Data-driven codon optimization in *Saccharomyces cerevisiae*

Alexey A. Gritsenko[1,3,4], Frank Koopman[2,3,4], Marcel J.T. Reinders[1,3,4,5], Jean-Marc Daran[2,3,4] and Dick de Ridder[1,3,4,5]

[1]Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands; [2]Industrial Microbiology Group, Dept. of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, The Netherlands; [3]Platform Green Synthetic Biology, P.O. Box 5057, 2600 GA Delft, The Netherlands; [4]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, The Netherlands and [5]Netherlands Bioinformatics Centre, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

Email: a.gritsenko@tudelft.nl

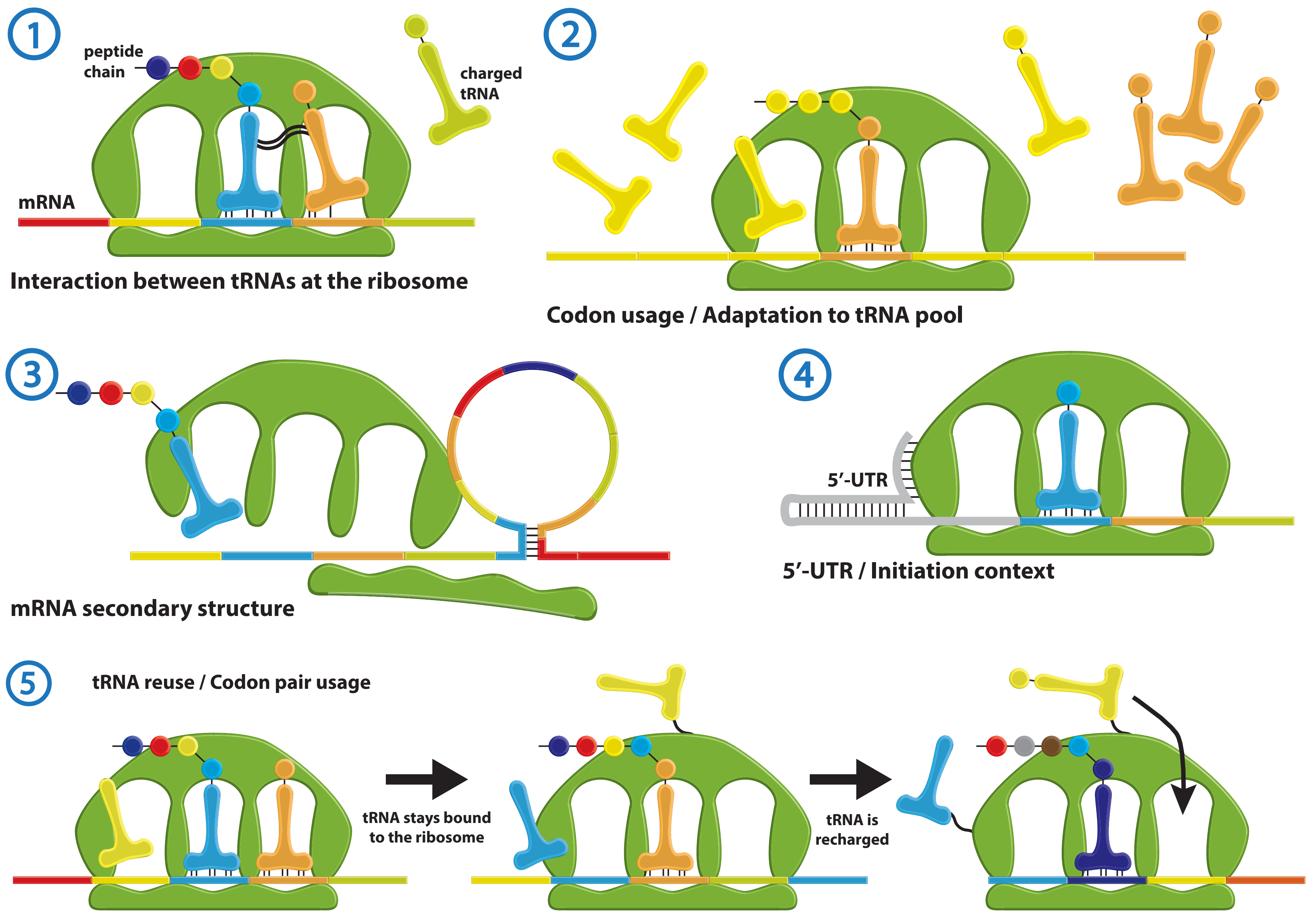
Choice of codons influences gene translation rates

Most of the 20 amino acids are encoded by 2 to 6 synonymous codons, which are rarely used at equal frequencies. Biased choice of synonymous codons is more evident in highly expressed genes, which can exclusive use a single codon per amino acid. This observation is often employed to prove that synonymous codons are not translated at equal rates.

The difference in translation rates between organisms presents a challenge for heterologous protein expression, when a donor gene can have little or no expression in a new host organism. The process of adapting gene's sequence for (more) efficient translation, called *codon optimization*, is instrumental for heterologous expression of single genes or complete pathways.

