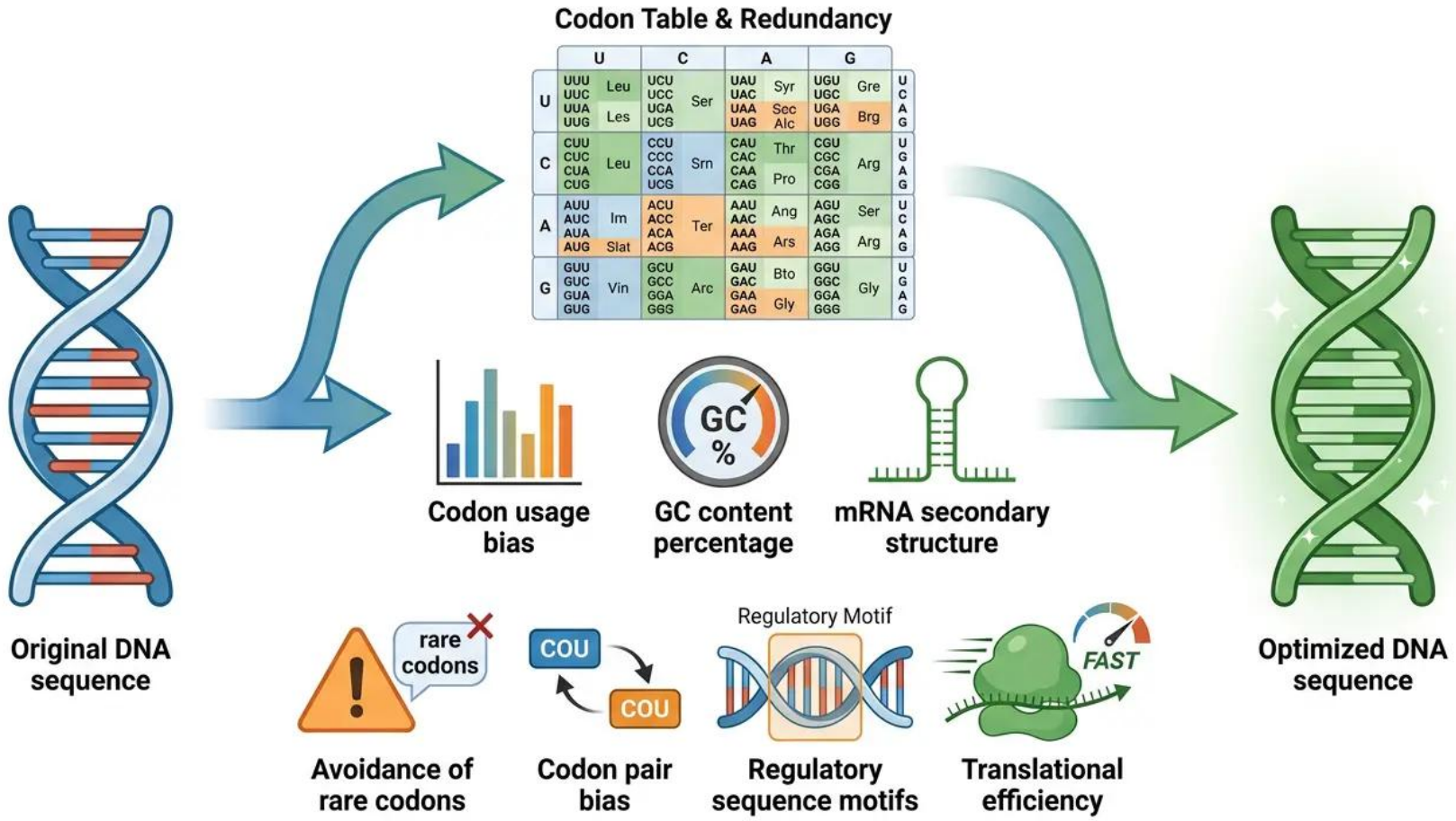
Codon optimization:
Why & how to design DNA sequences for optimal soluble protein expression

PARAMETRI CHIAVE NELL'OTTIMIZZAZIONE DEI CODONI PER LA SEQUENZA DI DNA

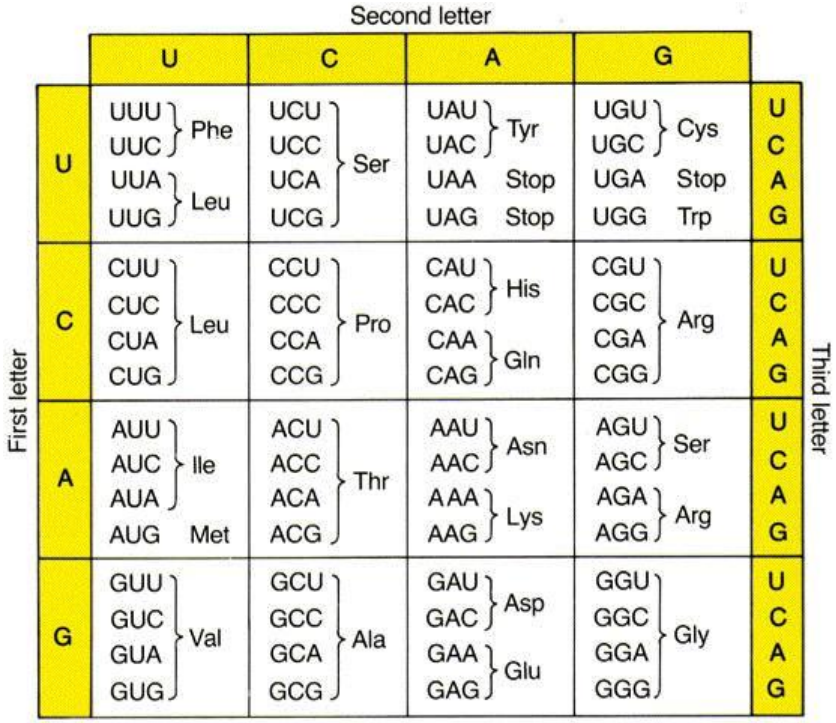
SEQUENZA DNA DI PARTENZA (WILD-TYPE) → **ALGORITMO DI OTTIMIZZAZIONE DEI CODONI** → **SEQUENZA DNA OTTIMIZZATA**



