International Scholars Journals

Advanced Journal of Microbiology Research ISSN 2241-9837 Vol. 14 (3), pp. 001-009, March, 2020. Available online at www.internationalscholarsjournals.org © International Scholars Journals

Full Length Research Paper

Theoretical analyses of gene expression for five human pathogens

Lin Ning and Feng-Biao Guo*

School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China.

Accepted 27 September, 2019

Predicted highly expressed (PHX) genes were analyzed in five human pathogens with significant translation selection measured by within-group correspondence analysis (WCA). Functional analysis showed that in the five genomes PHX genes involved in protein synthesis, constitute the largest functional category, followed by categories of energy metabolism and protein fate. The gene encoding pyruvate kinase involved in glycolysis was PHX gene in most genomes except *Bacillus anthracis* Ames, which preferred strictly aerobic environment. Special PHX genes were also found, such as genes encoding phosphotransferase (PTS) system mainly appeared in *Streptococcus pneumoniae* genome. The analysis of virulence factors indicates that only a few pathogenicity-related genes were predicted as PHX. This is contrary to previous observations of phytopathogens, where most virulence related genes were PHX. The PHX genes may provide potential drug targets for the design of new bactericide. Specially, PHX virulence factors may help for the understanding of the crucial mechanism of virulence of the bacterial pathogens when invading human body.

Key words: Highly expressed genes, human pathogen, *E*(*g*) measure, CAI; virulence related genes, inorganic pyrophosphatase.

INTRODUCTION

It has been revealed that in unicellular organisms, highly expressed genes always show a more biased codon usage than moderate or low expressed genes and a clear positive correlation between the extent of synonymous codon usage bias and cellular mRNA expression level has been observed (Ikemura, 1981). According to the observed significant positive correlation, Sharp and Li (1987) proposed one index, namely CAI, to quantify the codon usage bias towards optimal codons, which are typically determined from the reference set of ribosomal protein genes (Sharp and Li, 1987). And those genes with the top CAI values among a species are treated as highly expressed ones. Karlin and Mrázek (2000) also proposed one measure, namely *E*(*g*), which used three classes of genes as reference sets and was formulated differently from CAI index, to predict highly expressed genes. Currently, CAI and *E*(*g*) measure have become two widely accepted methods for theoretically recognizing highly expressed genes from codon usage in prokaryotic genomes (Jansen et al., 2003; Roymondal et al., 2009).

Anthrax, bacillary dysentery, cholera, meningitis, plague, pneumonia, syphilis and tuberculosis are eight prevalent and infectious diseases that human often suffers from. *Bacillus anthracis*, *Shigella flexneri*, *Vibrio cholerae*, *Neisseria meningitidis*, *Yersinia pestis*, *Chlamydophila pneumoniae, Mycoplasma pneumoniae, Streptococcus pneumoniae, Treponema pallidum* and *Mycobacterium tuberculosis* are ten of the pathogenic bacteria for the eight human epidemics. There are considerable experimental data of mRNA expression for some of the above human pathogens. However, significant conflicts exist in different micro-array experiments because of the different technologies and different culturing conditions. Anyway, computational analysis and predictions of gene expression in the ten human pathogens are of importance. Most importantly, comparison and analysis of PHX genes may contribute to

*Corresponding author. E-mail: fbguo@uestc.edu.cn. Fax: +86-28-83208238.

the understanding the commonness and differences in the pathogenicity of the ten pathogenic bacteria given that they belong to different phylogenetic lineages.

In this work, both of CAI and *E*(*g*) measure were employed for predicting of highly expressed genes. PHX genes of the 10 human pathogens were firstly predicted by both CAI and *E*(*g*) methods, and then consistently predicted or shared PHX genes by the 2 approaches were picked out. According to results of within-group correspondence analysis (WCA), the 10 human pathogens were divided into two groups. Five genomes in group 1, which had significant translation selections, were chosen for further investigation. Functional distribution of the shared PHX genes in the chosen genomes was analyzed. Special PHX genes and pathogenicity-related factors in each species were also investigated.