Multiple mechanisms determine translation rates

Numerous hypothetical mechanisms (①-⑤) have been suggested for explaining the observed difference in translation rates between synonymous gene encodings. However, none of the mechanisms are completely understood and little is known about interactions between different mechanisms, making it difficult to rationally adjust gene translation.



① Interaction between tRNAs at the ribosome

② Codon usage / Adaptation to tRNA pool

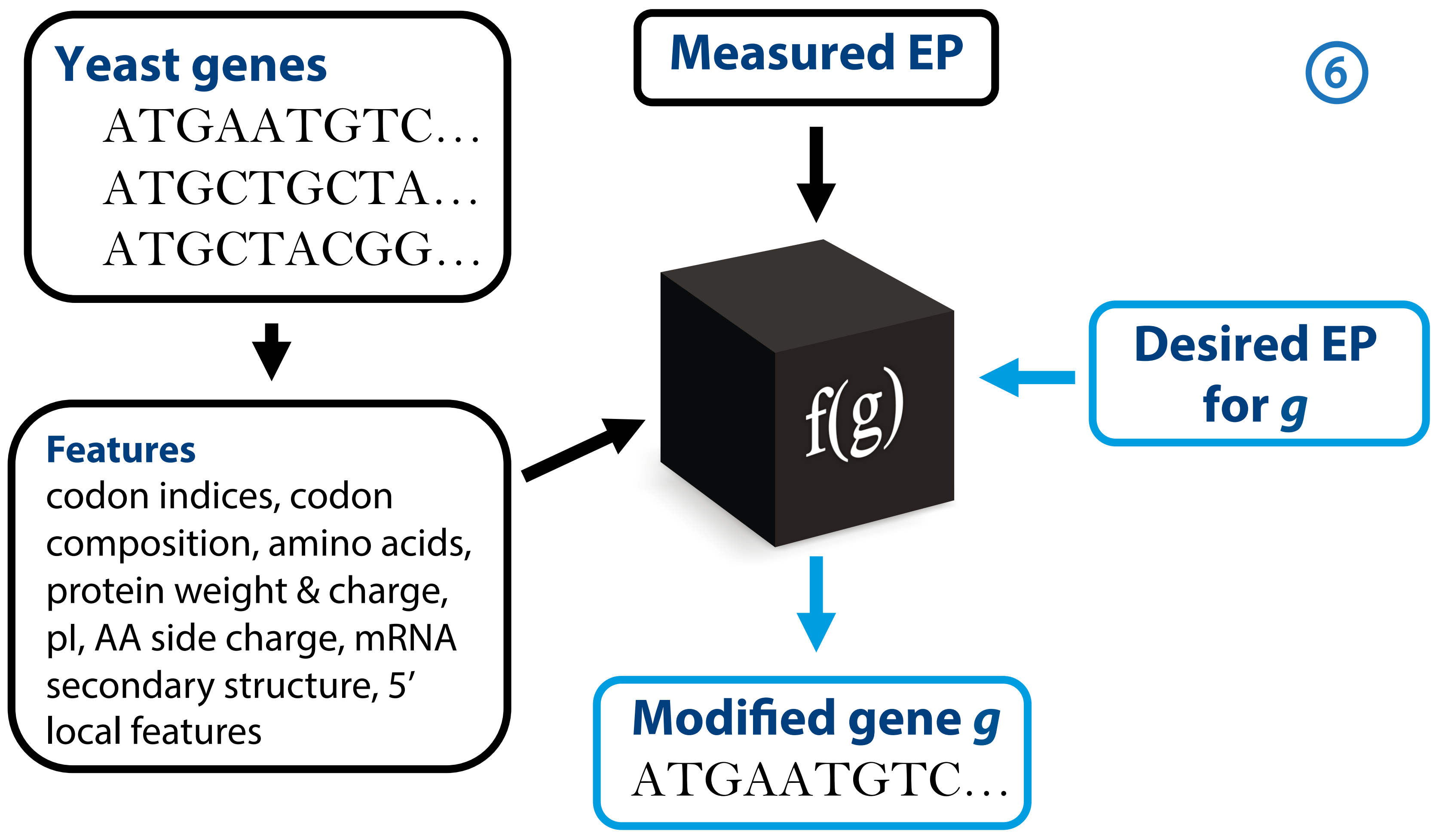
③ mRNA secondary structure

④ 5'-UTR / Initiation context

⑤ tRNA reuse / Codon pair usage

Sequence-based features can be used for predicting translation rates

Gene sequence-based features (such as CAI [1], tAI [2] or TPI [3]) have been devised for mechanisms ①-⑤ and have been shown to correlate with mRNA or protein levels. These correlations, capturing linear relationships between individual features and gene expression, have been used for codon optimization. However, such an approach does not take the potential interactions between different features (mechanisms) and cannot explain their combined effect on translation rates.



