Horizontal Gene Transfer in Glycosyl Hydrolases Inferred from Codon Usage in *Escherichia coli* and *Bacillus subtilis*

Santiago Garcia-Vallvé, Jaume Palau, and Antoni Romeu

Department of Biochemistry and Biotechnology, Rovira i Virgili University, Catalonia, Spain

Glycosyl hydrolase (GH) genes from *Escherichia coli* and *Bacillus subtilis* were used to search for cases of horizontal gene transfer. Such an event was inferred by G+C content, codon usage analysis, and a phylogenetic congruency test. The codon usage analysis used is a procedure based on a distance derived from a Pearson linear correlation coefficient determined from a pairwise codon usage comparison. The distances are then used to generate a distance-based tree with which we can define clusters and rapidly compare codon usage. Three genes (*yagH* from *E. coli* and *xynA* and *xynB* from *B. subtilis*) were determined to have arrived by horizontal gene transfer and were located in *E. coli* CP4-6 prophage, and *B. subtilis* prophages 6 and 5, respectively. In this study, we demonstrate that with codon usage analysis, the proposed horizontally transferred genes can be distinguished from highly expressed genes.

Introduction

O-glycosyl hydrolases (EC 3.2.1.–) are a large and diverse family of enzymes which hydrolyze the glycosidic bond between two or more carbohydrates or between a carbohydrate and a noncarbohydrate moiety (Davies and Henrissat 1995). The most extensive glycosyl hydrolase (GH) studies have been made on various prokaryotes (Moracci et al. 1994; Heinemann et al. 1996) and low eukaryotes (Stalbrand et al. 1995; Skory, Freer, and Bothast 1996) because of their ability to secrete considerable amounts of these industrially important enzymes (Teeri 1997; Bauer, Driskill, and Kelly 1998). The on-line release document (http://www.expasy.ch/cgi-bin/lists?glycosid.txt) provides an updated list of 70 GH families (Henrissat and Bairoch 1996).

GHs are present in all kinds of living organisms: eubacteria, archea, fungi, plants, and animals; but some particular GH subsets, such as cellulases, xylanases, amylases, and sialidases, are more representative of a particular kind. Thus, the biochemical diversity of these enzymes, together with their phylogenetic relationships, makes GH interesting for evolution studies. The possible role of horizontal gene transfer is one of the most debated questions in the field of molecular evolution (Smith, Feng, and Doolittle 1992; Syvanen 1994). According to the most striking conclusions of several studies which have compared multiple complete genomes from phylogenetically distant species, the number of universally conserved gene families is very small, and multiple events of horizontal gene transfer and genome fusion are major forces in evolution (Koonin and Galperin 1997). In the GH superfamily, some cases of horizontal transfers among genes have been reported (Grabnitz et al. 1989; Gilbert et al. 1992; Roggentin et al. 1993; Zhou et al. 1994).

In eubacteria, the mean genomic G+C content varies from 25% to 75% and is related to phylogeny (Osawa et al. 1992). This suggests that directional evolutionary pressure (constraint) has determined the specific G+C content in each phylogenetic line (Hori and Osawa 1987). Moreover, the relation between the G+C content of DNA and the amino acid composition of total protein in a wide variety of bacterial species has been described (Sueoka 1961). In addition, alternative synonymous codons are generally not used with equal frequency. It is apparent that genes from one species often share similarities in codon frequency, and under the genome hypothesis (Grantham et al. 1980, 1981), there is a species-specific pattern to codon usage (Sharp et al. 1988).

This report is an attempt to apply genomic analysis to a wide variety of prokaryotic GH genes from *E. coli* and *B. subtilis.* The evolutionary relatedness of GH codon usage and G+C content are compared with the defined clusters of the gene tree and the species tree. We used the Pearson linear correlation coefficient to test the similarity of GH codon usage patterns and thus characterize the role of horizontal gene transfer in the evolution of this protein superfamily. In this sense, the phylogenetic relationships of the GH system can form an experimental model from which speculations can be made about evolution.

