Genomics 98 (2011) 213–222

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Analysis of the genome and proteome composition of *Bdellovibrio bacteriovorus*: Indication for recent prey-derived horizontal gene transfer

Archana Pan [a,b,*], Ipsita Chanda [c], Jayprokas Chakrabarti [b]

[a] Centre for Bioinformatics, School of Life Sciences, Pondicherry University, Pondicherry – 605014, India
[b] Computational Biology Group, Theoretical Physics Department, Indian Association for the Cultivation of Science, Kolkata - 700032, India
[c] Department of Zoology, S.A. Jaipuria College, Kolkata - 700005, India

ARTICLE INFO

Article history:
Received 14 October 2010
Accepted 14 June 2011
Available online 22 June 2011

Keywords:
Translational selection
Local GC-bias
Horizontal gene transfer
Alcoholicity
Hydropathy
Aromaticity

ABSTRACT

The genome/proteome composition of *Bdellovibrio bacteriovorus*, the predatory microorganism that preys on other Gram-negative bacteria, has been analyzed. The study elucidates that translational selection plays a major role in genome compositional variation with higher intensity compared to other deltaproteobacteria. Other sources of variations having relatively minor contributions are local GC-bias, horizontal gene transfer and strand-specific mutational bias. The study identifies a group of AT-rich genes with distinct codon composition that is presumably acquired by *Bdellovibrio* recently from Gram-negative prey-bacteria other than deltaproteobacteria. The proteome composition of this species is influenced by various physico-chemical factors, viz, alcoholicity, residue-charge, aromaticity and hydropathy. Cell-wall-surface-anchor-family (CSAPs) and transporter proteins with distinct amino acid composition and specific secondary-structure also contribute notably to proteome compositional variation. CSAPs, which are low molecular-weight, outer-membrane proteins with highly disordered secondary-structure, have preference toward polar-uncharged residues and cysteine that presumably help in prey-predator interaction by providing particular bonds of attachment.

1. Introduction

Bdellovibrio bacteriovorus HD100 is a highly motile Gram-negative deltaproteobacterium marked with a number of conspicuous characteristics [1]. It is a predatory microorganism that preys on other Gram-negative bacteria including pathogenic ones [1]. It is ubiquitous in nature, having been discovered in a wide variety of environments that include fresh water, sewage, soil and even mammalian intestines [1]. This organism is obligatorily dependent upon a prey cell for its growth [1]. Though it is a tiny bacterium (about 0.2–0.5 μm wide and 0.5–2.5 μm long), its genome is not reduced like other obligatory intracellular bacteria [2,3]. It rather carries a quite large genome consisting of 37,82,950 bp, encoding 3584 proteins. The organism is especially unique in a sense that it undergoes morphogenetic changes as a part of its biphasic life-cycle [1]. Such intriguing characteristics of *B. bacteriovorus* tempt one to analyze its genome/proteome composition with a view to identify the different selection forces that shape it.

Bdellovibrio is also important in respect to its bacteriolytic nature and may be used as a bio-control agent as it can attack plant, animal, and human pathogens [1]. Elucidation of compositional characteristics of different groups of genes/proteins of the organism may give an insight into the evolutionary strategies taken by this predatory bacterium for attacking its prey cells.

Another interesting aspect of *B. bacteriovorus* is that in spite of having easy access to prey's genetic pool, no cases of recent prey-derived horizontal gene transfer (HGT) were detected so far [1,4]. Nearly 20% of the coding genes of diverse functions are believed to be horizontally acquired long ago and ameliorated to the bulk of the sequence [4]. It is, therefore, tempting to investigate if there could be any prey-derived horizontally transferred genes in *Bdellovibrio* in recent past.

Using multivariate statistical analysis, the present study attempts to address the following interrelated issues: (i) elucidation of the selection forces causing intra- genomic/proteomic variations in sequence composition, (ii) identification of the recent prey-derived horizontally transferred genes, if any, and (iii) identification of groups of genes/proteins characterized by distinct compositional characteristics which may play important role in predatory mode of life of *B. bacteriovorus*.

Abbreviations: CAI, Codon adaptation index; COA, Correspondence analysis; RSCU, Relative synonymous codon usage; RAAU, Relative amino acid usage; HEG, Highly expressed genes; LEG, Lowly expressed genes; GRAVY, Average hydropathy; HGT, Horizontal gene transfer; CSAP, Cell-wall surface anchor family protein; AA, Amino acid.

* Corresponding author at: Centre for Bioinformatics, School of Life Sciences, Pondicherry University, Pondicherry-605014, India. Fax: +91 413 2655211.
E-mail addresses: archana@bicpu.edu.in, archanpan@gmail.com (A. Pan).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.06.007

2. Materials and methods

2.1. Genome sequence data

All protein-coding sequences of *B. bacteriovorus HD100* were extracted from NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). To reduce the sampling error, genes with internal stop codons, untranslated codons and the annotated genes with fewer than 100 codons were excluded from the analysis, resulting 3331 annotated gene sequences in the final dataset. Also, the protein-coding sequences of other 10 deltaproteobacteria namely *Desulfatibacillum alkenivorans AK-01, Desulfobacterium autotrophicum HRM2, Geobacter lovleyi SZ, Geobacter uraniireducens Rf4, Pelobacter carbinolicus DSM 2380, Syntrophus aciditrophicus SB, Desulfococcus oleovorans Hxd3, Geobacter sp. FRC-32, Desulfotalea psychrophila LSv54, Desulfovibrio salexigens DSM 2638* were downloaded from NCBI FTP site.

2.2. Parameters used to identify selection forces in gene/protein composition

For each protein-coding gene, frequencies of nucleotides at the 3rd codon position (A_{3S}, T_{3S}, G_{3S}, and C_{3S}) and GC-content at 3rd codon position (GC_{3S}) were evaluated. This excludes the singlet codons (i.e., ATG for Met and TTG for Trp) from the calculation and considers only the third codon sites of the other 18 residues, which are encoded by more than one codon. Moreover, effective number of codons (Nc) [5], relative synonymous codon usage (RSCU) and codon adaptation index (CAI) [6] for each protein coding gene were evaluated and for corresponding gene products, relative amino acid usage (RAAU), average hydropathy (GRAVY) [7], aromaticity [8] and alcoholicity [9] were estimated. CAI values of individual genes were calculated taking a reference gene set comprising of genes encoding ribosomal proteins, which are known to be expressed at high level in most bacterial organisms [10–12].

2.3. Multivariate statistical analysis of codon and amino acid usage

Correspondence analysis (COA) was performed in order to identify the sources of intra-genomic and proteomic variations in *B. bacteriovorus* genome, using the program CodonW 1.4.2 (written by John Peden and available from ftp://molbiol.ox.ac.uk/Win95.codonW.zip). COA on RSCU was also performed for other 10 deltaproteobacteria under study. For COA on RSCU, the data were plotted in a multidimensional space of 59 axes (excluding AUG, UGG, and stop codons), while in COA on RAAU, the variables were the frequencies of 20 amino acids, which sum up to 1 for each gene.

2.4. Datasets for different groups of genes/proteins

The datasets for putatively highly and lowly expressed genes were generated taking 50 genes for each group respectively from extreme left and right of axis1, which is strongly negatively correlated with CAI. Two datasets of AT-rich genes and the reported HGT [4] were prepared. Datasets for cell-wall surface anchor family proteins (CSAPs) and transporter proteins (Transporters) were formed for amino acid usage comparison.

