W526–W531 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki376

JCat: a novel tool to adapt codon usage of a target gene to its potential expression host

Andreas Grote[1,2], Karsten Hiller[1], Maurice Scheer[1,3], Richard Münch[1], Bernd Nörtemann[2], Dietmar C. Hempel[2] and Dieter Jahn[1,]*

[1]Institut für Mikrobiologie, Spielmannstraße 7, [2]Institut für Bioverfahrenstechnik, Gaußstraße 17, Technische Universität Braunschweig, D-38106 Braunschweig, Germany and [3]Fachbereich für Informatik, Am Exer 2, Fachhochschule Wolfenbüttel, D-38302 Wolfenbüttel, Germany

Received February 11, 2005; Revised and Accepted March 7, 2005

ABSTRACT

A novel method for the adaptation of target gene codon usage to most sequenced prokaryotes and selected eukaryotic gene expression hosts was developed to improve heterologous protein production. In contrast to existing tools, JCat (Java Codon Adaptation Tool) does not require the manual definition of highly expressed genes and is, therefore, a very rapid and easy method. Further options of JCat for codon adaptation include the avoidance of unwanted cleavage sites for restriction enzymes and Rho-independent transcription terminators. The output of JCat is both graphically and as Codon Adaptation Index (CAI) values given for the pasted sequence and the newly adapted sequence. Additionally, a list of genes in FASTA-format can be uploaded to calculate CAI values. In one example, all genes of the genome of *Caenorhabditis elegans* were adapted to *Escherichia coli* codon usage and further optimized to avoid commonly used restriction sites. In a second example, the *Pseudomonas aeruginosa exbD* gene codon usage was adapted to *E.coli* codon usage with parallel avoidance of the same restriction sites. For both, the degree of introduced changes was documented and evaluated. JCat is integrated into the PRODORIC database that hosts all required information on the various organisms to fulfill the requested calculations. JCat is freely accessible at http://www.prodoric.de/JCat.

INTRODUCTION

The genetic code is degenerated, e.g. there are multiple codons, up to six different ones, coding for the same amino acid. Depending on the employed codon usage of each organism, various codons are employed differentially (1,2). For example, GC-rich organisms prefer GC-containing codons over AT-containing ones. Therefore, according to the non-randomness of the codon usage, there are optimal and non-optimal codons for each organism. This so-called 'dialect' of codon choice is not universal and unique for each organism. The codon usage of different genes of one organism also relates to their specific rate of expression (3). Currently, one of the central techniques used in biomedical research and biotechnological production processes is the heterologous gene expression. In order to succeed in the expression of genes outside their natural context, several problems were experienced (4). Codon usage, one major factor among others, has a significant impact on heterologous gene expression. Rarely employed codons of the expression host found in the target gene can lead to poorly translated mRNAs, decreased mRNA stability and sometimes to premature termination of translation (5,6). Finally, the usage of rare codons for arginine in *Escherichia coli* can provoke misincorporation of amino acids (7). Currently, there are two ways to prevent the problems described above. The first method is to supplement the gene expression host with the tRNAs that are otherwise infrequent in this organism (8). The second method is to adapt the codon usage of the gene that should be expressed in the host organism (9). Sharp and Li (10) developed the Codon Adaptation Index (CAI). This index permits the calculation of a comparable value for codon usage. The calculation of the CAI requires the definition of highly expressed genes. The codon usage of other genes from this organism is calculated with respect to this subset of genes. The CAI is the prevailing empirical measure of expressivity.

To our knowledge, there are currently only two bioinformatic tools available for the adaptation of codon usage. The first tool is UpGene (11). This tool is truly valuable for the codon optimization of SIV/HIV coding sequences in order to maximize their expression in eukaryotic cells. However, with this

*To whom correspondence should be addressed. Tel: +49 531 391 5801; Fax: +49 531 391 5854; Email: d.jahn@tu-bs.de

tool it is not possible to adapt codon usage to other organisms. Another disadvantage is the fact that UpGene is only available as a stand-alone application that makes its immediate use complicated. The second tool is Codon Optimizer (12). This tool is again only available as a stand-alone application. Since this tool uses CAI as the basic value for codon adaptation, it requires a set of highly expressed genes to be defined manually in order to get an optimal codon optimization. Here, simplicity stands against universality.

