CodonO: codon usage bias analysis within and across genomes

Michael C. Angellotti, Shafquat B. Bhuiyan, Guorong Chen and Xiu-Feng Wan*

Systems Biology Laboratory, Department of Microbiology, Miami University, Oxford, OH 45056, USA

Received January 31, 2007; Revised April 4, 2007; Accepted May 1, 2007

ABSTRACT

Synonymous codon usage biases are associated with various biological factors, such as gene expression level, gene length, gene translation initiation signal, protein amino acid composition, protein structure, tRNA abundance, mutation frequency and patterns, and GC compositions. Quantification of codon usage bias helps understand evolution of living organisms. A codon usage bias pipeline is demanding for codon usage bias analyses within and across genomes. Here we present a CodonO webserver service as a user-friendly tool for codon usage bias analyses across and within genomes in real time. The webserver is available at http//www.sysbiology.org/CodonO. Contact: wanhenry@yahoo.com.

INTRODUCTION

Within the standard genetic codes, all amino acids except Met and Trp are coded by more than one codon, which are called synonymous codons. DNA sequence data from diverse organisms clearly show that synonymous codons for any amino acid are not used with equal frequency, and these biases are as the consequence of natural selection during evolution. Extensive studies have shown that synonymous codon usage biases are associated with various biological factors, such as gene expression level, gene length, gene translation initiation signal, protein amino acid composition, protein structure, tRNA abundance, mutation frequency and patterns, and GC compositions (1–11). Quantification of codon usage bias, especially at genomic scale, helps understand evolution of living organisms.

Many different approaches have been developed in the past few decades. These methods may be grouped into two categories: (i) methods based on the statistical distribution, such as codon-usage preference bias measure (CPS) based on χ^2 (12) and scaled χ^2 analyses (13); (ii) methods using a group of gene sequences as reference, which can be 'optimal codons' [e.g. codon bias index (14)], a defined set of highly expressed genes [e.g. codon preference statistics (15) and codon adaptation index (16)], a defined gene class [e.g. Codon Bias (7)], or all genes in the entire genome [e.g. the Shannon Information Method (17)]. Most of existing computational approaches are only suitable for the comparison of codon usage bias within a single genome. In order to overcome these limitations, we developed a new informatics method based on Shannon informational theory, referred to as synonymous codon usage order (SCUO), which enables a measurement of synonymous codon usage bias within and across genomes (3,12). The review and comparison of SCUO and current available methods are detailed in Wan *et al.* (18). Several computational software packages or webservers, for instance, CodonW (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html) and JCAT (19), have been developed to measure Codon Adaptation Index (CAI) for genes. JCAT also integrates intrinsic terminators and enzyme digestion sites into their analyses.

Codon usage analyses within and across genomes will facilitate the understanding of evolution and environmental adaptation of living organisms. GC compositions have been shown to drive codon and amino-acid usages thus affect codon usage bias (20). Thus, it will be critical to study the correlation between GC compositions and codon usage bias. Previously, we have developed an analytical model to quantify synonymous codon usage bias by GC compositions based on SCUO (11). However, it is still laborious to perform codon usage analyses within and across genomes based on our knowledge, there is not any available tool designed for these purposes. The CodonO webserver described here is a pipeline for codon usage bias analyses within and across genomic sequences as well as a tool for studying the correlation between codon usage bias and GC compositions, especially for microbial species. Different from the standalone CodonO we developed earlier (10,11,18), CodonO webserver has the following additional functions: (i) besides allowing the users to compare their submissions, it connects genomic database and perform analyses in real time; (ii) it can be used to study the correlation between

*To whom correspondence should be addressed. Tel: +1-513-529-0426; Fax: +1-513-529-2431; Email: wanhenry@yahoo.com
Present address:
Xiu-Feng Wan, Molecular Virology and Vaccine Branch, Influenza Division, CDC, Atlanta, GA 30333, USA

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

SCUO and GC compositions; (iii) it performs statistical comparison of SCUO within and across genomes; (iv) besides SCUO values, it extracts and displays codon usage frequency table as well as the gene attribute for each gene from the genomic database; and (v) it provides a user-friendly interface.

