# THE EFFECT OF GENETIC CODE EVOLUTION ON PROTEIN STRUCTURE UNIVERSE.

V. Tretyachenko[1,2], J. Vymětal[2], J. Vondrášek[2], and K. Hlouchová[1,2]
[1]Faculty of Science, Charles University, Hlavova 8, Prague 2, Czech Republic (email address: klara.hlouchova@natur.cuni.cz)
[2]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo n. 2, 16610 Prague, Czech Republic.

**Introduction:** Biological evolution produced a system by which genetic instructions are decoded into protein products using a library of just 20 amino acids. This canonical amino acid alphabet has remained largely unchanged ever since it arose, more than 3 billion years ago. Diverse lines of evidence indicate that earlier life built proteins using a smaller alphabet of approximately 10 members, and that other amino acids were available for use from different potential prebiotic sources while the remaining canonical amino acids developed through pathways of biochemical synthesis, after the genetic code originated [1,2]. The question of why evolution "chose" the specific set of 20 canonical amino acids has been explored from several angles. Simple factors, such as prebiotic abundance, solubility, ease of biosynthesis, racemization stability, un/reactivity with tRNA and potential peptide product stability seem to explain some "choices" but not others [2].

Here we report a study of the relationship between the evolution of the amino acid alphabet, and the consequent repertoire of protein structures. The Uniprot database contains more than 50 million sequences of predicted proteins while the protein structure database (PDB) contains around 100 thousand protein structures that represent roughly 1300 structural fold families. These numbers are negligible compared to the astronomical sequence space of non-existing polypeptides that can be formed using the 20 genetically coded amino acids. These thoughts have led to assumptions that structured proteins are extremely rare in the protein sequence universe. In order to investigate this relationship and extend it to contexts of the genetic code evolution, we analysed random polypeptide sequences for the occurrence of secondary structure formation. We generated libraries of 10^6 random sequences of 100 residues from (i) the canonical amino acid alphabet (following the composition statistics of PDB-deposited proteins), (ii) the hypothetical early amino acid alphabet (Gly, Ala, Asp, Glu, Val, Ile, Leu, Pro, Thr, Ser), (iii) randomly selected 10 amino acids from the canonical alphabet, and (iv) 10 amino acids that best represent the repertoire of physicochemical properties of the canonical alphabet (Leu, Cys, Ala, Gly, Ser, Pro, Phe, Glu, Lys, His). Two control libraries were constructed from 100 residues protein fragments from PDB-deposited proteins and Disprot database of protein disorder. The content of secondary structure was predicted using the following secondary structure predictors: GOR4, Jnet, Predator, Simpa, and Psipred; disorder predictors: Disopred, DisEmbl, VSL2 and IUpred; and empirical indexes predicting solubility: CVsol, and Gravy. In addition to bioinformatical analyses, several peptides were selected from the random libraries for experimental characterization and verification of prediction accuracy.

Surprisingly, the occurrence of secondary structure within our random libraries from canonical amino acid alphabet is only slightly smaller (approximately 40%) than within the control library of structurally characterized proteins (approximately 50%). Our preliminary data also suggest that similar extent of secondary structure can be formed from early amino acids and that the early alphabet has a shifted potential to occupy alpha helices versus beta sheet motifs when compared with the canonical alphabet.

It is still not clear how different amino acid repertoires would affect the scope of protein structures/functions. Besides the implications for genetic code evolution and life's origin studied here, this area touches upon the very basic link of protein sequence-structure-function that lies at the core of many biotechnological and biomedicine problems.

Reference list:
[1] Zaia DA, Zaia CT, De Santana H. Which amino acids should be used in prebiotic chemistry studies? (2008) *Orig Life Evol Biosph.*, 38(6), 469-88.
[2] Freeland SJ. "Terrestrial" amino acids and their evolution. (2009) In: *Amino Acids, Peptides and Proteins in Organic Chemistry. Vol.1 – Origins and Synthesis of Amino Acids.* Hughes AB ed., Wiley-VCH, 43-75.