Here, we describe a novel and easy-to-use program, called JCat (Java Codon Adaptation Tool), for the adaptation of codon usage to most prokaryotic and some eukaryotic organisms of biotechnological interest. JCat is presented as a web service that offers immediate program usage. The web interface is easy to understand and the calculations are performed in real time. Furthermore, the algorithm does not require a definition of highly expressed genes. Calculations are made in advance with the aid of an algorithm proposed by Carbone *et al*. (13). The results of the calculation are stored in the PRODORIC database, where data of most freely available genomes of sequenced prokaryotes are hosted (14). The PRODORIC database is frequently updated; therefore, JCat is always up to date. Additionally, JCat features the possibility of avoiding cleavage sites for certain restriction enzymes and includes the avoidance of Rho-independent transcription terminators in the codon-optimized DNA sequence. The algorithm for the prediction of Rho-independent transcription terminators was previously proposed by Ermolaeva *et al*. (15).

MATERIALS AND METHODS

Algorithm for the codon optimization

The algorithm is based on the calculation of the CAI (10). Each codon is given a weight with respect to the subset of highly expressed genes defined for the considered organism. The so-called relative adaptiveness of a codon is defined as:

$$w_i = \frac{f_i}{f_{\max(i)}} \qquad \mathbf{1}$$

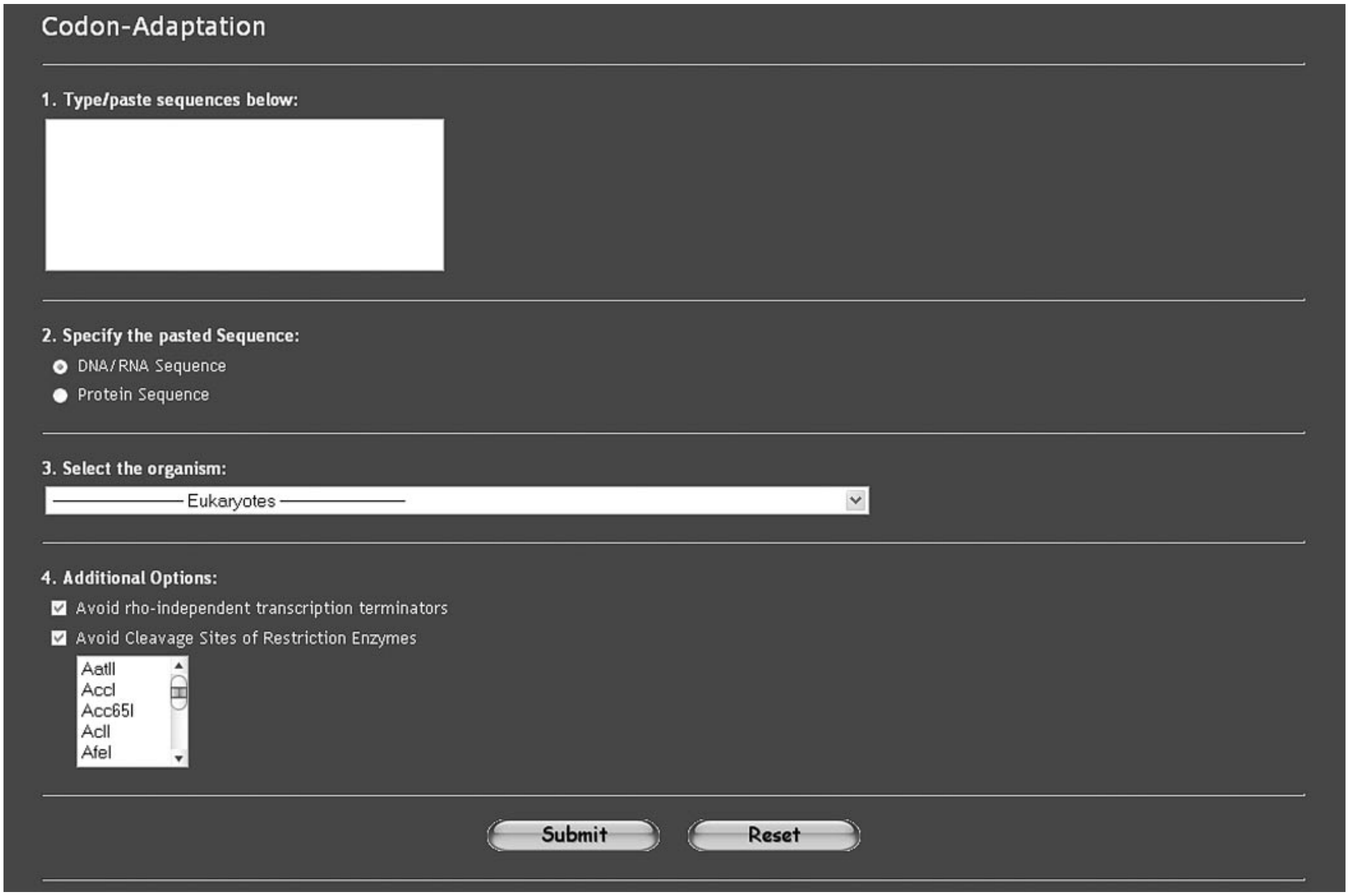
where f_i is the frequency of a codon (i) and $f_{\max(i)}$ is the frequency of the codon most often used to code for the considered amino acid in the subset of highly expressed genes.