MATERIALS AND METHODS

Synonymous codon usage order measurement

CodonO webserver employs the synonymous codon usage order (SCUO) measurement as the method to calculate synonymous codon usage biases. The details about the SCUO concept and method have been described previously (10,11,18). Simply, we calculate the entropy of the i-th amino acid in a sequence

$$H_i = -\sum_{j=1}^{n_i}\left(\frac{x_{ij}}{\sum_{j=1}^{n_i} x_{ij}}\right)\log\left(\frac{x_{ij}}{\sum_{j=1}^{n_i} x_{ij}}\right)$$

Where $1 \leqslant i \leqslant 18$, j is the codon for the i-th amino acid, $1 \leqslant j \leqslant 6$ for leucine, $1 \leqslant j \leqslant 2$ for tyrosine, etc. If the synonymous codons for the i-th amino acid were used at random, one would expect a uniform distribution of them as representatives for the i-th amino acid. Thus, the maximum entropy for the i-th amino acid in each sequence is

$$H_i^{\max} = -\log\frac{1}{n_i}$$

Thus, we can calculate SCUO for the i-th amino acid in each sequence.

$$SCUO_i = \frac{H_i^{\max} - H_i}{H_i^{\max}}$$

Then the average SCUO for each sequence can be represented to summarize the SCUO from each amino acid.

$$SCUO = \sum_{i=1}^{n_i}\left(\frac{\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{18}\sum_{j=1}^{n_i} x_{ij}}\right)SCUO_i$$

The $SCUO$ represents the synonymous codon usage bias for the entire sequence, and j is the codon for the i-th amino acid. Thus, $0 \leqslant SCUO \leqslant 1$, and a larger $SCUO$ denotes a higher codon usage bias in the sequence.

Statistical methods

CodonO webserver can perform codon usage bias analyses within genomes using Tukey statistical analysis (21) and across genomes using Wilcoxon Two Sample Test (22). Tukey statistical analysis is a simple and powerful method for estimating outliers for a population, which can be either a normal distribution or a non-normal distribution. We adapted the percentile calculation from JMP method (SAS, Inc., Cary, NC USA).

$$\frac{q}{100(n+1)} = R = IR + FR$$

where n is the number of data points; IR is the integer part of R while FR is the fraction part of R. Then,

q-th percentile $= IR$-th observation $+ FR[(IR+1)$-th observation $- IR$-th observation$]$

The Tukey outliers are genes with SCUO values less than Q1 − 1.5IQR or greater than Q3 + 1.5IQR, where IQR represent Interquartile range. IQR is the difference between 75th percentile and 25th percentile SCUO.

The Wilcoxon Two Sample Test (22) is utilized to test null hypothesis that the distributions of SCUO from two groups of sequences (e.g. genomes) are the same. The Wilcoxon Two Sample Test is a sensitive test in two groups even their values are not Normal distributed.

