# Improvement of translational efficiency by N-terminal codon optimization in the reconstituted cell-free protein synthesis system

## 再構成型無細胞タンパク質合成系（PURE*frex*®）におけるN末端コドン最適化による翻訳効率向上


○Rena Matsumoto, Satoshi Murakami and Takashi Kanamori (GeneFrontier Corporation)

○ 松本 令奈、村上 智史、金森 崇（ジーンフロンティア株式会社）

GeneFrontier

for more infomation, please contact us.
URL: www.genefrontier.com
E-mail: purefrex@genefrontier.com



## < Abstract >

大腸菌で組換えタンパク質を発現させる際、使用するコドンは同義コドンの中から高使用頻度のものを選択することが一般的である。一方、大腸菌タンパク質では、N末端部分（開始～数アミノ酸の範囲）にATリッチなコドンが多いものは発現レベルが高いといういくつかの報告がある[a,b,c]。しかし、これらの報告は、in vivoのデータを基に行われたものが多く、翻訳以外の反応（分解、代謝など）が最終的なタンパク質量に影響を与えている可能性がある。本研究では、大腸菌由来の再構成型無細胞タンパク質合成系 (PURE system) を用いることで翻訳以外の反応の影響を排除し、N末端のコドン選択がタンパク質合成量自体に与える影響、および低合成量タンパク質の合成効率向上に有効なコドン選択について検証した。

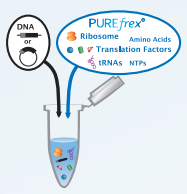
抗体断片（Fab）の重鎖などの元来合成量の低いタンパク質について、N末端部分のコドンを同義置換した種々の鋳型DNAを作製し、PURE systemを改良したPURE*frex*®を用いてタンパク質合成を行った。その結果、高使用頻度のコドンよりATリッチなコドンを選択した方が、タンパク質合成量が数倍増加する傾向がみられた。この結果は、N末端のATリッチなコドン配列が翻訳反応自体を促進することを示唆するものである。従って、合成量の低いタンパク質を効率よく合成するためには、無細胞合成系、大腸菌発現系のいずれにおいてもN末端にはATリッチなコドンを使用することが推奨される。



a) Allert *et al.* (2010) *J. Mol. Biol.*, vol.402, p. 905.
b) Bentele *et al.* (2013) *Mol. Sys. Biol.*, vol.9, p. 675.
c) Boel *et al.* (2016) *Nature*, vol.529, p.358.


## 1. PURE*frex*®; based on the PURE system technology

The PURE system is a reconstituted cell-free protein synthesis system, which consists of only purified factors necessary for transcription, translation and energy




**Advantage**
- Low level of contamination
- Easy adjustment of the reagent composition
- PCR products usable as a temlplate DNA

**Application**
- High throughput preparation of proteins (including Fab, scFv, protein toxin etc.)
- Protein science research
- in vitro display
  - Ribosome display
  - mRNA display

(Ref; Shimizu Y. *et al.* (2001) *Nat. Biotechnol.*, vol. 19, p. 751)

**PURE*frex*®1.0 • PURE*frex*®2.0**
a regular kit for the synthesis of proteins without disulfide bonds

**DS supplement**
a supplement for the synthesis of proteins containing disulfide bonds

**DnaK Mix • GroE Mix**
a supplement for the synthesis of aggregate-prone proteins

## 2. Experiments

| Protein | Reaction Mixture | Template DNA | Incubation |
|---|---|---|---|
| 1. Fab Heavy chain (Herceptin ) 2. scFv (HyHEL10) | PURE*frex* 2.0 + DS supplement + DnaK Mix | PCR product (20 ng/μL reaction) | 30°C, 4 h |
| 3. α-Synuclein 4. human hematopoietic prostaglandin D synthase (hPTGDS) | PURE*frex* 2.0 + DS supplement | PCR product (20 ng/μL reaction) | 37°C, 4 h |
| 5. Citrate Synthase | PURE*frex* 2.0 | PCR product (50 ng/μL reaction) | 30°C, 4 h |
| 6. Luciferase | PURE*frex* 2.0 + DnaK Mix | PCR product (50 ng/μL reaction) | 30°C, 4 h |