Materials and Methods

Genes were imported from the EMBL database. The EMBL accession numbers of the GH sequences of *E. coli* K-12 and *B. subtilis* strain 168 that were analyzed are summarized in table 1, along with the codes used in this study, gene names and genome positions, descriptions of the coded proteins, the GH families, and the numbers of base pairs. Gene positions were taken from on-line releases available in current *E. coli* (http://www.pasteur.fr/Bio/Colibri.html) and *B. subtilis* (Moszer, Glaser, and Danchin 1995) (http://www.pasteur.fr/Bio/SubtiList.html) complete genome internet servers. In the *E. coli* GH set, the genes *bglA, bglB, celF, ebgA,*

Abbreviations: CAI, codon adaptation index; GH, glycosyl hydrolase.

Key words: glycosyl hydrolases, horizontal gene transfer, codon usage, Pearson linear correlation coefficient, *Escherichia coli, Bacillus subtilis,* complete genome.

Address for correspondence and reprints: A. Romeu, Facultat Química, Departamento Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Pl. Imperial Tàrraco, 1, E-43005 Tarragona, Catalonia, Spain. E-mail: romeu@quimica.urv.es.

Mol. Biol. Evol. 16(9):1125–1134. 1999

Table 1
Bacterial Glycosyl Hydrolases Used

Code	Gene Name	Position (kb)	Product	Family	Accession No.	Length (bp)
Bacillus subtilis						
bs1	*yckE*	369.8	Similar to beta-glucosidases	1	D30762	1,434
bs2	*ydhP*	628.3	Similar to beta-glucosidases	1	D88802	1,398
bs3	*bglH*	4032.6	Beta-glucosidase (3.2.1.21)	1	D31856	1,410
bs4	*bglA*	4121.8	6-phospho-beta-glucosidase (3.2.1.86)	1	L19710	1,440
bs5	*ybbD*	188.4	Similar to beta-hexosaminidase	3	L19954	1,929
bs6	*lplD*	782.5	Hydrolytic enzyme	4	L19165	1,341
bs7	*glvA*	889.5	6-phospho-beta-glucosidase (3.2.1.86)	4	D50543	1,350
bs8	*melA*	3100.0	Alpha-galactosidase (3.2.1.22)	4	Z99119	1,299
bs9	*licH*	3958.6	6-phospho-beta-glucosidase (3.2.1.86)	4	Z49992	1,329
bs10 ...	*bglC*	1939.9	Endo-1,4-beta-glucanase (3.2.1.4)	5	Z20076	1,500
bs11 ...	*xynA*	2054.5	Endo-1,4-beta-xylanase (3.2.1.8); belongs to prophage 6	11	M36648	642
bs12 ...	*ycdG*	306.0	Similar to oligo-1,6-glucosidases	13	AB000617	1,686
bs13 ...	*amyE*	327.2	Alpha-amylase (3.2.1.1)	13	J01547	1,983
bs14 ...	*treA*	851.4	Trehalose-6-phosphate hydrolase (3.2.1.93)	13	X80203	1,686
bs15 ...	*amyX*	3062.9	Pullulanase	13	Z99119	2,157
bs16 ...	*glgB*	3170.7	1,4-alpha-glucan branching enzyme (2.4.1.18)	13	Z25795	1,884
bs17 ...	*yugT*	3215.1	Similar to exo-alpha-1,4-glucosidases	13	Z99120	1,665
bs18 ...	*yvdL*	3548.3	Similar to oligo-1,6-glucosidases	13	Z94043	1,686
bs19 ...	*yvdF*	3556.7	Similar to glucan 1,4-alpha-maltohydrolases	13	Z99121	1,770
bs20 ...	*bglS*	4011.4	Endo-beta-1,3-1,4 glucanase (3.2.1.73)	16	X00754	729
bs21 ...	*ydhT*	632.5	Beta-mannanase (3.2.1.25)	26	D37964	1,089
bs22 ...	*sacC*	2759.3	Levanase (3.2.1.65)	32	Y00485	2,034
bs23 ...	*yveB*	3536.6	Similar to levanases	32	Z94043	1,551
bs24 ...	*sacA*	3902.4	Sucrase-6-phosphate hydrolase (3.2.1.26)	32	M15662	1,443