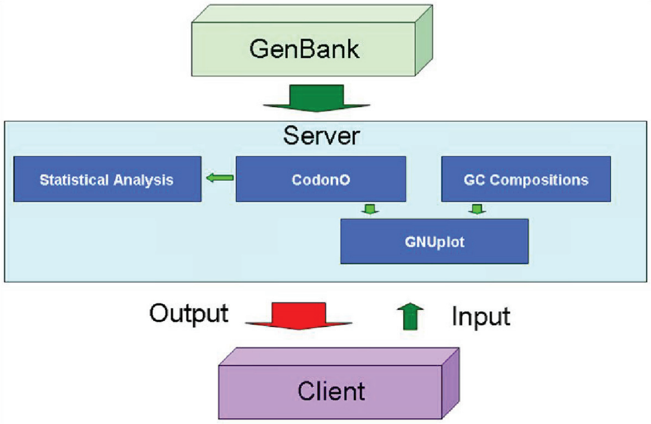
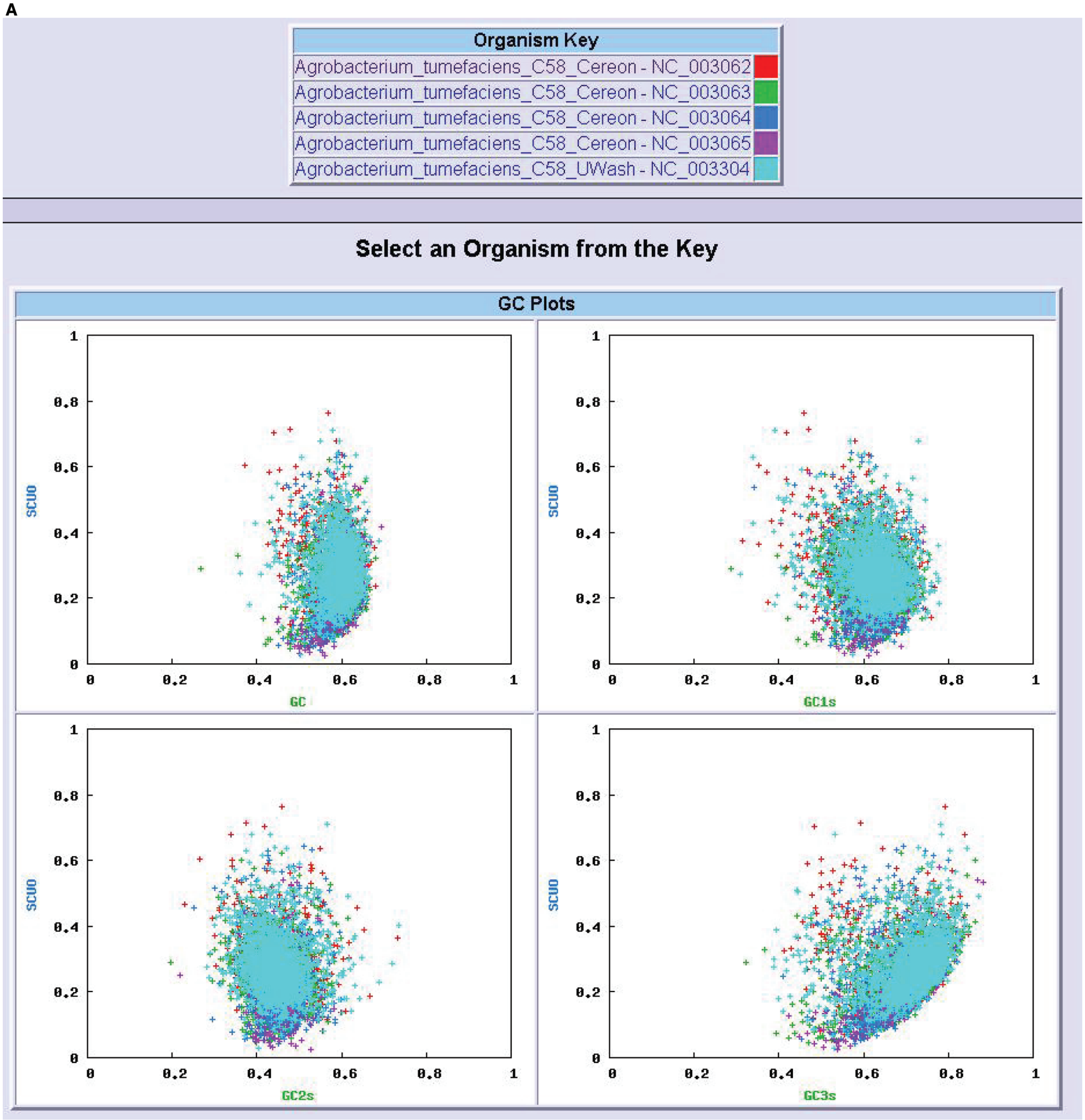


Figure 1. Simplified CodonO webserver infrastructure.

Features

As shown in Figure 1, CodonO server is directly connected and updated with GenBank genomic database daily. The user can define and select one or multiple genomes for analyses at the same time. The users can upload their own datasets as well. The underlying computations include synonymous codon usage order (SCUO) and GC composition measurements, and the latter includes GC, GC1s, GC2s and GC3s, where GC is the overall GC composition, GC1s is the GC composition at the first site of a codon, GC2s is the GC composition at the second site of a codon, and GC3s is the GC composition at the third site of a codon. The results will be plotted in a two-dimensional graph, by which the clients can visualize and compare the results. The webserver can display the results for multiple genomes in the same plots, by which, the users can analyse the two dimensional differences (GC/GC1s/GC2s/GC3s versus SCUO) between genes within and across genomes (Figure 2A) (11). Generally, a very low or very high GC composition is associated with a large codon usage bias. It has been shown that codon usage bias in some bacteria and archaea were affected by GC composition and environment condition (e.g. temperature) (23). Thus, the users can perform these types of analyses based on their own preferences.

