[BIOINFO 2025]

Multi-factor mRNA coding sequence optimization for enhanced RNA stability and protein expression

Sohyeon Ju[1, 2], Chae Young Kwon[1], and Hyeshik Chang[1,2*]

[1] *Center for RNA Research, Institute for Basic Science (IBS), Seoul National University*
[2] *School of Biological Sciences, Seoul National University*
** Corresponding author: hyeshik@snu.ac.kr*

mRNA therapeutics hold strong potential, but effective coding sequence design is required to overcome RNA instability, prevent excessive innate immune activation, and improve translation efficiency. We present VaxPress, a multi-factor codon optimizer that uses a modified genetic algorithm to iteratively maximize a linear combination function of dozens of fitness functions representing biological and biochemical properties of sequences. The scoring functions aim to increase RNA stability through highly folded structures, favor frequently used codons and codon pairs, decrease uridine content, and reduce repetitive sequences and extreme local GC content. We also incorporated two predictive scores from existing models (iCodon and DegScore) designed to improve RNA stability. To validate whether this multi-factor optimization strategy improves protein expression, we used VaxPress to design 18 nanoluciferase coding sequences with various weight combinations. For comparison, we selected eight previously reported sequences representing high, moderate, and low expression levels. We transfected all 26 mRNAs into HCT116 cells and quantified protein expression after 24 hours by measuring nanoluciferase signal intensity. The top VaxPress sequence achieved a 3.6-fold increase in expression compared to the reference sequence and a 1.6-fold increase over the best-performing benchmark sequence. Factor-wise analysis revealed that metrics indicating weakly folded structures and uridine content correlated negatively with expression (Pearson's $r \approx -0.16$ to -0.20 and $r = -0.27$, respectively), while codon adaptation index correlated positively ($r = 0.12$). These correlations were weaker but consistent with previous research. Notably, stability prediction models showed either limited correlation (DegScore: $r = -0.32$) or unexpected inverse correlation (iCodon-predicted stability: $r = -0.52$) with cellular expression. These results demonstrate that while biological and biochemical properties of coding sequences profoundly affect protein expression, their relationships are complex. VaxPress enables multi-factor design space

screening by integrating various scoring objectives and has successfully designed nanoluciferase sequences with enhanced cellular expression.