Because of the degeneracy of all genetic codes, **18/20 amino acids** are encoded by **more than one codon (2, 3, 4, or 6)**.

If synonymous codons are **strictly neutral**,

they should be used **randomly**



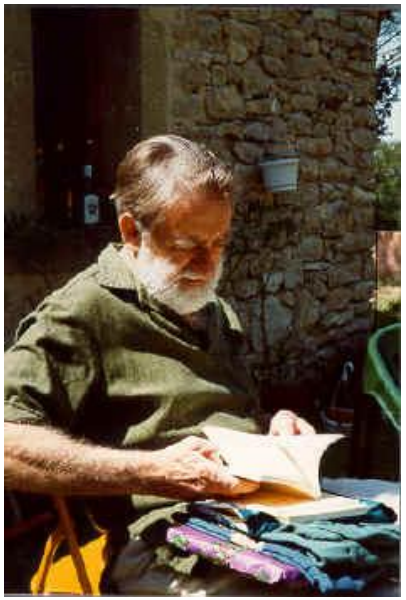
The genome hypothesis

- **All genes** in a genome tend to have the same coding strategy.

- That is, **they employ the codon catalog similarly** and show similar choices between synonymous codons.

- Different taxa have **different coding strategies.**

An example:
21 of the 23 leucine residues in the *E. coli* outer membrane protein II (ompA) are encoded by the codon CUG, although 5 other codons for leucine are available.



Richard Grantham

Amino Acid	Codon	*Escherichia coli*	
		High	Low
Leucine	UUA	1%	20%
	UUG	1%	15%
	CUU	2%	12%
	CUC	3%	11%
	CUA	1%	5%
	CUG	**92%**	37%
Valine	GUU	**60%**	27%
	GUC	2%	25%
	GUA	28%	16%
	GUG	10%	32%
Isoleucine	AUU	16%	46%
	AUC	**84%**	37%
	AUA	0%	17%
Phenylalanine	UUU	17%	67%
	UUC	**83%**	33%

Codon usage bias

intended as:
the non-random usage of synonymous codons in the protein translation process,
can be observed in virtually all organisms.
This phenomenon widely varies across different species and it is expected to significantly influence molecular genome evolution

One can locate **“optimal” codons** which are expected to be translated more efficiently than others.

We can define the **“codon bias”** of a specific gene, relative to the optimal codons…

Universal and species-specific patterns of codon usage

Universal patterns:

Codons that contain the **CG dinucleotide** are universally avoided (**low-usage codons**). This phenomenon is particularly notable as far as the arginine codons **CGA** and **CGG** are concerned.

Two selective factors have been convincingly invoked to explain codon usage bias.

(1) translation optimization

(2) **folding stability of the mRNA**

Gouy and Gautier (1982) and Bennetzen and Hall (1982) found **positive correlation between degree of codon bias and level of gene expression**.

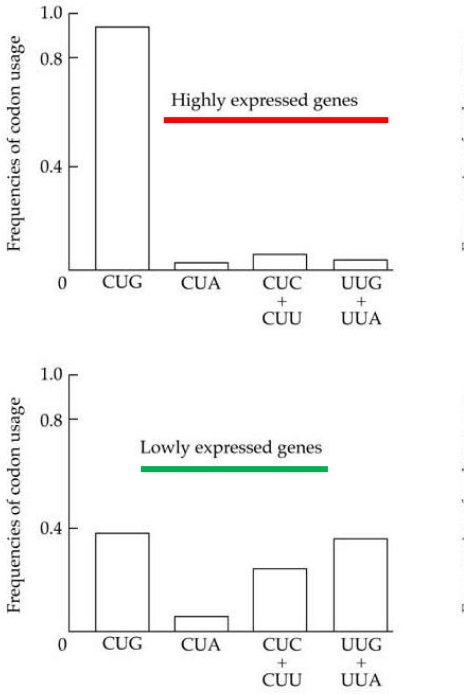
(b)

The **translation efficiency** of a codon

is related to

the **relative quantity of tRNA molecules**

That recognize the particular codon.



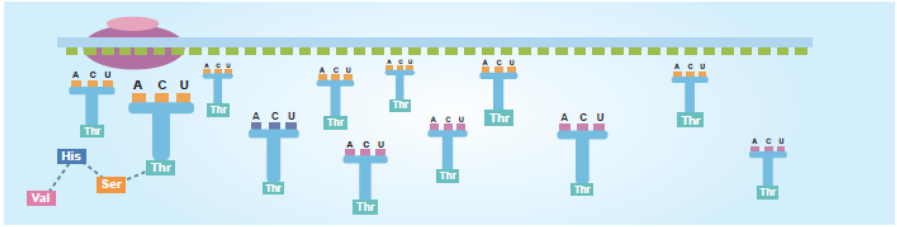
1. tRNA Abundance:

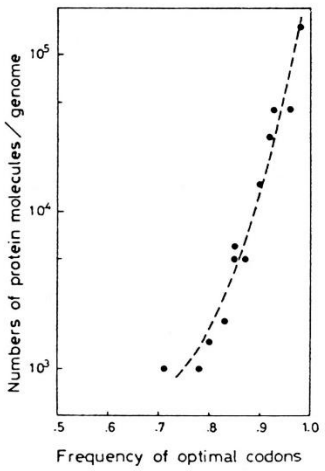
Translational efficiency depends on availability of tRNAs.

tRNA content is the major determinant of the bias between groups of codons recognized by different synonymous tRNAs.