The CAI for a gene ‘g’ can be calculated according to Equation 2:

$$\mathrm{CAI_g} = \left(\prod_{i=1}^{N} w_i \right)^{1/N} \qquad \mathbf{2}$$

where N is the number of codons in a gene ‘g’ without the initiation and stop codons.

The calculation of the relative adaptiveness for all genomes in the PRODORIC database was made in advance. The subset



Codon-Adaptation

1. Type/paste sequences below:

2. Specify the pasted Sequence:

◉ DNA/RNA Sequence

○ Protein Sequence

3. Select the organism:

———————— Eukaryotes ————————

4. Additional Options:

☑ Avoid rho-independent transcription terminators

☑ Avoid Cleavage Sites of Restriction Enzymes

AatII
AccI
Acc65I
AclI
AfeI

Figure 1. Screen shot of the web interface for the codon adaptation tool JCat.

of highly expressed genes for each organism was defined by applying the algorithm proposed by Carbone *et al.* (13). The algorithm is based on the assumption that in each genome there is a set of genes with high codon bias. The algorithm is iterative and reduces the set of genes (initially all genes of an organism) during each iteration until only 1% of genes remain with the highest codon bias of the initial set of genes.

The optimization of a given sequence splits into two parts. First, the sequence is examined whether it is either a correct gene sequence or a correct amino acid sequence. Subsequently, depending on the type of sequence, it is translated into an amino acid sequence. The second step is to translate the amino acid sequence into a gene sequence by using the codons that got the highest relative adaptiveness for the amino acid in question. In this way, every amino acid of the sequence is replaced until the whole sequence is retranslated.

Algorithm for the prevention of Rho-independent transcription terminators

The algorithm for the prevention of Rho-independent transcription terminators consists of two parts. The first part recognizes Rho-independent transcription terminators. The second part is dedicated to the improvement of the determined sequence in order to avoid these structures.

The algorithm proposed by Ermolaeva *et al.* (15) is based on the assumption that Rho-independent transcription terminators have two major characteristics. The first one is a defined mRNA stem–loop (hairpin) that can be predicted with the help of an energy scoring function. The second point is a poly-U strand close to the hairpin to allow translation termination (16).

After recognition of stem–loop-like structures and poly-U stretches, codons that are involved with both parameters are adapted separately by replacing optimal codons in the sequence by codons that are not optimal. Once transcription terminators are no longer being found in the adapted sequence, the sequence is stored. Each codon that is in the area of the loop and the corresponding poly-U transcription termination is exchanged. The decision as to which sequence has to be displayed is made with respect to the CAI that is calculated for each adapted sequence without transcription terminators.

Algorithm to prevent cleavage sites of specified restriction endonucleases

The algorithm for the prevention of certain cleavage sites resembles the algorithm described above. For restriction enzymes, cleavage sites are derived from the REBASE database (17). In a first step, cleavage sites are predicted. In a