2.5. Phylogenetic analysis

The orthologs for each AT-rich gene product were extracted using blastX [13]. The homologs with e-values <e-30, identity >35% and 80% of the query length coverage were considered as orthologs. In case of several hits of homologs for a single species, the best hit sequence that produced alignment with the lowest e-value was considered as ortholog. Multiple sequence alignments between the orthologs were carried out using clustalW (with defaults settings) [14]. Phylogenetic analysis was performed for each of the AT-rich genes by neighbor-joining algorithm and bootstrap analysis was carried out with 500 trials using MEGA program (Version 4) [15]. Phylogenetic trees were constructed only for functionally known AT-rich gene products. These trees were compared with a reference tree which was generated using DNA-directed RNA polymerase (RpoB), a core-genome marker. It is to be noted that for type I restriction–modification system S subunit, orthologs could not be determined as none of the blast hits of the query met 80% of the query length coverage. Thereby, phylogenetic tree for that protein could not be constructed. Also, as the hypothetical proteins in the AT-rich cluster have no significant homologs in the database, their phylogenetic trees could not be constructed.

2.6. Statistical analysis

In order to detect the significant differences between the highly and lowly expressed genes, if any, codon usage abundances were compared by chi-square in a 2×2 contingency table. For each codon, the first and second rows of the contingency table represented, respectively, the number of occurrences of the codon being analyzed and the total number of alternative synonyms for the corresponding residue in the two classes under comparison. For amino acids, rows in the contingency table were the count of a particular amino acid residue and the total amino acid counts for all 20 residues in the respective datasets under study.

2.7. Protein secondary structure prediction

Prediction of transmembrane domains of each protein of CSAP and Transporter clusters were made using TMHMM Server (http://www.cbs.dtu.dk/services/TMHMM/) and the ordered/globularity and disordered regions within proteins were predicted using GlobPlot (http://globplot.embl.de).

3. Results and discussion

3.1. Intra-species variations in gene composition

The overall codon usage pattern of 3331 protein-coding genes of *B. bacteriovorus* exhibits a strong bias toward the usage of the G-/C-ending codons (Table 1) although the average GC-content of its genome is 50.7%. The wide range of variations in GC_{3S} (19.3%–70.9%), effective number of codon, Nc (26.68–61) and CAI values (0.25–0.82) of the genes indicate the presence of intra-genomic variations in codon usage pattern.

In order to identify the trends of variations in gene composition, COA on RSCU in 3331 predicted protein-coding genes of *B. bacteriovorus* was performed. The first four axes generated by COA on RSCU account for 36.53% of the total variations. Among the four axes, first 3 axes explain 14.09%, 8.52% and 6.96% variations, respectively. Fig. 1A shows the position of genes among the first two major axes. It depicts that the most of the genes are clustered around the origin (0,0), while two small groups of potentially highly expressed genes and genes with atypical base composition are appeared at extreme left of axis1 (red diamonds) and upper end of axis2 (pink squares), respectively.

3.1.1. Influence of translational selection on synonymous codon usage

Axis1 shows strong negative correlation with the expression level of genes, as measured by their respective CAI values ($r=-0.94$, $p<0.001$) (Fig. 1B). The genes with CAI values higher than 0.65 encode potentially highly expressed genes (henceforth referred to as HEG) such as different ribosomal proteins, translational elongation factors Tu/Ts/EF-2, chaperonin proteins, DNA-directed RNA polymerase alpha/beta subunits etc. (red diamond, Fig. 1B). The comparative analysis of codon usage pattern of the genes lying at two extreme ends of axis1 of Fig. 1B reveals that there are marked differences in synonymous codon

Table 1
Codon usage of highly and lowly expressed genes positioned on extreme ends in axis1, total genes, reported HGT [4] and AT-rich genes in *B. bacteriovorus*.

AA[a]	Codon	RSCU[b]		RSCU		
		HEG[c]	LEG[d]	Total	Reported HGTs	AT-rich genes
Phe	UUU	0.20	1.30[e]	0.87	0.87	1.43[f]
	UUC	1.80[g]	0.70	1.13[h]	1.13[h]	0.57
Leu	UUA	0.02	0.33[e]	0.24	0.20	1.16[f]
	UUG	2.72[g]	1.45	1.46	1.48	1.46
	CUU	1.92[g]	0.78	1.04	1.08	1.38[f]
	CUC	0.04	0.40[e]	0.23	0.18	0.53[f]
	CUA	0.26[g]	0.15	0.16	0.14	0.78[f]
	CUG	1.04	2.89[e]	2.86[h]	2.93[h]	0.68
Ile	AUU	0.35	1.40[e]	1.10	1.15	1.64[f]
	AUC	2.65[g]	1.32	1.80[h]	1.77[h]	0.71
	AUA	0.00	0.28[e]	0.11	0.08	0.65[f]
Val	GUU	2.34[g]	0.82	1.02	0.90	1.65[f]
	GUC	0.12	1.16[e]	0.86	0.83	0.65
	GUA	1.00[g]	0.12	0.26	0.21	0.91[f]
	GUG	0.54	1.90[e]	1.86[h]	2.06[h]	0.79
Ser	UCU	3.23[g]	0.65	1.16	1.08	1.38[f]
	UCC	1.98[g]	1.37	1.72[h]	1.80[h]	0.71
	UCA	0.20	0.85[e]	0.56	0.57	1.34[f]
	UCG	0.04	1.05[e]	0.70	0.76	0.67
	AGU	0.09	1.14[e]	0.71	0.67	1.17[f]
	AGC	0.46	0.93[e]	1.15[h]	1.12[h]	0.72
Pro	CCU	1.38[g]	0.58	0.86	0.84	1.27[f]
	CCC	0.04	1.69[e]	0.73[h]	0.71[h]	0.55
	CCA	2.03[g]	0.30	0.73	0.65	1.45[f]
	CCG	0.55	1.43[e]	1.68[h]	1.79[h]	0.72
Thr	ACU	2.71[g]	0.60	0.94	0.86	1.36[f]
	ACC	0.20	1.92[e]	1.53[h]	1.57[h]	0.76
	ACA	0.78[g]	0.58	0.67	0.64	1.24[f]
	ACG	0.31	0.89[e]	0.85[h]	0.93[h]	0.64
Ala	GCU	2.07[g]	0.59	0.89	0.72	1.25[f]
	GCC	0.19	1.82[e]	1.46[h]	1.59[h]	0.70
	GCA	0.99[g]	0.49	0.58	0.54	1.32[f]
	GCG	0.76	1.10[e]	1.07[h]	1.16[h]	0.73
Tyr	UAU	0.28	1.36[e]	1.08	1.11	1.32[f]
	UAC	1.702[g]	0.64	0.92[h]	0.89[h]	0.68
His	CAU	0.24	1.14[e]	0.83	0.74	1.36[f]
	CAC	1.76[g]	0.86	1.17[h]	1.26[h]	0.64
Gln	CAA	1.27[g]	0.66	0.69	0.65	1.12[f]
	CAG	0.73	1.34[e]	1.31[h]	1.35[h]	0.88
Asn	AAU	0.38	1.02[e]	0.87	0.91	1.40[f]
	AAC	1.62[g]	0.98	1.13[h]	1.09[h]	0.60
Lys	AAA	1.59[g]	1.06	1.26	1.25	1.22
	AAG	0.41	0.94[e]	0.74	0.75	0.78
Asp	GAU	0.76	1.13[e]	1.03	1.08	1.41[f]
	GAC	1.24[g]	0.87	0.97[h]	0.92[h]	0.59
Glu	GAA	1.43[g]	1.21	1.37[h]	1.40[h]	1.17
	GAG	0.57	0.79[e]	0.63	0.60	0.83[f]
Cys	UGU	0.60	1.23[e]	0.86	0.81	1.18[f]
	UGC	1.40[g]	0.77	1.14[h]	1.19[h]	0.82
Arg	CGU	3.75[g]	0.63	1.88[h]	1.91[h]	0.73
	CGC	1.39	2.27[e]	2.56[h]	2.69[h]	0.72
	CGA	0.00	0.69[e]	0.38	0.34	1.12[f]
	CGG	0.00	1.57[e]	0.45	0.43	0.42
	AGA	0.84[g]	0.42	0.57	0.51	1.93[f]
	AGG	0.02	0.43[e]	0.16	0.13	1.08[f]
Gly	GGU	2.55[g]	0.67	1.24	1.23	1.24
	GGC	1.22	1.20	1.50[h]	1.53[h]	0.84
	GGA	0.15	0.96[e]	0.66	0.63	1.20[f]
	GGG	0.08	1.16[e]	0.61	0.61	0.72[f]

