Technology Offer

Novel protein expression optimization method combining machine learning with insights from mechanistic modelling of mRNA translation

Ref.-No.: 0401-5011-LC-WA

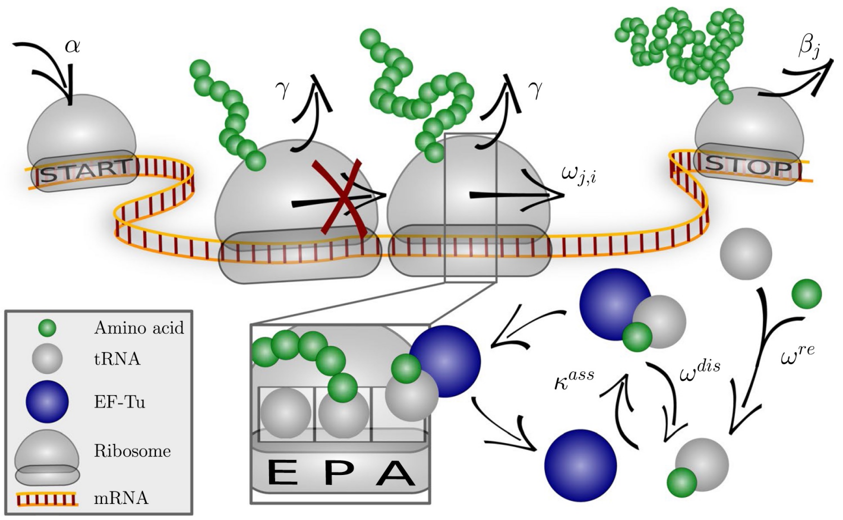
Researchers of the Max-Planck-Institute for Colloids and Interfaces and the Paul-Ehrlich-Institute (PEI) have developed a novel codon-specific elongation model (COSEM), that is based on the question how codon bias affects protein expression in combination with a deep understanding of protein synthesis.

Background

Todays standard codon optimization procedures, that are implemented in software tools such as GeneOptimizer, JCat, Optimizer, Synthetic Gene Designer, Codon Optimization OnLine (COOL) and EuGene, are based on codon adaptation to biases seen in highly expressed genes. This purely heuristic approach does not provide a deeper understanding of the underlying processes and does not answer the question of optimality in a context-dependent and mechanistic manner. As a consequence, these heuristic methods repeatedly cause unexpected or suboptimal outcomes, which triggered the search for further heuristic covariates such as length of genes, GC3 content and more complex mRNA sequence motifs as well as mRNA secondary structure. A codon-specific elongation model (COSEM), that is based on the question how codon bias affects protein expression in combination with a deep understanding of protein synthesis, offers new opportunities to overcome the limitations of the heuristic approaches.

Technology

Researchers of the Max-Planck-Institute for Colloids and Interfaces and the Paul-Ehrlich-Institute (PEI) have developed a mathematical model which allows more accurate forecasts and improved output in the biotechnology-based protein synthesis in host organism. Their codon-specific elongation model (COSEM) makes predictions about the translation rate per mRNA transcript (COSEM current) and is sketched in Fig. 1:



The translation process is initiated by ribosome attachment to the mRNA sequence j with the initiation rate α. Subsequently, ribosomes translate the mRNA with codon-specific elongation rates ωj,i, where i labels the codon position on the codon sequence j. Finally, ribosomes finish translation with the termination rate βj, corresponding to the elongation rate of the last codon, or leave the mRNA with the drop-off rate γ before reaching the last codon. When several ribosomes translate the same mRNA sequence, they cannot overtake each other. Furthermore, COSEM takes into account that each ribosome covers several codons and that each codon can be covered by only one ribosome at a time, where we take the ribosomal footprint to have a size of ten codons.

COSEM's codon-specific elongation rates ωj,i for *E. coli, S.cerevisiae* and the human HEK293 cell line were calculated from a detailed Markov model reflecting the current biochemical knowledge of translation elongation. In particular, the elongation rates depend on the concentrations of cognate, near-cognate, and non-cognate tRNAs and their competitive binding to the ribosomes. Machine learning approaches are used to integrate predictions made by COSEM on the translation rate per mRNA transcript (COSEM current) with additional covariates into the protein expression score which is used to predict protein expression.

