids in a polypeptide chain determines its uniquely folded three-dimensional conformation (Anfinsen *et al.*, 1961), which is largely determined by the distribution of polar and non-polar side chains (Cordes *et al.*, 1996). Proteins typically possess a shell of hydrophilic amino acid residues that surround an inside-core of hydrophobic residues, which have been buried so as to avoid contact with water. The peptide bonds hydrogen bond with each other to give a so-called "secondary structure," which is classified as α-helices and β-sheets. A structure of multiple adjacent elements of secondary structure is known as a supersecondary structure or motif, and includes α-helix hairpins, β-hairpins, and β-α-β motifs. When different motifs pack together they form domains, which comprise the fundamental units of a protein's tertiary structure. When several polypeptide chains associate into an oligomeric molecule, it is referred to as the protein's quaternary structure. For example, the protein Haemoglobin, which is essential for the transport of oxygen in the blood, is comprised of two α chains and two β chains (these chains are also

referred to as "subunits"). Figure 1 shows the three domains of Pyruvate kinase, one of the enzymes involved in glycolysis.

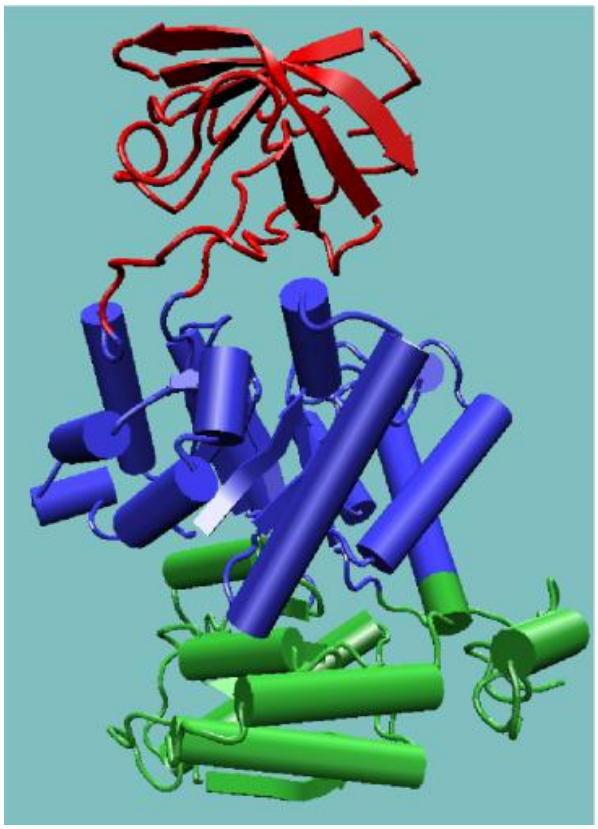

Figure 1: A graphical representation of the three domains of Pyruvate kinase, one of the enzymes involved in glycolysis.
Source: *http://upload.wikimedia.org/wikipedia/commons/6/67/1pkn.png*

Since there are 20 different species of amino acid used in organic proteins, the number of possible ways of arranging the amino acids in a polypeptide chain is given by 20^x where 'x' is the number of amino acid subunits in the chain. A relatively modestly-lengthed polypeptide consisting of 150 amino acids could be arranged in $20^{150} = 1.427 \times 10^{195}$ different ways. The mean protein length in bacteria is reported as 267 amino acids; in eukaryotes 361 amino acids; and in archaea 247 amino acids (Brocchieri and Karlin, 2005), although some protein sizes greatly exceed this mean. The longest known polypeptide is Titin (Labeit and Kolmerer, 2005; Labeit *et al.*, 1990), the giant protein (consisting of 244 different domains) which confers elasticity to muscle. Its human variant is comprised of 34,350 amino acids. As the polypeptide chain grows in size, the number of combinatorial possibilities increases exponentially.

A Prediction of ID

Since no more than a small fraction of Dembski's estimate of the maximum number of physical events in the visible universe are of relevance to proteome evolution, it is evident that only a small sample of conceptually possible polypeptide sequences have been actualised during the evolution of life on earth. It is also known that only a fraction of conceptually possible polypeptide sequences will fold into a functional protein.