MATERIALS AND METHODS

Sequences

The complete genome sequences and the annotation information of the ten human pathogenic bacteria were downloaded from NCBI RefSeq ftp site (ftp.ncbi.nih.gov/genomes/Bacteria/). These pathogens are *B. anthracis* Ames (NC_003997), *C. pneumoniae* AR39 (NC_002179), *M. tuberculosis* CDC1551 (NC_002755), *M. pneumoniae* M129 (NC_000912), *N. meningitidis* Z2491 (NC_003116), *S. flexneri* strain 301 (NC_004337), *S. pneumoniae* TIGR4 (NC_003028), *T. pallidum* strain Nichols (NC_000919), *V. cholerae* strain N16961 chromosome I (NC_002505) and *Y. pestis* CO92 (NC_003143). Function categories of these pathogens were derived from Comprehensive Microbial Resource (CMR) of the JCVI (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi). Virulence factor related data were extracted from VFDB (http://www.mgc.ac.cn/VFs/) (Yang et al., 2008). All of the bacterial strains involved in this work are virulent and clinical.

Analysis

In this work, two widely accepted methods are used together for theoretically recognizing highly expressed genes:

E(*g*) measure and CAI

***E*(*g*) measure:** According to Karlin and Mrázek (2000) *E*(*g*) measure, a gene is predicted as PHX if it is sufficiently similar in codon usage to ribosomal protein genes (RPs), the chaperone/degradation proteins (TFs) and the translation and transcription processing factors (CHs), but deviates strongly from the collection of all genes (C). The calculation of the codon bias relative to RPs, TFs, CHs and C is based on *E*(*g*) measure, measured by the ratios of *B* (*g*|*RP*), *B* (*g*|*TF*), *B* (*g*|*CH*) and *B* (*g*|*C*). Generally speaking, a gene will be deemed as PHX if *B* (*g*|*RP*), *B* (*g*|*TF*) and *B* (*g*|*CH*) are low while *B* (*g*|*C*) is high (Karlin and Mrázek, 2000; Karlin et al., 2001).

CAI: Sharp and Li's CAI method is also used to determine PHX genes (Sharp and Li, 1987). CAI is calculated for each gene by using the 'CodonW' software, which was downloaded from http://codonw.sourceforge.net/. When calculating CAI, dozens of ribosomal protein genes are chosen as reference set of highly expressed genes for each genome. The CAI value varies from 0 to 1.0 and a higher value indicates that the codon usage bias of the gene is more similar to that of optimal codons measured by the reference genes, and genes with top CAI values are recognized as PHX (Sharp and Li, 1987; Wu et al., 2005).

In this work, both of *E*(*g*) and CAI are employed. The details for generating and choosing PHX gene are described as follows. Firstly, the gene expression levels are predicted by either of the two methods. The PHX genes predicted by *E*(*g*) measure are picked out, and the number of these PHX genes is counted, and then based on this number the same amount of genes with top CAI values are chosen as PHX genes predicted by CAI. At the end, only PHX genes consistently predicted by both *E*(*g*) measure and CAI are chosen for further analysis.

RESULTS

Statistics of PHX genes in ten human pathogens

Table1 presents the statistics of shared or consistently predicted PHX genes in the ten human pathogens, and Table 2 lists the 10 genes with the highest expression levels based on *E*(*g*) measure, among the shared PHX genes in the five chosen genomes. According to Henry and Sharp (2007), codon usage based methods could be used to predict highly expressed genes only in genomes where evidences of translation selection are significant. Here, WCA is used to measure the strengths of translation selections in the ten genomes (Suzuki et al., 2008). Plots of the two most important axes after WCA for the total genes in the ten pathogens are shown in Figure 1. Each sub-figure corresponds to one pathogen. In each sub-figure, red circles denote RP genes whereas black circles denote the other genes. Such kind of figure is used to detect whether a genome is under translation selection and roughly compare the relative strength of the selection. According to the WCA analysis, the ten human pathogens are divided into two groups. And only the five pathogenic genomes in group 1 are chosen for the following analysis because of their significant translation selections from the direct view in Figure 1. They are *B. anthracis*, *S. flexneri*, *S. pneumoniae*, *V. cholerae* and *Y. pestis*, respectively.