[a] Amino acid.
[b] Relative synonymous codon usage.
[c] Highly expressed genes.
[d] Lowly expressed genes.
[e] Codons occur in significantly higher amount in LEGs ($p<0.001$) compared to HEGs.
[f] Codons occurring in significantly higher amount in AT-rich genes ($p<0.001$) compared to all genes under study.
[g] Codons occurring in significantly higher amount in HEGs ($p<0.001$) compared to LEGs.
[h] Codons occurring in significantly higher amount in total genes and reported HGTs under study ($p<0.001$) compared to AT-rich genes.

usage between these two groups of genes (Table 1). It is found that the highly expressed genes use a subset of 26 synonymous codons, known as optimal codons, with significantly ($p<0.001$) higher frequencies than the lowly expressed genes. Among these 26 optimal codons, 17 are A-/U-ending and the rest are G-/C-ending codons (Table 1). The preference of putatively highly expressed genes toward A/T base at their synonymous codon positions is supported by the positive correlation of axis1 and GC_{3S} ($r=0.41$, $p<0.001$) (Fig. 1C). Thus, it is supposed that translational selection [16] operates in generating codon usage bias in highly expressed genes. The most important evidence that translational selection is operative for this species is the fact that all universal optimal codons (mainly those C-ending codons belonging to two-fold degenerate triplets) are incremented among HEGs. Moreover, the tRNAscan-SE analysis (the genomic tRNA database) of this genome reveals a good correspondence between preferred (optimal) codons and the tRNA abundance of the species (Table 2). Table 2 shows that about 66% of the available tRNAs in this species complement to the optimal codons of the highly expressed genes. This correlation is remarkable, particularly in two- and three-fold degenerate codons. Thus, all these findings suggest that the translational selection plays a major/primary role in synonymous codon usage of *B. bacteriovorus*.

In order to examine whether the translational selection is a general characteristic of deltaproteobacteria or whether it is a specific feature of *B. bacteriovorus* only, synonymous codon usage pattern in deltaproteobacterial genomes (GC-content comparable to *B. bacteriovorus*) sequenced so far was analyzed in the present study. Table 3 represents the comparison of codon usage pattern (as determined by COA on RSCU values of genes) of 11 completely sequenced deltaproteobacteria. In all organisms except *B. bacteriovorus* and *Desulfovibrio salexigens*, the position of each gene along the first axis exhibits a much higher correlation with the respective GC_{3S}-content than with CAI. In *Desulfatibacillum alkenivorans, Desulfobacterium autotrophicum* three species of *Geobacter* and *Desulfotalea psychrophila*, strong correlations of axis1 are observed with GC_{3S} as well as GT_{3S}-content of the genes. Furthermore, Table 3 shows that the correlation of CAI with axis1 is strongest in *B. bacteriovorus* in comparison to other genomes, suggesting that translational selection is operating in higher intensity in this predatory organism. This phenomenon is important in view of fact that it may explain the rapid amelioration of horizontally acquired genes, which make up about 20% of the coding genes of the genome of *B. bacteriovorus* [4].

3.1.2. Influence of local GC-bias

Axis2 has strong negative correlation with GC_{3S} ($r=-0.70$, $p<0.001$) (Fig. 1D). In order to check whether this could be due to variation in local GC-bias, GC_{3S} values of individual genes were compared with the GC-content of the respective flanking regions. The flanking regions represent the upstream and downstream non-coding regions of individual genes. According to the mutational hypothesis [17], the GC_{3S}-content of genes is expected to vary with the GC-content of the neighboring non-coding sequences. About two-hundred *B. bacteriovorus* genes and their corresponding flanking regions (>100 bases long) were extracted from its genomic sequence manually. The GC_{3S}-content of these genes shows significant positive correlation with GC-content of the respective flanking regions ($r=0.32$, $p<0.001$), suggesting that the local GC-bias does play a secondary role in imparting variations in codon usage pattern in this organism.

3.1.3. Influence of recent prey-derived horizontally transferred genes

A distinct cluster of genes appears toward the upper end of axis2, as detected in the axis1 versus axis2 plot (pink square, Fig. 1A). Table 1 shows that the usage of A and T is significantly higher at the synonymous codon positions of these genes (henceforth referred to as AT-rich genes) than the rest of the genes. The strong negative correlation of axis2 with GC_{3S} ($r=-0.70$, $p<0.001$) (Fig. 1D) also supports that the genes in this cluster are AT-rich. Furthermore, this should be noted that although the highly expressed genes of *B.*

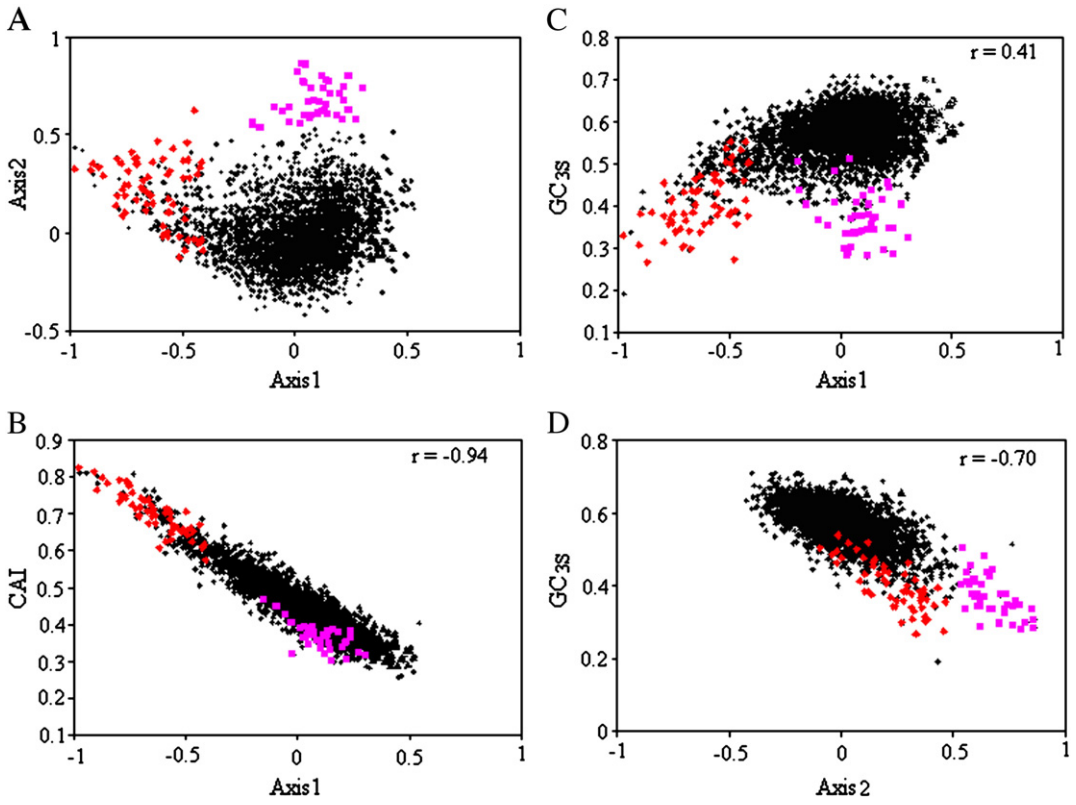

Fig. 1. Position of genes along axis1 generated by COA on RSCU has been plotted against axis2 (A), CAI (B), GC_{3S} (C) and that of axis2 against GC_{3S} (D). Red and pink represent HEG and AT-rich genes, respectively.

bacteriovorus prefer to use A-/U-ending codons, the majority of the codons overrepresented in the genes of pink cluster (Fig. 1A) are different from the optimal codons used by the HEGs (Table 1) and the CAI values of these genes are relatively low (Fig. 1B), suggesting that these genes are different from the highly expressed genes in codon usage pattern and potential level of expression.

