# A statistical model for codon optimization

D. Luna Cerralbo[1,2], I. Blasco Machn[3], E. Broset[3], J. Martinez[3] and P. Bruscolini[1,2]
[1]Department of Theoretical Physics, University of Zaragoza.
[2]Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza.
[3]Department of RNA Design, Certest Pharma.

The degeneracy of the codon alphabet allows different codons to translate to the same amino acid, and it is well-known that different species show different statistics of codon usage. However, the reasons why a species adopts a particular statistics are not clear, even if protein yield, production speed, and RNA stability are believed to play a role in the choice. In this context, codon optimization involves adjusting the codon sequence for a target protein, to mimic the natural choice a given species would make, to produce that protein. However, conventional methods used for codon optimization are often simplistic (e.g., resorting just to the importance of each codon), or phenomenological, using the observed average frequency of codon pairs as input, instead of obtaining it as a result. Using large databases of human proteins, we propose a statistical-physics model, where the probability of any codon sequence is related to the interactions between neighboring codons. We have adjusted the model's parameters to maximize the dataset's probability. We have applied the method to the case of Luciferase, as a simple test protein, optimizing the codon sequence by Simulated Annealing, and comparing the results to those obtained by conventional methods.

---

[1] Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13)

[2] National Center for Biotechnology Information (NCBI) `https://www.ncbi.nlm.nih.gov/`

[3] ViennaRNA Package 2.0 doi:10.1186/1748-7188-6-26