Rationale for 'EPC codon optimization'

Gene synthesis is available from many suppliers nowadays and is often more effective and affordable than 'traditional' cloning approaches. This raises the question of how to deal with codon usage. Commercial suppliers of gene synthesis usually offer their own algorithms for codon optimization of genes for a limited number of species. Given that this optimization process requires information of the species (usually in the form of a drop-down menu), codon optimization likely relies on species-specific codon bias. These algorithms certainly use additional criteria for 'gene optimization'. This probably includes the avoidance of repeats and secondary structures so synthesis is feasible, as well as user-defined constraints such as lack of certain restriction sites. Besides the obvious codon bias, suppliers generally remain rather secretive about the 'biological' criteria that are taken into consideration. Some suppliers state that they try to avoid ribosome binding site (RBS)-like sequences, which would cause ribosome pausing/stalling or translation of a truncated gene if positioned in the right context of a nearby start codon (ATG, GTG).

Importantly, codon optimization using these tools only 'optimizes' the gene for the one selected organism. However, it might be desirable to use the gene in several different organisms which differ substantially in codon usage; for example, a fluorescent protein. This means that using common 'codon optimization' tools, one has to create a species-specific protein for each organism one is working with. An alternative to 'codon optimization' of a gene for a single species would be to 'generally' codon-optimize it to a number of select species by selecting codons that are favored in all species. In addition, from a practical point of view, some codons optimized for high-GC bacteria might be toxic in *Escherichia coli* and thus fail to yield the intended construct during cloning. E.g. AAG (Lys) is much rarer in *E. coli* compared to *Pseudomonas aeruginosa*, and TTC is used >95% for Phe in *Pseudomonas*, whereas in *E. coli* the usage of TTC and TTT for Phe is roughly equal. Similarly, ATC is used in 91% of cases for Ile in *Pseudomonas*, but roughly equally with ATT in *E. coli*.

Here, we analyze codon usage of three bacteria: *Escherichia coli*, *Pseudomonas aeruginosa* and *Caulobacter crescentus*. We provide a simple Python-based script that 'universally codon-optimizes' ORFs taking into consideration shared codon bias of all three species, and prioritizing *Pseudomonas*. We refer to this optimization method as 'EPC codon optimization', according to the three organisms it is optimized for. The code is simple and intuitive: codon usage can be adjusted by simply varying the numbers of each codon allowed for one given amino acid.

Of note, although GTG is a favorable codon for Val in all three organisms, we refrained from using it since it might serve as a start codon, although this, of course, heavily depends on the context (RBS-like sequence present). As a stop codon we use a 'double codon' (TGATAA) since the opal codon is strongly enriched in high-GC bacteria, but ochre is clearly preferred in *E. coli*.

In general, the code doesn't take into consideration anything but shared codon usage, so it is only the first step to design the desired gene, and most often it is necessary to modify the sequence to meet the individual requirements, e.g. eliminate unwanted restriction sites. Especially for ATG encoding Met internally we recommend to carefully check the 20 bp upstream and, if necessary, diverge the sequence away from a possible RBS (AGGAGG, anything AG-rich). GTG can be used for Val if no RBS-like sequence is upstream. And for many amino acids alternative codons are not extremely rare, but simply not favored.

The accompanying file 'epc_codon_freq_summary.csv' can give some guidance towards which alternative codons can be used in addition to the 'EPC-optimal' codons. Codon usage frequencies for *P. aeruginosa*

PAO1, *C. vibrioides* NA1000, *E. coli* MG1655 were calculated from data retrieved from the Codon Statistics Database (http://codonstatsdb.unr.edu/)[1].

References

1. Subramanian, K., Payne, B., Feyertag, F. & Alvarez-Ponce, D. The Codon Statistics Database: A Database of Codon Usage Bias. *Mol Biol Evol* **39**, msac157 (2022).

By

Anne Francez-Charlot
Andreas Kaczmarczyk

2024