As mentioned in the 'Statistical and methods' section, the webserver can identify the outliers for a genome or a



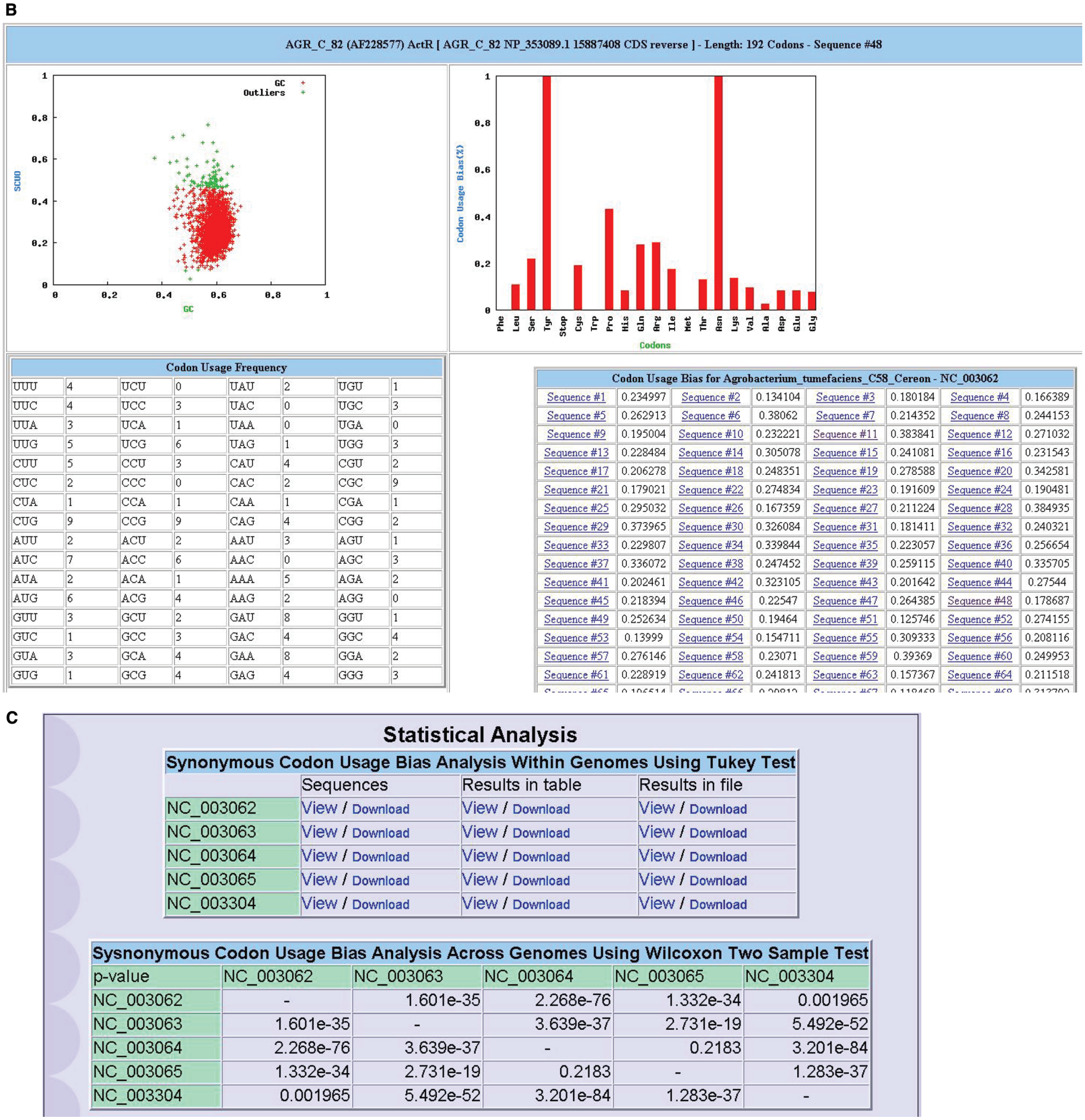
Organism Key	
Agrobacterium_tumefaciens_C58_Cereon - NC_003062	🟥
Agrobacterium_tumefaciens_C58_Cereon - NC_003063	🟩
Agrobacterium_tumefaciens_C58_Cereon - NC_003064	🟦
Agrobacterium_tumefaciens_C58_Cereon - NC_003065	🟪
Agrobacterium_tumefaciens_C58_UWash - NC_003304	🟦

Select an Organism from the Key

GC Plots

Figure 2. (**A**) Visualization of the correlation between synonymous codon usage bias and GC compositions; (**B**) Visualization of synonymous codon usage bias for each gene in a specific genome; (**C**) Statistical analysis of synonymous codon usage bias.

group of sequences based on Tukey statistical analysis (21). The clients can pick and select the ‘outlier’ from the plot and find associated information for each codon and annotation information of a specific gene (Figure 2B), in which the outliers are marked in different color from the other members in the SCUO population. To compare the statistical analyses across genomes, the CodonO webserver applys the Wilcoxon Two Sample Test (22) to compare whether the SCUO populations are the same or not between different genomes. The *P*-values from statistical comparison between genomes are listed in table (Figure 2C), and a *P*-value less than 0.05 informs a significant difference between two SCUO populations compared.