Protein Synthesis
↓
SDS-PAGE (0.5 μL of reaction/lane)
↓
Gel stain with Oriole (Bio-Rad)
↓
Analysis with LAS4000 (FujiFilm)

## 3. Results 1: N-terminal codon and protein synthesis

### 3-1. Fab Heavy chain (Herceptin)

Herceptin Fab Hc

Met-Glu-Val-Gln-Leu-Val- … FLAG

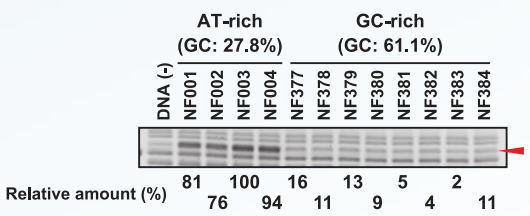
| 2 | | 3 | | 4 | | 5 | | 6 | |
|---|---|---|---|---|---|---|---|---|---|
| Glu (E) | | Val (V) | | Gln (Q) | | Leu (L) | | Val (V) | |
| codon | freq (%) | codon | freq (%) | codon | freq (%) | codon | freq (%) | codon | freq (%) |
| gaa | 70 | gtt | 25 | caa | 30 | ttg | 15 | gtt | 25 |
| gag | 30 | gtc | 18 | cag | 70 | tta | 12 | gtc | 18 |
| | | gta | 17 | | | ctt | 12 | gta | 17 |
| | | gtg | 40 | | | ctc | 10 | gtg | 40 |
| | | | | | | cta | 5 | | |
| | | | | | | ctg | 46 | | |

Frequency (E, Q) 65% 50% 35%
Frequency (V, L) 35% 25% 15%

*Frequency is calculated from Codon Usage Database in Kazusa DNA Res. Inst. (*E.coli* K-12 strain)

**All clones: 384**
**Tested clones: 56**

#### AT-rich codon vs GC-rich codon




↓

**AT-rich codon > GC-rich codon**

#### AT-rich codon vs the most frequently used codon

Summary (Herceptin Hc)