Fig 2 shows protein abundances predicted by this protein expression score in comparison with measured protein abundances in *E. coli, S. cerevisiae* and HEK293 cells using protein and transcript abundance data from public databases.

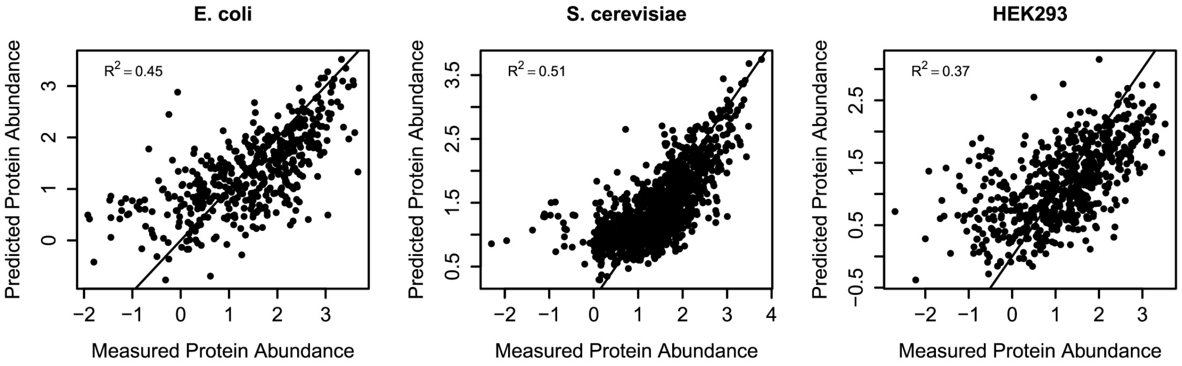


Figure 2: Comparing measured protein abundance with predicted protein expression for all *E. coli, S.cerevisiae*, and HEK293 genes where proteome data are available. The coefficients of determination R^2 for E. coli, S. cerevisiae, and HEK293 are 0.45 (95% CI 0.39–0.52), 0.51 (95% CI 0.47–0.54) and 0.37 (95% CI 0.32–0.43), and the number of coding sequences are 1563; 4479; 2136, respectively. Measured protein abundances are log-transformed values from PaxDb database's common abundance metric in ppm, for *E. coli* and *S. cerevisiae* there is a noticeable cutoff at 0 caused by a lower resolution limit of the measurement methods used.

It was shown that COSEM current, i.e., the translation rate per mRNA transcript, in combination with transcript abundance has already high predictive power for protein expression which confirms the relevance of COSEM current in combination with mRNA levels for understanding total protein expression.

COSEM can naturally be adapted to alternative target organisms based on their tRNA pools for exploratory analyses. The full algorithm can then be trained and validated for these new organisms with a dataset of measured protein expression levels.

We expect that the understanding of protein expression in codon optimization schemes will substantially improve the current state of the art in the field. The presented tools have in particular the potential to advance the design of precisely tailored genes for a wide range of applications in synthetic biology.

Literature

S. Rudorf, R. Lipowsky: "Protein Synthesis in E. coli: Dependence of Codon-Specific Elongation on tRNA Concentration and Codon Usage", PLoS ONE 10(8): e0134994. doi:10.1371/journal.pone.0134994

J.-H. Trösemeier, S. Rudorf, H. Loessner, B. Hofner, A. Reuter, T. Schulenborg, I. Koch, I. Bekeredjian-Ding, R. Lipowsky, Ch. Kamp: "Optimizing the dynamics of protein expression", Scientific Reports (2019) 9:7511

Patent Information

EP priority application filed 07.12.2016.
PCT application WO2018/104385A1 filed 06.12.2017, nationalized in EP and US.

Contact

Dr Lars Cuypers

Senior Patent- & License Manager
Chemist

Phone: +49 (0)89 / 29 09 19 - 21
eMail: cuypers@max-planck-innovation.de