Function distributions of shared PHX genes in the five chosen genomes

In order to obtain a global view of the functions of PHX genes in the five chosen human pathogens, all the function-known PHX genes are categorized based on the functional groups designated by CMR at JCVI. In Table 3, there are a total of 16 functional categories according to CMR annotations. It is also shown that the PHX genes in protein synthesis constitute the largest functional class, followed by energy metabolism and protein fate. Other functional classes, such as transport and binding proteins, amino acid biosynthesis, central intermediary metabolism, DNA metabolism, purine, pyrimidine, cell envelope and so on, also include several PHX genes in each genome. In addition, there have special functional

Table 1. Statistics of PHX genes in ten human pathogen.

Genome	Group 1					Group 2				
	B. anthracis	*S. flexneri*	*S. pneumoniae*	*V. cholerae*	*Y. pestis*	*C. pneumoniae*	*M. pneumoniae*	*M. tuberculosis*	*N. meningitidis*	*T. pallidum*
Chromosome size (kb)	5227	4607	2160	4033	4830	1230	816	4400	1780	1138
GC content (%)	35.37	50.88	39.70	47.48	47.64	40.60	40.00	65.60	51.80	52.80
Number of shared PHX genes	230	192	154	117	144	48	22	270	61	35
Percentage of shared PHX genes (%)	4.33	4.59	7.32	4.28	3.71	4.32	3.19	6.45	3.20	3.38
Average $E(G)$ value of all genes	0.96	0.8	0.87	0.87	0.86	0.97	0.96	0.97	0.86	0.96
Average E(g) value of shared PHX genes	1.34	1.44	1.44	1.43	1.33	1.11	1.12	1.13	1.21	1.12
Average CAI values of all genes	1.34	1.44	1.44	1.43	1.33	0.60	0.67	0.59	0.78	0.67
Average CAI values of shared PHX genes	0.68	0.64	0.65	0.62	0.66	0.68	0.74	0.71	0.85	0.74

Table 2. Top 10 genes in the five chosen human pathogens.

Gene	Genomes *E(g)*/CAI				
	B. anthracis	*S. pneumoniae*	*S. flexneri*	*V. cholerae*	*Y. pestis*
Protein synthesis					
ribosomal protein L1	1.68/0.832	1.71/0.849	1.96/0.776	2.04/0.746	1.80/0.739
ribosomal protein L2	1.95/0.762	1.90/0.776	2.31/0.756	2.09/0.729	1.92/0.75
ribosomal protein L3	1.48/0.701	1.56/0.71	1.82/0.747	1.84/0.705	1.53/0.707
ribosomal protein S1	1.02/0.644	1.80/0.76	1.92/0.802	2.00/0.789	1.77/0.788
ribosomal protein S2	1.96/0.794	1.81/0.772	2.24/0.792	1.92/0.765	1.84/0.795
ribosomal protein S3	1.78/0.752	1.70/0.807	2.09/0.774	1.90/0.653	1.95/0.727
ribosomal protein S4	1.89/0.774	1.83/0.804	1.58/0.617	1.69/0.627	1.75/0.683
ribosomal protein S7	2.07/0.764	1.73/0.767	1.67/0.688	1.82/0.745	1.60/0.72
ribosomal protein S9	1.60/0.661	1.56/0.775	1.83/0.814	1.91/0.762	1.83/0.735
ribosomal protein S13	1.99/0.765	1.68/0.75	1.13/0.53	1.29/0.56	1.51/0.684
translation elongation factor G	2.15/0.763	1.88/0.76	2.05/0.774	1.98/0.684	1.94/0.743
elongation factor Tu	1.76/0.818	1.71/0.807	2.03/0.807 2.04/0.804	1.67/0.69 1.65/0.68	1.90/0.728
glycyl-tRNA synthetase	2.00/0.794	(0.99/0.529)	-	1.12/0.493	1.11/0.588
protein chain initiation factor IF-2	1.42/0.681	1.18/0.519	2.12/0.647	1.42/0.559	1.83/0.663
Energy metabolism					
formate acetyltransferase	1.14/0.7	2.02/0.761	(0.92/0.794)	2.09/0.629	-
pyruvate kinase	(0.92/0.428)	1.88/0.761	2.29/0.741	1.07/0.653	1.58/0.654