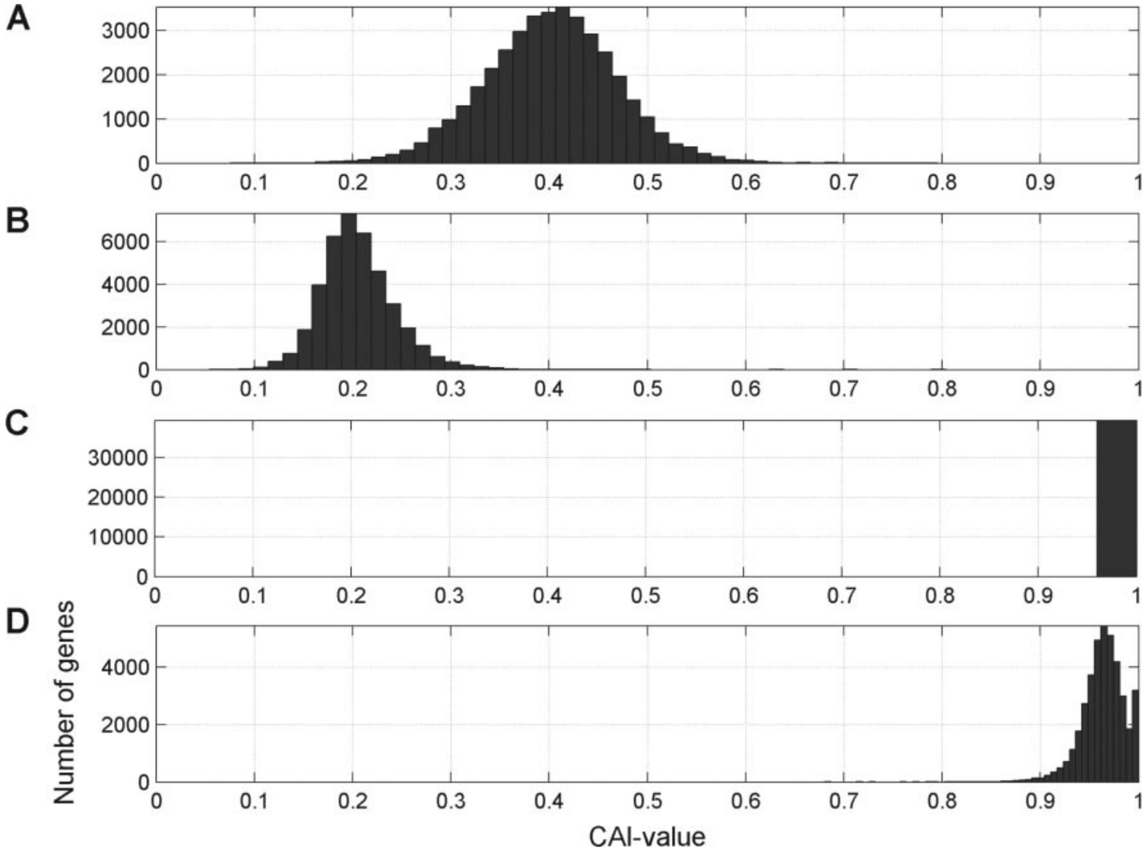


Figure 2. Distribution of CAI values of the *C.elegans* genome under different conditions. (**A**) Calculation of the CAI applying the relative adaptiveness of *C.elegans*. (**B**) Calculation of the CAI applying the relative adaptiveness of *E.coli*. (**C**) Distribution of CAI values of the genes from *C.elegans* after adaptation to the codon usage of *E.coli*. The CAI was calculated applying the relative adaptiveness of *E.coli*. (**D**) Same as shown in (C); however, restriction endonucleases cleavage sites for ApaI, AccI, BamHI, BstXI, ClaI, DraII, EcoO109I, EcoRV, EcoRI, EagI, HincII, HindIII, KpnI, NotI, PstI, SalI, SmaI, SpeI, SacII, SacI, XhoI and XbaI were eliminated, resulting in a decrease of the CAI.

second step, the detected restriction site is eliminated. The cleavage sites are removed using the same method as described above. The optimal codons in the area of the cleavage site are substituted by non-optimal codons. Then, the considered cleavage site is searched again in the DNA. Once no cleavage site is found in the sequence it is stored. After the substitution of all codons in the area of the cleavage site, the CAI, calculated for the various improved sequences, decides which adapted sequence is displayed.

Web interface and analyses

The algorithms were programmed in Java (http://java.sun.com) to take advantage of its object-oriented technology and to allow integration into dynamic web sites using Java Server Page (JSP) technology. Jakarta Tomcat runs as the servlet container and web server on a Linux machine. Tomcat is available as an open source tool at http://jakarta.apache.org/tomcat.

Graphical output of the data is carried out with the help of the Java package JFreeChart. JFreeChart is freely available at http://jfree.org/jfreechart.

Alignments were carried out with ClustalX that is freely available at http://www.biolinux.org/clustalx.html.

RESULTS

Rational of the approach

In basic biochemical research and for various biotechnological applications, heterologous protein production is of central importance. However, there is only a limited number of prokaryotic and eukaryotic production hosts for the wealth of organisms under investigation. The codon usage of the gene of interest and of its desired production host often differ significantly. Frequently, this results in low protein recoveries. In the age of commercial complete gene synthesis, a synthetic target gene with perfect host codon usage offers an attractive alternative to rare tRNA-supplemented strains, which are currently only available for *E.coli*.

In this context, JCat offers the possibility to adapt the codon usage of a gene of interest to the one of a selected organism. For this purpose, the CAI for a list of genes or for a pasted sequence is calculated and displayed. These values provide the basis for the codon optimization. They are also consulted during additional optimization steps, including unwanted restriction site and transcription terminator prevention. There is no need for the manual definition of codon usage of highly expressed genes for the organism of interest.

