

Origins of Ancient Life – Genomic Research Proposal

This is a summary background paper that outlines the basic science and logic supporting the concept of conducting a search for the origins of life program by examining details within a variety of terrestrial genomes. The essence of the argument for conducting this Genomic research program arises from the combination of several reasonable (some would say likely) assumptions.

First is that there are reasonable scientific grounds to support the origin of life predating the formation of the solar system, and indeed, the seeding of life in the solar system and Earth itself from the pre-solar reservoir. This idea, with long historic roots, is referred to as panspermia, and it is currently supported in several versions from soft (or pseudo) through hard panspermia. Soft panspermia argues that the organic building blocks of life originated in space and were incorporated in the solar nebula from which the planets condensed and were further (and continuously) distributed to planetary surfaces where life then emerged. The current conventional dogma is that Earth was the prime, if not the sole location for this genesis. Hard panspermia advocates that not only organic building blocks (various complex organics including amino acids) but also the "design" elements in the form of DNA and even bacteria or bacterial spores, were (and some believe still are) deposited on Earth. [Note: while no serious advocate of panspermia argues against natural selection and Darwinian evolution as the dominant factors shaping the diversity of life on Earth, some argue that at least some sudden major evolutionary jumps may well have been triggered by continuing incorporation of genetic material from other sources.] It is the potential validity of hard panspermia that comprises one element of the justification for a search for the origins of ancient life based on genomic analyses.

The second element underpinning the origins of ancient life genomic analysis program is the recent sequencing of the human genome, including the current and rapidly evolving technology for genomic sequencing in general. The proportion of the human genome dedicated to genes, *per se*, is estimated at between 1.5 and 5% of the total. While in all likelihood there are many essential functions yet to be discovered in the generation of living organisms that lie outside traditionally recognized coding sequences, especially within the introns, there is clearly a vast reservoir of genomic material whose function and/or significance is entirely unknown. What is known from the sequencing is that lateral (or horizontal) gene transfer (or more generally, DNA transfer) has occurred within human history, even from bacteria. Indeed, there are several examples of bacterial genes having been directly incorporated into the human genome with little or no alterations. Clearly genomic paleontological exploration is indicated to understand both the extent and the mechanics of historic DNA transfer into the human genome. The second element supporting a genomic search for extraterrestrial life is the potential of the genomic paleontological record for containing historic data preceding the emergence of life on Earth.

The final argument supporting a program in the origins of ancient life is in the form of a "Gedank" experiment. First one posits the existence, over 5 billion years ago, of a DNA-based, highly advanced intelligent and technological culture somewhere in our galaxy. The second premise is that intelligent life, wherever it may exist, is ultimately interested in legacy, especially if it perceives imminent extinction. Assuming the two conditions above, the question is, what form, or forms, might that legacy initiative take? The

Brian M. Sager and Russell L. Schweickart

argument for a Genomic search for origins of ancient life life program is that one possible form for such a legacy would be to robustly encapsulate DNA, perhaps in a variety of forms, and broadcast it physically out into the galactic medium. While the nature of such broadcast mechanisms is itself an interesting topic for discussion, it is clear from current astronomical knowledge that supernovae explosions are a major mechanism for spreading heavy elements throughout the galaxy. Successfully hitching a ride on such a vehicle would require robust encapsulation, but this is an engineering challenge, not a conceptual one.

Two fundamental implications arise: First, this DNA legacy might have been incorporated into the human genome relatively intact, and therefore contained within that fraction of the human genome not encoding for specific genes. If so, the discovery of a potentially non-terrestrial genetic legacy would indicate a high likelihood for the existence of non-terrestrial life.

Second, might not such an advanced civilization, having "donated" its evolutionary legacy to the future in the form of broadcast DNA, not also have left within this vast information reservoir, a "signature" or message of some kind? Recall that on all spacecraft destined to leave the solar system we humans have thus far placed a variety of plaques, records, diagrams, and other greetings to whomever or whatever might intercept them in some distant future.

The proposal is to search the genomic record of Earthly life forms for the possible inclusion of such a non-terrestrial genomic encoding scheme, and perhaps detect within that scheme a signature or message, either implicit or explicit. Given that "classic" SETI has spent millions of dollars listening for radio messages from space, and that it continues to be supported despite having received no signals to date, it would seem that for a genomic search, having a "signal" already at hand, and much cheaper to search, would be a worthy and complimentary research effort.

A potential for hidden information content has already been shown in a linguistic analysis of the non-coding portion of the human genome. This preliminary research suggested that this "silent" DNA may harbor a property shared by all written languages, namely that particular non-coding regions of the human genome nevertheless follow Zipf's law, which characterizes a descending logarithmic curve describing a correlation between the frequency with which a word is used, and the utility ranking it holds in a set of all words (1). That analysis shows that the "silent" DNA resembles a language when the linguistic analyses are run on short DNA segments whose putative codon reading framework ranges from three to eight nucleotides in length.

Accordingly, there are several possible approaches in which to frame a search for non-terrestrial genetic coding. Under the fundamental assumption that non-terrestrial life utilized similarly composed DNA as the sole basis for storing genetic information, it is possible that the three-nucleotide length, four-nucleotide base codon set is conserved between human and potentially non-terrestrial genetic codes. In this scenario, the earth-standard amino acid to codon relationships may or may not have been preserved, and it would be prudent to search for latent genetic information using a variety of amino acid to codon relationships. Second, if the three-nucleotide length, four-nucleotide base codon set is not conserved between human and potentially non-terrestrial genetic codes, it would be appropriate to search for latent genetic information in potential codons whose

Brian M. Sager and Russell L. Schweickart

basic word length may utilize four, five, six, seven, or eight nucleotides. This search range would correspond with the information content suggested by the Zipf's law analyses. In both cases, it would be important to search for information in all potential reading frames, in both the "forward" and "reverse" directions. Precedents for highly compressed information content reside in the compact genomes of terrestrial viruses, whose genomes often encode for genes in multiple reading frames and in both reading directions. It is possible that at least some compacted information content may similarly reside in the human genome.

The recent discovery of a surprising preponderance of L-biased amino acids found in extraterrestrial objects such as the Murchison meteorite lends support to the idea of a relatively complex genetic code. Stereochemically-biased amino acid chirality is both inherently anti-entropic and, in the L-form, a signature for all terrestrial-based life. Assuming that all such L-biased amino acids relate to a corresponding codon in a non-terrestrial genetic framework, the number of codons required for this full set of L-biased amino acids would exceed 64, the maximum permitted by a three-nucleotide length, four-nucleotide base codon set (e.g. 4^3). Given this fundamental capacity limitation, the potential existence of a four-nucleotide length genetic code is suggested (e.g. 4^4).

Alternatively, a three-nucleotide length codon set could code for more than sixty four amino acids if more than four nucleotide bases were present in a non-terrestrial genetic code (e.g. 5^3).

Since this form of genomic search would involve using advanced analytic and cryptographic techniques in deep analysis of those elements of the human genome that are the least understood, it seems highly likely that much valuable knowledge about genomics will be gained even if no non-terrestrial genetic framework or signal is found. Further, the cost for implementing this analytic strategy is relatively small, given the ubiquity of vast amounts of freely and publicly available DNA sequence information. The question then comes down to, simply, "Why not?"

(1) Mantegna, R.N., S.V. Buldryev, A.L. Goldberger, S. Havlin, Peng, C.-K., Simons M. and H.E. Stanley. 1994. Linguistic Features of Noncoding DNA Sequences. *Physical Review Letters*. **73** (23), pp. 3169-3172.