Table 2 Contd

phosphotransacetylase	-	(0.93/0.538)	2.22/0.713	-	-
pyruvate dehydrogenase (decarboxylase component)	-	-	2.19/0.706	-	-
Protein fate					
chaperone protein dnaK	1.90/0.746	1.78/0.699	2.25/0.757	1.79/0.661	1.80/0.751
trigger factor	1.90/0.817	1.88/0.766	1.87/0.764	(0.94/0.708)	1.66/0.732
Cellular processes					
pyruvate oxidase	-	1.88/0.749	(0.94/0.415)	-	-
Transcription					
RNA polymerase, beta subunit	1.65/0.611	1.73/0.596	2.56/0.68	1.89/0.574	1.93/0.663
RNA polymerase, beta` subunit	1.85/0.632	1.72/0.563	2.38/0.729	1.64/0.585	2.04/0.714
Regulatory functions					
catabolite control protein A	1.86/0.608	(0.90/0.428)	-	-	-
Cell envelope					
phosphoglucomutase	1.01/0.624	1.87/0.670	(1.01/0.594)	(0.88/0.392)	0.58/0.499
Transport and binding proteins					
ABC transporter, substrate-binding protein.	(0.89/0.58)	1.96/0.714	(1.00/0.267)	(0.66/0.358)	(0.58/0.478)
Amino acid biosynthesis					
5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase	(0.75/0.576)	1.93/0.655	-	(1.06.0.399)	(0.89/0.581)
Signal transduction					
PTS system, IIABC components	(0.94/0.648)	1.91/0.691	1.48/0.581	(0.49/0.364)	(0.93/0.539)
Unknown function					
putative GTP-binding factor	-	-	2.26/0.672	-	-
putative GTPase	-	-	-	-	2.00/0.695

[a]Top 10 PHX genes are those underlined. Numbers in parentheses indicate the gene is not PHX.