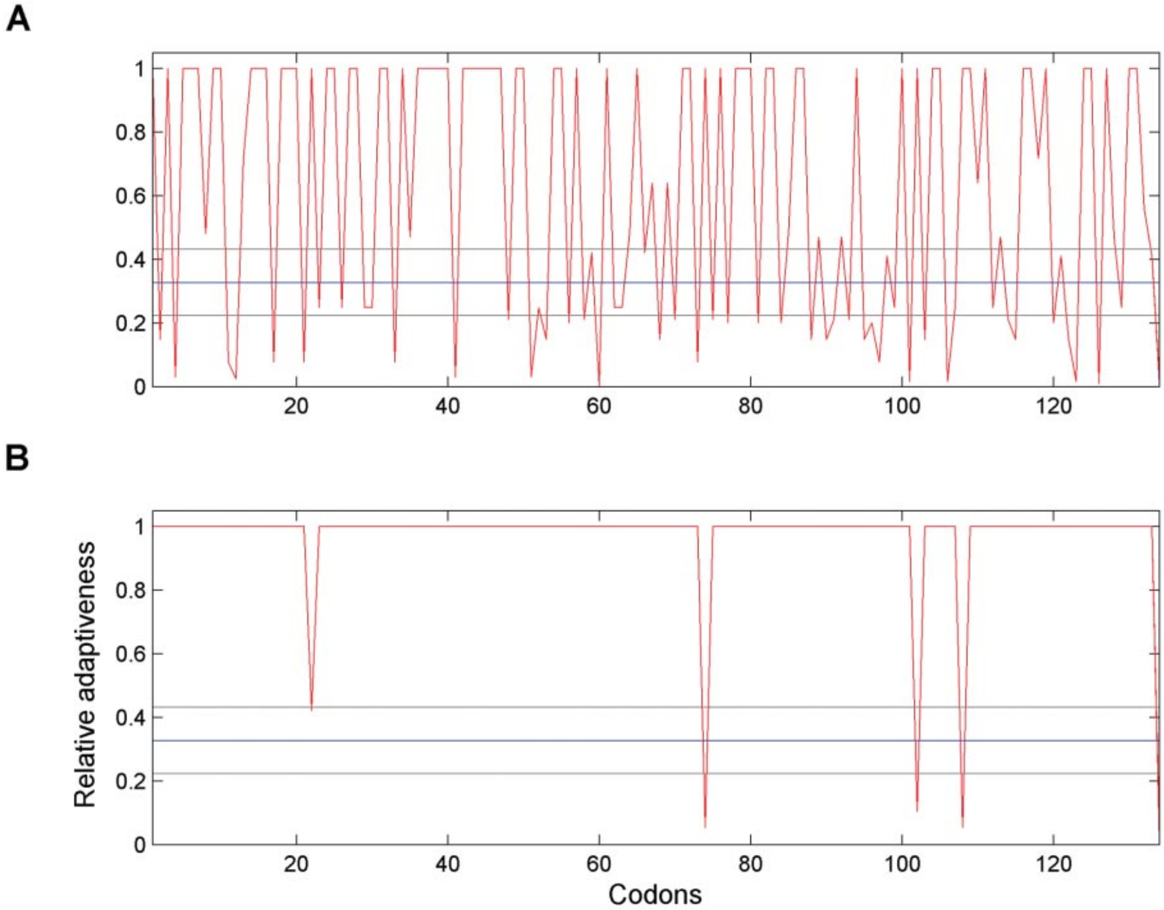


Figure 3. Graphical presentation of the relative adaptiveness (w_{ij}) of the codon usage of the gene for the transport protein ExbD from *P.aeruginosa* to the codon usage of *E.coli*. The mean codon usage of *E.coli* is presented as a blue line. The gray lines above and below the mean codon usage mark the standard deviation of the codon usage. (**A**) Unadapted codon usage. (**B**) *E.coli* adapted codon usage including the elimination of various restriction enzyme cleavage sites.

These have been precalculated for all published prokaryotic and various eukaryotic organisms and are an integral part of JCat.

USE OF THE JCAT WEB INTERFACE

Input

There are several options offered by JCat for codon optimization of DNA sequences (Figure 1). First, the sequence is pasted into a text field. In a second step, the pasted sequence is specified as a DNA, an RNA or an amino acid sequence. Further options for the codon adaptation are offered. Options include the possibility to avoid Rho-independent transcription terminators and certain cleavage sites of defined restriction enzymes in the adapted sequence. Finally, the organism to which the codon usage should be adapted is selected.

