Bioinformation by Biomedical Informatics Publishing Group **open access**
www.bioinformation.net **Hypothesis**

Study of codon bias perspective of fungal xylanase gene by multivariate analysis

Smriti Shrivastava, Raju Poddar, Pratyoosh Shukla*, Kunal Mukhopadhyay

Department of Biotechnology, Birla Institute of Technology (Deemed University), Mesra, Ranchi, India;
Dr. Pratyoosh Shukla - E-mail: pshukla@bitmesra.ac.in; * Corresponding author

Received February 1, 2009; revised March 2, 2009; accepted March 8, 2009; published July 27, 2009

Abstract:
Fungal xylanases has important applications in food, baking, pulp and paper industries in addition to various other industries. Xylanases are produced extensively by both bacterial and fungal sources and has tremendous potential of being active at extremes of temperature and pH. In the present study an effort has been made to explore the codon bias perspective of this potential enzyme using bioinformatics tools. Multivariate analysis has been used as a tool to study codon bias perspectives of xylanases. It was further observed that the codon usage of xylanases genes from different fungal sources is not similar and to reveal this phenomenon the relative synonymous codon usage (RSCU) and base composition variation in fungal xylanase genes were also studied. The codon biasing data like GC content at third position (GC_{3S}), effective codon number (N_C), codon adaptive index (CAI) were further analyzed with statistical softwares like Sigma1plot 9.0 and Systat 11.0. Furthermore, study of translation selection was also performed to verify the influences of codon usage variation among the 94 xylanase genes. In the present study xylanase gene from 12 organisms were analyzed and codon usages of all xylanases from each organism were compared separately. Analysis indicates biased codon among all 12 fungi taken for study with *Aspergillus nidulans*, *Chaetomium globosum*, *Aspergillus terreus* and *Aspergillus clavatus* showing maximum biasing. N_C plot and correspondence analysis on relative synonymous codon usage indicate that mutation bias and translation selection influences codon usage variation in fungal xylanase gene. To reveal the relative synonymous codon usage and base composition variation in xylanase, 94 genes from 12 fungi were used as model system.

Keywords: relative synonymous codon usage, correspondence analysis, translation selection, multivariate statistical analysis, xylanases.

Background:
The analysis of codon usage patterns can be traced back to when the first molecular sequence databases were being collated **[1]**. Since then, a great many different causes and consequences of codon usage variation have been identified **[2]**. Amino acid usage also varies between proteins and this variation has also been shown to correlate with the properties of the proteins **[3]**. Codon usage data are used as a guide to direct back-translation of protein sequences to their probable DNA sequences, to identify protein-coding regions of DNA **[4]** and to identify regions that probably do not encode a protein.

In order to evaluate codon and amino acid usage variation, multivariate analysis options are available. Correspondence analysis (CA) **[5]** is the most popular and appropriate multivariate analysis method for contingency table data such as codon usage values. The program also has some principal components analysis (PCA) methods implemented for comparative purposes. CA can identify the major sources of variation in the dataset. The output from a CA can be used to evaluate other aspects of the genes, such as base composition, expressivity, aromaticity, location on the genome, etc. In the event that the user wishes to perform a more in-depth multivariate analysis, there is an option of writing the data to disk in ADE-compatible format **[6]**.

A synonymous codon usage of an organism is not random **[7]** and varies not only among genomes but also among genes of a given genome. Mutational bias **[8]**, translational selection **[1]** and replication-transcriptional selection **[9]** are responsible for codon usage variation.

RSCU is dominantly used as one of the best indicators of bias. To investigate major trends in codon usage variation among genes, correspondence analysis on RSCU values have been used widely **[10]**; **[11]**; **[12]**. One distinct advantage of RSCU values is that when correspondence analysis is applied on RSCU values, optimal codons can be selected for high translation efficiency **[11]**. RSCU is defined as the ratio of observed frequency of codons to the expected frequency if all the synonymous codons for those amino acids are used equally **[13]**. RSCU values greater than 1.0 indicate that the corresponding codon is more frequently used than expected, whereas the reverse is true for RSCU values less than 1.0. GC_{3S} is the frequency of (G + C) and A_{3S}, T_{3S}, G_{3S} and C_{3S} are the frequencies of A, T, G and C at the synonymous third positions of codons. N_C, the effective number of codons used by a gene, is generally used to measure the bias of synonymous codons and independent of amino acid compositions and codon number **[14]**. The values of *Nc* range from 20 (when one codon is used per amino acid) to 61 (when all codons are used with equal probability). Highly biased genes are generally highly expressed **[15]**.