| ID | Glu | Val | Gln | Leu | Val | GC content[a] | ΔG[b] | Product amount[c] |
|---|---|---|---|---|---|---|---|---|
| NF001 | g a a g t t | c a a t t a | g t t | | | 27.8% | 0.00 | 81% |
| NF002 | g a a g t t | c a a t t a | g t a | | | 27.8% | 0.00 | 76% |
| **NF003** | **g a a g t a** | **c a a t t a** | **g t t** | | | **27.8%** | **0.00** | **100%** |
| NF004 | g a a g t a | c a a t t a | g t a | | | 27.8% | 0.00 | 94% |
| NF013 | g a a g t t | c a g t t a | g t t | | | 33.3% | 0.00 | 65% |
| NF014 | g a a g t t | c a g t t a | g t a | | | 33.3% | 0.00 | 81% |
| NF015 | g a a g t c | c a a t t a | g t t | | | 33.3% | -1.60 | 78% |
| NF016 | g a a g t c | c a a t t a | g t a | | | 33.3% | -1.60 | 74% |
| NF017 | g a a g t a | c a a t t a | g t c | | | 33.3% | 0.00 | 68% |
| NF018 | g a a g t a | c a a t t a | g t g | | | 33.3% | 0.00 | 62% |
| NF019 | g a a g t a | c a a t t g | g t t | | | 33.3% | -5.00 | 51% |
| NF021 | g a a g t a | c a a c t t | g t t | | | 33.3% | -0.80 | 56% |
| NF023 | g a a g t a | c a a c t a | g t t | | | 33.3% | 0.00 | 57% |
| NF025 | g a a g t a | c a g t t a | g t t | | | 33.3% | 0.00 | 71% |
| NF026 | g a a g t a | c a g t t a | g t a | | | 33.3% | -0.10 | 88% |
| NF027 | g a a g t g | c a a t t a | g t t | | | 33.3% | 0.00 | 85% |
| NF028 | g a a g t g | c a a t t a | g t a | | | 33.3% | 0.00 | 88% |
| NF029 | g a g g t t | c a a t t a | g t t | | | 33.3% | -1.60 | 43% |
| NF030 | g a g g t t | c a a t t a | g t a | | | 33.3% | -1.60 | 57% |
| NF031 | g a g g t a | c a a t t a | g t t | | | 33.3% | -3.00 | 63% |
| NF032 | g a g g t a | c a a t t a | g t a | | | 33.3% | -3.00 | 79% |
| NF065 | g a a g t a | c a a c t c | g t t | | | 38.9% | -2.60 | 63% |
| NF069 | g a a g t a | c a a c t g | g t t | | | 38.9% | -3.10 | 30% |
| NF072 | g a a g t a | c a g t t a | g t g | | | 38.9% | 0.00 | 63% |
| NF080 | g a a g t g | c a a t t a | g t g | | | 38.9% | 0.00 | 65% |
| NF101 | g a g g t a | c a a t t a | g t c | | | 38.9% | -3.00 | 51% |
| NF102 | g a g g t a | c a a t t a | g t g | | | 38.9% | -3.00 | 46% |
| NF103 | g a g g t a | c a a t t g | g t t | | | 38.9% | -5.00 | 27% |
| NF105 | g a g g t a | c a a c t t | g t t | | | 38.9% | -4.70 | 32% |
| NF107 | g a g g t a | c a a c t a | g t t | | | 38.9% | -3.00 | 44% |
| NF148 | g a a g t a | c a a c t g | g t g | | | 44.4% | -3.90 | 49% |
| NF170 | g a a g t g | c a g t t a | g t g | | | 44.4% | 0.00 | 32% |
| NF203 | g a g g t c | c a g t t a | g t t | | | 44.4% | -1.70 | 32% |
| NF209 | g a g g t a | c a a c t c | g t t | | | 44.4% | -6.50 | 28% |
| NF213 | g a g g t a | c a a c t g | g t t | | | 44.4% | -3.10 | 22% |
| NF236 | g a a g t t | c a g c t g | g t g | | | 50.0% | -4.30 | 19% |
| NF254 | g a a g t a | c a g c t g | g t g | | | 50.0% | -5.00 | 23% |
| NF258 | g a a g t g | c a a c t g | g t g | | | 50.0% | -3.60 | 39% |
| NF260 | g a a g t g | c a g t t g | g t g | | | 50.0% | -5.50 | 24% |
| NF262 | g a a g t g | c a g c t t | g t g | | | 50.0% | -0.50 | 19% |
| NF266 | g a a g t g | c a g c t a | g t g | | | 50.0% | 0.00 | 16% |
| NF267 | g a a g t g | c a g c t g | g t t | | | 50.0% | -3.90 | 13% |
| NF268 | g a a g t g | c a g c t g | g t a | | | 50.0% | -5.00 | 21% |
| NF336 | g a a g t c | c a g c t g | g t g | | | 55.6% | -7.10 | 14% |
| NF337 | g a a g t g | c a g c t c | g t c | | | 55.6% | -2.90 | 16% |
| NF338 | g a a g t g | c a g c t c | g t g | | | 55.6% | -2.90 | 23% |
| NF339 | g a a g t g | c a g c t g | g t c | | | 55.6% | -3.90 | 6% |
| **NF340** | **g a a g t g** | **c a g c t g** | **g t g** | | | **55.6%** | **-4.70** | **16%** |
| NF377 | g a g g t c | c a g c t c | g t c | | | 61.1% | -5.50 | 16% |
| NF378 | g a g g t c | c a g c t c | g t g | | | 61.1% | -5.50 | 11% |
| NF379 | g a g g t c | c a g c t g | g t c | | | 61.1% | -6.30 | 13% |
| NF380 | g a g g t c | c a g c t g | g t g | | | 61.1% | -7.10 | 9% |
| NF381 | g a g g t g | c a g c t c | g t c | | | 61.1% | -6.80 | 5% |
| NF382 | g a g g t g | c a g c t c | g t g | | | 61.1% | -6.80 | 4% |
| NF383 | g a g g t g | c a g c t g | g t c | | | 61.1% | -3.90 | 2% |
| NF384 | g a g g t g | c a g c t g | g t g | | | 61.1% | -4.70 | 11% |