Alternatively, for the calculation of the CAI for a list of genes, the web interface provides a menu item called 'CAICalculation'. The selection of this item opens a new webpage with a new form. This form includes a field for file upload. The file must be in valid FASTA-format. Sequences with other letters than A, G, C, T and U are ignored. After selection of the target organism for the calculation, all necessary settings are made.

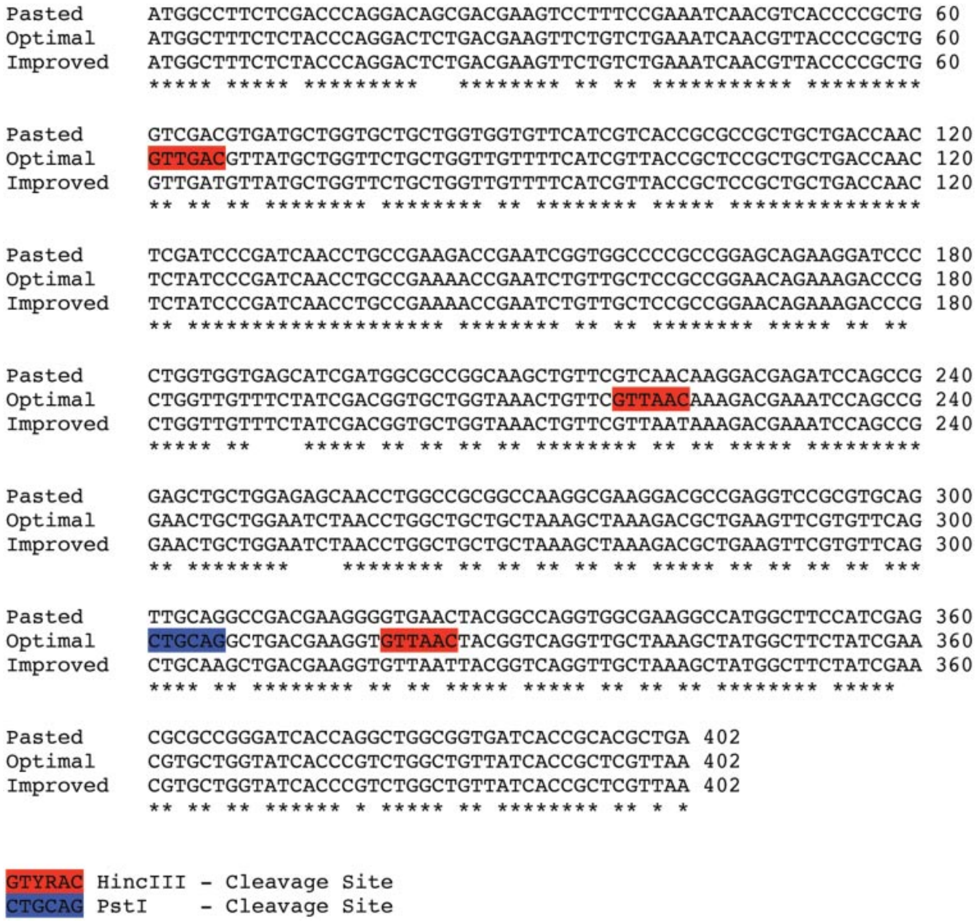
Output

The detailed nature of the output depends on the input options. If a DNA/RNA sequence is to be analyzed, the CAI values are displayed for the pasted sequence and for the optimized sequence. Furthermore, a graphical representation of the relative adaptiveness for each codon of the sequence is available (Figure 3). Again, data for the analyzed and the adapted sequence are shown.

If an amino acid sequence is analyzed, the results are only presented for the adapted sequence.

The output of the calculated CAIs for a list of genes is in the form of a two-column table containing the identifier of the FASTA sequence and individual calculated CAIs.