This raises an interesting question: How common (or rare) are the functional amino acid sequences with respect to the vast combinatorial space of possibilities? Both neo-Darwinian evolution and intelligent design make a prediction on the answer to this question.

What does the evolutionary model predict? To understand this, it is helpful to think about the problem in terms of a fitness landscape. A fitness landscape is a concept used in the field of evolutionary biology to visualise the relationship between an organism's phenotype and its reproductive success. Adaptive fitness is represented by mountain ranges ('peaks'). An example of a fitness landscape is given in figure 2:

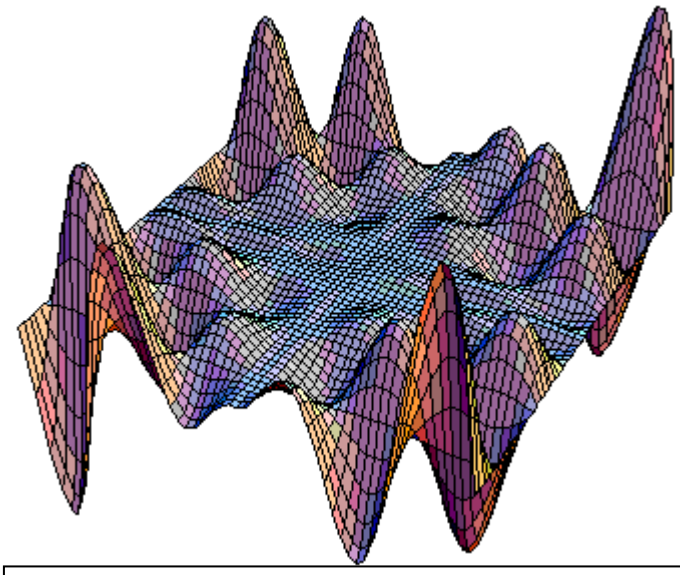

Figure 2: A fitness landscape.
Source: http://classes.yale.edu/fractals/CA/GA/Fitness/Fitness.html

Imagine a fitness landscape with 2000-3000 peaks, each one representing a different protein fold. The peaks are dispersed in sequence space and exceedingly rare. Now, how likely is it that a blind search will be able to navigate to the base of a fitness peak? If, by some fluke of

luck, you landed at the base of a perfectly smooth fitness peak, then it is feasible that natural selection could, in time, facilitate your ascension up the peak, thereby optimising the protein to a high level of function. But suppose you landed somewhere miles from any peak on the flat plain of non-functionality. In that case, a blind search would never be able to navigate its way to a functional peak. If, on the other hand, these fitness peaks were very common or clustered together, a blind search may well stumble upon the base of a fitness peak. The evolutionary model thus predicts that the proportion of combinatorial space that is functional will be relatively high (so that locating the bases of functional fitness peaks by mutation and selection is feasible).

What does the theory of intelligent design predict? Since the criteria for justifying an inference to design are (i) complexity and (ii) specification, the hypothesis that protein folds exhibit evidence of design predicts that protein folds will meet both of those criteria. We have already established that protein folds are functionally specific – the precise sequence of arrangement of the amino acids will determine the conformation of the polypeptide chain. But are functional protein folds sufficiently improbable to exhaust the available probabilistic resources? ID predicts that the answer to this question will be "Yes," and thus that the prevalence of functional folds in combinatorial sequence space will be astronomically low. We are able to attribute written sentences in the English language to an intelligent agent precisely because meaningful – semantically significant – sequences of alphabetic characters are so astronomically rare with respect to combinatorial sequence space (Denton, 1986). The ratio of meaningful 12-letter words to total sequence space is 1 in 10^{14}. Furthermore, most meaningful sentences are highly isolated with respect to one another, meaning that random substitution is far more likely to degrade meaning rather than enhance it. Is the same true here? Let's see.

Testing the Prediction

Several important studies have shed light on the answer to this question. One such study examined a 153-residue domain of a 263-residue TEM-1 β-lactamase (Axe, 2004). A β-lactamase (shown in figure 3) is an enzyme responsible for bacterial resistance to beta-lactam antibiotics including penicillins (Abraham and Chain, 1940). The enzyme confers resistance to the bacteria that possess it by breaking open the β-lactam ring of the antibiotic by hydrolysing its peptide bond, thereby deactivating the antibiotic (Majiduddin *et al.,* 2002).