Codons recognized by abundant tRNAs are used more often than those recognised by rare tRNAs, particularly in highly expressed genes, probably owing to selection at the level of translation against codons recognized by rare tRNAs.

(**Reference:** M. Bulmer. Coevolution of codon usage and transfer RNA abundance. **Nature.** 1987. 325:728-730)





Toshimishi Ikemura

Measures of codon-usage bias

The **relative synonymous codon usage** (***RSCU***) was first suggested by Sharp et al. (1986).

RSCU is

the **number of times a codon appears in a gene**

divided by

the number of **expected occurrences** under equal codon usage.

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

n = number of synonymous codons ($1 \leq n \leq 6$) for the amino acid under study,
X_i = number of occurrences of codon i.

If the synonymous codons of an amino acid are used with equal frequencies,
their *RSCU* values will equal 1.

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

n = number of synonymous codons ($1 \leq n \leq 6$) for the amino acid under study,
X_i = number of occurrences of codon i.

the **number of times a codon appears in a gene**

divided by

the number of **expected occurrences** under equal codon usage.

In una Proteina di **200 aa** trovo il seguente uso di **Isoleucine**

AUU = 7
AUC = 18
AUA = 2

Qual è il valore di RSCU per il codone **AUC** ?

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

n = number of synonymous codons ($1 \leq n \leq 6$) for the amino acid under study,
X_i = number of occurrences of codon i.

AUU = 7 RSCU= 7/9 = 0,77
AUC = 18 RSCU= 18/9= 2
AUA = 2 RSCU= 2/9 = 0,22

compile **a table of *RSCU* values for highly expressed genes.**

From this table, it is possible to identify the codons that are most frequently used for each amino acid.
The **relative adaptiveness** of a codon (w_i) is computed as

$$w_i = \frac{RSCU_i}{RSCU_{max}}$$

where $RSCU_{max}$ = the *RSCU* value for the most frequently used codon for an amino acid.

Codon Adaptation Index (CAI)

Codon adaptation index is a measurement of the **relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes**

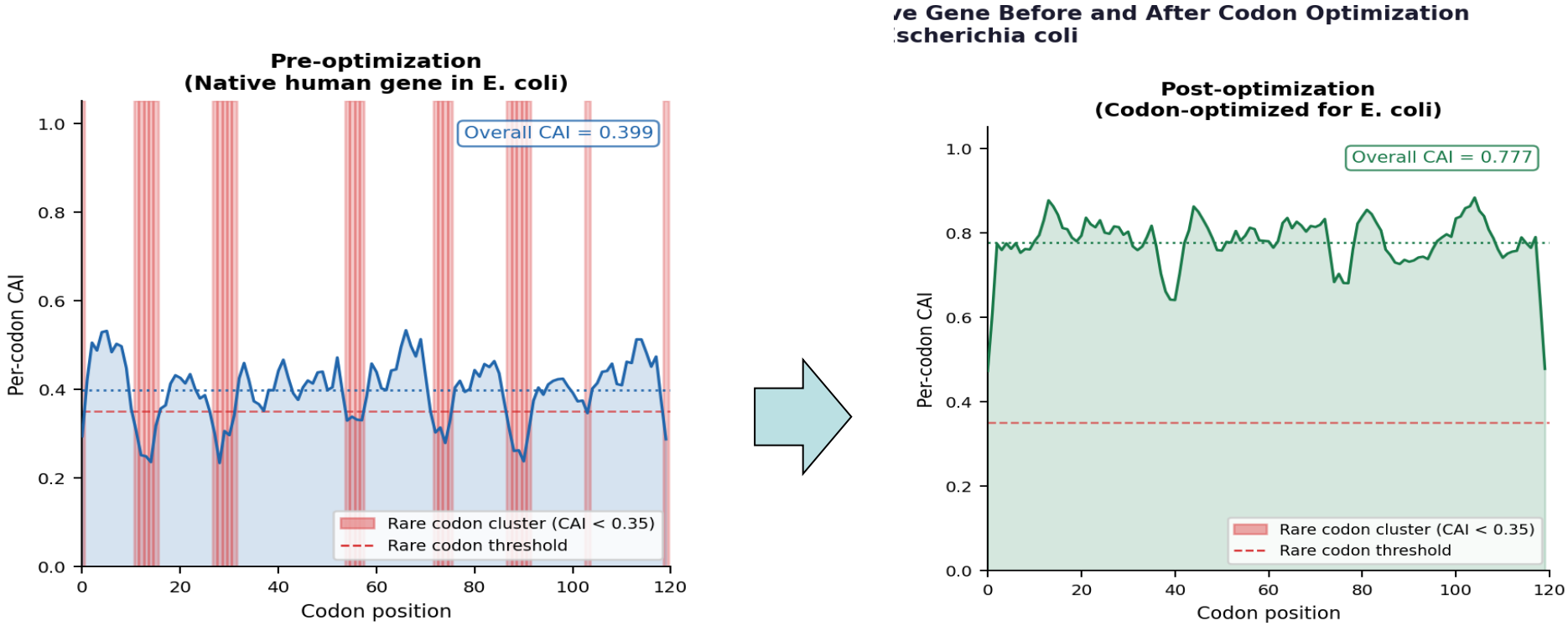
The CAI index is defined as the geometric mean of these relative adaptiveness values.

