Journal of Bioinformatics and Sequence Analysis Vol. 2(2), pp. 25-29, June 2010
Available online at http://www.academicjournals.org/JBSA
ISSN 2141-2464 © 2009 Academic Journals

Full length Research Paper

QPSOBT: One codon usage optimization software for protein heterologous expression

Yujie Cai[1]*, Jun Sun[2], Jie Wang[2], Yanrui Ding[2], Xiangru Liao[1] and Wenbo Xu[2]

[1]Key Laboratory of Industrial Biotechnology, School of Biotechnology, JiangNan University, 1800 Lihu road, Wuxi, Jiangsu 214122, China.
[2]School of Information Technology, JiangNan University, 1800 Lihu road, Wuxi, Jiangsu 214122, China.

Accepted 17 May, 2010

QPSOBT is a codon usage optimization software based on the Quantum-behaved Particle Swarm Optimization (QPSO) algorithm. It can design synthetic genes of multikilobase sequences for protein heterologous expression rapidly. The program runs on .NET platform. Compared to the existing codon optimization software and web services, QPSOBT is able to generate better results when DNA/RNA sequence length is less than 6 kb which is a commonly-used range, especially when some restriction sites need to be removed. QPSOBT is freely available (www.sigcib.org/qpsobt.html).

Keywords: Codon usage, quantum-behaved particle swarm, optimization, protein expression.

INTRODUCTION

It has been suggested by biologists that the synonymous codons are not able to be used randomly in all organisms, since their usage patterns vary with species and even genes in the same species so that the balance between natural mutation and selection can be achieved (Grantham et al., 1980; Sharp et al., 1993). As has been demonstrated, natural selection shapes codon usage in both unicellular and multicellular organisms (Duret and Mouchiroud, 1999). Optimal codons in fast-growing micro-organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, help to achieve faster translation rates and higher accuracy. In other organisms that do not show high growing rates (such as *Homo sapines*) or that present small genomes (such as *Helicobacter pylori*), codon usage optimization is normally absent, and thus codon preferences are determined by the characteristic mutational biases seen in that particular genome (Hershberg and Petrov, 2008).

*Corresponding author. E-mail: yu_jie_cai@yahoo.com.cn. Tel: 86-510-85916372.

Abbreviations: GC, Guanine-cytosine; **CAI,** codon adaptation index.

Genetic engineering has become one of the most powerful tools in modern biochemistry and biology. Recombinant DNA technology is widely used in both research and industry. Several hosts, such as *E. coli*, *Pichia pastoris*, plants, animals, are used to express the foreign protein. The desired protein sequence should be reversely translated into a nucleotide sequence. However, there are enormous possible synthetic sequences that can be made. Some researchers have proved that codon optimization is vital to establish high heterologous gene expression. For the purpose of stability maintenance and high expression, the codon usage of the foreign genes should be optimized (Akashi, 1994; Bulmer, 1991). For example, Tokuoka investigated expression levels of native and optimized Der f 7 genes in *Aspergillus oryzae*. They found codon optimization markedly increased protein and mRNA production levels (Tokuoka et al., 2008). Another example involves the study of transgene expression in *Tetrahymena* (Ngumbela et al., 2008). It also emphasizes the importance of codon optimization which resulted in about ten times more drug resistant transformants than a cassette containing the non-codon-optimized original neo gene. In addition, the transcript level of optimized glycoside hydrolase family 45 endoglucanase genes from termite-gut symbionts in *A. oryzae* was 1.8-fold higher than that of native gene (Sasaguri et

