microbial biotechnology

Open Access

Crystal Ball

The new strategies to overcome challenges in protein production in bacteria

Anna Lipońska,[1] Farès Ousalem,[1] Daniel P. Aalberts,[2] John F. Hunt[3] and Grégory Boël[1,*]

[1] *Institut de Biologie Physico-Chimique, Sorbonne Paris Cité, UMR 8261 CNRS-University Paris Diderot, 13 rue Pierre et Marie Curie, 75005 Paris, France.*

[2] *Physics Department, Williams College, Williamstown, MA 01267, USA.*

[3] *Department of Biological Sciences, Columbia University, New York, NY 10027, USA.*

Introduction

Protein production has been of great interest to industry for a long time: first for the food industry and household products, then bio-production, now for medicine and biotech, tomorrow for the development of synthetic biology and protein nanomachines. Nowadays, market demand for proteins not only concerns chemical and food industries, but also pharmaceuticals (Palomares *et al*., 2002). From the time of the first commercialized pharmaceutical recombinant protein, human insulin (Gentech/Eli Lilly in 1982), the protein therapeutics market has been steadily increasing. From 2011 to 2016, 62 new biologics were approved by the FDA (Lagassé *et al*., 2017). Today, this production is centralized and large-scale, but in the future, small-scale manufacturing adapted to individual needs of smaller patient populations may become the standard (Crowell *et al*., 2018). An aim of the biological revolution will be to produce functional protein in a cost-efficient manner.

In past decades, most proteins were extracted from the living organisms that produce them. This process was time-consuming and resulted in low quantities of desired proteins. Along with science and biotechnology development, this problem was solved by heterologous protein overproduction in model organisms. The gene encoding the protein of interest is over-expressed in another organism than the native one, such as in bacteria, yeasts, insect and human cell lines, each with advantages and disadvantages. The bacteria *Escherichia coli* is widely used because it is less time-consuming and often more cost-efficient than other systems, moreover it benefits from all the knowledge, genetic tools and new methods of protein production optimization.

Protein expression is a complex task; the whole process from transcription to translation involves hundreds of components and many variables that are cross-correlated. Consequently, the optimization of the production can be performed by influencing different stages and changing different parameters. For the purpose of this short review, we focus mainly on optimization directly related to translation. We have divided this discussion into cis and trans-optimization. Cis-optimization concentrates on nucleotide sequence improvement whereas trans-optimization will focus on the use of the right, optimized bacterial strain.

Cis-optimization

Sequence optimization consists of designing a DNA sequence that is optimal for expressing a protein. The DNA sequence is transcribed into mRNA, which is the template for protein synthesis catalysed by the ribosome. The synthesis of the protein starts with the binding of the small subunit of the ribosome upstream of the coding sequence at the Shine and Dalgarno (SD) sequence. This initiation can be modulated by modifying the SD sequence complementarity with the ribosomal anti-SD sequence and its distance from the start codon (Schurr *et al*., 1993; Chen *et al*., 1994). The Salis group has developed an algorithm to optimize the SD site (Espah Borujeni *et al*., 2014).

Choosing the right codons – frequent doesn't mean better!

The ribosome reads the mRNA sequence in three base groups called codons. The first codon read is the initiator codon, of which AUG is the most efficient of three possibilities. Each codon encodes an amino acid with the

Received 29 October, 2018; accepted 30 October, 2018.
*For correspondence. E-mail: boel@ibpc.fr; Tel. +33 1 58 41 51 21; Fax +33 1 58 41 50 25.
Microbial Biotechnology (2019) **12**(1), 44–47
doi:10.1111/1751-7915.13338

exception of the three stop codons that signal the end of the message. It is important to remember that we have 61 codons and only 20 amino acids, so most amino acids are encoded by more than one codon (called synonymous codons). The frequencies of use for each codon are not equal, some of them occur more often (frequent codons), whereas others rarely (rare codons). Since in *E. coli* the most frequent codons are decoded by the most abundant tRNAs, this codon usage is considered to correlate with the availability of some tRNAs, the most limiting step in translation elongation (Ikemura, 1981). Logically, it has been postulated that rare codons translate slower and therefore reduce protein production.

Native *E. coli* proteins that are highly expressed often use frequent codons. Hence, codon metrics based on codon frequency have been used for optimization of poor genes with low expression. Ikemura calculated the frequency of optimal codons in a gene, but the most widely used metric was a Codon Adaptation Index (CAI), also based on codon usage (Sharp and Li, 1987). These conventional observations led to the concept that the more frequent codons are the 'good' ones whereas the rare codons the 'bad' ones.

