Bornberglab
Molecular Evolution and Bioinformatics

In silico evolution of newly emerging proteins

Background: New protein coding genes frequently evolve from existing protein coding genes, via various mechanisms. Recent studies have shown that protein coding genes can also emerge "*de novo*" from genomic regions that did not previously encode any gene. It has been shown that *de novo* genes exist in a continuum, such that the properties of *de novo* genes are a function of their evolutionary age. For example, old *de novo* genes more closely resemble conserved protein coding genes in various aspects such as protein structure and composition. For most conserved proteins, their folding into a defined 3D structure is critical for their function. Therefore *de novo* genes that encode proteins that fold well, are likely to be functional, than genes that encode proteins with poor foldability. Furthermore, a *de novo* gene can gradually evolve such that it encodes proteins with better foldability.

Objective: The aim of the project is to understand how quickly random protein sequences evolve structures. To this end, the student will simulate protein evolution computationally (*in silico*) as summarized in the following points:

1. Choose a set of protein coding sequences which will include: random sequences generated computationally, known intergenic open reading frames, and putative *de novo* genes
2. Introduce random mutations in the DNA sequences at different mutation rates: 1, 2 and more than 2 mutations per iteration. The probability of these mutations will depend on mutation bias (for e.g. A→G mutations are 10 times more likely than A→C mutations)
3. Predict the structure of the translated protein sequences (post mutation) using structure prediction algorithms like ESMfold and OmegaFold. Also predict disorder using flDPnn.
4. Simulate evolutionary selection of mutants that have (a) higher "structure score", or (b) lower disorder. The mutants that have these properties will have more "offspring" in the next iteration, than other mutants.
5. Repeat steps 2 – 4 until structures become more apparent.
6. Estimate which kind of protein folds emerge more frequently.

Methods: Computational protein structure prediction, programming

Requirements: Basic programming experience, basic knowledge of protein biochemistry, Basic bioinformatics

Supervision: Dr. Bharat Ravi, b.ravi@uni-muenster.de, Molecular Evolution and Bioinformatics Group (https://bornberglab.org/)

Relevant literature:

- Van Oss SB, Carvunis AR. PLoS Genet.,2019
- I. Anishchenko *et al*., *Nature*, 2021
- Z. Lin et al., *Science*, 2023 (ESMfold)