The genes in the pink cluster (Fig. 1A) encode probable UDP-glucose 4-epimerase, putative aminotransferase, putative UDP-N-acetylglucosamine-2-epimerase NeuC, putative formyltransferase, putative N-acetylneuraminic acid synthetase, hexapeptide transferase family protein, Mannose-1-phosphate guanyltransferase, probable acylneuraminate cytidylyltransferase, three different type I restriction–modification system subunits and many hypothetical proteins (Supplementary Table 1). The genes in the AT-rich cluster (Fig. 1A) were compared with the genes in four AT-rich regions of *B. bacteriovorus* genome as reported by Rendulic et al.[1]. These four AT-rich regions code ribosomal, lipopolysaccharide (LPS) synthesis, prophage and restriction modification genes (R/M system) [1]. It is found that the genes in the AT-rich cluster (Fig. 1A) belong to the LPS cluster and R/M system of AT-rich regions of the genome [1]. Surprisingly, it is also noticed that neither all the genes of LPS cluster nor the prophage and the ribosomal proteins of AT-rich regions of the genome [1] are present in the pink cluster (Fig. 1A). These genes appear along with the main group of genes around the origin and highly expressed gene cluster, respectively (Supplementary Fig. 1). This phenomenon indicates that these genes seem to be almost ameliorated with the main group of genes and native highly expressed genes, respectively of *B. bacteriovorus* genome. On the contrary, genes in the pink cluster (Fig. 1A) have the synonymous codon usage pattern distinctly different from the rest of the genome and also from the ancient horizontally transferred genes [4] (Table 1). This indicates that these genes might have transferred recently.

Table 2
List of optimal codons and number of corresponding tRNAs in *B. bacteriovorus*.

Amino acids	Optimal codons	No. of corresponding tRNAs
Phe	UUC	1 (1)
Tyr	UAC	1 (1)
His	CAC	1 (1)
Gln	CAA	1 (1)
Asn	AAC	1 (1)
Lys	AAA	1 (1)
Asp	GAC	1 (1)
Glu	GAA	1 (1)
Cys	UGC	1 (1)
Ile	AUC	2 (2)
Arg	CGU, AGA	2 (3)
Leu	UUG, CUU, CUA	2 (5)
Ser	UCU, UCC	1 (3)
Val	GUU, GUA	1 (2)
Pro	CCU, CCA	1 (2)
Thr	ACU, ACA	1 (2)
Ala	GCU, GCA	1 (2)
Gly	GGU	1 (2)

Number in parentheses indicates total number of isoacceptor tRNA available.

3.1.3.1. The phylogenetic analysis of AT-rich genes. The phylogenetic analysis was carried out on the genes (functionally known) that are present in the AT-rich cluster (pink dots, Fig. 1A) in order to identify their potential source organisms. A reference tree for RpoB (Supplementary Fig. 2) was generated which depicts that *B. bacteriovorus* is close to other deltaproteobacteria. This is not unexpected in view of the fact that the gene encoding RpoB is a house-keeping/native gene. On the other hand, the phylogenetic trees of Mannose-1-phosphate- guanyltransferase and UDP-glucose 4-epimerase illustrate that these gene products are branched with the organisms other than deltaproteobacteria like *Nitrococcus mobilis*, *Raphidiopsis brookii* (88% bootstrap value) and *Thiomicrospira crunogena*, *Pseudoalteromonas tunicate* (55% bootstrap value),

Table 3
Genome characteristics and correlations of CAI, GC_{3S} and GT_{3S}-content with first two axes of COA on RSCU values of genes in 11 completely sequenced deltaproteobacterial genomes.

				Variation explained by COA on RSCU (%)		Correlation (r)					
						Axis1			Axis2		
Organism with strain	Accession No.	Size (Mb)	GC (%)	Axis1	Axis2	GC_{3S}	CAI	GT_{3S}	GC_{3S}	CAI	GT_{3S}
Desulfatibacillum alkenivorans AK-01	CP001322.1	6.50	54.5	14.82	6.88	**−0.90**	−0.51	**−0.79**	0.08^{NS}	−0.50	0.30
Desulfobacterium autotrophicum HRM2	CP001087.1	5.66	48.8	13.23	4.88	**0.83**	0.25	**0.83**	-0.01^{NS}	-0.08^{NS}	0.51
Geobacter lovleyi SZ	CP001089.1	3.98	54.7	15.54	5.59	**0.83**	0.51	**0.77**	0.15	-0.08^{NS}	-0.02^{NS}
Geobacter uraniireducens Rf4	CP000698.1	5.14	54.2	18.39	4.18	**0.84**	0.38	**0.78**	-0.03^{NS}	-0.05^{NS}	0.09
Pelobacter carbinolicus DSM 2380	CP000142.2	3.67	55.1	17.01	7.89	**0.92**	0.65	−0.48	-0.03^{NS}	0.15	**−0.76**
Syntrophus aciditrophicus SB	CP000252.1	3.18	51.5	18.62	4.16	**−0.94**	−0.43	0.23	−0.12	0.01^{NS}	0.33
Desulfococcus oleovorans Hxd3	CP000859.1	3.94	56.2	22.97	3.70	**0.96**	0.57	−0.24	-0.01^{NS}	−0.16	0.50
Geobacter sp. FRC-32	CP001390.1	4.30	53.5	13.57	4.40	**−0.92**	−0.43	**−0.82**	0.12	−0.39	0.33
Desulfotalea psychrophila LSv54	CR522870.1	3.66	46.6	11.49	9.27	**−0.72**	-0.02^{NS}	**0.78**	−0.51	−0.29	−0.42
Desulfovibrio salexigens DSM 2638	CP001649.1	4.3	47.1	15.70	8.53	0.61	**0.76**	0.22	0.01^{NS}	0.15	0.31
Bdellovibrio bacteriovorus HD100	BX842601.2	3.8	50.6	14.09	8.52	0.41	**−0.94**	0.49	**−0.70**	−0.46	0.44

Note: All Correlations are significant at $p<0.001$ except the correlation represented by ^{NS}.
^{NS} indicates non-significant. Boldfaces highlight higher correlation values >0.7.

respectively (Fig. 2). These organisms are Gram-negative gamma-proteobacteria. Similarly, the phylogenetic trees of the other gene products from AT-rich cluster show that these are also branched with the different Gram-negative bacteria that belong to classes like gamma-, alpha-, beta-proteobacteria, cyanobacteria, fusobacteria etc. (Supplementary Fig. 3, Table 4). These findings are statistically significant with the bootstrap values ranging from 55 to 98%. *Bdellovibrio* species have a broad range of prey and any Gram-negative bacterium that *Bdellovibrio* can penetrate could be a host [4]. Hence, the Gram-negative bacteria that have closest homologs to the gene products of AT-rich genes of *B. bacteriovorus* are supposed to be its prey species. It has also been identified that many of these Gram-negative bacteria are AT-rich (Table 4). In order to support the evidence of this analysis, a counter-test on functionally known GC-rich genes (>56%) was carried out using the same procedure. The phylogenetic trees demonstrate that the gene products of GC-rich genes are also branched with Gram-negative bacteria other than deltaproteobacteria with bootstrap values varying between 61 and 98% (Supplementary Fig. 4). It should be mentioned here that although these GC-rich genes are supposed to be horizontally transferred from other Gram-negative prey-bacteria, their codon usage pattern is similar to that of majority genes of the genome, clustered around the origin (black dots, Supplementary Fig. 1). This finding leads us to assume that the AT-rich genes of the cluster, that are compositionally different from the rest of the genes of *B. bacteriovorus,* are acquired by horizontal transfer from different Gram-negative bacteria while preyed upon them in the recent past.