Optimization based on codon usage became routine, but its success has been variable, suggesting that it is not a rational optimization method. Indeed, several studies have recently shown that rare codons are not systematically correlated with low expression (Goodman *et al.*, 2013; Boël *et al.*, 2016). The concept that tRNA concentration controls elongation speed under normal physiological conditions has been challenged by a variety of different sources. The failure to observe a significant global correlation with ribosome dwell time and tRNA concentration in any prokaryotic ribosome profiling experiment (Mohammad *et al.*, 2016; Aalberts *et al.*, 2017) could reflect technical limitations in those methods, but their failure to provide support for the traditional model resonates with the failure to observe significant correlations in a variety of other global profiling studies conducted using orthogonal methods (Goodman *et al.*, 2013; Boël *et al.*, 2016). Overexpression could create stress on the tRNA pool that makes cognate tRNA concentration important under those conditions (Makrides, 1996), but there is no evidence of a systematic correlation even for expression because codons with similarly low frequency and cognate tRNA concentration have divergent influences on protein overexpression level (Boël *et al.*, 2016).

We now see some intricate relations between codon usage and other central pathways like protein folding, mRNA degradation and transcription/translation coupling. Comprehension of those relationships will guide us towards more rational optimization strategies (Fig. 1A).

Some codons or codon combinations can stall ribosomes and thus reduce protein synthesis. This reduction can be exacerbated by the fact that stalled ribosomes can expose mRNA to RNases or/and actively recruit the RNase machinery (Fig. 1A) (Boël *et al.*, 2016; Hanson and Coller, 2017).

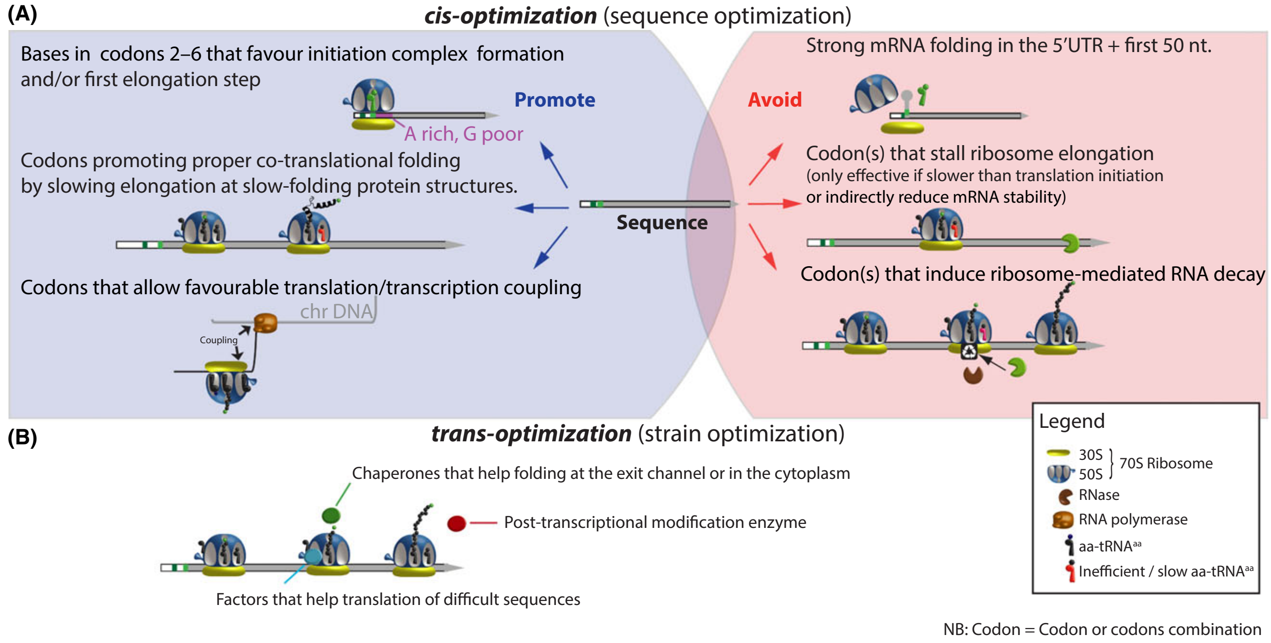
(A)
cis-optimization (sequence optimization)
Bases in codons 2–6 that favour initiation complex formation and/or first elongation step
Promote
A rich, G poor
Codons promoting proper co-translational folding by slowing elongation at slow-folding protein structures.
Codons that allow favourable translation/transcription coupling
chr DNA
Coupling
Sequence
Strong mRNA folding in the 5′UTR + first 50 nt.
Avoid
Codon(s) that stall ribosome elongation
(only effective if slower than translation initiation or indirectly reduce mRNA stability)
Codon(s) that induce ribosome-mediated RNA decay
(B)
trans-optimization (strain optimization)
Chaperones that help folding at the exit channel or in the cytoplasm
Post-transcriptional modification enzyme
Factors that help translation of difficult sequences
Legend
30S
50S
70S Ribosome
RNase
RNA polymerase
aa-tRNA^aa
Inefficient / slow aa-tRNA^aa
NB: Codon = Codon or codons combination