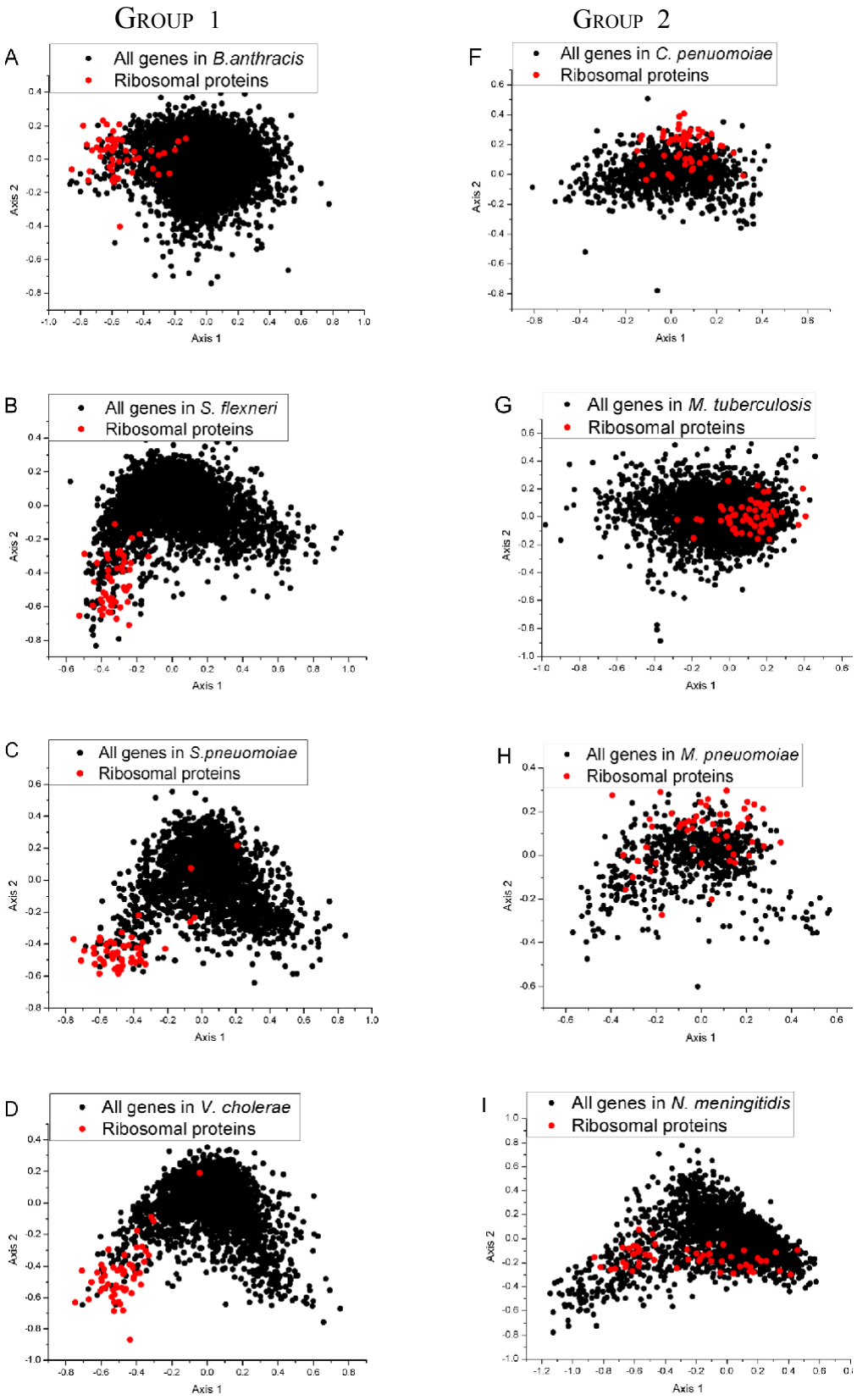

Figure 1. Plot of the two most important axes of WCA for all genes of the ten human pathogens. Each sub-figure corresponds to one pathogen. In each sub-figure, RP genes are denoted by red filled cycles and the other genes are denoted by black filled circles. Only the five genomes with significant translation selections in Group1 are chosen for the further investigation.

Table 3. Function distribution of consistently predicted PHX genes in five pathogenic genomes with significant translation selection.

Function category	Percentage of PHX genes				
	***B. anthracis* (%)**	***S. flexneri* (%)**	***S. pneumonia* (%)**	***V. cholerae* (%)**	***Y. pestis* (%)**
Protein synthesis	26.19	34.29	38.62	53.64	41.41
Energy metabolism	18.10	15.43	11.72	11.82	9.38
Protein fate	8.10	9.14	7.59	5.45	8.59
Transport and binding proteins	6.67	7.43	7.59	5.45	3.91
Cellular processes	8.10	4.57	5.52	0.91	7.81
Purines, pyrimidines, nucleosides, and nucleotides	6.67	6.29	6.21	2.73	2.34
Transcription	8.57	4.57	4.83	6.36	5.47
Cell envelope	4.76	4.00	3.45	4.55	7.81
Central intermediary metabolism	3.33	2.29	2.07	1.82	4.69
DNA metabolism	1.43	3.43	2.07	2.73	2.34
Amino acid biosynthesis	2.38	0.57	3.45	0.91	0.78
Signal transduction	1.43	0.57	6.90	0.00	0.78
Regulatory functions	0.95	2.29	0.00	2.73	2.34
Fatty acid and phospholipid metabolism	0.48	2.86	0.00	0.91	0.78
Biosynthesis of cofactors, prosthetic groups, and carriers	2.38	0.57	0.00	0.00	0.78
Mobile and extra chromosomal element functions	0.48	1.71	0.00	0.00	0.78

classes exclusively contained by some genomes, for example, signal transduction in *S. pneumoniae*, mobile and extra-chromosomal element in *S. flexneri*.