3.1.4. Influence of other sources of variations

Axis3 has strong negative correlation with GT_{3S} ($r=-0.56$, $p<0.001$). GT_{3S} is considered to be a parameter denoting the presence of strand-specific mutational bias in many eubacterial, viral and organelle genomes, where the leading strand is reported to be relatively GT-rich and the lagging strand is CA-rich [18–21]. Thus, it plays a minor/tertiary role in intra-genomic variation of codon usage. In order to detect the replication origin, Ori-Finder [22], an online tool was used followed by GC-skew analysis [23]. The probable replication origin of the species is found to be located near about 3781 kbps (Supplementary Fig. 5). The comparison of synonymous codon usage pattern between the leading and lagging strands on both sides of the replication origin depicts that the synonymous codon usage pattern does not vary significantly on both strands. Therefore, it supports that strand-specific mutational bias plays a minor role in codon usage variation.

3.2. Intra-species variations in protein composition

In order to identify the trends of variations in protein composition, COA on RAAU in 3331 predicted protein-coding genes of *B. bacteriovorus* was performed. The first four axes generated by COA on RAAU account for 45.65% of the total variations in amino acid composition of proteins, among these the first two axes explain 17.64% and 11.99% of variations, respectively. Fig. 3A shows the position of each protein within the first two major axes. Most of the proteins including the products of the HEGs (red dots) and those of the AT-rich genes (pink dots) formed together a large cluster (black dots) near the origin (0,0), indicating that the amino acid composition of the products of HEGs as well as AT-rich genes follow the general trend in amino acid usage in *B. bacteriovorus*. A small cluster of mainly cell-wall surface anchor family proteins and a few pili proteins along with some hypothetical proteins (blue dots, henceforth referred together as CSAP) lie toward the positive coordinate of axis1 and a small cluster of transporter proteins including hypothetical proteins (green dots, henceforth referred together as Transporter) are segregated toward the positive coordinate of axis2 (Fig. 3A) indicating that the proteins belong to these two clusters have distinct amino acid composition.

3.2.1. Influence of different physico-chemical factors

In course of determination of sources of variations in proteome composition, it is found that the first axis shows strong positive correlation with the alcoholicity ($r=0.62$, $p<0.001$) (Fig. 3B) and negative correlation with charged residues like Glu, Lys and Arg ($r=-0.73$, -0.57 and -0.47, respectively, $p<0.001$), whereas the second axis is positively correlated with aromaticity ($r=0.69$, $p<0.001$) and GRAVY ($r=0.60$, $p<0.001$) (Fig. 3C, D). This indicates that the different physico-chemical factors like alcoholicity and residue-charge primarily and aromaticity and hydropathy secondarily influence the proteome compositional variations in this predatory organism.

3.2.2. Distinct protein composition of cell-wall surface anchor family proteins and transporter proteins

To determine the amino acid composition in detail, CSAPs and Transporters were further analyzed. Analysis reveals that CSAPs are characterized by strikingly high frequencies of Ser, Thr, Asn, Gly, Ala, Cys and low frequencies of Asp, Glu, Lys and Arg (Table 5). This finding is supported by Fig. 4 showing the position of the amino acids in the plane defined by the two main axes. The result suggests that CSAPs prefer to use (i) sp3-hybridized hydroxyl group of Serine/Threonine

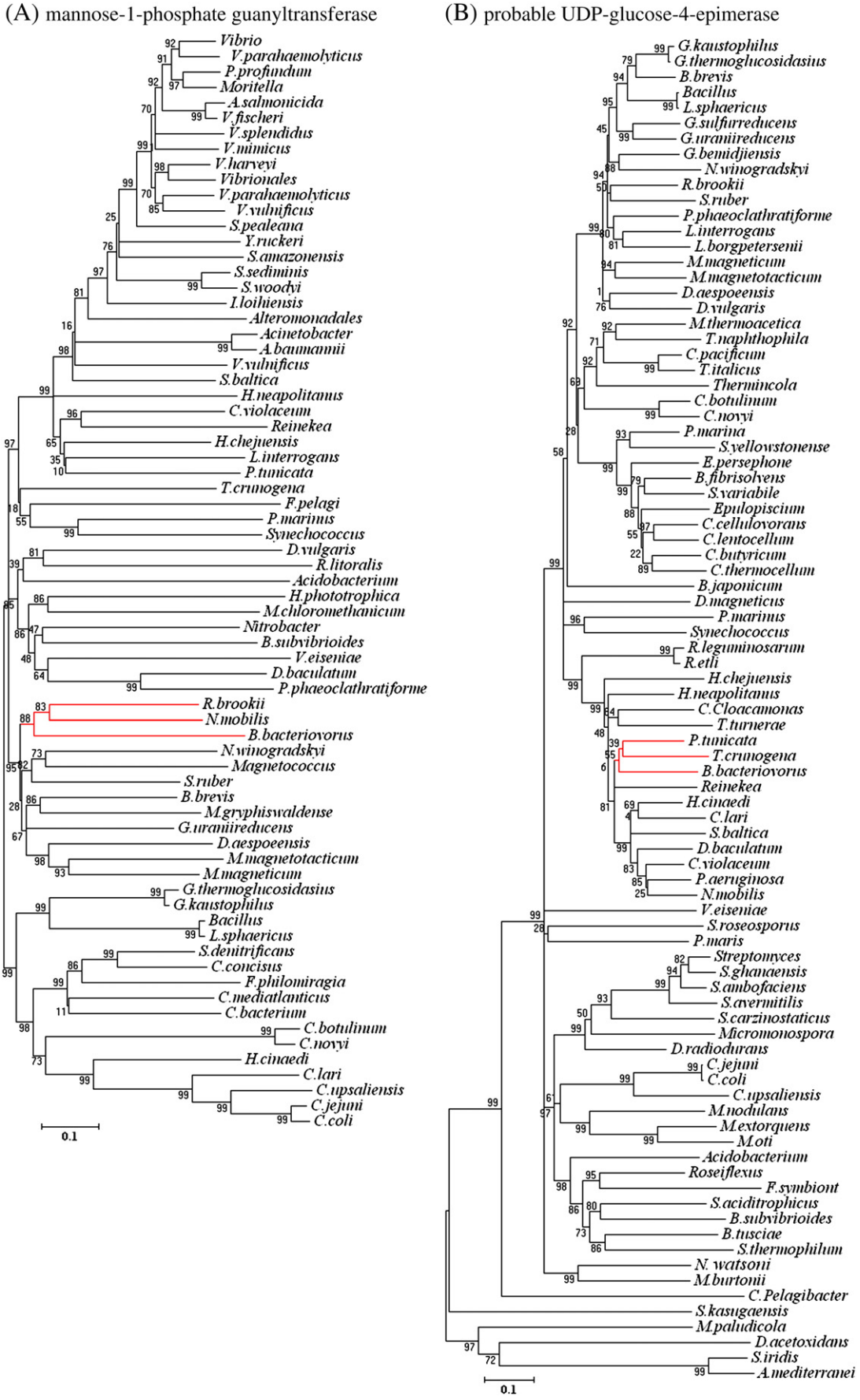

Fig. 2. Phylogenetic trees for (A) Mannose-1-phosphate- guanyltransferase, (B) probable UDP-glucose 4-epimerase. Red lines indicate *B. bacteriovorus* and its phylogenetically closest organisms.