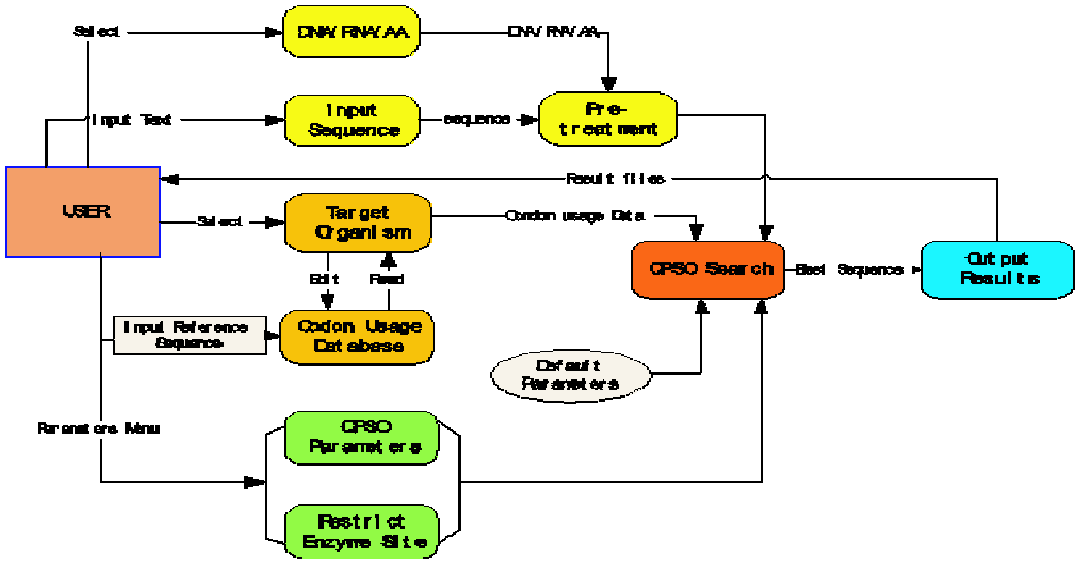

Figure 1. Workflow of QPSOBT.

al., 2008). To see this, many optimization methods of codon usage have been proposed (Wu et al., 2007). In our previous work, we presented a novel method to optimize codon usage (Cai et al., 2008). To facilitate others to employ this method for codon usage optimization, we programmed a stand-alone software based on the proposed method. The software is based on the quantum-behaved particle swarm optimization (QPSO) algorithm called QPSOBT. It is shown that by using QPSOBT, one can find the optimal gene sequence according to the hosts' codon usage frequency efficiently.

MATERIALS AND METHODS

The codon usage optimization algorithm

The task of codon usage optimization can be reduced to minimization of the following objective function of the relative squared error E defined as

$$E=\sum_{i=1}^{64}\left(\frac{F_{ri}-F_{ti}}{F_{ri}}\right)^2$$

Where E is the relative squared error between F_{ri}, the frequency of codon *i* in the reference or sequence (per 1000), and F_{ti}, the frequency of codon *i* in the synthetic sequence. The goal of minimization of E is to find the codon usage frequency in the synthetic sequence as similar as possible to that in the reference host.

Figure 1 visualizes the workflow of QPSOBT. It begins with importing a protein or gene sequence, followed by selection of a target system. The target system can be selected from a pre-defined database and, alternatively, can be self-defined by users. After that, QPSOBT runs the QPSO algorithm to search the optimization sequence. The user can set up the parameters of the algorithm and restrict enzyme sites before running the program; otherwise QPSOBT runs the algorithm with default setups. The optimized sequence and statistical results are outputted to the result files when the algorithm terminates.

The detailed procedure of the optimization algorithm was described in our previous work (Cai et al., 2008).

Programming

Microsoft Visual C# was used to develop QPSOBT on the .NET platform. NET is an environment not only for the Windows platform, but other operating systems. The .NET Framework can be downloaded from Microsoft's website freely.

RESULTS

Figure 2 visualizes the main interface of QPSOBT. The input sequence may be DNA/RNA or protein sequence, while the output sequence may be DNA or RNA. Several codon usage tables from the Codon Usage Database (Nakamura et al., 2000) were pre-set in the Microsoft Access database. The users can choose the codon usage (that is, target organism) table in the list box. There are two methods for self-defining codon usage data: editing the data directly or calculating through a DNA/RNA sequence (as shown in Figure 3). The high express genes can be downloaded from a synthetic gene database (Wu et al., 2007) and can be used as a reference sequence. The users can also select the avoiding restriction enzymes in the setup menu item (Figure 4). The optimized sequence will be shown in the text box with a DNA or RNA format. After the running of the optimization algorithm, the codons frequency and GC content of the optimized sequence are displayed in the main interface for comparison with the reference host or sequence. Two