bs25 ...	*yesZ*	774.3	Similar to beta-galactosidases	42	Z99107	1,992
bs26 ...	*lacA*	3504.1	Beta-galactosidase (3.2.1.23)	42	Z99121	2,064
bs27 ...	*xynB*	1888.0	Xylan beta-1,4-xylosidase; belongs to prophage 5	43	Z99113	1,602
bs28 ...	*csn*	2748.1	Chitosanase (3.2.1.132)	46	X92868	834
bs29 ...	*xsa*	2914.3	Beta-xylosidase and alpha-L-arabinosidase	51	Z75208	1,488
bs30 ...	*abfA*	2938.9	Alpha-L-arabinofuranosidase (3.2.1.55)	51	Z75208	1,503
bs31 ...	*yvfO*	3502.0	Similar to endo-1,4-beta-galactosidases	53	Z99121	1,278
bs32 ...	*yhfE*	1094.5	Similar to endo-1,4-glucanases	60	Y14083	1,041
bs33 ...	*ysdC*	2950.4	Similar to endo-1,4-glucanases	60	Z99118	1,086
bs34 ...	*ytoP*	3055.4	Similar to endo-1,4-glucanases	60	Z99119	1,074
bs35 ...	*yvdK*	3550.6	Similar to trehalases	65	Z94043	2,274
Escherichia coli						
ec1	*ascB*	2839.0	6-phospho-beta-glucosidase (3.2.1.86)	1	M73326	1,425
ec2	*bglA*	3041.7	6-phospho-beta-glucosidase (3.2.1.86)	1	AE000373	1,440
ec3	*bglB*	3901.3	Phospho-beta-glucosidase (3.2.1.86)	1	AE000449	1,413
ec4	*ebgA*	3220.2	Phospho-beta-D-galactosidase	2	M64441	3,096
ec5	*lacZ*	365.5	Beta-galactosidase (3.2.1.23)	2	J01636	3,072
ec6	*uidA*	1694.1	Beta-D-glucuronidase (3.2.1.31)	2	S69414	1,812
ec7	*bglX*	2220.0	Periplasmic beta-glucosidase (3.2.1.21)	3	U15049	2,298
ec8	*melA*	4339.5	Alpha-galactosidase (3.2.1.22)	4	X04894	1,356
ec9	*celF*	1816.5	Phospho-beta-glucosidase (3.2.1.86)	4	AE000268	1,353
ec10 ...	*yhjM*	3687.9	Similar to endoglucanases	8	U00039	1,107
ec11 ...	*malS*	3735.1	Alpha-amylase (3.2.1.1)	13	X58994	2,031
ec12 ...	*amyA*	2004.2	Cytoplasmic alpha-amylase (3.2.1.1)	13	L01642	1,488
ec13 ...	*glgB*	3571.1	1,4-alpha-glucan branching enzyme (2.4.1.18)	13	M13751	2,187
ec14 ...	*glgX*	3569.0	Glycogen operon protein	13	AE000419	1,974
ec15 ...	*malZ*	421.7	Maltodextrin glucosidase (3.2.1.20)	13	AE000147	1,818
ec16 ...	*treC*	4462.3	Trehalose-6-phosphate hydrolase (3.2.1.93)	13	U06195	1,656
ec17 ...	*yegX*	2181.7	Similar to lysozimes	25	AE000299	828
ec18 ...	*cscA*		Sucrose-6-phosphate hydrolase (3.2.1.26) strain EC3132	32	X81461	1,434
ec19 ...	*rafD*		Raffinose invertase (3.2.1.26) from plasmid PRSD2	32	M27273	1,431
ec20 ...	*rafA*		Alpha-galactosidase (3.2.1.22) from plasmid PRSD2	36	M27273	2,127
ec21 ...	*treA*	1246.6	Periplasmic trehalase (3.2.1.28)	37	X15868	1,698
ec22 ...	*treF*	3667.2	Probable cytoplasmic trehalase	37	U00039	1,650
ec23 ...	*yagH*	286.0	Similar to betaxylosidases; belong to the CP4-6 prophage	43	AE000135	1,611
ec24 ...	*frvX*	4088.5	Similar to endo-1,4-glucanases	60	L19201	1,071
ec25 ...	*sgcX*	4529.2	Similar to endo-1,4-glucanases	60	U14003	1,152
ec26 ...	*ycjT*	1375.9	Similar to trehalases	65	AE000229	2,268

and *ebgC* are described as cryptic genes; that is, they cannot be expressed and so belong to a part of the genome that seems to be silent, carried along from generation to generation without expression (Riley 1993).