Bioinformation by Biomedical Informatics Publishing Group
www.bioinformation.net
open access
Hypothesis

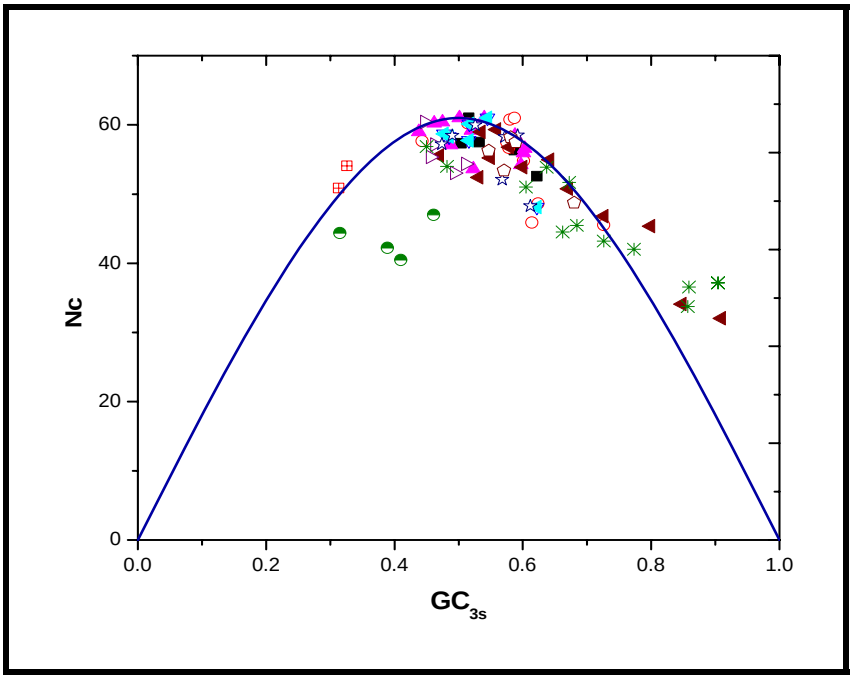

Figure 1: N_C plot of 94 xylanase from 12 fungi. The organisms are represented as follows: **:** *Aspergillus clavatus,* **:** *Aspergillus fumigatus,* **:** *Aspergillus nidulans,* **:** *Aspergillus niger,* **:** *Aspergillus terreus,* **:** *Chaetomium globosum,* **:** *Giberrella zeae,* **:** *Magnaporthe griesa,* **:** *Neosartorya fischeri,* **:** *Neurospora crasa,* **:** *Pistia stipitis,* **:** *Saccharomyces cereviseae.*