Table 4
List of AT-rich gene products, phylogenetically analyzed.

Location of gene in chromosome	Function	Closest organisms	GC% of closest organism	Gram-strain	Class
1615053-1616036	Probable UDP-glucose 4-epimerase	*Thiomicrospira crunogena*	43.1%	Negative	Gamma-proteobacteria
		Pseudoalteromonas tunicata	40%	Negative	Gamma proteobacteria
1616029-1617177	Putative aminotransferase	*Sphingomonas wittichii*	67%	Negative	Alpha-proteobacteria
1617215-1618387	Putative UDP-N-acetylglucosamine-2-epimerase NeuC	*Hahella chejuensis*	53%	Negative	Gamma-proteobacteria
1618384-1619271	Putative formyltransferase	*Campylobacter jejuni*	32.7%	Negative	Epsilon-proteobacteria
1619936-1621006	Putative N-acetylneuraminic acid synthetase	*Chromobacterium violaceum*	64%	Negative	Beta-proteobacteria
1621680-1622732	Mannose-1-phosphate- guanyltransferase	*Nitrococcus mobilis*	60%	Negative	Gamma-proteobacteria
		Raphidiopsis brookii	40%	Negative	Cyanobacteria
1622710-1623402	Probable acylneuraminate cytidylytransferase	*Idiomarina loihiensis*	47.4%	Negative	Gamma-proteobacteria
1624303-1625571	Putative polysaccharide biosynthesis protein CpsL	*Nitrococcus mobilis*	60%	Negative	Gamma-proteobacteria
		Deferribacter desulfuricans	31.1%	Negative	Other bacteria
c1634762-1633893	UDP-N-acetyl-D-quinovosamine 4-epimerase	*Fusobacterium nucleatum*	26.84%	Negative	Fusobacteria
c3576689-3573462	Type I restriction-modification system restriction subunit	*Vibrio harveyi*	47%	Negative	Gamma-proteobacteria
		Acetobacter pasteurianus	50.7%	Negative	Alpha-proteobacteria
c3579707-3577950	Type I restriction enzyme M protein	*Pseudomonas syringae*	58.4%	Negative	Gamma-proteobacteria

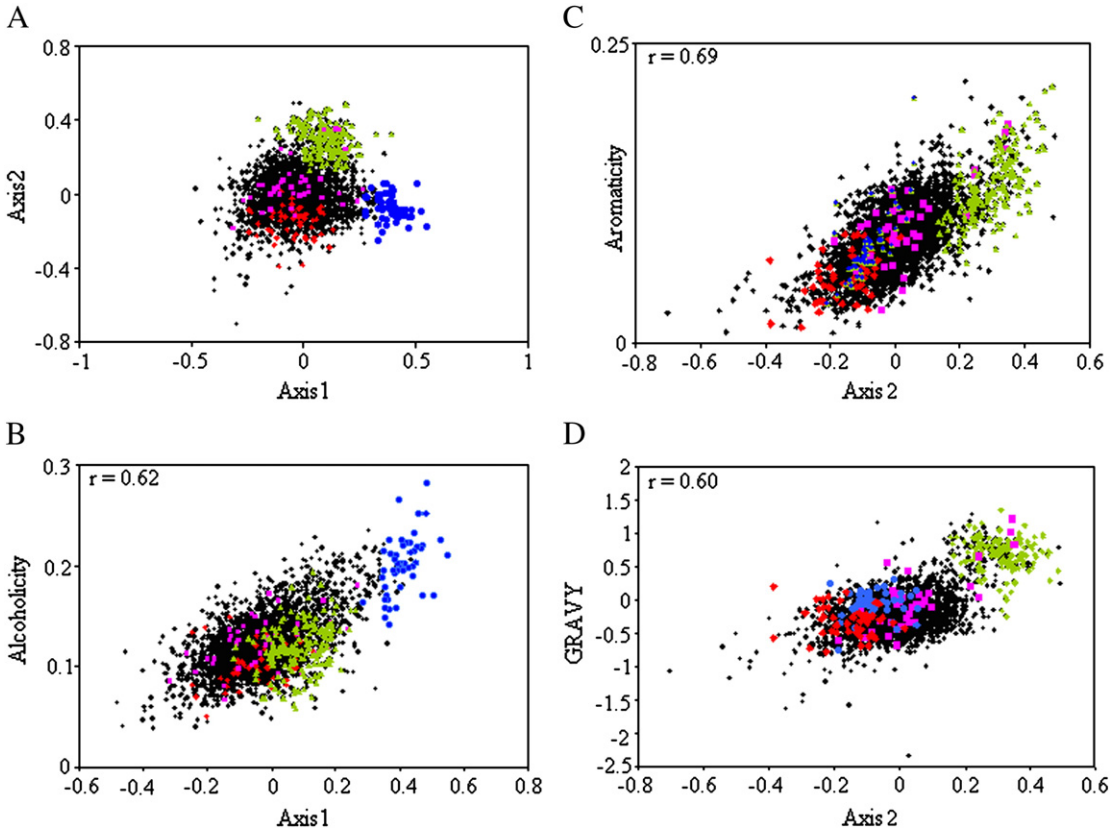

Fig. 3. Position of genes along axis1 generated by COA on RAAU has been plotted against axis2 (A) and alcoholicity (B) and that of axis2 against aromaticity (C) and GRAVY (D). Red, pink, blue, green represents HEG, AT-rich genes, CSAPs, Transporters, respectively.

that provides point of attachment for other molecules, particularly non-amino acid ones to the respective proteins [9], (ii) hydrogen bonds provided by amide group of Asn to the surrounding atoms during inter-molecular interactions [24] and (iii) disulfide bonds contributed by Cys residues probably for the irreversible anchor to the prey after the short period of recognition [1]. The significant decrease of the charged residues in CSAPs implies that these proteins have less inclination toward the use of inter-molecular electrostatic interaction. Furthermore, the occurrence of Gly and Ala at higher frequencies in CSAPs indicates that these proteins consist of residues of low molecular weight which is further supported by the negative correlation of axis1 generated by COA on RAAU and molecular weight of the proteins ($r = -0.62$, $p < 0.0001$). The prediction of transmembrane domains of known cell-wall surface anchor family proteins and hypothetical proteins in the cluster reveals that these proteins hardly have any transmembrane domain and are located in the outer membrane of the bacterial cells (Fig. 5). It is worth mentioning at this point that outer membrane proteins (OMPs) of *Bdellovibrio* are of low molecular weight [25]. Fig. 5 also depicts that the proteins in CSAP cluster have highly disordered structure indicating the lack of regular secondary structures and increased flexibility in the polypeptide chain [26] which probably help in protein–protein interaction [27] between prey and predator cells in this species. All these findings point toward the major role of CSAPs in prey-predator interaction by providing

Table 5
Amino acid composition of CSAPs, transporters and other genes in *B. bacteriovorus*.