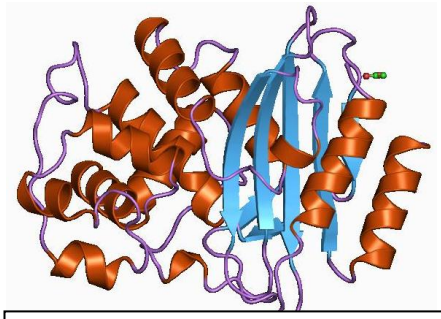

Figure 3: A graphical representation of the structure of a *Streptomyces albus* beta-lactamase.
Source: http://upload.wikimedia.org/wikipedia/commons/8/87/PDB_1bsg_EBI.jpg

Using a technique called insertion mutagenesis (involving alterations to the protein's amino acid sequence), Axe began his research with a sequence with a particular hydropathic signature, and set out to ascertain how many sequence variants with that signature (out of all possible variants) could form a functional structure. He began his work with a very weak variant, because he wanted to be able to detect variants operating at the lowest level – at the threshold of detectability – because an evolving new fold would be expected to be very poorly functional.

Axe's paper estimates "the prevalence of protein folds adopting functional enzyme folds" to be roughly 1 in 10^{77}. This figure was extrapolated from the number of variants that were able to perform TEM-1's function, even very weakly, compared to all of the possible sequences of TEM-1's length. This allowed him to estimate the rarity of functional folds in all of sequence space.

An additional study examined the AroQ family of chorismate mutase (Taylor *et al.*, 2001). Chorismate mutase is responsible for catalysing the chemical reaction that converts chorismate to prephenate – part of the pathway that leads to the production of phenylalanine and tyrosine. The structure of chorismate mutase is shown in figure 4.

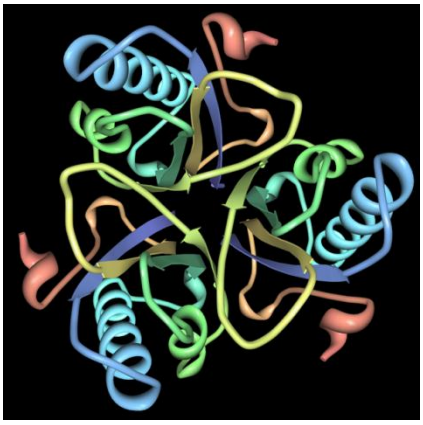

Figure 4: A graphical representation of the structure of Chorismate mutase.
Source: http://upload.wikimedia.org/wikipedia/commons/6/60/Chorismate-mutase-pdb- 2CHS.png

This paper arrived at a similarly low prevalence, giving a value of 1 in 10^{24}. This protein is much smaller in size, however, and possesses a much simpler fold than the TEM β-lactamase described above. When this value is adjusted to reflect a residue of the same length as the 150-residue section analysed from β-lactamase, it yields a result of 1 in 10^{53}.

A further study examined two α-helical regions of the λ repressor, and reported on the "the high level of degeneracy in the information that specifies a particular protein fold," concluding that although "there should be about 10^{57} different allowed sequences for the entire 92-residue domain," nonetheless "the estimated number of sequences capable of adopting the lambdha repressor fold is still an exceedingly small fraction, about one in 10^{63}, of the total number of possible 92-residue sequences. A similar result has been obtained for cytochrome c based on phylogenetic sequence comparisons," (Reidhaar-olson and Sauer, 1990). Another paper documented that more than a million million random sequences were required in order to stumble upon a functioning modestly-sized ATP-binding domain (Keefe and Szostak, 2001).

In addition, a review paper published in 2009 reported that "[t]he accepted paradigm that proteins can tolerate nearly any amino acid substitution has been replaced by the view that the deleterious effects of mutations, and especially their tendency to undermine the thermodynamic and kinetic stability of protein, is a major constraint on protein evolvability—the ability of proteins to acquire changes in sequence and function," (Tokuriki and Tawfik, 2009).