In order to illustrate the mode of operation of JCat, the CAI values for all genes of the whole *Caenorhabditis elegans* genome were calculated (Figure 2A). Subsequently, the CAI for the *C.elegans* genome was calculated with the relative adaptiveness of *E.coli* to simulate the case of unadapted codon usage (Figure 2B). One of the main differences between natural codon usage of the genes of a certain organism (Figure 2A) and unadapted codon usage of this organism (Figure 2B) is the degree of diversity. The natural codon usage of the *C.elegans* genome varied between 0.2 and 0.6. The unadapted codon usage of the *C.elegans* genome in *E.coli* showed variation



```
Pasted        ATGGCCTTCTCGACCCAGGACAGCGACGAAGTCCTTTCCGAAATCAACGTCACCCCGCTG 60
Optimal       ATGGCTTTCTCTACCCAGGACTCTGACGAAGTTCTGTCTGAAATCAACGTTACCCCGCTG 60
Improved      ATGGCTTTCTCTACCCAGGACTCTGACGAAGTTCTGTCTGAAATCAACGTTACCCCGCTG 60
              ***** ***** *********   ******** ** ** *********** *********

Pasted        GTCGACGTGATGCTGGTGCTGCTGGTGGTGTTCATCGTCACCGCGCCGCTGCTGACCAAC 120
Optimal       GTTGACGTTATGCTGGTTCTGCTGGTTGTTTTCATCGTTACCGCTCCGCTGCTGACCAAC 120
Improved      GTTGATGTTATGCTGGTTCTGCTGGTTGTTTTCATCGTTACCGCTCCGCTGCTGACCAAC 120
              ** ** ** ******** ******** ** ******** ***** ***************

Pasted        TCGATCCCGATCAACCTGCCGAAGACCGAATCGGTGGCCCCGCCGGAGCAGAAGGATCCC 180
Optimal       TCTATCCCGATCAACCTGCCGAAAACCGAATCTGTTGCTCCGCCGGAACAGAAAGACCCG 180
Improved      TCTATCCCGATCAACCTGCCGAAAACCGAATCTGTTGCTCCGCCGGAACAGAAAGACCCG 180
              ** ******************** ******** ** ** ******** ***** ** **

Pasted        CTGGTGGTGAGCATCGATGGCGCCGGCAAGCTGTTCGTCAACAAGGACGAGATCCAGCCG 240
Optimal       CTGGTTGTTTCTATCGACGGTGCTGGTAAACTGTTCGTTAACAAAGACGAAATCCAGCCG 240
Improved      CTGGTTGTTTCTATCGACGGTGCTGGTAAACTGTTCGTTAATAAAGACGAAATCCAGCCG 240
              ***** **    ***** ** ** ** ** ******** ** ** ***** *********

Pasted        GAGCTGCTGGAGAGCAACCTGGCCGCGGCCAAGGCGAAGGACGCCGAGGTCCGCGTGCAG 300
Optimal       GAACTGCTGGAATCTAACCTGGCTGCTGCTAAAGCTAAAGACGCTGAAGTTCGTGTTCAG 300
Improved      GAACTGCTGGAATCTAACCTGGCTGCTGCTAAAGCTAAAGACGCTGAAGTTCGTGTTCAG 300
              ** ********    ******** ** ** ** ** ** ***** ** ** ** ** ***

Pasted        TTGCAGGCCGACGAAGGGGTGAACTACGGCCAGGTGGCGAAGGCCATGGCTTCCATCGAG 360
Optimal       CTGCAGGCTGACGAAGGTGTTAACTACGGTCAGGTTGCTAAAGCTATGGCTTCTATCGAA 360
Improved      CTGCAAGCTGACGAAGGTGTTAATTACGGTCAGGTTGCTAAAGCTATGGCTTCTATCGAA 360
              **** ** ******** ** ** ***** ***** ** ** ** ******** *****

Pasted        CGCGCCGGGATCACCAGGCTGGCGGTGATCACCGCACGCTGA 402
Optimal       CGTGCTGGTATCACCCGTCTGGCTGTTATCACCGCTCGTTAA 402
Improved      CGTGCTGGTATCACCCGTCTGGCTGTTATCACCGCTCGTTAA 402
              ** ** ** ****** * ***** ** ******** ** * *

GTYRAC HincIII - Cleavage Site
CTGCAG PstI    - Cleavage Site
```

Figure 4. Nucleotide sequence comparison of *P.aeruginosa exbD* before (pasted) and after (optimal) adaptation to *E.coli* codon usage, and with the elimination of restriction sites (improved).

of the CAIs only between ~0.1 and 0.3, which is besides that of course lower.

Finally, the codons of the *C.elegans* genome were adapted to the genome of *E.coli*. If the optimization is executed without further options, the CAIs of all genes should be shifted to a value of 1.0 (Figure 2C). In our example, we further chose to optimize the analyzed DNA sequences of the *C.elegans* genome for the avoidance of the cleavage sites of different standard restriction enzymes (ApaI, AccI, BamHI, BstXI, ClaI, DraII, EcoO109I, EcoRV, EcoRI, EagI, HincII, HindIII, KpnI, NotI, PstI, SalI, SmaI, SpeI, SacII, SacI, XhoI and XbaI). The restriction enzymes whose cleavage sites were avoided are the restriction enzymes that are located on the multiple cloning site of the widely spread vector pBluescript.