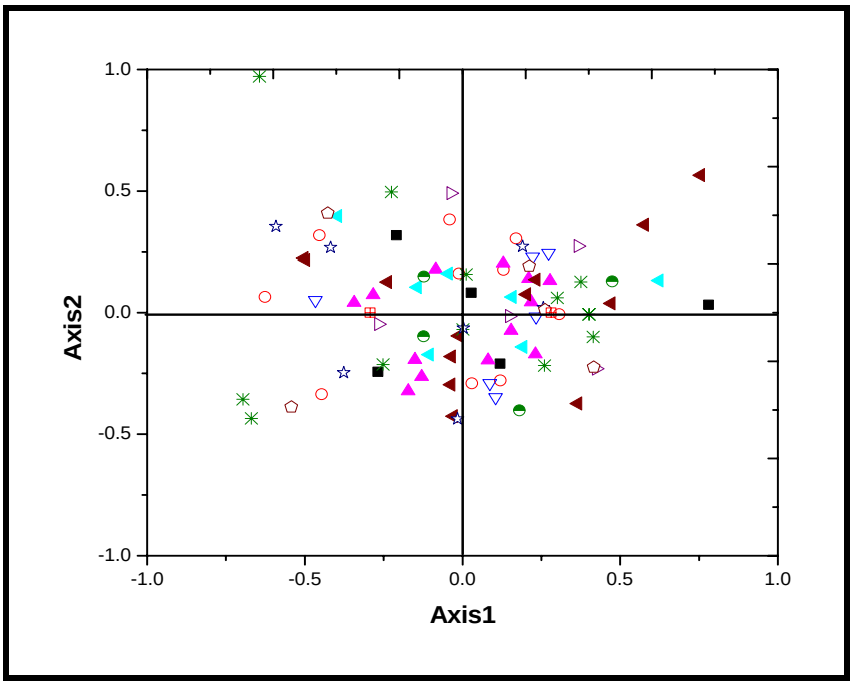

Figure 2: Position of 94 xylanase coding gene from 12 fungi along the two major axes of variation in the correspondence analysis on RSCU values. The organisms are represented as follows: **:** *Aspergillus clavatus,* **:** *Aspergillus fumigatus,* **:** *Aspergillus nidulans,* **:** *Aspergillus niger,* **:** *Aspergillus terreus,* **:** *Chaetomium globosum,* **:** *Giberrella zeae,* **:** *Magnaporthe griesa,* **:** *Neosartorya fischeri,* **:** *Neurospora crasa,* **:** *Pistia stipitis,* **:** *Saccharomyces cereviseae.*

Bioinformation by Biomedical Informatics Publishing Group
www.bioinformation.net
open access
Hypothesis

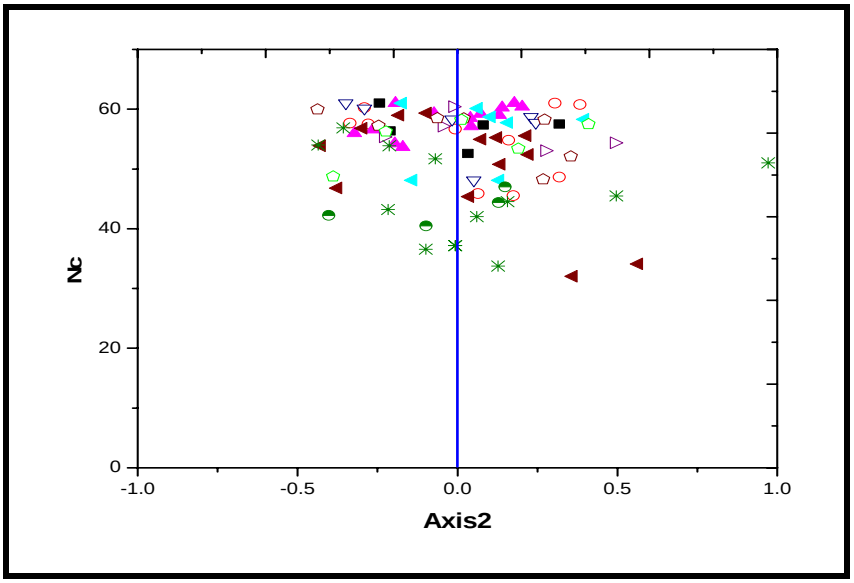

Figure 3: Scatter plot of 94 xylanase coding genes from 12 fungi. A plot of second major axis Vs N_C value. The organisms are represented as follows: ■ **:** *Aspergillus clavatus,* ✿ **:** *Aspergillus fumigatus,* ▲**:** *Aspergillus nidulans,* ▽**:** *Aspergillus niger,* ◀**:** *Aspergillus terreus,* ◀ **:** *Chaetomium globosum,* ▷ **:** *Giberrella zeae,* ✳ **:** *Magnaporthe griesa,* ★ **:** *Neosartorya fischeri,* ⬠**:** *Neurospora crasa,* ◓**:** *Pistia stipitis,* ■**:** *Saccharomyces cereviseae.*