The first multivariate codon usage analysis was carried out by Grantham et al. (1980). This procedure uses cluster analysis between absolute codon frequencies, or between relative values of synonymous codon usage, and considers protein-coding sequences as points in a 61-dimensional space that are finally projected into 2-dimensional space with a minimum loss of information. Finally, a general clustering method is necessary to define groups or classes without any a priori knowledge of the system. In this paper, instead of a multivariate codon usage analysis, we used the Pearson linear correlation coefficient (*r*) (Snedecor and Cochran 1995) between absolute codon frequencies. As an arbitrary measure of the pairwise distance (*D*) between each pair of genes on the basis of the similarity of their codon usage values, we used the following linear transformation of *r*: $D = (1 - r) \times 100$. Distances were located in a triangular matrix, and the output was usable as input for the GROWTREE program of the GCG sequence analysis software package (Genetics Computer Group, University of Wisconsin). Thus, it was possible to draw a tree using the UPGMA method based on these distances and show the corresponding clusters. Trees generated by this method do not necessarily reflect evolutionary histories; rather, they are clustered by virtue of similarity of codon usage. This method is an automated procedure to conveniently identify genes with a codon usage pattern that differs from that of the majority of genes. A similar mathematical approach has been used by Nesti et al. (1995) in a phylogenetic study inferred from codon usage patterns in 31 organisms from main taxa.

To distinguish potential horizontal transfers from codon usage differences associated with differences in expression levels, a set of highly expressed genes was defined in both organisms, and the 25 genes with highest CAI values (Sharp and Li 1987) that were longer than 300 bp were collected. Individual CAI values for all of the *E. coli* and *B. subtilis* genes were taken from the *E. coli* Genome Center (http://www.genetics.wisc.edu) and the NRSub database (Perriere, Gouy, and Gojobori 1998), respectively. The *E. coli* set of highly expressed genes included some of the genes defined as very highly expressed genes (Sharp and Li 1986) and included alkyl hydroperoxide reductase (*ahpC*); chaperones Hsp60 (*mopA*) and Hsp70 (*dnaK*); elongation factors G (*fusA*), Tu (*tufA* and *tufB*), and Ts (*tsf*); enolase (*eno*); formate acetyltransferase (*pflB*); fructose-bisphosphate aldolase (*fba*); glyceraldehyde-3-phosphate dehydrogenase (*gapA*); outer membrane proteins 3a (*ompA*), 1b (*ompC*), and X (*ompX*); phosphoglycerate kinase (*pgk*); ribosomal proteins (*rplA, rplC, rplD, rplI, rpsA, rpsB, rpsC,* and *rpsI*); superoxide dismutase (*sodA*); trigger factor, a molecular chaperone involved in cell division (*tig*), and triosephosphate isomerase (*tpiA*). All of the *B. subtilis* genes in the set of highly expressed genes belonged to *B. subtilis* class 2 (Moszer 1998) and included alkyl hydroperoxide reductase small subunit (*ahpC*); elongation factors G (*fus*), Tu (*tufA*), and Ts (*tsf*); enolase (*eno*); fructose-1,6-bisphosphate aldolase (*fbaA*); glyceraldehyde-3-phosphate dehydrogenase (*gap*); ribosomal proteins (*rplA, rplB, rplL, rplM, rplN, rplP, rplQ, rplT, rplU, rplX, rpsB, rpsD, rpsG, rpsI,* and *rpsM*); trigger factor (prolyl isomerase) (*tig*), and unknown-function proteins (*yocJ* and *yaaK*).