$$CAI = \left(\prod_{i=1}^{L} w_i \right)^{\frac{1}{L}}$$

The **CAI index ranges from 0 to 1**

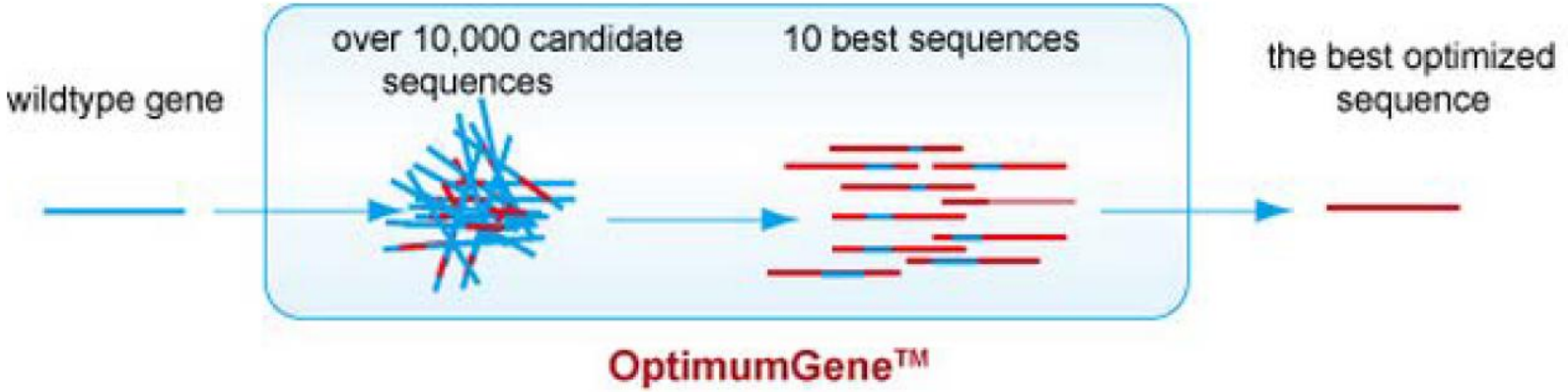
being **1** if a gene always uses, for each encoded amino acid, the most frequently used synonymous codon in the reference set.

Figure 1. Per-codon CAI Profile of a Representative Gene Before and After Codon Optimization for Expression in Escherichia coli



Codon Optimization Tools and Techniques

- Numerous online tools are available from providers like IDT, GenScript GenSmart™, and VectorBuilder.
- Advanced algorithms are employed to optimize codon usage, manage GC content, and avoid undesirable sequences like repeats or restriction sites.
- GenScript's GenSmart™ platform, for instance, has demonstrated remarkable success, increasing protein expression 8- to 18-fold in Chinese Hamster Ovary (CHO) cells.
- These tools also significantly improve gene synthesis success rates and enhance cloning efficiency, streamlining experimental workflows.



Review your Free Optimization Report

Codon Adaptation Index (CAI)

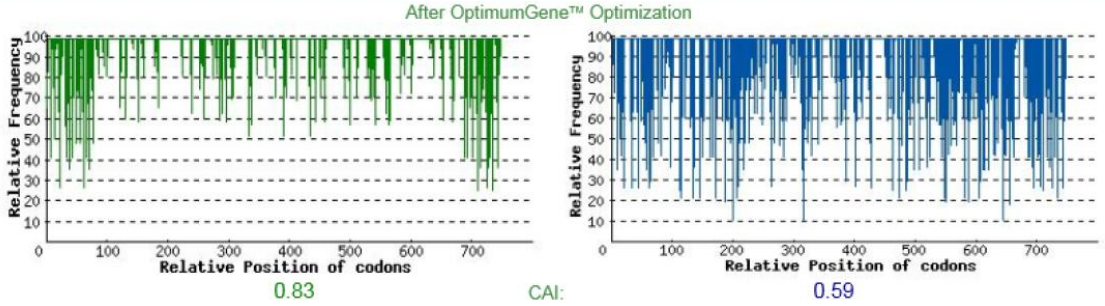


Figure 1a. The distribution of codon usage frequency along the length of the gene sequence. A CAI of 1.0 is considered to be perfect in the desired expression organism, and a CAI of > 0.8 is regarded as good, in terms of high gene expression level.