Protein synthesis and fate

Protein is the basic and essential component for living cells and genes coding for protein metabolism play an important role in bacteria (Wu et al., 2006). Table 2 shows that almost all of genes coding for ribosomal proteins, genes encoding translation factors including elongation factor G, Tu, Ts, P, etc., and peptide chain release factors, attain PHX levels in each genome, as well as some other factors involved in protein synthesis. Genes coding for protein secretion, trafficking, modification, folding and degradation, which are involved in protein fate after protein synthesis, also attain PHX levels. The major chaperone proteins such as DnaK and GroEL are PHX genes in the five human pathogens.

Energy metabolism

It is inevitable that genes coding for carbohydrate metabolism attain highly expressed levels because carbohydrates play a necessary role in microbes, serving as structural elements, energy sources, etc. (Fu et al., 2005). The genes coding for energy metabolism can be divided into four groups: glycolysis, pyruvate metabolism, pentose phosphate pathway and TCA cycle. Owing to the different habitat, lifestyle and nutrient sources, the PHX genes in energy metabolism are different in the five pathogenic genomes. The pyruvate kinase, which catalyzes the transfer of a phosphate from phosphoenolpyruyate (PEP) to ADP and hence yields one molecule of pyruvate as well as one molecule of ATP, is the key enzyme involved in glycolysis (Liapounova et al., 2006). It can be seen from Table 2 that the gene coding for pyruvate kinase is predicted as PHX in the four genomes including *S. pneumoniae, S. flexneri, V. cholerae* and *Y. pestis*, which are all facultative anaerobes and glycolysis is an important energy

metabolisms for these species. While in *B. anthracis*, which prefers strictly aerobic environment, 4 TCA genes attain highly expressed levels. The gene encoding formate acetyltransferase attains the highest expression level both in *S. pneumoniae* and *V. cholerae*, the two facultative anaerobes. The enzyme formate acetyltransferase catalyzes the chemical reaction of S-adenosyl-L-methionine and dihydroflavodoxin into 5'-deoxyadenosine, L-methionine and flavodoxin semiquinone. It is accompanied with the synthesis of ATP, which is the direct power source of living cells (Frey et al., 1994).

Overall, those PHX genes among the top functional classes are involved in primary metabolisms, which are necessary for bacterial growth, that is., house-keeping genes.

Common and special PHX genes

There are a total of 44 PHX genes commonly contained by all the five pathogenic genomes. Among them, 36 are found to encode ribosomal proteins and 7 codes for major translation and transcription factors. Major chaperone protein encoding gene dnaK constitutes the last one that attains high expression level in all the five species.

Another case should be noted is the pyrophosphatase (PPase) enzymes and its alternative PpaC enzyme. Inorganic PPase, which catalyzes the hydrolysis of inorganic pyrophosphate (PPi) into phosphate (Pi), provides a thermodynamic driving force for important biosynthetic reactions in almost all bacteria (Ko et al., 2007). There is one gene encoding inorganic PPase for each of the genomes of *S*. *flexneri V*. *cholerae* and *Y*. *pestis* and they are all consistently predicted as PHX. Whereas there are manganese-dependent inorganic PpaC, which perform similar functions with PPase enzymes and are thought to be a new class of PPase enzymes, rather than usual Ppase enzymes in genomes of *B*. *anthracis* and *S*. *pneumoniae*. The two genes encoding PpaC enzymes are consistently predicted as PHX, too. Therefore, there is one PHX gene, which encodes enzyme catalyzing the hydrolysis of inorganic pyrophosphate into phosphate, for each of the five pathogenic genomes. The fact indicates the consistent importance of the hydrolysis of inorganic pyrophosphate into phosphate in cellular cycle of the five pathogens.

The phosphotransferase system (PTS), which belongs to the functional category of signal transduction, exists in almost all kinds of bacteria. It is responsible for the transfer of the phosphoryl from phosphoenolpyruvate to the sugar substrates, and is concomitant with the translocation of over 20 kinds of carbohydrates through the bacterial membrane (Karlin et al., 2004). In this study, no PHX genes are involved in signal transduction among the five human pathogens except genes coding for PTS system. Among 47 PTS genes in *S. pneumoniae*, 10 attain highly expressed levels. The PTS system plays an important role in carbohydrates transportation pathways for *S. pneumoniae*. However, there are only a total of 9 PHX genes encoding PTS system in the other four pathogenic genomes. It can also be found specially that genes coding for fructose and lactose components have PHX levels, suggesting that both fructose and lactose are important energy sources for *S. pneumoniae*.