Yeast genes
ATGAATGTC…
ATGCTGCTA…
ATGCTACGG…

Features
codon indices, codon composition, amino acids, protein weight & charge, pI, AA side charge, mRNA secondary structure, 5' local features

Measured EP

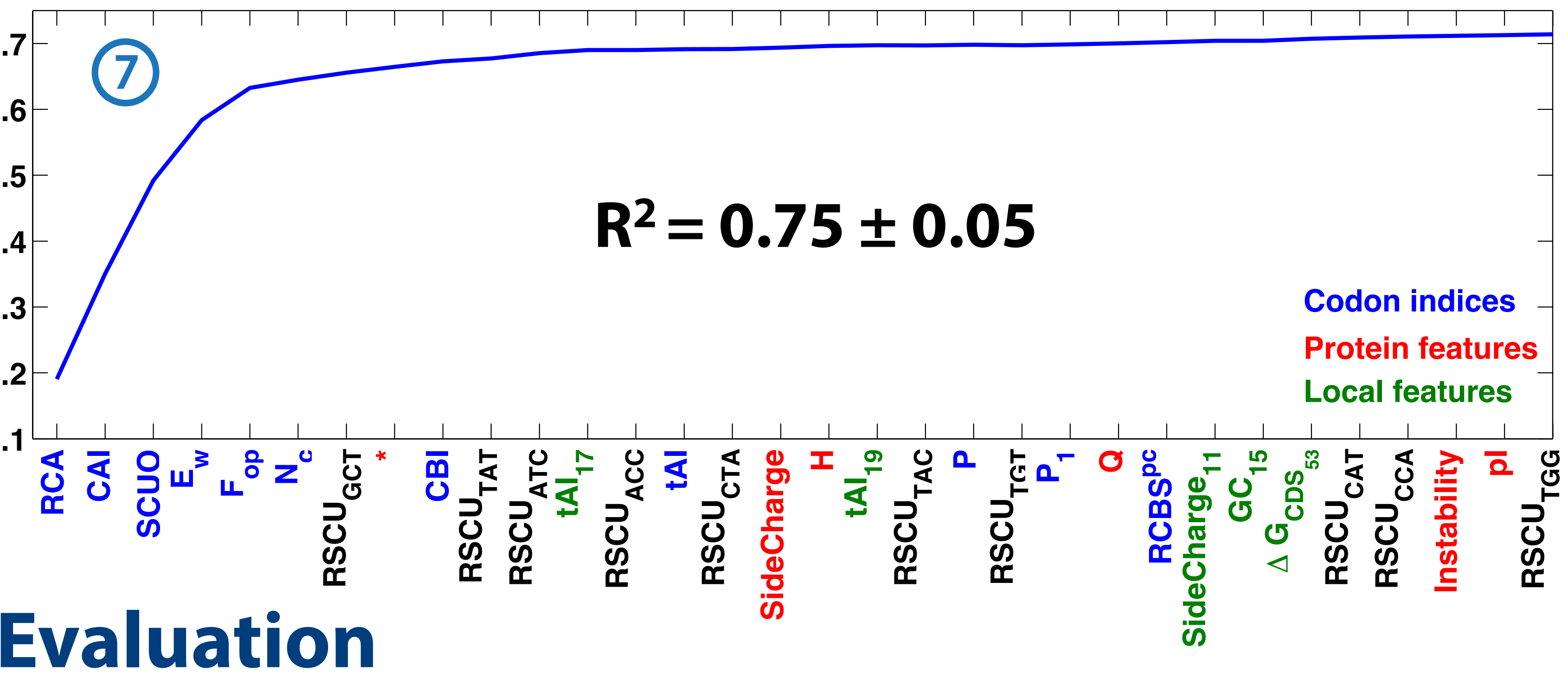
Desired EP for *g*

Modified gene *g*
ATGAATGTC…

We use a non-linear regression approach for learning a predictor of *Expected amount of Protein* (EP) in *S.cerevisiae*, defined as mRNA levels x Ribosome density [4], from multiple sequence features. We then invert this predictor to see which codon substitutions increase the prediction ⑥.

Predictor training and evaluation

Support Vector Regression (ν-SVR) and backward feature selection have been used to train a predictor with R^2=0.75±0.05 (10-fold CV) and 79 features (see ⑦ for the first 30 features). The regression combines features describing multiple translation mechanisms in a single predictor, allowing to make predictions about translation rates without understanding the underlying mechanisms.



Evaluation

We synonymously changed sequences of the PAL1 and 4CL genes involved in plant flavonoid biosynthesis such that they would give maximum predicted EP ⑧,⑨. We also codon optimized them using JCat [5].

⑧ **4CL**

Type	Codons changed	EP, folds
cDNA	0	1
JCat	338	1035
Optimized EP	344	1976

⑨ **PAL1**

Type	Codons changed	EP, folds
cDNA	0	1
JCat	414	403
Optimized EP	438	855

In silico predictions suggest an achievable 2-fold increase in translation rate compared to genes codon optimized using an existing method. These results are currently being verified in the lab by measuring mRNA levels, protein levels and enzymatic activity of original and optimized PAL1 and 4CL versions.

References

[1] Sharp & Li *NAR* (1987); [2] dos Reis *et al. NAR* (2003); [3] Friberg *et al. LNBI* (2003); [4] Ingolia *et al. Science* (2009); [5] Grote *et al. NAR* (2005).