GC Content Adjustment

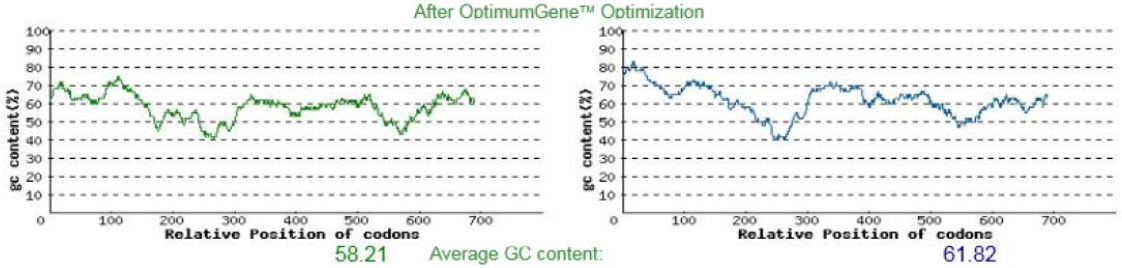
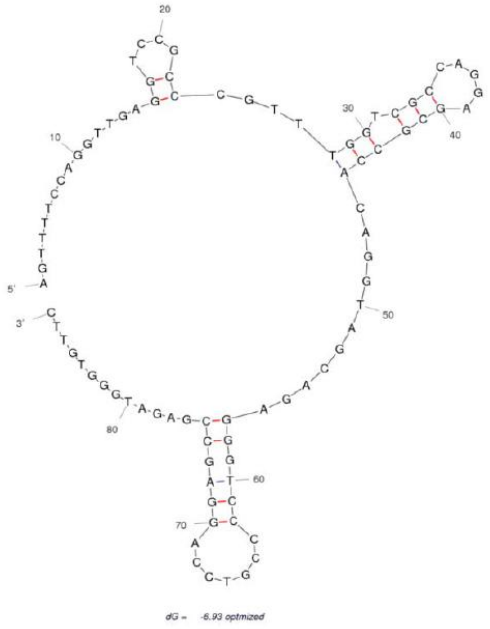


Figure 2. The ideal percentage range of GC content is between 30-70 %. Peaks of %GC content in a 60 bp window have been removed.



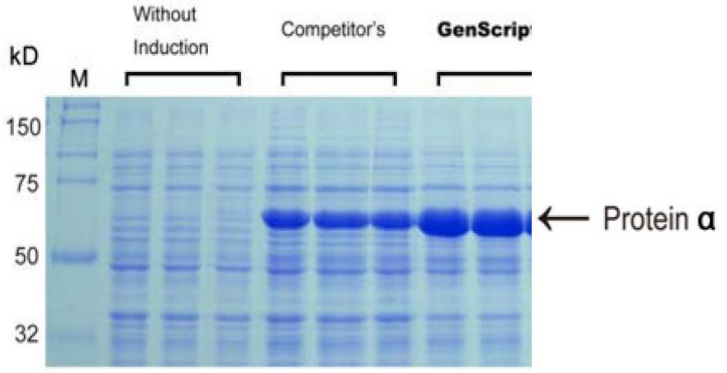
Helices in structure (all)

Helix	ΔG (kcal/mol)	Length	Position
1	-4.74	5	64-->68 ; 85<--81
2	-4.08	3	16-->18 ; 42<--40
3	-3.12	3	19-->21 ; 38<--36
4	-2.17	2	69-->70 ; 76<--75
5	-2.17	2	24-->25 ; 33<--32
6	-1.84	2	53-->54 ; 59<--58
7	-1.84	2	43-->44 ; 49<--48

Hairpins in structure (all)

Hairpin	ΔG (kcal/mol)	Length	Position
1	3.10	10	24-->...<--33
2	2.50	8	69-->...<--76
3	1.50	7	53-->...<--59
4	1.50	7	43-->...<--49

OptimumGene™ Improves Protein Expression Better that Competitors' Optimization



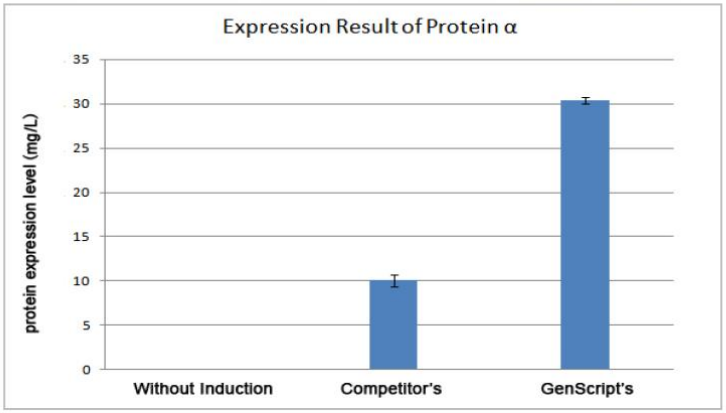
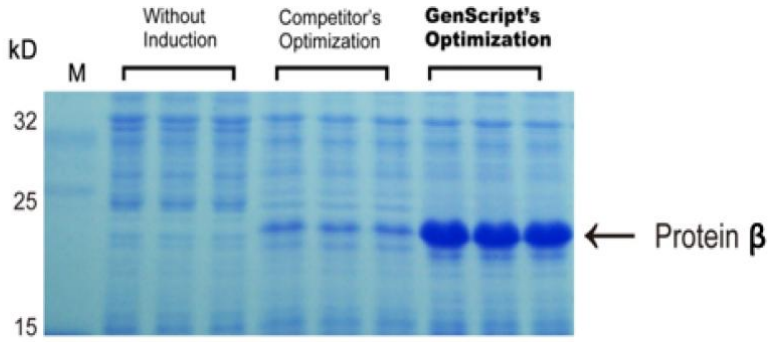


Fig. 1: Expression Result of Protein α after Codon Optimization. The expression level of Protein α using GenScript's OptimumGene™ Codon Optimization is **3** times more than that of competitor's.



