A Simple Program to Calculate Codon Bias Index

Tsung-Tsan Wang,[1] Wen-Chi Cheng,[2] and Byong H. Lee[*,1]*

Abstract

A computer program (PCBI) was developed to quickly calculate codon bias index (CBI). PCBI can analyze a gene containing introns. The 22 preferred codons defined from *Saccharomyces cerevisiae* were used in PCBI as the standard to measure the CBI values. However, users can modify the preferred codons to suit each organism. The data PCBI provides include DNA sequence of open reading frame without introns, amino acid sequence of gene product, a table of amino acid composition, a table of codon usage and (G + C) content, parameters for calculating CBI, and the value of CBI. PCBI runs on a DOS or Windows environment, but results can be saved in ASCII text format.

Index Entries: Codon bias index; computer program; gene information.

1. Introduction

Although the genetic code is degenerate, there is often an unequal use of synonymous codons. For instance, in both prokaryotic and eukaryotic genes, the selection of the synonymous codons is strongly biased *(1–6)*. Codon usage is an important factor that affects gene expression *(7,8)*. For evaluating the selection of synonymous codons by a specific organism, several measurements have been developed. These are codon adaptation index *(9)*, codon bias index (CBI) *(6)*, intrinsic codon deviation index *(10)*, and effective number of codons *(11)*. After analyzing *S. cerevisiae* genes encoding alcohol dehydrogenase isozyme I (*ADH*I) and glyceraldehyde-3-phosphate dehydrogenase, Bennetzen and Hall *(6)* arbitrarily chose 22 codons encoding for 17 amino acids as preferred codons for this species. Based on these 22 preferred codons, the value of CBI was defined as

$$\mathrm{CBI} = (N_{\mathrm{pfr}} - N_{\mathrm{ran}}) \div (N_{\mathrm{tot}} - N_{\mathrm{ran}})$$

where N_{pfr} is the total number of occurrences of preferred codons, N_{ran} is the expected number of the preferred codons if all synonymous codons were used equally, and N_{tot} is the total number of the 17 amino acids encoded by the preferred codons. A strongly expressed gene has a higher value of CBI and a more biased codon usage than a weakly expressed gene *(6)*. Thus, CBI is one criterion to evaluate gene expression. However, it is tedious and time-consuming to calculate the CBI manually. The aim of this research is to develop a computer program (PCBI) to measure the value of CBI.

2. Materials and Methods

The PCBI program was written with Turbo Basic language and compiled as an executable program. It can run on DOS or Windows environment. The flowchart of PCBI performance is summarized in **Fig. 1**. First, the user must give PCBI a gene name, gene source, and file name to retrieve the DNA sequence, the interval of open reading frame as well as introns, if present, and the modification of preferred codons, if necessary. Next, the user must specify a file name in which PCBI will save the executed results and choose the executed results to save. After retrieving and performing DNA sequence, PCBI shows the

*Author to whom all correspondence and reprint requests should be addressed. E-mail: BLEE@AGRADM.LAN.MCGILL.CA. [1]Department of Food Science and Agricultural Chemistry, McGill University, Macdonald Campus, Ste. Anne de Bellevue, Quebec, Canada H9X 3V9, Dr. Lee is also affiliated with the Food Research and Development Center, Agriculture Canada, Saint-Hyacinthe, Quebec, Canada; [2]Kaohsiung District Agricultural Improvement Station (KDAIS), Pingtung, Taiwan, Republic of China.

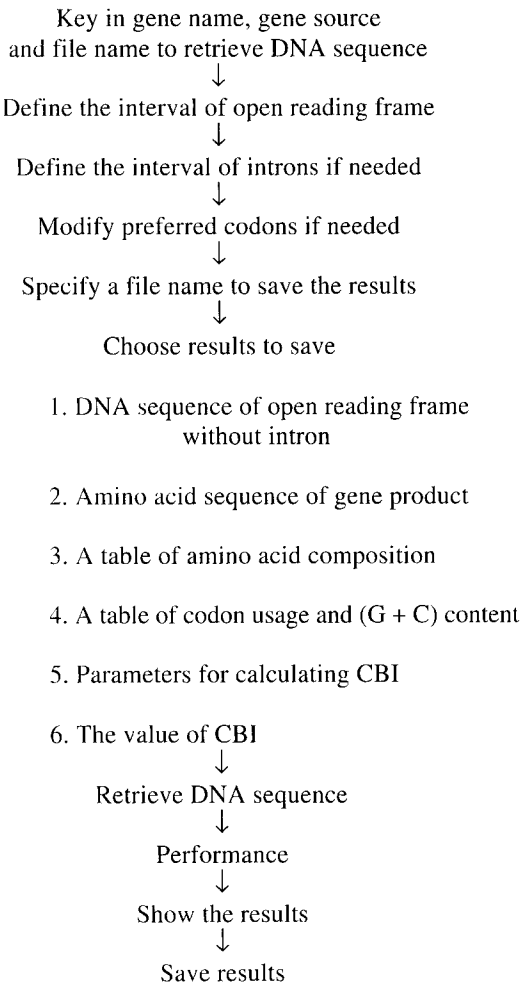

Fig. 1. The flow chart of PCBI performance.