To apply the phylogenetic congruency test to investigation of the possible horizontal transfer of the *B. subtilis xynA* gene, we defined a set of ortholog sequences for this gene collecting all the bacterial GHs in family 11 defined in the CAZy (Carbohydrate-Active enZYmes) server (http://afmb.cnrs-mrs.fr/~pedro/CAZY). These ortholog sequences include *xynA* from *Aeromonas caviae* (with SwissProt/TrEmbl accesion number Q43993); *xlnA* from *Bacillus circulans* (P09850); *xynA* from *Bacillus pumilus* (P00694); *xynY* from *Bacillus* sp. (Q59257); *xynA* from *Bacillus stearothermophilus* (P45705); *xynA* from *Bacillus subtilis* (P18429); *xynD* from *Caldicellulosiruptor* sp. (O52375); *xynD* from *Cellulomonas fimi* (P54865); *xyn* from *Cellvibrio mixtus* (Q59300); *xynA* from *Clostridium acetobutylicum* (P17137); *xynA* from *Clostridium stercorarium* (P33558); *xynU* and *xynV* from *Clostridium thermocellum* (O52780 and O52779); *xynB* from *Dictyoglomus thermophilum* (P77853); *xynE* from *Pseudomonas fluorescens* (Q59674); *xynA* from *Ruminococcus albus* (Q52644); *xynA, xynB,* and *xynD* from *Ruminococcus flavefaciens* (P29126, Q52753, and Q53317); *xyn1* from *Ruminococcus* sp. (Q59790); *xynB* and *xynC* from *Streptomyces lividans* (P26515 and P26220); *xyn* and *xyl1* from *Streptomyces* sp. (Q56013 and Q59962); *stxiI* from *Streptomyces thermoviolaceus* (O08346); and *tfxA* from *Thermomonospora fusca* (Q56265). The sequence alignment of orthologs and phylogenetic analysis were performed using the CLUSTAL W software package (Jeanmougin et al. 1998). Bootstrap values were calculated in 1,000 replicates.

Results

The total G+C contents of *E. coli* and *B. subtilis* GH genes were 52.6 ± 3.2% and 45.5 ± 2.7%, respectively, which correlates well with the average of the total G+C genomic contents of these organisms; that is, 50.8% for *E. coli* (Blattner et al. 1997) and 43.5% for *B. subtilis* (Kunst et al. 1997). In addition, the G+C content at the different codon positions was as follows: *E. coli*— first, 59.1 ± 2.6%; second, 40.6 ± 3.0%; third, 58.2 ± 7.9%; and *B. subtilis*—first 51.7 ± 3.9%; second, 37.7 ± 3.9%, third, 47.1 ± 5.8%. The statistical data are shown in figure 1. The genes which display G+C content values significantly different from the mean value are highlighted. In terms of G+C content, statistical differences were found for the *E. coli yagH* gene, which had a total value of 64.4% and a third-codon position value of 87.9%; for *B. subtilis* GH genes, differences were detected in *xynB* (total: 37.5%; third codon position: 27.3%) and in *xynA* (first codon position: 38.3%; second codon position: 52.8%).

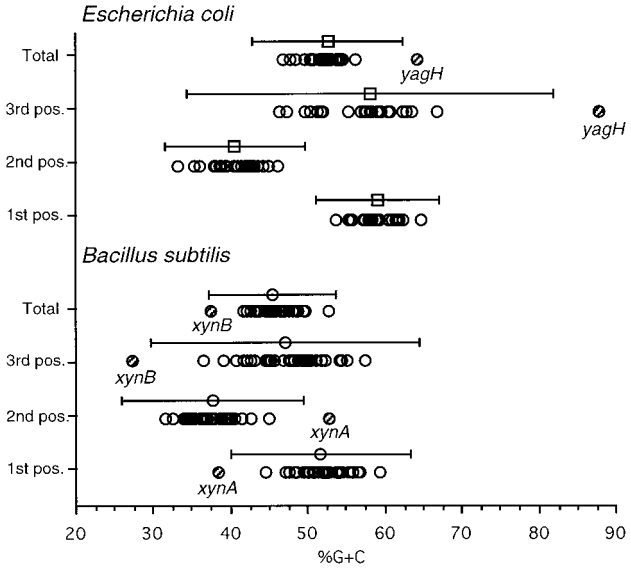

FIG. 1.—Total and codon-position G+C content of the bacterial GH genes. The data consist of sequences described in table 1. The horizontal lines show the mean value ± 3 SD. Those points whose G+C contents differ from the mean by more than three times the standard deviation are shaded and called by their respective gene names.