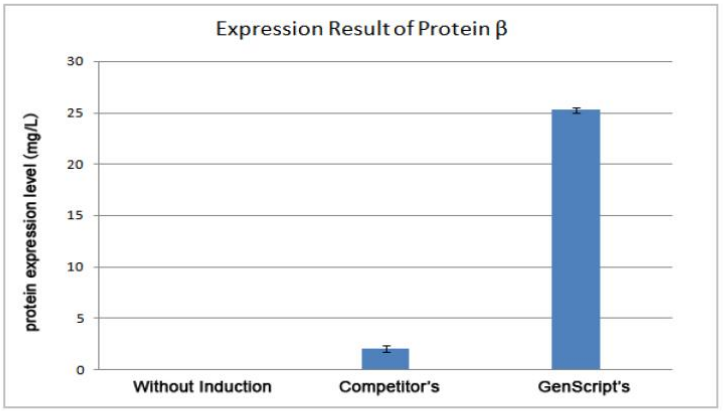


Fig. 2: Expression Result of Protein β after Codon Optimization. The expression level of Protein β using GenScript's OptimumGene™ Codon Optimization is **13** times more than that of competitor's.

PARAMETRI CHIAVE NELL'OTTIMIZZAZIONE DEI CODONI PER LA SEQUENZA DI DNA

SEQUENZA DNA DI PARTENZA (WILD-TYPE) → ALGORITMO DI OTTIMIZZAZIONE DEI CODONI → SEQUENZA DNA OTTIMIZZATA

1 (1) OTTIMIZZAZIONE DELL'ESPRESSIONE PROTEICA

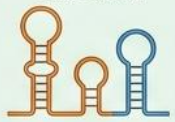
BIAS DEI CODONI DELL'OSPITE

INDICE DI ADATTAZIONE DEI CODONI (CAI)

$$CAI = \left(\sum w_i\right)^{1/L}$$

2 (2) STABILITÀ DELL'mRNA

STRUTTURE SECONDARIE DELL'mRNA



ENERGIA LIBERA MINIMA (MFE) — ENERGIA LIBERA MINIMA (MFE)

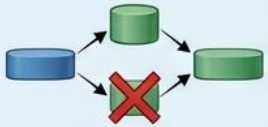
CONTENUTO GC

MOTIVI DI DESTABILIZZAZIONE

AUUUA, etc.

