SEARCH INTO THE UNEVOLVED PROTEIN SEQUENCE SPACE.

Tretyachenko Vyacheslav[1,2] and Hlouchova Klara[1,2]

[1] *Charles University in Prague, Faculty of Science, Department of Biochemistry, Hlavova 2030, 128 00 Praha 2;*
[2] *Institute of Organic Chemistry and Biochemistry, Flemingovo nám. 2, 166 10 Praha 6*

klara.hlouchova@natur.cuni.cz

The protein world as we know it today is a result of more than 3 billion years of evolutionary processes. The sequence space that current proteins occupy is negligible when compared to the vast sequence space that can be obtained by random combinations of the 20 proteinogenic amino acids and that Nature could have been sampling throughout the history. A majority of functinonal proteins own their properties to specific secondary/tertiary structures. Is that a rare evolved property or is the unevolved sequence space also potent with secondary/tertiary protein structure? What biophysical properties have shaped proteins to evolve into the efficient functional biomolecules and stand out from the vast space of possible alternatives?

To address these questions, we performed a systematic computational and experimental exploration of the canonical amino acid alphabet structural consequences. A library (10^4) composed of 100 amino acid long sequences was generated *in silico* and occurrence of secondary structure was evaluated by 5 different prediction algorithms and compared to properties of natural proteins. Two experimental approaches were then employed: (i) scarce-sampling of the library, where 3x15 candidate proteins were selected based on the predicted properties (high, low or random secondary structure occurrence) and characterized individually; (ii) high-throughput analysis of a synthetic library that mimicked the properties of the *in silico* library. Because stemming from identical input parameters, all the outcomes of this study can be directly compared. Therefore, we simultaneously provide a test of the prediction algorithm accuracy when applied to unevolved sequence space.

Our study suggests that there is no significant difference between the secondary structure formation potential between extant proteins and unevolved (*i.e.* random) sequences. The overall distribution of secondary structure properties is however broader than for natural proteins and reflects work of evolutionary constraints. Evolution seems to have played a significant role in optimization of solubility and anti-aggregation properties. Overall, this implies that even the earliest polypeptides could be sufficiently equipped to form structured scaffolds.