AGR_C_82 (AF228577) ActR [AGR_C_82 NP_353089.1 15887408 CDS reverse] - Length: 192 Codons - Sequence #48

Codon Usage Frequency

UUU	4	UCU	0	UAU	2	UGU	1
UUC	4	UCC	3	UAC	0	UGC	3
UUA	3	UCA	1	UAA	0	UGA	0
UUG	5	UCG	6	UAG	1	UGG	3
CUU	5	CCU	3	CAU	4	CGU	2
CUC	2	CCC	0	CAC	2	CGC	9
CUA	1	CCA	1	CAA	1	CGA	1
CUG	9	CCG	9	CAG	4	CGG	2
AUU	2	ACU	2	AAU	3	AGU	1
AUC	7	ACC	6	AAC	0	AGC	3
AUA	2	ACA	1	AAA	5	AGA	2
AUG	6	ACG	4	AAG	2	AGG	0
GUU	3	GCU	2	GAU	8	GGU	1
GUC	1	GCC	3	GAC	4	GGC	4
GUA	3	GCA	4	GAA	8	GGA	2
GUG	1	GCG	4	GAG	4	GGG	3

Codon Usage Bias for Agrobacterium_tumefaciens_C58_Cereon - NC_003062

Sequence #1	0.234997	Sequence #2	0.134104	Sequence #3	0.180184	Sequence #4	0.166389
Sequence #5	0.262913	Sequence #6	0.38062	Sequence #7	0.214352	Sequence #8	0.244153
Sequence #9	0.195004	Sequence #10	0.232221	Sequence #11	0.383841	Sequence #12	0.271032
Sequence #13	0.228484	Sequence #14	0.305078	Sequence #15	0.241081	Sequence #16	0.231543
Sequence #17	0.206278	Sequence #18	0.248351	Sequence #19	0.278588	Sequence #20	0.342581
Sequence #21	0.179021	Sequence #22	0.274834	Sequence #23	0.191609	Sequence #24	0.190481
Sequence #25	0.295032	Sequence #26	0.167359	Sequence #27	0.211224	Sequence #28	0.384935
Sequence #29	0.373965	Sequence #30	0.326084	Sequence #31	0.181411	Sequence #32	0.240321
Sequence #33	0.229807	Sequence #34	0.339844	Sequence #35	0.223057	Sequence #36	0.256654
Sequence #37	0.336072	Sequence #38	0.247452	Sequence #39	0.259115	Sequence #40	0.335705
Sequence #41	0.202461	Sequence #42	0.323105	Sequence #43	0.201642	Sequence #44	0.27544
Sequence #45	0.218394	Sequence #46	0.22547	Sequence #47	0.264385	Sequence #48	0.178687
Sequence #49	0.252634	Sequence #50	0.19464	Sequence #51	0.125746	Sequence #52	0.274155
Sequence #53	0.13999	Sequence #54	0.154711	Sequence #55	0.309333	Sequence #56	0.208116
Sequence #57	0.276146	Sequence #58	0.23071	Sequence #59	0.39369	Sequence #60	0.249953
Sequence #61	0.228919	Sequence #62	0.241813	Sequence #63	0.157367	Sequence #64	0.211518

Statistical Analysis

Synonymous Codon Usage Bias Analysis Within Genomes Using Tukey Test

	Sequences	Results in table	Results in file
NC_003062	View / Download	View / Download	View / Download
NC_003063	View / Download	View / Download	View / Download
NC_003064	View / Download	View / Download	View / Download
NC_003065	View / Download	View / Download	View / Download
NC_003304	View / Download	View / Download	View / Download

Sysnonymous Codon Usage Bias Analysis Across Genomes Using Wilcoxon Two Sample Test

p-value	NC_003062	NC_003063	NC_003064	NC_003065	NC_003304
NC_003062	-	1.601e-35	2.268e-76	1.332e-34	0.001965
NC_003063	1.601e-35	-	3.639e-37	2.731e-19	5.492e-52
NC_003064	2.268e-76	3.639e-37	-	0.2183	3.201e-84
NC_003065	1.332e-34	2.731e-19	0.2183	-	1.283e-37
NC_003304	0.001965	5.492e-52	3.201e-84	1.283e-37	-

Figure 2. Continued.