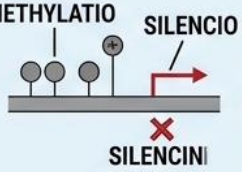
3 SICUREZZA E CONTROLLO DI QUALITÀ

SITI DI SPLICING CRIPTICI



INCORRECT PSUTIORI DI LICOLIGIO

ISOLE CpG

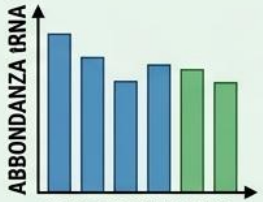


POTENZIAL SILENCING

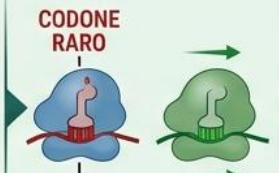
SITI DI RESTRIZIONE NON DESIDERATI

4 EFFICIENZA DELLA TRADUZIONE E RIPIEGAMENTO

DISPONIBILITÀ DI tRNA



CODONI RARI



CO-TRADUZIONALE RIPIEGAMENTO PROTEICO

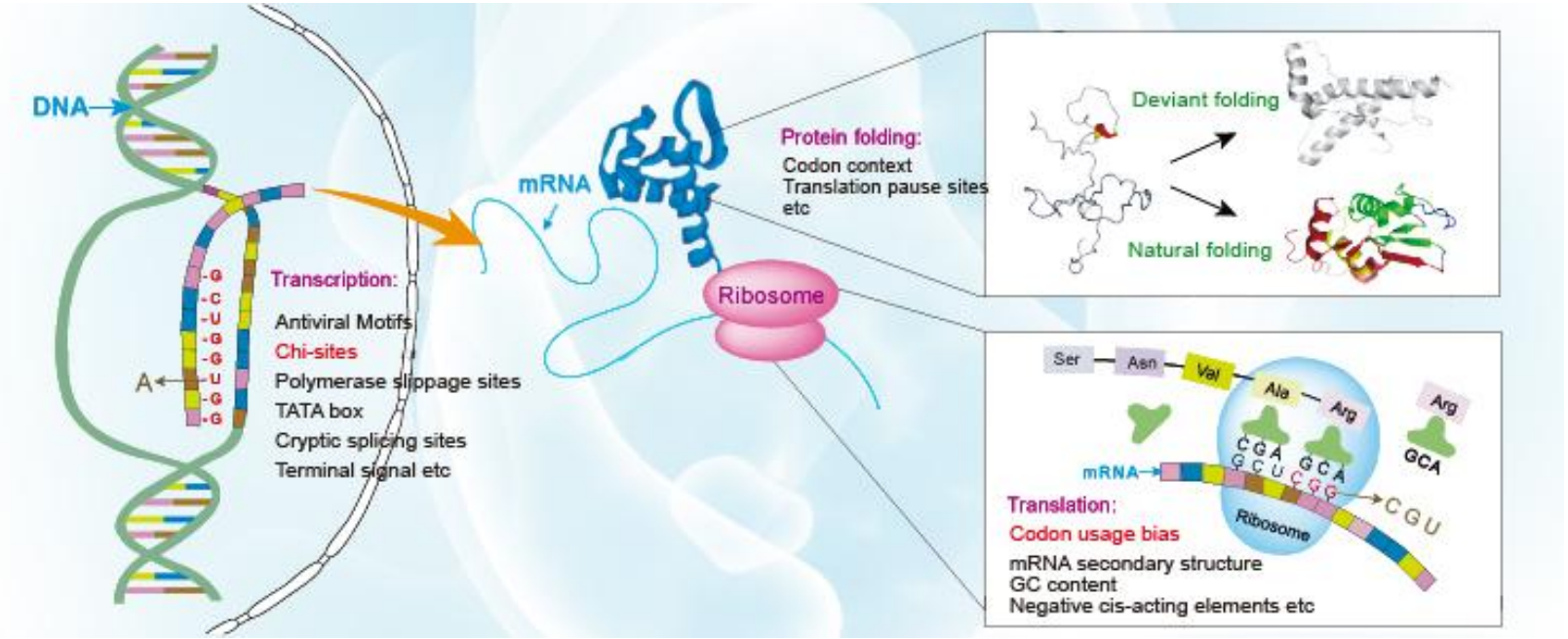


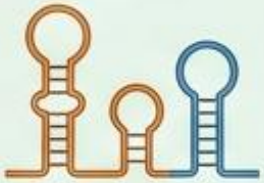
Table 1. Parameters that influence protein expression.

Transcriptional level	mRNA level	Translational level
• GC content • Consensus splice sites • Cryptic splice sites • SD sequences • TATA boxes • Termination signals • Artificial recombination sites	• RNA instability motifs • Ribosomal entry sites • Repetitive sequences	• Codon usage • Premature poly(A) sites • Ribosomal entry sites • Secondary structures

(2) STABILITÀ DELL'mRNA

STRUTTURE SECONDARIE DELL'mRNA

mRNA ORIGINALE



ENERGIA LIBERA MINIMA (MFE)

mRNA OTTIMIZZATO

ENERGIA LIBERA MINIMA (MFE)

CONTENUTO GC

ORIGINALE

OTTIMIZZATO

GC/AT 40-60%

MOTIVI DI DESTABILIZZAZIONE

AUUUA, etc.

Coding-Sequence Determinants of Gene Expression in *Escherichia coli*

Grzegorz Kudla,[1]* Andrew W. Murray,[2] David Tollervey,[3] Joshua B. Plotkin[1]†

Synonymous mutations do not alter the encoded protein, but they can influence gene expression. To investigate how, we engineered a synthetic library of 154 genes that varied randomly at synonymous sites, but all encoded the same green fluorescent protein (GFP). When expressed in *Escherichia coli*, GFP protein levels varied 250-fold across the library. GFP messenger RNA (mRNA) levels, mRNA degradation patterns, and bacterial growth rates also varied, but codon bias did not correlate with gene expression. Rather, the stability of mRNA folding near the ribosomal binding site explained more than half the variation in protein levels. In our analysis, mRNA folding and associated rates of translation initiation play a predominant role in shaping expression levels of individual genes, whereas codon bias influences global translation efficiency and cellular fitness.

www.sciencemag.org SCIENCE VOL 324 10 APRIL 2009 255

Science

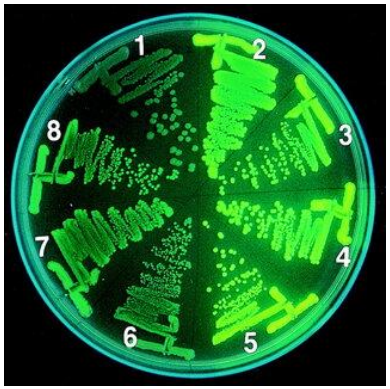
Synonymous mutations do not alter the encoded protein, but they can influence gene expression.

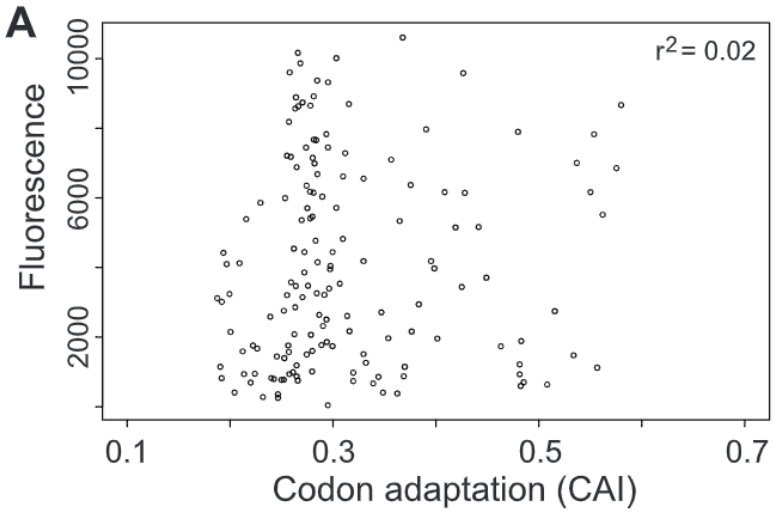
To investigate how, we engineered a specific gene:

- all encoded the same **green fluorescent protein (GFP).**

- a synthetic library of **154 genes** that varied randomly at **synonymous sites**,

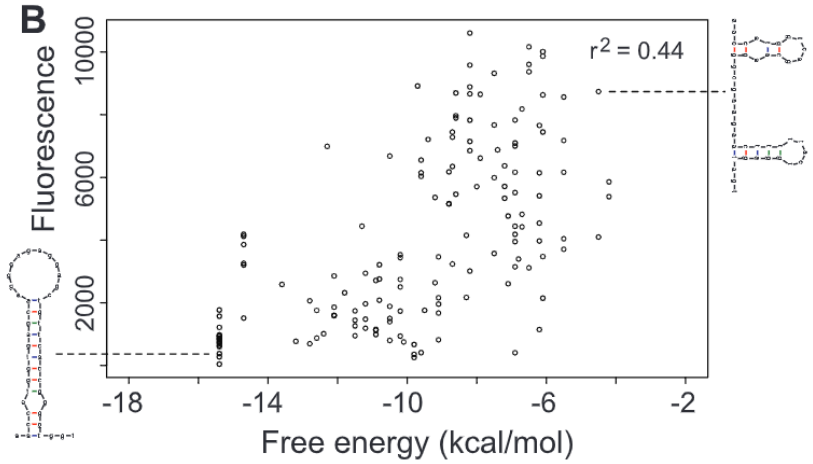
When expressed in *Escherichia coli*, **GFP protein levels varied** -fold across the library.