Figure 2 shows the dendrogram of *E. coli* and *B. subtilis* GH genes based on the matrix of correlation distances described in *Materials and Methods.* It is significant that genes of both species cluster separately, except for three *E. coli* genes (*yagH, yegX,* and *sgcX*) and three *B. subtilis* genes (*xynA, xynB, bglS*) that cluster anomalously. Interestingly, the *E. coli yagH* gene (whose encoded protein is a β-xylosidase from GH family 43) belongs to the cryptic prophage which Blattner et al. (1997) named CP4-6 (positions 262122–296489). This prophage includes the *argF* gene, a known duplicate gene in the arginine biosynthesis pathway that was acquired through a transduction event (Van Vliet, Boyen, and Glansdorff 1988). The prophage also includes the IS911A complex, a partial IS30 copy, two copies of IS1, and one copy of IS5 (Blattner et al. 1997). The *B. subtilis xynA* and *xynB* GH genes also belong to prophages 6 (positions 2046237–2077878) and 5 (positions 1879331–1899765), respectively. Table 2 shows the different patterns of codon usage for *E. coli* and *B. subtilis* GH and prophage genes.

Figure 3 shows the dendrogram based on the Pearson linear correlation constructed with *E. coli* GH genes, together with genes from the above-mentioned prophage CP4-6 and the set of *E. coli* highly expressed genes. The three *E. coli* genes (*yagH, yegX,* and *sgcX*) which deviate in figure 2 cluster anomalously and are distinguishable from *E. coli* GH genes and from the set of highly expressed genes. The *yagH* gene clusters with the viral genes, and the *yegX* gene (whose encoded protein is a GH member of 25 that includes lysozymes) and the *sgcX* gene are both located in a genome region near a prophage insertion site. In figure 3, it is also worth noting that genes *cscA, rafD,* and *rafA,* which do not belong to the *E. coli* K-12 chromosome, cluster in a separate branch. The *rafD* and *rafA* genes are involved in plasmids, but the *cscA* gene is present in the *E. coli* strain EC3132 chromosome.

Results of applying the Pearson correlations to *B. subtilis* codon usage are shown in figure 4. The three *B. subtilis* genes (*xynA, xynB,* and *bglS*) that deviate in fig-

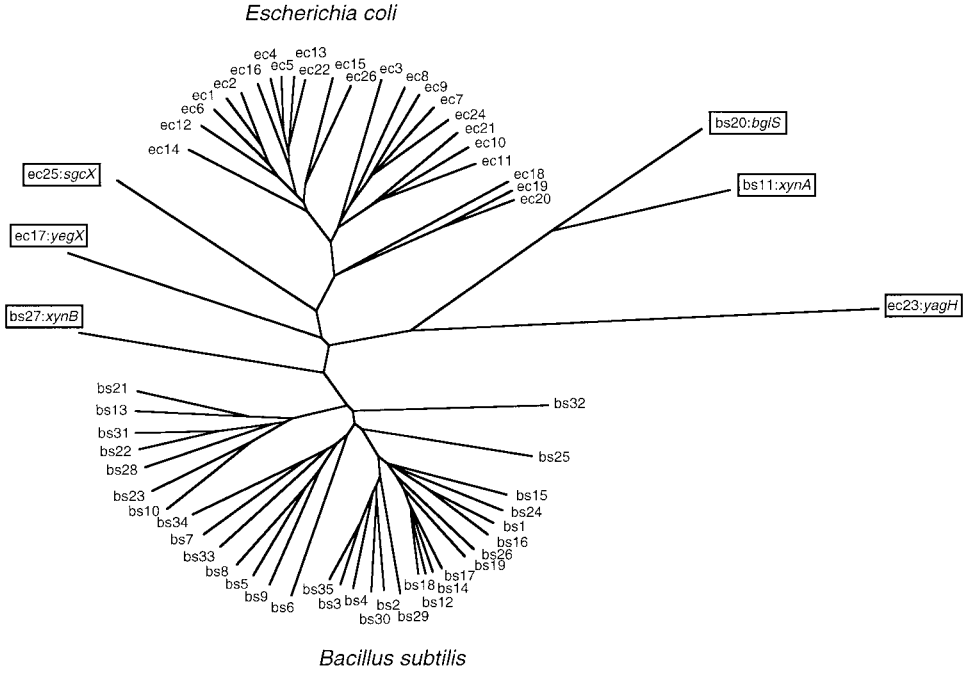

FIG. 2.—Dendrogram of the relationships between the codon usage of bacterial GHs constructed from the matrix of correlation distances by the program GROWTREE from the GCG sequence analysis software package, using the UPGMA method (see *Materials and Methods*). The majority of *E. coli* and *B. subtilis* genes cluster separately. However, the genes that cluster anomalously are called by their names and boxed.