results on screen one by one. During the entire performance, PCBI will inform users if any problem occurs.

The entire DNA sequence must be in text format with no redundant nucleotides, and to be retrieved must start with a translation initiation codon (ATG) and stop with a translation termination codon (TAA, TAG, or TGA). PCBI uses the 22 preferred codons determined by Bennetzen and Hall *(6)* from *S. cerevisiae* as standard.

3. Results and Discussion

PCBI provides other useful information in addition to the CBI value, enabling the user to gather further information from the genes. After analyzing a gene, PCBI can show:

Table 1
An Example of the Amino Acid Composition of the Gene Product of *Schw. occidentalis SCR2* Gene

AA	Amount	%
Ala	6	5.66
Arg	7	6.60
Asn	1	0.94
Asp	2	1.89
Cys	5	4.72
Gln	10	9.43
Glu	3	2.83
Gly	9	8.49
His	4	3.77
Ile	1	0.94
Leu	7	6.60
Lys	24	22.64
Met	1	0.94
Phe	4	3.77
Pro	1	0.94
Ser	2	1.89
Thr	8	7.54
Trp	0	0.00
Tyr	4	3.77
Val	7	6.60

1. The DNA sequence of the open reading frame without intron;
2. The amino acid sequence of the gene product;
3. A table showing the amino acid composition of the protein encoded from the gene (**Table 1**);
4. A table displaying codon usage as well as (G + C) content of the gene (**Table 2**);
5. Each parameter for calculating the CBI value of the gene; and
6. The CBI value of the gene (**Table 3**).

The preferred codons are indicated in **Table 2**. In addition, **Table 3** summarizes each factor given by users to calculate the CBI value. The results are saved in ASCII text format.

PCBI uses preferred codons based on *S. cerevisiae*. These may be different in other organisms *(12,13)*. If necessary, users can modify the preferred codons to suit each specific organism. For example, to calculate the CBI value of *Schwanniomyces occidentalis* genes, TTG has to be substituted with TTA for Leu, and AAG replaced with AAA for Lys *(13)*.

Table 2
An Example of the Codon Usage and (G + C) Dontent of *Schw. occidentalis SCR2* Gene

Codon	Amt	%	Codon	Amt	%
TTT-Phe	1	0.94	TCT-Ser[b]	0	0.00
TTC-Phe[b]	3	2.83	TCC-Ser[b]	2	1.89
TTA-Leu[b]	7	6.60	TCA-Ser	0	0.00
TTG-Leu	0	0.00	TCG-Ser	0	0.00
CTT-Leu	0	0.00	CCT-Pro	0	0.00
CTC-Leu	0	0.00	CCC-Pro	0	0.00
CTA-Leu	0	0.00	CCA-Pro[b]	1	0.94
CTG-Leu	0	0.00	CCG-Pro	0	0.00
ATT-Ile[b]	1	0.94	ACT-Thr[b]	3	2.83
ATC-Ile[b]	0	0.00	ACC-Thr[b]	5	4.72
ATA-Ile	0	0.00	ACA-Thr	0	0.00
ATG-Met	1	0.94	ACG-Thr	0	0.00
GTT-Val[b]	6	5.66	GCT-Ala[b]	6	5.66
GTC-Val[b]	1	0.94	GCC-Ala[b]	0	0.00
GTA-Val	0	0.00	GCA-Ala	0	0.00
GTG-Val	0	0.00	GCG-Ala	0	0.00
TAT-Tyr	1	0.94	TGT-Cys[b]	4	3.77
TAC-Tyr[b]	3	2.83	TGC-Cys	1	0.94
TAA-	0	—	TGA-	0	—
TAG-	1	—	TGG-Trp[b]	0	0.00
CAT-His	1	0.94	CGT-Arg	2	1.89
CAC-His[b]	3	2.83	CGC-Arg	0	0.00
CAA-Gln[b]	10	9.43	CGA-Arg	0	0.00
CAG-Gln	0	0.00	CGG-Arg	0	0.00
AAT-Asn	0	0.00	AGT-Ser	0	0.00
AAC-Asn[b]	1	0.94	AGC-Ser	0	0.00
AAA-Lys	17	16.03	AGA-Arg[b]	4	3.77
AAG-Lys	7	6.60	AGG-Arg[b]	1	0.94
GAT-Asp	1	0.94	GGT-Gly[b]	9	8.49
GAC-Asp[b]	1	0.94	GGC-Gly	0	0.00
GAA-Glu[b]	3	2.83	GGA-Gly	0	0.00
GAG-Glu	0	0.00	GGG-Gly	0	0.00

[a](G + C) content = 34.9%.
[b]The preferred codons used in the DNA analysis.