Xylanases are hydrolytic enzymes which randomly cleaves the ß-1, 4 backbone of complex plant cell wall polysaccharide, xylan. Diverse forms of these enzymes exist, displaying varying folds, mechanisms of action, substrate specificities, hydrolytic activities and physicochemical characteristics. Research has mainly focused on only two of the xylanase containing glycoside hydrolase families, namely families 10 and 11, yet enzymes with xylanase activity belonging to family 5, 7, 8 and 43 have also been identified. Driven by industrial demands for enzymes that can work at process conditions, a number of extremophilic xylanases have been isolated. The adaptation strategies of extremophilic xylanases have contributed a lot to there potential industrial application. Our analysis showed that the codon usage pattern of almost all xylanase gene were very similar and among all fungi *Aspergillus nidulans, Chaetomium globosum, Aspergillus terreus* and *Aspergillus clavatus* were maximum biased.

Methodology:

All xylanase gene sequence of *Aspergillus clavatus, Apsergillus nidulans, Aspergillus terreus, Aspergillus niger, Chaetomium globosum, Pistia stipitis, saccharomyces cereviseae, Aspergillus fumigatus, Giberrella zeae, Magnaporthe grieasa, Neosartorya fischeri, Neurospora crasa* were downloaded from NCBI (www.ncbi.nlm.nih.gov) and EMBL (www.ebi.ac.uk). A total of 94 xylanase coding genes from 12 different fungi were studied. All information regarding the xylanase gene taken for study was extracted from gene bank (NCBI and EMBL) records. The relative synonymous codon usage had been determined to study the overall codon usage variation among the xylanase gene. RSCU is dominantly used as one of the best indicators of bias **[16]**. The parameters RSCU, GC_{3S} (frequency of G+C at synonymous third position of codon), A_{3S} (frequency of A at synonymous third position of codon), T_{3S} (frequency of T at synonymous third position of codon), G_{3S} (frequency of G at synonymous third position of codon), C_{3S} (frequency of C at synonymous third position of codon) were determined and correspondence analysis were carried out by program codonW v1.3 (available at http://www.molbiol.ox.ac.uk/cu).

Discussion:

Synonymous codon usage variation in xylanase:

To study the codon usage bias in xylanase the overall RSCU in 94 xylanase coding gene in 12 fungi were determined. The value of N_C ranged from 33.74 to 61 with a mean of 47.37 and the value of GC_{3S} ranges from 0.315 to 0.904 with a mean of 0.6045.

The effect of mutational bias on codon usage variation of xylanase:

To determine the determinants of codon usage variation N_C plots (a plot of N_C Vs GC_{3S}) and correspondence analysis was used. The N_C plot of the genes of xylanase suggests that maximum points lie on the expected curve with almost none towards GC poor region and very few above the line towards GC rich region. Points demonstrating xylanases from *Aspergillus nidulans, Chaetomium globosum, Aspergillus terreus, Aspergillus clavatus, Neosartorya fischeri, Aspergillus niger* and *Neurospora crasa* in majority lied towards or almost on the expected curve whereas points demonstrating xylanase genes from *Pistia stipitis, Magnaporthe griesa* lied below the expected line (**Figure 1**). This suggests that effect of mutational bias on codon usage

variation in *Pistia stipitis* and *Magnaporthe griesa* is weak as compared to other organisms whose gene demonstrating points lie towards or on the curve. Some xylanase demonstrating points from *Chaetomium globosum* lie away from the expected line which suggests that effect of mutational bias on codon usage variation of this organism is not uniform and it varies from one xylanase gene to other of the same organism (**Figure 1**). The correspondence analysis of RSCU values of 94 xylanase coding genes (of the above 12 fungi) states that there is very little effect of mutational bias and other factors as most of points demonstrating xylanase gene lie clustered to the principal line. Correspondence analysis is a multivariate statistical analysis tool to study codon usage variation among genes **[14]**. CA is a sophisticated technique in which the codon usage data (59 codon) are plotted in a multidimensional space of 59 axes (Met, Trp and stop codons are excluded) and that it identifies an axis which represents the most prominent factors contributing the variance among genes. The positions of the genes along the first two major axes are shown in (**Figure 2**). The position of gene of the first two major axes (**Figure 2**) shows that almost all genes from *Aspergillus nidulans, Aspergillus terreus, Chaetomium globosum, Giberrella zeae and Pistia stipitis* are clustered. The number and occurrence of each codon and its RSCU values for two groups of genes with maximum codon biased are shown as **Table 1 (see supplementary material)**. The two groups of genes taken are xylanase coding genes from *Aspergillus nidulans* and *Aspergillus terreus.*