More recent work has shown that even a seemingly trivial switch from the function Kbl_2 (which is involved in the metabolism of the amino acid threonine) to the function of BioF (which is involved in the biotin synthesis pathway) requires at least seven co-ordinated mutations (Gauger and Axe, 2011), putting the transition well beyond the reach of a Darwinian process within the time allowed by the age of the earth (Axe, 2010a). Other studies have yielded similar results. For example, one study reported that "Interchanging reactions catalyzed by members of mechanistically diverse superfamilies might be envisioned as 'easy' exercises in (re)design: if Nature did it, why can't we? … Anecdotally, many attempts at interchanging activities in mechanistically diverse superfamilies have since been attempted, but few successes have been realized," (Gerlt and Babbitt, 2009).

The proteins examined in the aforementioned studies are relatively small relative to many much larger ones, and – as with English sentences – the prevalence of functional sequences declines as sequence-length increases.

For an excellent critical review paper on this subject, I refer readers to Douglas Axe's 2010 paper, "The Case Against the Darwinian Origin of Protein Folds" (Axe, 2010b).

Conclusions

The "sampling problem", described above, provides not only a serious challenge to neo-Darwinian evolution. It also offers a compelling positive argument for intelligent design. In all of our experience, only one type of cause – one category of explanation – is known to be causally sufficient to produce the complex specified information (CSI) necessary for the origin of novel protein folds. That cause is conscious deliberative activity – also known as intelligence. What we are increasingly finding is that evolving novel functions involves discontinuous 'jumps' in complexity, which cannot be scaled by a blind search.

Literature Cited

Abraham, E., & Chain, E. (1940). An enzyme from bacteria able to destroy penicillin. *Nature*, *46*(3713), 837-837.

Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. (1961). The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *Proceedings of the National Academy of Sciences*, *47*(9), 1309-1314.

Axe, D. D. (2004). Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds. *Journal of Molecular Biology*, *341*(5), 1295-1315.

Axe, D. D. (2010a). The Limits of Complex Adaptation: An Analysis Based on a Simple Model of Structured Bacterial Populations. *Bio-Complexity*, *2010*(4), 1-10.

Axe, D. D. (2010b). The Case Against a Darwinian Origin of Protein Folds. *Bio-Complexity*, *2010*(1), 1-12.

Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, *33*(10), 3390-3400.

Cordes, M. H. J., Davidsont, A. R., & Sauer, R. T. (1996). Sequence space, folding and protein design. *Current Opinion in Structural Biology*, *6*(1), 3-10.

Dembski, W. (1998). *The Design Inference Eliminating Chance Through Small Probabilities.* Cambridge University Press.

Dembski, W. (2001). *No Free Lunch : Why Specified Complexity Cannot Be Purchased Without Intelligence*. Rowman & Littlefield.

Denton, M. (1986). *Evolution: A Theory in Crisis* (pp. 309-311). Adler & Adler.

Gauger, A. K., & Axe, D. D. (2011). The Evolutionary Accessibility of New Enzyme Functions : A Case Study from the Biotin Pathway. *Bio-Complexity*, *2011*(1), 1-17.

Gerlt, J. A., & Babbitt, P. C. (2009). Enzyme (re) design : lessons from natural evolution and computation. *Current Opinion in Chemical Biology*, *13*(1), 10-18.

Keefe, A. D., & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, *410*, 715-718.

Labeit, S., Barlow, D. P., Gautel, M., Gibson, T., Holt, J., Hsieh, C. L., Francke, U., et al. (1990). A regular pattern of two types of 100-residue motif in the sequence of titin. *Nature*, *345*, 273-276.

Labeit, S., & Kolmerer, B. (1995). Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science*, *270*(5234), 293-296.

Majiduddin, F. K., Materon, I. C., & Palzkill, T. G. (2002). Molecular analysis of beta-lactamase structure and function. *International Journal of Medical Microbiology*, *292*(2), 127-137.

Meyer, S. (2009). *Signature in the Cell: DNA and the Evidence for Intelligent Design*. HarperOne.

Reidhaar-olson, J. F., & Sauer, R. T. (1990). Functionally Acceptable Substitutions in Two a-Helical Regions of λ Repressor. *PROTEINS: Structure, Function and Genetics*, *7*, 306-316.

Taylor, S. V., Walter, K. U., Kast, P., & Hilvert, D. (2001). Searching sequence space for protein catalysts. *Proceedings of the National Academy of Sciences*, *98*(19), 11-16.

Tokuriki, N., & Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology*, *19*(5), 596-604.