AT-rich ↑ … ↓ GC-rich

**6-fold**

The most frequently codons

a) GC contents of N-terminus of ORF(1-6 amino acids)
b) ΔG (Minimum Free Energy; MFE) from Shine-Dalgarno sequence to N-terminus of ORF (1-6 amino acids) was calculated on RNAfold Web Server (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).
c) NF003 was used as a positive control of each experiments to calculate relative synthesized protein.
(Synthesized amount from NF003: **600-700 ng/μL reaction**)

| ID | Glu | Val | Gln | Leu | Val | GC content[a] |
|---|---|---|---|---|---|---|
| NF340 | g a a g t g | c a g c t g | g t g | | | 55.6% |
| NF170 | g a a g t g | c a g t t a | g t g | | | 44.4% |
| NF258 | g a a g t g | c a a c t g | g t g | | | 50.0% |
| NF148 | g a a g t a | c a a c t g | g t g | | | 44.4% |
| NF080 | g a a g t g | c a a t t a | g t g | | | 38.9% |
| NF072 | g a a g t a | c a g t t a | g t g | | | 38.9% |
| NF018 | g a a g t a | c a a t t a | g t g | | | 33.3% |
| **NF003** | **g a a g t a** | **c a a t t a** | **g t t** | | | **27.8%** |

Convert to AT-rich codon
The most frequently codons
+1
+2
+3
+4

a) ΔG (Minimum Free Energy; MFE) from Shine-Dalgarno sequence to N-terminus of ORF (1-6 amino acids) was calculated on RNAfold Web Server (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

Based on the NF340, we generated a constructs with increased AT rich codons one by one.

The most frequently codons
Convert to AT-rich codon

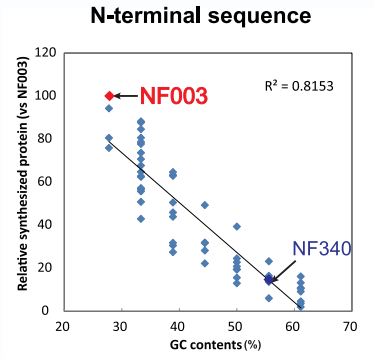
16 32 39 44 65 63 62 100
Relative amount (%)

**6-fold**

↓

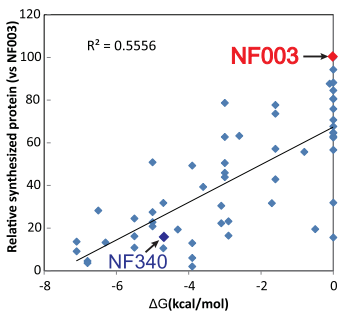
**AT-rich codon > the most frequently**

#### Relationship between protein synthesis and GC contents of N-terminal sequence




**There is a correlation between GC contents (1-6 aa) and yield of synthesized protein.**

#### Relationship between protein synthesis and mRNA secondary structure




**mRNA secondary structure (SD sequence + 1-6 aa) has a tendency to relate protein synthesis.**

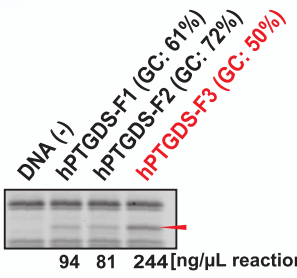
### 3-2. hPTGDS

hPTGDS ORF

Met-Ala-Pro-Glu-Ala-Gln-

| ID | Ala | Pro | Glu | Ala | Gln | GC content |
|---|---|---|---|---|---|---|
| hPTGDS-F1 | gca | ccg | gaa | gca | cag | 61% |
| hPTGDS-F2 | gcg | ccg | gaa | gcg | cag | 72% |
| hPTGDS-F3 | gca | cct | gaa | gct | caa | 50% |

