Amino Acid Alphabet Distribution Unveils Nature's Laws of Protein Design

Davinder Kaur Dhalla

Department of Chemistry & Biochemistry, University of Lethbridge, AB, Email: d.kaur@uleth.ca

Proteins are made up of a common set of 20 amino acids known as the standard amino acid alphabet (AAA). Conventional protein engineering approaches for applications in biofuel production, pharmaceuticals and gene therapy generally utilize a design space limited to this standard AAA, thus restricting the incorporation of unnatural amino acids and the novel chemical functions that can be harnessed from them. Reducing the standard AAA is an approach that could free up the codon space for this and accelerate computational protein design. A functional protein with a reduced AAA will cost less for its synthesis *in vivo* and *ex vivo*. Previously-reported reduced AAA proteins (RAPs) have shown little to no activity relative to their wild-type counterparts. We predict that this is due to the over-reliance on substitution rules based on the physico-chemical properties of amino acids, neglecting the importance of protein dynamics on structure and function. The aim of our research is to develop a generalizable computational approach to design RAPs to facilitate efficient forward-engineering of proteins. Towards that aim, we first investigate the AAA of the proteins found in nature to understand the amino acid composition and preference in different classes of proteins from various domains of life. This includes performing scripting and data mining on the protein sequences available in protein sequence databases. Further downstream analysis and binning of the data shows interesting design principles and preferences followed by nature for different classes of proteins. These design preference will be utilized in the rational design of RAPs in the next step, where we combine *in silico* approaches and *in vitro* studies to investigate and validate the dynamic and functional properties of RAPs. In our two-pronged approach of RAP design, we utilize both the distinct conservation scores of residues within a protein and the physico-chemical properties of amino acids for designing RAPs. Finally, the developed novel computational framework will enable prediction and design of reduced alphabet proteins, providing powerful tools to other researchers for protein engineering. These tools will also have applications in understanding and modifying dynamic properties of disease-causing protein variants (e.g. HRAS), widening the accessibility of individualized therapies and personalized medications.