AA[a]	Others	CSAP[b]	Transporter
Phe	4.93[c]	3.14	8.04[d]
Leu	10.16[c]	6.64	15.10[d]
Ile	5.75[c]	4.60	6.91[d]
Met	2.50[c]	1.49	3.27[d]
Val	6.72	7.33[e]	7.79[d]
Ser	6.58	9.79[e]	6.79
Pro	4.72[c,f]	3.82	4.11
Thr	5.03[f]	11.25[e]	4.77
Ala	7.61	11.08[e]	7.91[d]
Tyr	3.29	3.17	3.56[d]
His	2.38[c,f]	0. 95	1.89
Gln	3.79[c,f]	2.96	2.87
Asn	3.74[f]	6.63[e]	2.39
Lys	6.86[c,f]	3.18	3.65
Asp	5.21[c,f]	4.72	2.34
Glu	6.27[c,f]	2.37	2.97
Cys	1.17[f]	1.43[e]	0.92
Trp	1.36	1.25	3.18[d]
Arg	5.01[c,f]	2.58	4.19
Gly	6.90	11.62[e]	7.35[d]

[a] Amino acid.
[b] Cell-wall surface anchor family protein.
[c] Amino acids occurring in significantly higher amount in the rest of genes compared to CSAPs ($p<0.001$).
[d] Amino acids occurring in significantly higher amount in transporters compared to the rest of the genes ($p<0.001$).
[e] Amino acids occurring in significantly higher amount in CSAPs compared to the rest of the genes ($p<0.001$).
[f] Amino acids occurring in significantly higher amount in the rest of genes compared to transporters ($p<0.001$).

both reversible and irreversible bonds of interaction during short period of recognition and thereafter irreversible period of anchor and invasion to the prey [1], respectively.

Moreover, the Transporters that are found to be clustered in upper end of axis2 are characterized by significantly higher occurrence of hydrophobic and aromatic residues (Fig. 3C, D and Table 5). This finding is relevant from the fact that Transporters are integral membrane protein (Fig. 5) and thus, contain higher amount of hydrophobic and aromatic residues in the transmembrane regions. The known transporters as well as hypothetical proteins in the cluster are highly organized/ordered and contain stable secondary structures (Fig. 5). This group of proteins comprises a large proportion of the proteome of *B. bacteriovorus* and contributes significantly in solute transport particularly from prey's cytosol soon after its entry to prey's periplasm [1]. Therefore, the cell-wall surface anchor family proteins and transporter proteins with their distinct amino acid composition, potential disordered structures and intra-membrane location contribute significantly to proteome compositional variation.

4. Conclusion

In the present study, the genome and proteome composition of the predatory bacterium, *B. bacteriovorus* have been analyzed using multivariate statistical approaches. The analysis reveals that translational selection plays a major role in shaping synonymous codon composition of the organism, whereas local GC-bias and horizontal gene transfer play secondary role. Among the other sources of variations, the strand-specific mutational bias contributes minor role in synonymous codon usage preference. Comparative analysis of codon usage pattern of 11 different deltaproteobacterial genomes having GC-content comparable to *B. bacteriovorus* reveals that in this predatory bacterium the strength of translation selection is more compared to others. The study identifies a group of AT-rich genes whose codon usage pattern is significantly different from the rest of the genes of the genome. These genes are located at the AT-rich regions of the genome reported by Rendulic et al.[1]. Phylogenetic and genome compositional analysis of these genes indicate that *Bdellovibrio* might have acquired these genes recently from Gram-negative bacteria other than deltaproteobacteria, while preyed upon them. The study of proteome composition of *B. bacteriovorus* shows that it varies due to various physico-chemical properties of individual proteins, primarily alcoholicity and residue-charge and secondarily aromaticity and hydropathy. The proteins of CSAP and Transporter cluster with distinct amino acid composition and specific secondary structure also contribute significantly to variation in proteome composition. Moreover, the study reveals that CSAPs are of

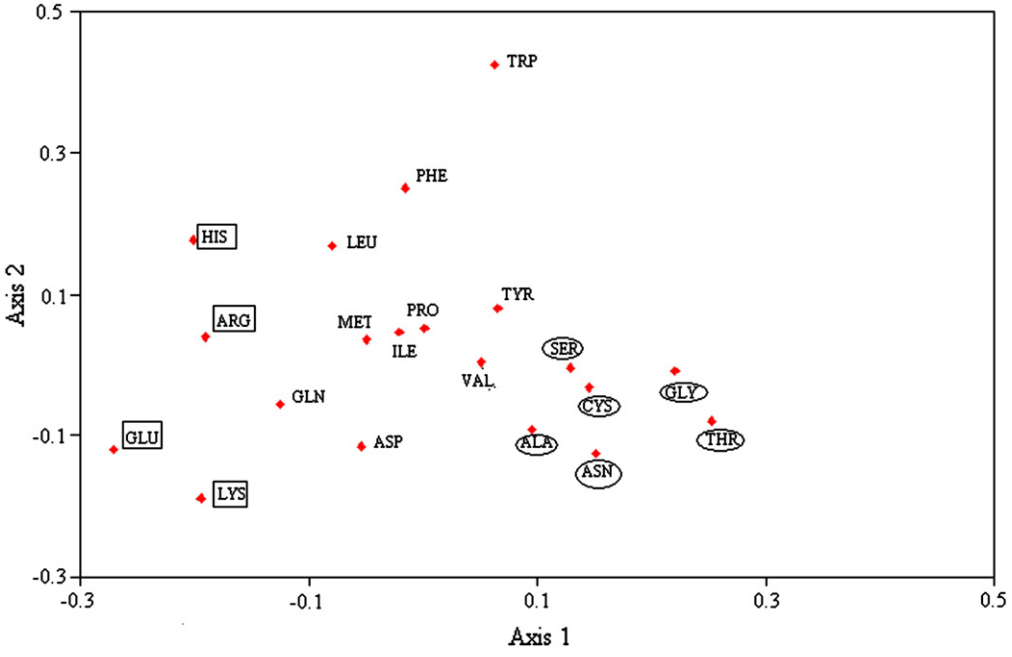

Fig. 4. The position of amino acids along axis1 and axis 2 generated by COA on amino acid usage. Circles represent the amino acids incremented in CSAPs, which present at one end of axis1. The decremented amino acids in CSAPs are marked with box at other end of axis1.