Fig. 1. Strategies to optimize protein production.
A. Factors to promote or to avoid for increased protein expression.
B. Cellular factors that can help specific proteins to be properly synthesized.

Codon usage may also influence protein folding introducing a context dependency to codon choice. A change to a synonymous 'faster' codon, which locally speeds up translation, may allow the nascent peptide chain to rapidly and negatively influence protein folding (Komar *et al.*, 1999). The use of some specific codons or codon combinations that slow the elongation process may allow pauses for the proper folding of the protein to occur. Sequence optimization based on harmonization of the codon frequency usage of the expression host to match the frequency used in the native host helps protein folding (Siller *et al.*, 2010; Buhr *et al.*, 2016). An evolutionist view of codon usage also shows that rare codons can be used to direct tRNA specificity during translation. Some rare codons are less prone to error than the frequent ones; therefore, they are more used to encode key amino acids of the protein (Drummond and Wilke, 2008). The challenge is to take all those parameters in account to generate the best sequence. The future will possibly be tailored optimization methods that account for protein specificity.

Choosing the right codons – mRNA folding and base composition effects

mRNA secondary structures in the 5′ untranslated transcribed region (UTR) of the mRNA and the beginning of the coding sequence strongly influence gene expression. Folding of the mRNA can prevent the binding of the ribosome small subunit to the SD (Geissmann *et al.*, 2009). Limiting the folding of this part of the mRNA is crucial for good sequence optimization. It has been shown that a higher amount of adenosine (A) in the first 18 nucleotides of the coding sequence increases the probability of higher protein expression, whereas G decreases it (U has an intermediate positive and C intermediate negative effect) (Boël *et al.*, 2016). A synonymous codon substitution makes many changes simultaneously: the codon usage frequency, the base composition, the mRNA folding. All have a strong impact on translation. With this taken into account, we have to use more accurate tools for sequence optimization, one that can integrate multiparameter optimization.

Transcription/translation coupling

In most biotechnological applications, protein expression in *E. coli* occurs by use of T7/IPTG system. IPTG induces synthesis of bacteriophage T7 RNA polymerase, which then can recognize the T7 promotor controlling expression of the desired protein. The T7 RNA polymerase is much faster than *E. coli* RNA polymerase; these kinetic differences limit the coupling of the translation with the transcription. Therefore, T7 RNA polymerase activity results in a mass production of mRNA, that is not protected by transcribing ribosomes, which occurs normally with the *E. coli* RNA polymerase (Iost and Dreyfus, 1995). Evolution of sequence optimization has to take those parameters into account. It is possible that the coupling with the RNA polymerase can be improved algorithmically in the future. In the case of the T7 expression, the best optimization may differ from the one used for *E. coli* endogenous RNA polymerase.

Trans-optimization

As discussed, cis-optimization methods can help expression of proteins, but to get the best results, these should be combined with trans-optimization methods. Optimization of growth media and temperature, the right concentration of inducer or use of protein fusion can play a big role as well. However, selecting of the right bacterial strain is particularly important to get the best results. When dealing with a protein prone to misfolding and aggregation, like membrane proteins, a strain co-expressing molecular chaperones can be used (Fig. 1B).

Proteins with disulphide bonds are difficult to express because bacterial cytoplasm is typically not suitable for sulphide bond formation; however, *E. coli* strains have been successfully engineered to oxidize cysteines in the cytoplasm (Anton *et al.*, 2016). Moreover, there are now *E. coli* strains that can perform post-translational modifications like N-glycolyzation, a modification generally occurring only in eukaryotic cells (Wacker *et al.*, 2002) or acetylation (Johnson *et al.*, 2010). These strains co-express heterologous enzymes that can catalyse those modifications. Another challenge is the expression of membrane proteins which can create toxicity during their overexpression and can be misfolded. This effect can be reduced by using *E. coli* strains that use a more reduced T7 RNA expression than regular ones (Angius *et al.*, 2018).