Table 3
An Example of a Summary to Calculate the CBI of *Schw. occidentalis SCR2* Gene

Gene name and source: *SCR2* gene of *Schw. occidentalis*
Open reading frame: from 1004 to 1777
Intron number: 1
Intron interval: 1007–1459
Total codons: 106 not including stop codon
Calculation of CBI: CBI = $(N_{pfr} - N_{ran}) \div (N_{tot} - N_{ran}) = (89 - 38.25) \div (103 - 38.25) = 0.78$
Where N_{pfr} = the total number of occurrences of preferred codons; N_{ran} = the expected number of the preferred codons if all synonymous codons were used equally; and N_{tot} = the total number of the amino acids encoded by the preferred codons.

Table 4
The CBI Value of Some Yeast Genes

Gene	Source[a]	Gene product	CBI[b]	Reference
*ADH*I	Sc	Alcohol dehydrogenase I	0.91	***6***
*ARG*4	Sc	Argininosuccinate lyase	0.31	***14***
*THR*4	Sc	Threonine synthase	0.51	***14***
$CYC1_0$	So	Cytochrome c protein	0.74	***13***
HXK	So	Hexokinase	0.62	***13***
INV	So	Invertase	0.25	***13***

[a]Sc: *S. cerevisiae*; So; *Schw. occidentalis*.
[b]To calculate the CBI value of *Schw. occidentalis* genes, the two standard preferred codons were modified. The codon of TTA substitutes for TTG encoding leucine and AAA replaces AAG for lysine.

An *SCR2* gene was used in a demonstration of PCBI performance, because the gene is from *Schw. occidentalis*, and contains an intron. The results, without DNA and amino acid sequences, are given in **Tables 1**, **2**, and **3**. **Table 4** shows a PCBI analysis of three genes of *S. cerevisiae* and *Schw. occidentalis*. The CBI values obtained by PCBI matched those previously published (**Table 4**). PCBI can read a DNA sequence up to 15 kb, which is much larger than an average gene. Besides, the performance is without a speed limitation. For example, in an 80486 100 MHz personal computer, PCBI took less than 1 s to analyze a 15-kb DNA with a putative gene of 6 kb.

4. Availability

PCBI can be obtained via Wen-Chi Cheng (E-mail: w54001@wind.cc.ntnu.edu.tw).

References

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8,** 49–63.
2. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9,** 43–74.
3. Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146,** 1–21.
4. 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.
5. Ikemura, T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **158,** 573–597.
6. Bennetzen, J. L., and Hall, B. D. (1982) Codon selection in yeast. *J. Biol. Chem.* **257,** 3026–3031.
7. Kinnaird, J. H., Burns, P. A., and Fincham, J. R. (1991) An apparent rare-codon effect on the rate of translation of a *Neurospora* gene. *J. Mol. Biol.* **221,** 733–736.
8. Solomovici, J., Lesnik, T., and Reiss, C. (1997) Does *Escherichia coli* optimize the economics of the translation process? *J. Theor. Biol.* **185,** 511–521.
9. 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.
10. Freire-Picos, M. A., Gonzalez-Siso, M. I., Rodriguez-Belmonte, E., Rodriguez-Torres, A. M., Ramil, E., and Cerdan, M. E. (1994) Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes. *Gene* **139,** 43–49.
11. Wright, F. (1990) The effective number of codons used in a gene. *Gene* **87,** 23–29.
12. Nagasu, T. and Hall, B. D. (1985) Nucleotide sequence of the *GDH* gene coding for the NADP-specific glutamate dehydrogenase of *Saccharomyces cerevisiae*. *Gene* **37,** 247–253.
13. Wang, T. T., Lee, C. F., and Lee, B. H. (1998) The molecular biology of *Schwanniomyces occidentalis* Klocker. *CRC Crit. Rev. Biotechnol.* (Submitted).
14. Sharp, P. M. and Cowe, E. (1991) Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7,** 657–678.