Frequency (A, P) 35% 25% 15%
Frequency (E, Q) 65% 50% 35%




94 81 244 [ng/μL reaction]

**AT-rich codon in N-terminal sequence facilitates the protein synthesis in hPTGDS.**

## 4. Result 2: Additional sequence in N-terminus of ORF

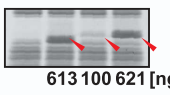
### 4-1. Codon optimization of Hisx6-tag

#### Conversion from GC-rich codon of Hisx6-tag to AT-rich codon

Met-Gly-Ser- (His)x6 -Gly-Ser- α-Synuclein

| Gly | Ser | His | His | His | His | His | His | Gly | Ser | GC content |
|---|---|---|---|---|---|---|---|---|---|---|
| GC-rich Hisx6-tag | | | | | | | | | | |
| ggc | agc | cac | cac | cac | cac | cac | cac | ggt | agc | 67% |
| AT-rich Hisx6-tag | | | | | | | | | | |
| ggt | tct | cat | cat | cat | cat | cat | cat | gga | agt | 39% |

Frequency (Gly, Ser) 35% 25% 15%
Frequency (His) 50% 30% 10%




613 100 621 [ng/μL reaction]

↓

**AT-rich tag > GC-rich tag**

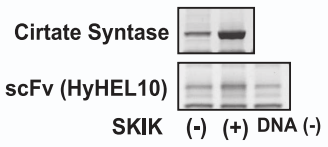
### 4-2. Effect of additional sequence of N-terminus to improve translational efficiency

Met-Ser-Lys-Ile-Lys (SKIK) | ORF

tct aaa ata aaa

AT-rich sequence

(Ref. Bivona et al.(2010), *Protein Expr. and Purif.* 74, p. 248)



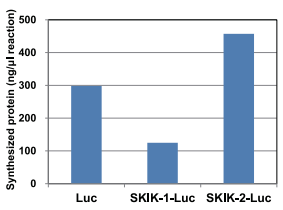

**SKIK sequence in N-terminus enhanced protein synthesis.**

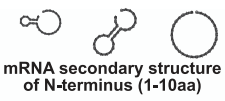
SKIK-Luciferase

| ID | Ser | Lys | Ile | Lys | Glu | Asp | Ala | Lys | Asn | ΔG[a] |
|---|---|---|---|---|---|---|---|---|---|---|
| Luc | | | | | gaa | gac | gcc | aaa | aac | -1.60 |
| SKIK-1-Luc | tct | aaa | atc | aaa | gaa | gac | gcc | aaa | aac | -2.10 |
| SKIK-2-Luc | agc | aaa | att | aaa | gaa | gac | gcc | aaa | aac | 0.00 |

a) ΔG (Minimum Free Energy; MFE) of N-terminal sequence (1-10 amino acids) was calculated on RNAfold Web Server (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

Frequency (Ser, Ire, Ala, Asn) 35% 25% 15%
Frequency (Lys, Glu, Asp) 65% 50% 35%



mRNA secondary structure of N-terminus (1-10aa)

**mRNA secondary structure:Rigid (additional sequence + N-terminus of ORF)**

↓

**The amount of protein synthesis: Low**

## 5. Conclusion

**In N-terminal sequence...**

- AT-rich codon facilitates the protein synthesis.
- No correlation between codon usage and protein synthesis.
- AT-rich additional sequence enhance the protein synthesis.

↓

**We recommend the AT-rich codons in N-terminal sequence for high protein expression level.**

### On-going

- Further analysis of Herceptin Hc codon optimization in N-terminal sequence
- Investigation of suitable additional sequence for high protein expression