Recently identified translation factors assist the synthesis of sequences difficult to translate; for example, the factor Ef-P, which suppresses translation inhibition at poly-Proline stretches (Ude *et al.*, 2013). These factors and others that remain to be discovered can be overexpressed in specific strains to assist the synthesis of proteins that require their help. It is important to note that some trans-optimization could change the influence of synonymous codons, making it possible that cis-optimization and trans-optimization cannot be done independently.

The future of optimization will integrate all those parameters and will fine-tune them according to the nature of the protein to be synthesized. Translation speed will be encoded to facilitate protein folding, localization and post-translational modifications. This will be coupled with an expression strain adapted for the specific protein.

Acknowledgements

D.P.A. is supported by NIH grant GM106372. G.B. is supported by ANR grant EZOtrad, CNRS-UMR8261, Université Paris Diderot and LABEX program (DYNAMO ANR-11-LABX-0011).

References

Aalberts, D.P., Boël, G., and Hunt, J.F. (2017) Codon clarity or conundrum? *Cell Syst* **4:** 16–19.

Angius, F., Ilioaia, O., Amrani, A., Suisse, A., Rosset, L., Legrand, A., *et al.* (2018) A novel regulation mechanism of the T7 RNA polymerase based expression system improves overproduction and folding of membrane proteins. *Sci Rep* **8:** 8572.

Anton, B.P., Fomenkov, A., Raleigh, E.A., and Berkmen, M. (2016) Complete genome sequence of the engineered *Escherichia coli* SHuffle strains and their wild-type parents. *Genome Announc* **4:** e00230-16.

Boël, G., Letso, R., Neely, H., Price, W.N., Wong, K.-H., Su, M., *et al.* (2016) Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* **529:** 358–363.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., *et al.* (2016) Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol Cell* **61:** 341–351.

Chen, H., Bjerknes, M., Kumar, R., and Jay, E. (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* m RNAs. *Nucleic Acids Res* **22:** 4953–4957.

Crowell, L.E., Lu, A.E., Love, K.R., Stockdale, A., Timmick, S.M., Wu, D., *et al.* (2018) On-demand manufacturing of clinical-quality biopharmaceuticals. *Nat Biotechnol* **36:** 988–995.

Drummond, D.A., and Wilke, C.O. (2008) Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **134:** 341–352.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013) Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **342:** 475–479.

Espah Borujeni, A., Channarasappa, A.S., and Salis, H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res* **42:** 2646–2659.

Geissmann, T., Marzi, S., and Romby, P. (2009) The role of mRNA structure in translational control in bacteria. *RNA Biol* **6:** 153–160.

Hanson, G., and Coller, J. (2017) Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* **19:** 20–30.

Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* **151:** 389–409.

Iost, I., and Dreyfus, M. (1995) The stability of *Escherichia coli* lacZ mRNA depends upon the simultaneity of its synthesis and translation. *EMBO J* **14:** 3252–3261.

Johnson, M., Coulton, A.T., Geeves, M.A., and Mulvihill, D.P. (2010) Targeted amino-terminal acetylation of recombinant proteins in *E. coli*. *PLoS ONE* **5:** e15801.

Komar, A.A., Lesnik, T., and Reiss, C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett* **462:** 387–391.

Lagassé, H.A.D., Alexaki, A., Simhadri, V.L., Katagiri, N.H., Jankowski, W., Sauna, Z.E., and Kimchi-Sarfaty, C. (2017) Recent advances in (therapeutic protein) drug development. *F1000Research* **6:** 113.

Makrides, S.C. (1996) Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol Rev* **60:** 27.

Mohammad, F., Woolstenhulme, C.J., Green, R., and Buskirk, A.R. (2016) Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell Rep* **14:** 686–694.

Palomares, L.A., Kuri-Breña, F., and Ramírez, O.T. (2002) *Industrial Recombinant Protein Production*. Oxford, UK: EOLSS Publishers.

Schurr, T., Nadir, E., and Margalit, H. (1993) Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. *Nucleic Acids Res* **21:** 4019–4023.

Sharp, P.M., and Li, W.-H. (1987) The codon adaptation index- a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:** 1281–1295.

Siller, E., DeZwaan, D.C., Anderson, J.F., Freeman, B.C., and Barral, J.M. (2010) Slowing Bacterial Translation Speed Enhances Eukaryotic Protein Folding Efficiency. *J Mol Biol* **396:** 1310–1318.

Ude, S., Lassak, J., Starosta, A.L., Kraxenberger, T., Wilson, D.N., and Jung, K. (2013) Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science* **339:** 82–85.

Wacker, M., Linton, D., Hitchen, P.G., Nita-Lazar, M., Haslam, S.M., North, S.J., *et al.* (2002) N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science* **298:** 1790–1793.