ABSTRACT

To understand protein structure emergence is to comprehend the evolutionary transition from messy chemistry to the first heritable molecular systems. Early proteins were probably flexible in structure, promiscuous in activity and ambiguous in sequence. Moreover, first sequences were presumably composed of prebiotically plausible amino acids from endogenous and exogenous sources which form only a subset of the extant protein alphabet. Here we investigate the effect of most recent additions to the amino acid alphabet on protein structure/function relationship and the properties of random proteins as the evolutionary point-zero for the earliest sequences as well as for proteins emerging *de novo* from the non-coding parts of the genome. Random or never born proteins are of a special interest for the contemporary biology as they unveil the unexposed side of the protein sequence space. We constructed an *in silico* library of random proteins with the natural amino acid alphabet, analyzed its structure/disorder/aggregation content and selected 45 sequences for subsequent experimental preparation and biophysical characterization. We observed that structure content in random sequence space does not differ significantly from the natural proteins. However, the analyses of the aggregation propensity showed a significant level of optimization in natural protein space. Experimental characterization led to the surprising discovery of random disordered proteins being the most tolerated sequences upon the *in vivo* expression. Next, we designed a high throughput pipeline for experimental library preparation with proteins composed either of canonical 20 amino acids as well as of prebiotically plausible set of 10 amino acids. In order to implement this design experimentally we built CoLiDe – COmbinatorial Library Design tool based on degenerate codon composition optimization. We designed the libraries using CoLiDE, prepared them in a cell free expression system, and tested their properties by means of chaperone interaction analysis and selective proteolysis. Preliminary results suggest structure formation in prebiotic amino acid library and higher disorder content in canonical amino acid library of random proteins. Subsequently, as a case study we analyzed structure and function of contemporary protein dephospho coenzyme A kinase upon substitution of its aromatic amino acids by their prebiotically plausible counterparts. This analysis showed that protein function can be maintained in the absence of aromatic amino acids although structure is inevitably destabilized. Moreover, we observe significant structural changes upon ligand binding in aromatic-less mutants foreshadowing the essential effects of ancient cofactors on early protein stabilization.

Overall, this thesis represents one of the first windows into properties of evolutionary early proteins, with respect to prebiotically plausible amino acids. Its results imply that even proteins composed of prebiotically early amino acids have structural and functional propensities and could play an important role in the early biosphere.