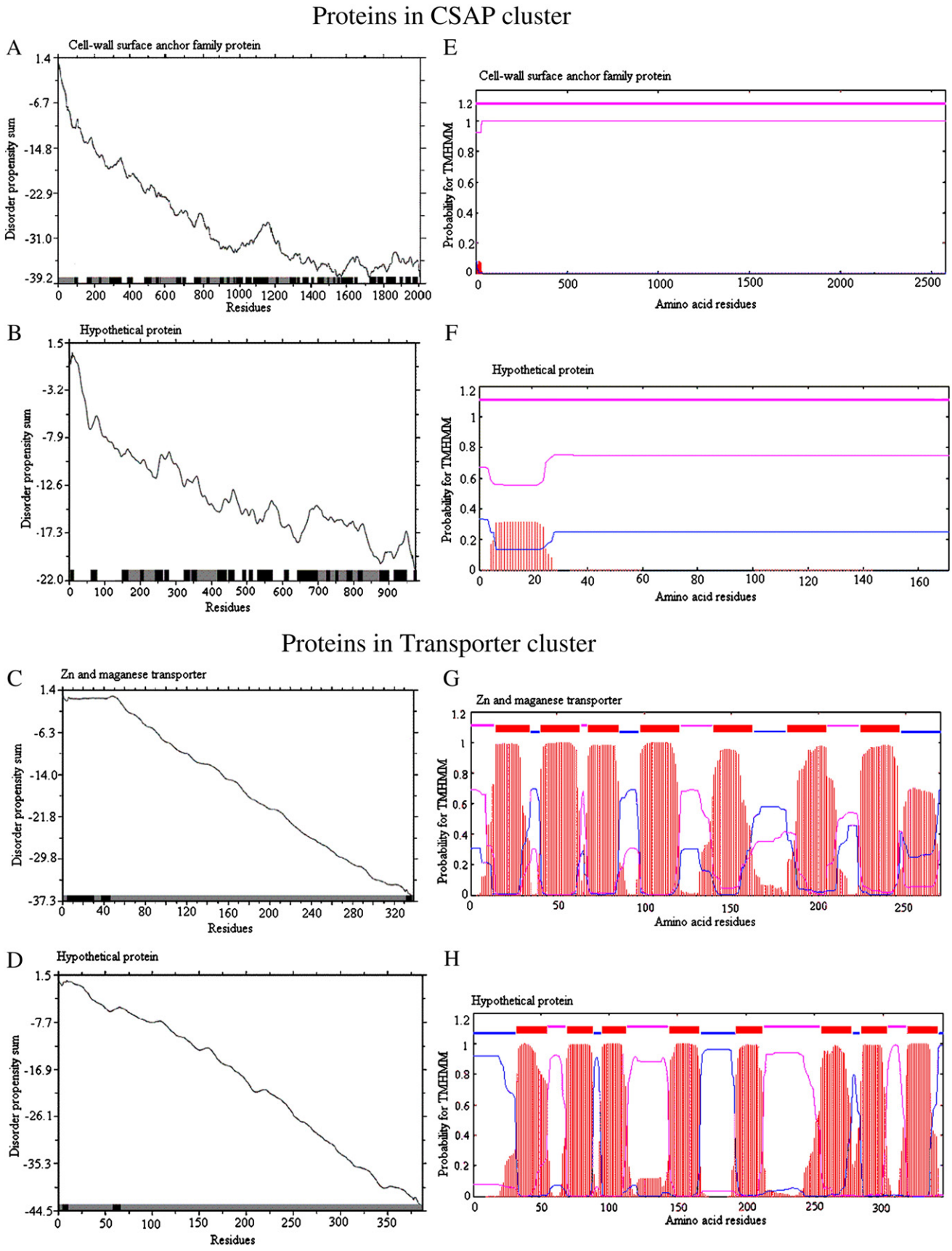

Fig. 5. GlobPlot of known cell-wall surface anchor family protein (A), hypothetical protein of CSAP cluster (B), known transporter protein (C), hypothetical protein of Transporter cluster (D) Black color indicates disordered regions (lack of regular secondary structures) and gray color indicates ordered regions (globular). TMHMM plot of known cell surface anchor family protein (E), hypothetical protein of CSAP cluster (F), known transporter protein (G), hypothetical protein of Transporter cluster (H). Red, blue and pink colors indicate the transmembrane, outside and inside regions of the protein, respectively.

low molecular weight, outer-membrane proteins with highly disordered secondary structure and have preference toward polar-uncharged residues (Ser, Thr, and Asn) and cysteine. This indicates that CSAPs participate in prey–predator interaction by providing particular reversible and irreversible bonds of interaction between them during the time of recognition, anchor and invasion to prey [1]. Thus, this study not only indicates, for the first time, the possibility of recent horizontal acquisition of AT-rich genes by *Bdellovibrio* from Gram-negative prey-bacteria other than deltaproteobacteria, but also sheds light on the distinct amino acid composition and structure of some hypothetical proteins, presumably involved in the important physiological functions like prey–predator interactions and solute transport of *B. bacteriovorus*.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.06.007.

Acknowledgments

This work was supported by the Department of Biotechnology, Government of India. AP and IC are indebted to Bioinformatics Centre, IICB, Kolkata for providing infrastructure facility during initial stages of this work. We thank Dr. Chitra Dutta, Sandip Paul, IICB, Kolkata and Dr. Murali Ayaluru, Pondicherry University, Pondicherry for critical reading of this manuscript and providing valuable suggestions.

References

[1] S. Rendulic, P. Jagtap, A. Rosinus, M. Eppinger, C. Barr, C. Lanz, H. Keller, C. Lambert, K.J. Evans, A. Goesmann, et al., A predator unmasked: life cycle of *Bdellovibrio bacteriovorus*, Science 303 (2004) 689–692.
[2] R.S. Stephens, S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R.L. Tatusov, Q. Zhao, E.V. Koonin, R.W. Davis, Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*, Science 282 (1998) 754–759.
[3] R. Gil, F.J. Silva, E. Zientz, F. Delmotte, F. Gonzalez-Candelas, A. Latorre, C. Rausell, J. Kamerbeek, J. Gadau, B. Hsolldobler, et al., The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 9388–9393.
[4] U. Gophna, R.L. Charlebois, W.F. Doolittle, Ancient lateral gene transfer in the evolution of *Bdellovibrio bacteriovorus*, Trends Microbiol. 4 (2006) 64–69.
[5] F. Wright, The effective number of codons used in a gene, Gene 87 (1990) 23–29.
[6] P.M. Sharp, W.H. Li, The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295.
[7] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[8] J.R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res. 22 (1994) 3174–3180.
[9] S. Das, S. Ghosh, A. Pan, C. Dutta, Compositional variation in bacterial genes and proteins with potential expression level, FEBS Lett. 579 (2005) 5205–5210.
[10] S. Karlin, J. Mrázek, Predicted highly expressed genes of diverse prokaryotic genomes, J. Bacteriol. 182 (2000) 5238–5250.
[11] S. Karlin, J. Mrázek, Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 5240–5245.
[12] S. Karlin, J. Mrázek, A. Cambell, D. Kaiser, Characterization of highly expressed genes of four fast-growing bacteria, J. Bacteriol. 183 (2001) 5025–5040.
[13] W. Gish, D.J. States, Identification of protein coding regions by database similarity search, Nat. Genet. 3 (1993) 266–272.
[14] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[15] K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.
[16] A. Pan, C. Dutta, J. Das, Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias, Gene 215 (1998) 405–413.
[17] F. Alvarez, C. Robello, M. Vignali, Evolution of codon usage and base contents in kinetoplastid protozoans, Mol. Biol. Evol. 11 (1994) 790–802.
[18] S. Das, S. Paul, S. Chatterjee, C. Dutta, Codon and amino acid usage in two major human pathogens of genus *Bartonella*—optimization between replication–transcriptional selection, translational control and cost minimization, DNA Res. 12 (2005) 79–90.
[19] J.O. McInerney, Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 10698–10703.
[20] S. Das, S. Paul, C. Dutta, Evolutionary constraints on codon and amino acid usage in two strains of human pathogenic actinobacteria *Tropheryma whipplei*, J. Mol. Evol. 62 (2006) 645–658.
[21] S. Das, S. Paul, C. Dutta, Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydropathy, Virus Res. 117 (2006) 227–236.
[22] F. Gao, C.T. Zhang, Ori-Finder: a web-based system for finding *oriCs* in unannotated bacterial genomes, BMC Bioinforma. 9 (2008) 1–6.
[23] J.R. Lobry, Origin of replication of *Mycoplasma genitalium*, Science 272 (1996) 745–746.
[24] C.X. Weichenberger, M.J. Sippl, NQ-Flipper: recognition and correction of erroneous asparagine and glutamine side-chain rotamers in protein structures, Nucleic Acids Res. 35 (2007) W403–W406 (Web Server issue).
[25] G. Barel, A. Sirota, H. Volpin, E. Jurkevitch, Fate of predator and prey proteins during growth of *Bdellovibrio bacteriovorus* on *Escherichia coli* and *Pseudomonas syringae* prey, J. Bacteriol. 187 (2005) 329–335.
[26] R. Linding, R.B. Russell, V. Neduva, T.J. Gibson, GlobPlot: exploring protein sequences for globularity and disorder, Nucleic Acids Res. 31 (2003) 3701–3708.
[27] C.R. Kissinger, A.K. Dunker, E. Shakhnovich, Disorder in protein structure and function, Pac. Symp. Biocomput. 4 (1999) 517–519.