Effect of translational selection on codon usage variation in xylanase:

Along with mutational bias, translational selection also influences codon usage variation in xylanase coding gene from fungi. A scatter plot is drawn between the position of genes along the second major axis and N_C values (**Figure 3**). It was seen that xylanase gene from almost all fungi has a higher N_C value except for genes *Magnaporthe griesa* and *Pistia stipitis* which had comparatively lower N_C values. Genes from *Chaetomium globosum* showed mixed response with two genes which were not biased having the least N_C values in the entire model system.

Conclusion:

A synonymous codon bias of xylanase gene was studied in the present work. Codon usage of 94 xylanase gene from 12 different fungi was analyzed and a comparative analysis of codon usage and RSCU values of xylanase coding gene was done. Study of effect of mutational bias on codon usage variation of the gene suggested weak biasing in case of *Pistia stipitis* and *Magnaporthe griesa* as compared to gene from other organisms taken for study. Moreover effect of mutation bias in xylanase genes from *Chaetomium globosum* mixed characteristics with few genes showing weak biasing and remaining with strong biasing effect. The position of gene on the first two major axes (**Figure 2**) shows that almost all genes from *Aspergillus nidulans, Aspergillus terreus, Chaetomium globosum, Giberrella zeae* and *Pistia stipitis* are clustered. RSCU values of two groups of genes located at extreme ends of first major axis determined by CA for xylanase gene from two fungi *Aspergillus nidulans* and *Aspergillus terreus* suggested that none of the codon occurred significantly higher in the highly expressed genes. The occurrence of codon was not influenced by expression level of the gene. A comparative analysis of codon usage and RSCU values of xylanase encoding gene from the 12 fungi was done. From the analysis (**Figure 3**) it is observed that xylanase gene from *Aspergillus nidulans* and *Aspergillus niger* are most expressed genes compared with other fungal xylanase gene. Based on this data it can also be suggested that translational selection also influences codon usage variation in case of fungal xylanase gene expression.

Acknowledgements:

The authors are thankful to the Sub-Distributed Information Center (BTISnet SubDIC), Department of Biotechnology (No. BT/BI/04/065/04), New Delhi, India. Our thanks are also to Jharkhand government for the lab infrastructure.

References:

[1] R. Grantham et al., *Nucleic Acids Res.*, 9: r43-r74 (1981) [PMID: 7208352]
[2] P. M. Sharp and E. Cowe., *Yeast*., 7: 657-678 (1991) [PMID: 1776357].
[3] J. R. Lobry and C. Gautier., *Nucleic Acids Res*., 22: 3174–3180 (1994) [PMID: 8065933]
[4] J. W. Fickett., *Nucleic Acids Res.*, 10: 5303–5318 (1982) [PMID: 7145702]
[5] M. J. Greenacre., *Academic Press, London.,* (1984)
[6] J. Thioulouse et al., *Stat. Comput.,* 7: 75–83 (1997)
[7] T. Ikemura., *Mol. Biol. Evol.* 2: 13-34 (1985) [PMID: 3916708]
[8] D. B. Levin and B. Whittome., *J. Gen. Virol.*, 81: 2313-2325 (2000) [PMID: 10950991]
[9] J. O. McInerney., *Proc. Natl. Acad. Sci. USA.,* 95: 10698-10703 (1998) [PMID: 9724767]
[10] B. R. Morton., *Proc. Natl. Acad. Sci. USA* 96: 5123-5128 (1999) [PMID: 10220429]
[11] H. Musto et al., *J. Mol. Evol.*, 49: 27-35 (1999) [PMID: 10368431]
[12] G. A. Singer and D. A. Hickey., *Gene* 317: 39-47 (2003) [PMID: 14604790]
[13] P. M. Sharp and W. H. Li., *Nucleic Acids Res*., 15, 1281-1295 (1987) [PMID: 3547335]
[14] F. Wright., *Gene*., 87: 23-29 (1990) [PMID: 2110097]
[15] P. M. Sharp et al., *Nucleic Acids Res*., 16: 8207-11 (1988) [PMID: 3138659]
[16] A. Ranjan et al., *In Silico Biology*., 7: 0030 (2007) [PMID:18391235]