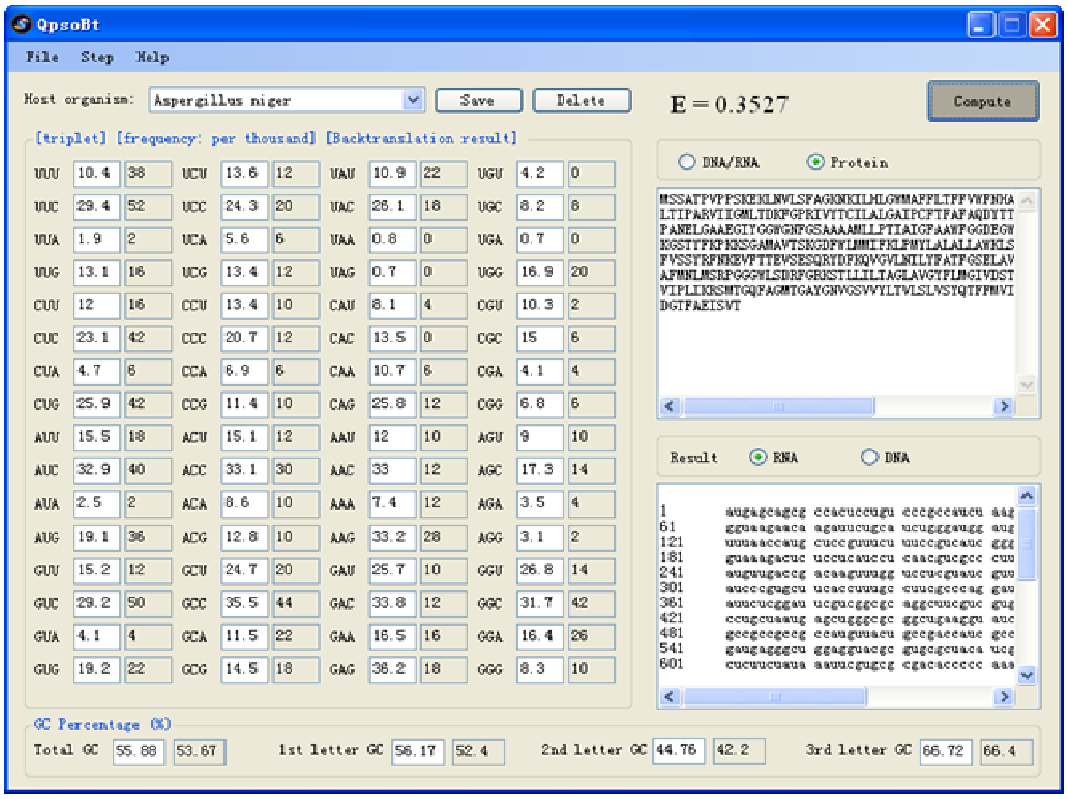

Figure 2. Main interface of QPSOBT.

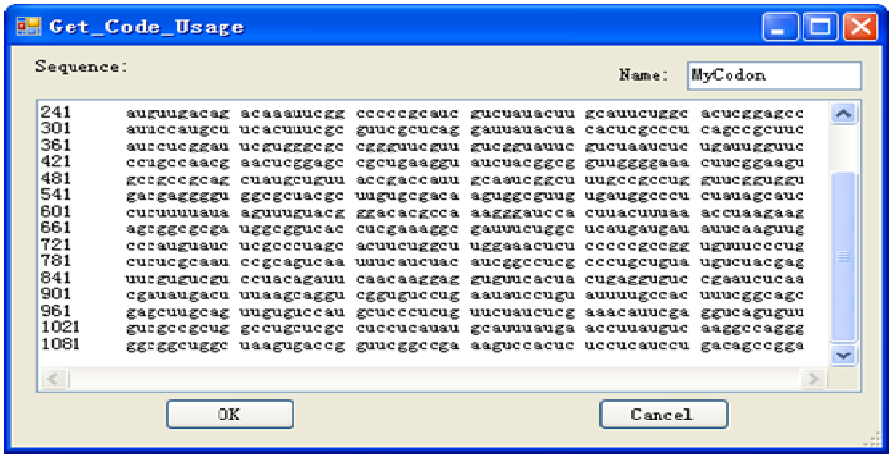

Figure 3. Interface of computing codon usage from the self-defined sequence.

output files are used to record statistical results and the process of optimization. QPSOBT is freely available (www.sigcib.org/qpsobt.html).