Figure 2D shows the efficiency of the algorithm. The codon usage was optimized for each gene of the *C.elegans* genome to *E.coli* codon usage with the exclusion of certain restriction sites, demonstrating the efficient optimization of codon usage. Most CAI values lay between 0.9 and 1.0.

However, the usual laboratory application of the JCat tool might rather be codon optimization of single genes of interest. Therefore, in a second example we adapted the codon usage of the ExbD transport protein from *Pseudomonas aeruginosa* to the codon usage of *E.coli* (Figure 3). The unadapted codon usage of the *P.aeruginosa exbD* gene in *E.coli* is given in Figure 3A. The codon usage adaptation again avoided the restriction enzyme cleavage sites of the multiple cloning site of pBluescript. In a first step, every codon was substituted by the corresponding optimal codon of *E.coli*. In a second step, the algorithm searched for cleavage sites. In this example, cleavage sites were found in position 60 (HincII), 216 (HincII), 300 (PstI) and 318 (HincII) of the DNA. In a third step, the codons in the area of these sites were substituted by non-optimal codons. Figure 3B gives the graphical presentation of the codon optimization. Obviously, the codons with the number 22, 74, 102 and 108 have been exchanged by non-optimal codons. These codons were in the area of the cleavage sites. Figure 4 compares the original *P.aeruginosa* DNA sequence with that optimized for *E.coli* demonstrating the degree of optimization.

In addition to adaptation to *E.coli* codon usage demonstrated here, every known gene can be adapted to the codon usage of all prokaryotic organisms with completed genome, as well as to those of selected eukaryotes, including *Saccharomyces cerevisae*, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *C.elegans*. Finally, the codon usage of homologous genes of a gene expression host might be subject to optimization since optimal codon usage is not used throughout the complete genome.

ACKNOWLEDGEMENTS

The authors thank Dr Barbara Schulz for critical proofreading of the manuscript. This work was funded by the German Bundesministerium für Bildung und Forschung (BMBF) for the Bioinformatics Competence Center Imtergenomics and Sonderforschungsbereich 578 'From gene to product', Deutsche Forschungsgemeinschaft; Fonds der Chemischen Industrie. Funding to pay the Open Access publication charges for this article was provided by BMBF.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Ikemura,T. (1985) Codon usage and tRNA content on unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.
3. Gouy,M. and Gaultier,C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–7074.
4. Gustafsson,C., Govindarajan,S. and Minshull,J. (2004) Codon bias and heterologous protein expression. *Trends Biotechnol.*, **22**, 346–353.
5. Sorensen,M.A., Kurland,C.G. and Pedersen,S. (1989) Codon usage determine translation rate in *Escherichia coli*. *J. Mol. Biol.*, **207**, 365–377.
6. Goldman,E., Rosenberg,A.H., Zubay,G. and Studier,F.W. (1995) Consecutive low-usage leucine codons block translation only when near the 5′ end of a message in *Escherichia coli*. *J. Mol. Biol.*, **245**, 467–473.
7. Calderone,T.L., Stevens,R.D. and Oas,T.G. (1996) High-level misincorporation of lysine for arginine at AGA codon in a fusion protein expressed in *Escherichia coli*. *J. Mol. Biol.*, **262**, 407–412.
8. Brinkmann,U., Mattes,R.E. and Buckel,P. (1989) High-level expression of recombinant genes in *Escherichia coli* is dependent on the availability of the dnaY gene product. *Gene*, **85**, 109–114.
9. Hernan,R.A., Hui,H.L., Andracki,M.E., Noble,R.W., Sligar,S.G., Walder,J.A. and Walder,R.Y. (1992) Human hemoglobin expression in *Escherichia coli*: importance of optimal codon usage. *Biochemistry*, **31**, 8619–8628.
10. 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 Acid Res.*, **15**, 1281–1295.
11. Gao,W., Rzewski,A., Sun,H., Robbins,P.D. and Gambotto,A. (2004) UpGene: Application of a web-based DNA codon optimization algorithm. *Biotechnol. Prog.*, **20**, 443–448.
12. Fuglsang,A. (2003) Codon optimizer: a freeware tool for codon optimization. *Protein Expr. Purif.*, **31**, 247–249.
13. Carbone,A., Zinovyev,A. and Kepes,F. (2003) Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, **19**, 2005–2015.
14. Munch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.*, **31**, 266–269.
15. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
16. d'Aubenton Carafa,Y., Brody,E. and Thermes,C. (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem–loop structures. *J. Mol. Biol.*, **216**, 835–858.
17. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2003) REBASE: restriction enzymes and methyltransferases. *Nucleic Acids Res.*, **31**, 418–420.