Edited by P. Kangueane

Citation: Shrivastava *et al*, Bioinformation 3(10): 425-429 (2009)

Bioinformation by Biomedical Informatics Publishing Group **open access**
www.bioinformation.net **Hypothesis**

Supplementary material

Table 1: RSCU values of two groups of genes located at extreme ends of first major axis determined by CA.
Group A: (Genes from *Aspergillus nidulans*)

AA[a]	Codon	RSCU[b]	N[b]	AA	Codon	RSCU[b]	N[b]
Phe	UUU	1.14	8	Ser	UCU	0.77	5
	UUC	0.086	6		UCC	1.69	11
Leu	UUA	0.67	3		UCA	1.23	8
	UUG	2.67	12		UCG	0.31	2
	CUC	0.44	2		CCC	0.71	3
	CUA	0.22	1		CCA	0.71	3
	CUG	1.33	6		CCG	1.65	7
Ile	AUU	0.83	5	Thr	ACU	0.93	7
	AUC	1.33	8		ACC	0.53	4
	AUA	0.83	5		ACA	0.40	3
Met	AUG	1.00	11		ACG	2.13	16
Val	GUU	0.57	2	Ala	GCU	0.77	5
	GUC	1.14	4		GCC	1.08	7
	GUA	1.14	4		GCA	0.77	5
	GUG	1.14	4		GCG	1.38	9
Tyr	UAU	0.57	2	Cys	UGU	0.67	2
	UAC	1.43	5		UGC	1.33	4
TER	UAA	0.46	2	TER	UGA	2.08	9
	UAG	0.46	2	Trp	UGG	1.00	6
His	CAU	1.00	2	Arg	CGU	0.51	4
	CAC	1.00	2		CGC	1.28	10
Gln	CAA	0.86	6		CGA	0.77	6
	CAG	1.14	8		CGG	0.26	2

Group B: (Genes from *Aspergillus terreus*)

AA[a]	Codon	RSCU[b]	N[b]	AA	Codon	RSCU[b]	N[b]
Ile	AUU	1.07	5	Thr	ACU	0.55	3
	AUC	1.71	8		ACC	1.27	7
	AUA	0.21	1		ACA	1.64	9
Met	AUG	1.00	5		ACG	0.55	3
Val	GUU	1.00	5	Ala	GCU	1.56	7
	GUC	1.00	5		GCC	1.56	7
	GUA	1.40	7		GCA	0.44	2
	GUG	0.60	3		GCG	0.44	2
Tyr	UAU	1.43	5	Cys	UGU	0.93	7
	UAC	0.57	2		UGC	1.07	8
TER	UAA	0.60	1	TER	UGA	1.20	2
	UAG	1.20	2	Trp	UGG	1.00	8
His	CAU	1.13	9	Arg	CGU	0.49	3
	CAC	0.88	7		CGC	0.97	6
Gln	CAA	1.36	17		CGA	1.14	7
	CAG	0.64	8		CGG	1.46	9
Asn	AAU	1.09	6	Ser	AGU	1.55	8
	AAC	0.91	5		AGC	0.97	5
Lys	AAA	1.33	2	Arg	AGA	1.46	9
	AAG	0.67	1		AGG	0.49	3
Asp	GAU	1.67	10	Gly	GGU	1.08	10
	GAC	0.33	2		GGC	1.51	14
Glu	GAA	1.33	8		GGA	0.86	8
	GAG	0.67	4		GGG	0.54	5