DISCUSSION

Among the approaches of codon usage optimization in existing stand-alone softwares and web servers, the two most frequently referred are the methods based on the “CAI = 1” theory and on codon usage probabilities. By the former method, the protein sequence is back-translated using the “one amino acid-one codon” approach according to the “CAI = 1” theory (Pesole et al., 1988). During back-translation, the same amino acids are only encoded by the most commonly-used synonymous codon in the reference set. There are two main shortcomings in this method. One is that the method can only optimize

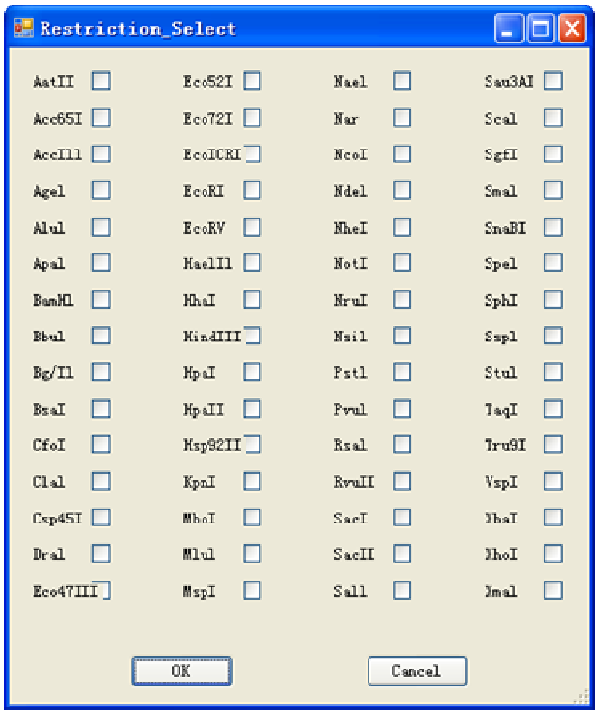

Figure 4. Interface of selecting avoided restriction enzymes.

highly-expressed genes due to its measurement based on codon usage bias. The “CAI = 1” method cannot consider tRNAs, ribosomal RNAs, and other non-coding RNAs. The other shortcoming is that the methods may result in low efficiency for organisms with low translation bias (Willenbrock et al., 2006).

The codon-usage-probability-based approach is currently widely used in applications such as DNAWorks (Hoover and Lubkowski, 2002), GeMs (Jayaraj et al., 2005), SGD (Wu et al., 2006), Gene Designer (Villalobos et al., 2006), GeneDesign (Richardson et al., 2006), and OPTIMIZER (Puigbo et al., 2007). When the process of optimization in the application is executed, a random number distributed uniformly within [0, 100] is generated, and a specific triplet is picked out using a roulette selection (a Monte Carlo method (Matousek, 2009)) according to the random number and the probability distribution. The process is executed across all the codons in the sequence to fulfill optimization of the codon usage, followed by removing restriction sites from the resulting sequence to get a modified one. According to the large number law and central limitation law, it can be inferred that with more samples, the higher precision it will acquire. Our previous work had proved it could only generate a good result when protein lengths were more than 2000 (Cai et al., 2008).

The QPSO algorithm employed in QPSOBT was proposed in our previous work (Jun et al., 2004, 2005). The algorithm is a global convergent and has a stronger global search ability than its predecessor - the Particle Swarm Optimization (PSO) algorithm. Our previous experiment results show that it can efficiently find out the optimal gene sequence according to hosts’ codon usage frequency. QPSOBT provides the most basic functions for optimization process which makes it easily understood and used. Users can use other more professional bioinformatics tool to analyze and operate the optimized sequences.

ACKNOWLEDGEMENT

This work was supported by the innovation team-building project of Jiangnan University, JNIRT0702.

REFERENCES

Akashi H (1994). Synonymous codon usage in Drosophila melanogaster, natural selection and translational accuracy. Genet. 136: 927-935.

Bulmer M (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics, 129: 897-907.