PHX genes encoding virulence factors

According to related data deposited in VFDB for the five pathogens, there are 387 genes encoding virulence factors for the five human pathogens. Statistics of these virulence related genes are listed in Table 4 and the names of all of them are listed in Supplementary Table, which is freely available at the website of the authors (http://cobi.uestc.edu.cn/resource/ning/). There are only 28 and 17 virulence related genes which are independently predicted as PHX by *E*(*g*) or CAI method. Among them, only 7 are consistently predicted as PHX by both methods. The other virulence related genes are predicted as moderately or poorly expressed. This is contrary with previous observation in phytopathogens, where most of virulence related genes are predicted as PHX (Fu et al., 2005). Furthermore, no common PHX virulence related genes are found among the five human pathogens, indicating that their virulence mechanism when invading human body may be different.

Among the five human pathogens, in *B. anthracis* the gene coding for one of the suppressor of groEL mutation proteins, sugE-1 (BA_4435), and the gene encoding an outer surface protein, adhesion lipoprotein (BA_2035) attain highly expressed levels. In *S. pneumoniae*, gene SP0117 that encodes surface protein PspA, as well as the lipoprotein psaA (SP1650) and the chaperone protein trigger factor are predicted as PHX. In *V. cholerae* one hcp protein, which is in the type VI secretion system (T6SS), VC_1415, is predicted as PHX. The flagella related gene flic is predicted as PHX in *Y*. *pestis*, with *E*(*g*) value and CAI values of 1.68 and 0.683, respectively. In *S. flexneri*, no virulence related genes are found to attain highly expressed level based on both *E*(*g*) and CAI methods. Only the gene encoding ATP-binding component of ferric enterobactin transport, fepC, is predicted as PHX level by *E*(*g*), and no virulent PHX genes are found based on CAI.

DISCUSSION

The ‘codon adaptation index’ (CAI) and *E*(*g*) measure are both based on the codon usage bias and the two methods have been widely used to evaluate and analyze the gene expression levels in various prokaryotes. In this

Table 4. Statistics of genes encoding virulence factors in five pathogenic genomes with significant translation selection.

Genome	Number of all virulence genes	Number of PHX virulence genes predicted by E(g)	Number of PHX virulence genes predicted by CAI	No. of PHX virulence genes consistently predicted by two methods	Average *E*(*g*) of virulence genes	Average CAI of virulence genes
B. anthracis	89	2	7	2	0.79	0.46
S. flexneri	31	1	0	0	0.76	0.34
S. pneumoniae	45	4	6	3	0.74	0.41
V. cholerae	137	13	2	1	0.77	0.31
Y. pestis	85	8	2	1	0.82	0.44

work, both of the two methods are employed, and only those shared PHX genes consistently predicted by the two methods are investigated. It is obvious that choosing only shared PHXs will make the prediction results more reliable and credible. Here, *E*(*g*) and CAI methods are only used to identify the PHX genes, and comparison between them is not involved in this study.

With the sequencing of hundreds of bacterial genomes, it has been found that synonymous codon bias exerted by translation selection varies distinctly among species (Sharp et al., 2005). And even in some microbes, selection seems to have been ineffective. Henry and Sharp (2007), once proposed a general way to predict highly expressed genes in a genome: the codon usage based method can be used only in the genome with significant evidence of translation selection (Sharp et al., 2005). In order to obtain a direct view of the strength of translation selection, WCA is used in this work. Compared with the other correspondence analysis such as CA-AF, CA-RF or CA-RSCU, WCA uses a different calculation, that is., it adjusts the value for each codon by the average value of all the codons that encode the same amino acid. WCA is an alternative method, which can directly remove the influences of different amino acid compositions from the effects of synonymous codon usage (Suzuki et al., 2008). In this work, based on WCA analysis, five human pathogens in group1 including *B. anthracis*, *S. flexneri*, *S. pneumoniae*, *V. cholerae* and *Y. pestis* are chosen for the further investigation due to their more significant translation selections compared with the other five genomes.