Codon bias

did not correlate with gene expression.



Stability of mRNA folding

near the ribosomal binding site explained more than half the variation in protein levels.

In our analysis, mRNA folding and associated rates of translation initiation play a predominant role in shaping expression levels of individual genes, whereas codon bias influences global translation efficiency and cellular fitness.

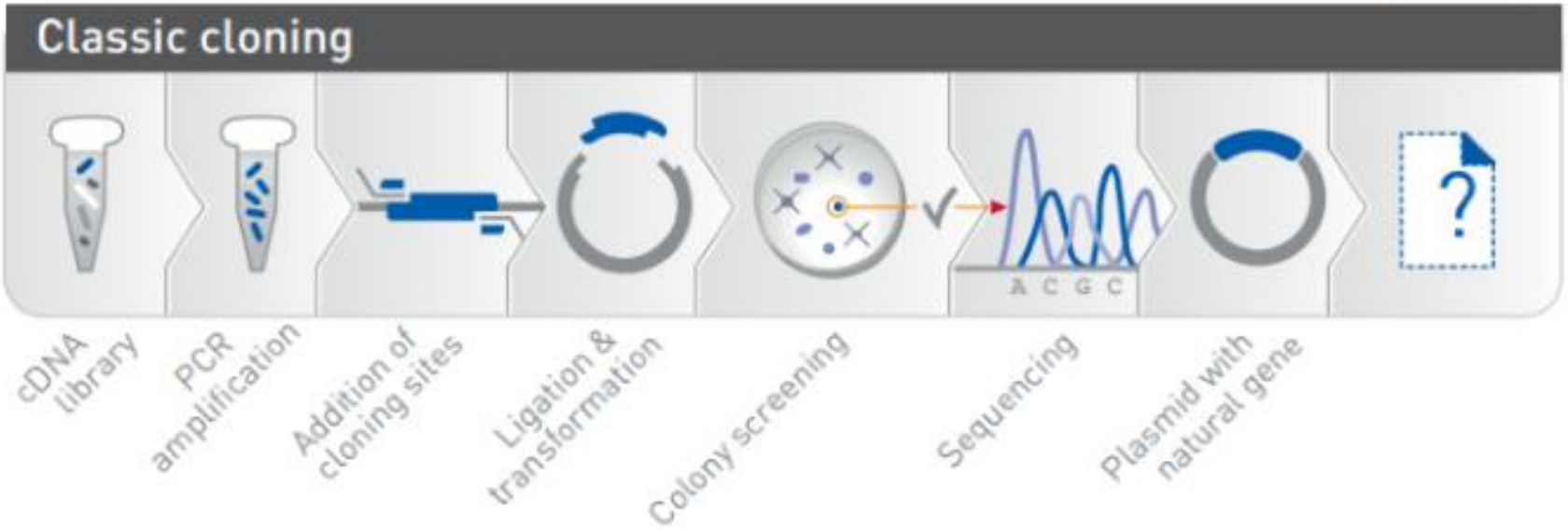
Codon Adaptation is not the most important factor for protein yield

Coding-sequence determinants of gene expression in Escherichia coli.
Kudla G, Murray AW, Tollervey D, Plotkin JB. *Science.* 2009 Apr 10;324(5924):255-8.

- 154 synthetic GFP genes with random synonymous mutations
- 250-fold variation in fluorescence
- 44% of variation explained by 5' mRNA free energy (nt −4 to +37)

The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria.
Li GW, Oh E, Weissman JS. *Nature.* 2012 Mar 28;484(7395):538-41.

- Variation in Translation Rate does not correlate with rare codon use
- Orthogonal ribosomes with altered anti-SD sequences: pausing results from hybridization between 16s rRNA and SD-like sequences in mRNA


Classic cloning
A C G C
?
cDNA library
PCR amplification
Addition of cloning sites
Ligation & transformation
Colony screening
Sequencing
Plasmid with natural gene


GeneArt® Gene Synthesis
WWW
GeneOptimizer®
optimization algorithm
GeneAssembler™
gene synthesis platform
Expression
Wild type
DNA sequence
GeneArt® GeneOptimizer®
improved sequence

ORDERING

Complete synthetic genes with 100% sequence verification are delivered in a cloning or expression vector and are ready to use for a variety of applications.

- Never any reoccurring or hidden fees for custom vectors
- Transfection grade and endotoxin free options now available
- Guaranteed delivery dates

Genes in tubes

ORDER TUBES

- No order minimum for tubes
- Delivered dry, normalized to 4 µg with scale up options available*
- Glycerol stocks available

Product	Amount	Price	Typically Shipped†
MiniGene 25-500 bp	4 µg	$125.00 USD	8 business days
Gene 501-1500 bp	4 µg	$0.25 USD / bp	12 business days
Gene 1501-3000 bp	4 µg	$0.25 USD / bp	12 business days
Gene 3001-5000 bp	4 µg	$0.30 USD / bp	12 business days

Genes in plates