Cai YJ, Sun J, Wang J, Ding YR, Tian N, Liao XR, Xu WB (2008). Optimizing the codon usage of synthetic gene with QPSO algorithm. J. Theor. Biol., 254: 123-127.

Duret L, Mouchiroud D (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. 96: 4482-4487.

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980). Codon catalog usage and the genome hypothesis. Nucleic Acids Res., 8: R49-R62.

Hershberg R, Petrov DA (2008). Selection on codon bias. Annu. Rev. Genet., 42: 287-299.

Hoover DM, Lubkowski J (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res., 30: e43.

Jayaraj S, Reid R, Santi DV (2005). GeMS: an advanced software package for designing synthetic genes. Nucleic Acids Res., 33: 3011-3016.

Jun S, Bin F, Wenbo X (2004) Particle swarm optimization with particles having quantum behavior. Evolutionary Computation, 2004. CEC2004. Congress on. 321: 325-331

Jun S, Wenbo X, Bin F (2004) A global search strategy of quantum-behaved particle swarm optimization. Cybernetics and Intelligent Systems, 2004 IEEE Conference on. 111-116 vol.111.

Jun S, Wenbo X, Bin F (2005) Adaptive parameter control for quantum-behaved particle swarm optimization on individual level. Systems, Man and Cybernetics, 2005 IEEE International Conference on., 3044: 3049-3054

Matousek R (2009) Genetic Algorithm and Advanced Tournament Selection Concept. In Krasnogor, N., MelianBatista, M.B., Perez, J.A.M., MorenoVega, J.M. and Pelta, D.A. (eds), Nicso 2008: Nature Inspired Cooperative Strategies for Optimization. pp. 189-196.

Nakamura Y, Gojobori T, Ikemura T (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res., 28: 292-292.

Ngumbela KC, Ryan KP, Sivamurthy R, Brockman MA, Gandhi RT, Bhardwaj N, Kavanagh DG (2008). Quantitative effect of suboptimal codon usage on translational efficiency of mRNA encoding HIV-1 gag in intact T cells. PLoS One, 3: e2356.

Pesole G, Attimonelli M, Liuni S (1988). A backtranslation method based on codon usage strategy. Nucleic Acids Res. 16: 1715-1728.

Puigbo P, Guzman E, Romeu A, Garcia-Vallve S (2007). OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. 35: W126-W131.

Richardson SM, Wheelan SJ, Yarrington RM, Boeke JD (2006). Gene

Design: rapid, automated design of multikilobase synthetic genes. Genome Res., 16: 550-556.
Sasaguri S, Maruyama J, Moriya S, Kudo T,Kitamoto K , Arioka M (2008). Codon optimization prevents premature polyadenylation of heterologously-expressed cellulases from termite-gut symbionts in Aspergillus oryzae. J. Gen. Appl. Microbiol. 54: 343-351.
Sharp PM, Stenico M, Peden JF, Lloyd AT (1993). Codon usage: mutational bias, translational selection, or both? , 647th Meeting of the Biochemical-Society. Sheffield, U .K, pp. 835-841.
Tokuoka M, Tanaka M, Ono K, Takagi S,Shintani T, Gomi K (2008). Codon optimization increases steady-state mRNA levels in Aspergillus oryzae heterologous gene expression. *Appl*. Environ. Microbiol., 74: 6538-6546.
Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S (2006). Gene Designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinformatics, 7: 285.
Willenbrock H, Friis C, Juncker AS, Ussery DW (2006). An environmental signature for 323 microbial genomes based on codon adaptation indices. Genome Biol., 7: R114.
Wu G, Bashir-Bello N, Freeland SJ (2006). The Synthetic Gene Designer: a flexible web platform to explore sequence manipulation for heterologous expression. Protein Exp. Purif. 47: 441-445.
Wu G, Dress L, Freeland SJ (2007). Optimal encoding rules for synthetic genes: the need for a community effort. Mol. Syst. Biol., 3: 134.
Wu G, Zheng Y, Qureshi I, Zin HT, Beck T, Bulka B , Freeland SJ (2007). SGDB: a database of synthetic genes re-designed for optimizing protein over-expression. Nucleic Acids Res., 35: D76-79.