Among the five human pathogens, except *S. flexneri* with GC content exceeding 50%, the others are all GC-poor genomes. Besides the different GC content, the diversities in habitat, lifestyle, nutrient sources, and invasion mode also exist among the five human pathogens. Consequently, the differences of the gene expression levels are apparent among the five human pathogens. Function analysis shows that although genes among the top functional classes are all involved in primary metabolism ('house-keeping' function) for each genome, the least function classes are much species-specific. The gene coding for pyruvate kinase is predicted as PHX in four facultative anaerobes (*S. pneumoniae, S. flexneri, V. cholerae* and *Y. pestis*), but not attains high expressed level in *B. anthracis*, which prefers district aerobic lifestyle. Much more PTS genes are found to attain highly expressed levels in *S. pneumoniae* than in the other four pathogens. The virulence factors predicted as PHX are also different among the five pathogens, indicating the different virulence mechanism when invading human. In a word, comparative results of the five human pathogens would reflect to some extent their differences in habitat, lifestyle, nutrient sources, invasion mode, etc. We hope this work could benefit the further research for the five important human pathogens.

ACKNOWLEDGMENTS

Authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. The present study was supported by National Natural Science Foundation of China (grant 60801058 and 31071109), and the Fundamental Research Funds for the Central Universities of China (grant ZYGX2009J082).

REFERENCES

Frey M, Rothe M, Wagner AF, Knappe J (1994). Adenosylmethionine-dependent synthesis of the glycyl radical in pyruvate formate-lyase by abstraction of the glycine C-2 pro-S hydrogen atom. Studies of [2H]glycine-substituted enzyme and peptides homologous to the glycine 734 site. J. Biol. Chem., 269(17): 12432–12437.

Fu QS, Li F, Chen LL (2005). Gene expression analysis of six GC-rich Gram-negative phytopathogens. Biochem. Biophys. Res. Commun., 332(2): 380–387.

Henry I, Sharp PM (2007). Predicting gene expression level from codon usage bias. Mol. Biol. Evol., 24(1): 10-12.

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(3): 389-409.

Jansen R, Bussemaker HJ, Gerstein M (2003). Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res., 31(8): 2242-51.

Karlin S, Mrázek J (2000). Predicted highly expressed genes of diverse prokaryotic genomes. J. Bacteriol., 182(18): 5238-5250.

Karlin S, Mrázek J, Campbell A, Kaiser D (2001). Characterizations of highly expressed genes of four fast-growing bacteria. J. Bacteriol., 183(17): 5025-5040.

Karlin S, Theriot J, Mrázek J (2004). Comparative analysis of gene expression among low G+C gram-positive genomes. Proc. Natl. Acad. Sci. U.S.A., 101(16): 6182-6187.

Ko KM, Lee W, Yu JR, Ahnn J (2007). PYP-1, inorganic pyrophosphatase, is required for larval development and intestinal function in *C. elegans*. FEBS Lett., 581(28): 5445-5453.

Liapounova NA, Hampl V, Gordon PM, Sensen CW, Gedamu L, Dacks JB (2006). Reconstructing the mosaic glycolytic pathway of the anaerobic eukaryote *Monocercomonoides*. Eukaryot. Cell, 5(12): 2138–2146.

Roymondal U, Das S, Sahoo S (2009). Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome. DNA Res., 16(1): 13-30.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res., 33(4): 1141-1153.

Sharp PM, Li WH (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res., 15(3): 1281–1295.

Suzuki H, Brown CJ, Forney LJ, Top EM (2008). Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res., 15(6): 357–365.

Wu G, Nie L, Zhang W (2006). Predicted highly expressed genes in *Nocardia farcinica* and the implication for its primary metabolism and nocardial virulence, Antonie. Van. Leeuwenhoek., 89(1): 135-146.

Wu G, Culley DE, Zhang W (2005). Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. Microbiology, 151(7): 2175-2187.

Yang J, Chen L, Sun L, Yu J, Jin Q (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res., 36: D539-42.