LIFE'S LIMITED EXPLORATION OF PROTEIN SPACE. T. D. Ely[1]
[1]School of Earth and Space Exploration Arizona State University PSF-209 550 E. Tyler Mall Tempe, AZ 85287, USA

Introduction: The functional complexity of life is a direct result of the diversity of the informational structures it is built upon (DNA → Amino Acids (AA) → Proteins). Life, however, appears to use but a tiny fraction of the potential diversity implied by the structure at each level, specifically the ordering of the 20 common AA in protein sequences. For example, all possible AA combinations (known as *protein space*) to a length x is 20^x. This is an immense potential when considered the average lengths of proteins in extant species (5,000 proteins to an average length 300AA for prokaryotes and 20,000 unique proteins to an average length 300 AA for Eukaryotes) [1]. Although it is highly unlikely that many of the potential sequences within protein space are useful, it is likely that there exist many orders of magnitude more than is currently displayed by life.

Life's exploration of protein space: Limitations as to the usefulness of any given combination of amino acids is mostly the result of life's necessity for proteins with well-defined tertiary geometries. These geometries are required in order to carry out specific functions among specific reactants. Stable, beneficial geometries are the result of the thermodynamic energy minima (large $-\Delta G$) associated with a transition from an unfolded state to a folded, or, native state ($U \Leftrightarrow N$). This can present life with difficulties as environments shift beneath them. A protein capable of some specified function in one environment is either useless or greatly reduced in its activity in another. This, in addition or resource competition, would have driven primitive life on Earth to seek out novel functional proteins in protein space better adapted to their conditions at any given time.

The Model: A python-based model is currently being constructed that quantifies the dynamics of primitive replicators constrained to limited proteomes in time and space, as they respond to the evolutionary pressures associated with environmental change. The initial proteome utilized by the model is merely a list of required functions that will serve as a *jumping off point* from which duplication and mutation can create novel sequences. These functions are meant to mimic capabilities believed to be essential the early replicative precursors thought to precede the Last Unified Common Ancestor (LUCA) [2].

The big idea here is simple that, given typical rates of duplication and mutation which generate novel AA sequences within a population under selective pressure, the total inventory of functionally useful proteins within protein space can be explored. This exploration has a characteristic rate on geologic timescales, and, once determined, will allow for a quantitative assessment of the proficiency with which life seeks out the useful complexities of its biochemistry.

References: [1] Tiessen, A., Perez-Rodriguez, P., and Delaye-Arredondo, L. J. (2012) *BMC Res. Notes. 85.* [2] Wolf-Watz, M., et al. (2004) *Nat. Struct. Mol. Bio.*, *11,* 945-949. [2] Glansdroff, N., et al. (2008) *Biol. Dir. 29.*