Implementation

The programs in this solution package are written in C/C++ or Java. The shell scripts are written in korn shell script in order to achieve high performance. GNUPlot is used for visualization. Cascading style sheets (CSS) are used for a consistent look across the pages. This also enables to change the overall design just by replacing the CSS definition file. PHP has been used as server side scripting and is written in C. In order to achieve high performance for computing in a genomic scale, we apply hash function or a binary tree, which enables that the codon usage analyses have a time complexity of *O*(nlog(n)) or *O*(n). The webservers have also designed special functions targeting the security and concurrency issues.

ACCESS

CodonO has been tested on Microsoft Internet Explorer, Netscape and Mozilla Firefox. The users need JavaScript to obtain full function of CodonO server. The webserver is available at http//www.sysbiology.org/CodonO/. This webserver can be run in a real time manner. The users can compare the maximum of 16 genomes for comparative analyses at the same time.

CONCLUSIONS

In summary, CodonO webserver has three major computational features for codon usage bias analyses: (i) it calculates the codon usage bias for one or more genomes; (ii) it compares and visualizes the correlation between codon usage bias and GC compositions; (iii) it performs statistical analyses for codon usage bias within and across genomes. Thus, CodonO provides an efficient user friendly web service for codon usage bias analyses across and within genomes using SCUO in real time.

ACKNOWLEDGEMENTS

We are grateful to Dr Steven Hutcheson from University of Maryland for his critical suggestion. Funding to pay the Open Access publication charges for this article was provided by the start-up funds of Miami University.

Conflict of interest statement. None declared.

REFERENCES

1. Bains,W. (1987) Codon distribution in vertebrate genes may be used to predict gene length. *J. Mol. Biol.*, **197**, 379–388.
2. D'Onofrio,G., Ghosh,T.C. and Bernardi,G. (2002) The base composition of the genes is correlated with the secondary structures of the encoded proteins. *Gene*, **300**, 179–187.
3. Bernardi,G. and Bernardi,G. (1986) Compositional constraints and genome evolution. *J. Mol. Evol.*, **24**, 1–11.
4. Gouy,M. and Gautier,C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–7074.
5. Gu,W., Zhou,T., Ma,J., Sun,X. and Lu,Z. (2004) The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. *Biosystems*, **73**, 89–97.
6. 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.
7. Karlin,S. and Mrazek,J. (1996) What drives codon choices in human genes? *J. Mol. Biol.*, **262**, 459–472.
8. Lobry,J.R. and Gautier,C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Res.*, **22**, 3174–3180.
9. Ma,J., Campbell,A. and Karlin,S. (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.*, **184**, 5733–5745.
10. Wan,X.F., Xu,D. and Zhou,J. (2003) In Dagli, (ed.), *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME Press, New York, Vol. 13, pp. 1101–1118.
11. Wan,X.F., Xu,D., Kleinhofs,A. and Zhou,J. (2004) Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.*, **4**, 19.
12. McLachlan,A.D., Staden,R. and Boswell,D.R. (1984) A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Res.*, **12**, 9567–9575.
13. Shields,D.C. and Sharp,P.M. (1987) Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. *Nucleic Acids Res.*, **15**, 8023–8040.
14. Bennetzen,J.L. and Hall,B.D. (1982) Codon selection in yeast. *J. Biol. Chem.*, **257**, 3026–3031.
15. Gribskov,M., Devereux,J. and Burgess,R.R. (1984) The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.*, **12**, 539–549.
16. 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.
17. Zeeberg,B. (2002) Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome Res.*, **12**, 944–955.
18. Wan,X.F., Xu,D. and Zhou,J. (2006) CodonO: a new informatics method measuring synonymous codon usage bias. *Int. J. General Syst.*, **35**, 109–125.
19. Grote,A., Hiller,K., Scheer,M., Munch,R., Nortemann,B., Hempel,D.C. and Jahn,D. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.*, **33**, W526–W531.
20. Knight,R.D., Freeland,S.J. and Landweber,L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.*, **2**, RESEARCH0010.
21. Tukey,J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Inc.
22. Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics*, **1**, 80–83.
23. Lynn,D.J., Singer,G.A. and Hickey,D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.*, **30**, 4272–4277.