Table 2
Codon Usage for Bacterial Glycosyl Hydrolase and Prophage Genes

Amino acid	Codon	*Bacillus subtilis*	*Escherichia coli*	Prophage CP4-6 from *E. coli*	Prophages 5 and 6 from *B. subtilis*
A	GCT.....	15.6 (4.8)	10.6 (4.6)	2.8 (2.1)	19.8 (8.3)
A	GCC.....	15.7 (7.4)	27.4 (9.0)	36.9 (16.7)	11.8 (9.0)
A	GCG.....	19.6 (7.1)	26.9 (8.7)	45.2 (22.6)	8.7 (6.7)
A	GCA.....	19.5 (7.7)	15.4 (6.0)	5.9 (5.3)	24.7 (11.9)
C	TGT.....	2.6 (2.3)	5.8 (2.9)	1.8 (2.4)	4.9 (4.4)
C	TGC.....	5.1 (4.5)	6.5 (4.7)	11.1 (5.7)	3.5 (3.6)
D	GAT.....	41.1 (10.0)	42.0 (7.6)	14.4 (8.1)	42.3 (12.1)
D	GAC.....	25.2 (8.5)	23.5 (6.8)	38.6 (10.6)	14.7 (9.5)
E	GAG.....	24.2 (8.5)	20.7 (6.4)	34.7 (14.5)	18.9 (5.8)
E	GAA.....	44.3 (10.5)	42.3 (9.2)	17.1 (9.7)	54.7 (15.7)
F	TTT.....	30.4 (11.7)	24.7 (6.7)	12.1 (4.6)	36.9 (11.0)
F	TTC.....	13.0 (6.6)	18.2 (4.8)	27.0 (15.9)	12.7 (3.9)
G	GGT.....	9.7 (4.7)	21.1 (5.6)	6.4 (5.1)	15.6 (6.5)
G	GGC.....	28.6 (8.9)	33.9 (9.4)	56.9 (13.4)	15.7 (13.9)
G	GGG.....	12.9 (6.6)	13.6 (5.5)	19.5 (5.2)	11.1 (5.9)
G	GGA.....	22.4 (6.0)	8.5 (5.5)	4.9 (5.4)	23.0 (16.5)
H	CAT.....	19.6 (9.1)	18.6 (8.1)	8.8 (6.2)	20.2 (8.8)
H	CAC.....	9.7 (4.2)	13.4 (5.9)	21.1 (8.4)	6.4 (4.4)
I	ATT.....	29.4 (10.1)	24.3 (11.1)	16.8 (5.6)	38.4 (10.9)
I	ATC.....	23.6 (8.4)	18.7 (6.6)	33.4 (15.4)	16.2 (6.7)
I	ATA.....	7.1 (4.3)	3.0 (3.0)	0.7 (1.3)	21.1 (13.3)
K	AAG.....	20.0 (7.2)	8.7 (4.8)	20.0 (10.5)	18.0 (15.8)
K	AAA.....	44.8 (12.9)	32.2 (13.8)	20.9 (9.2)	52.5 (19.0)
L	TTG.....	13.2 (5.0)	11.1 (5.1)	4.0 (2.4)	13.0 (5.8)
L	TTA.....	13.3 (5.5)	12.1 (7.3)	3.4 (1.8)	35.3 (15.8)
L	CTT.....	16.9 (7.4)	8.3 (5.8)	6.4 (3.7)	18.2 (7.6)
L	CTC.....	9.1 (5.4)	9.5 (4.7)	21.1 (7.3)	4.7 (3.4)
L	CTG.....	18.5 (6.2)	43.9 (10.6)	72.0 (13.1)	8.8 (7.5)
L	CTA.....	3.2 (2.6)	3.7 (2.9)	0.8 (1.5)	9.1 (6.9)
M	ATG.....	26.9 (7.8)	27.7 (6.4)	26.5 (11.5)	22.4 (9.9)
N	AAT.....	28.9 (9.6)	17.5 (5.5)	4.0 (2.9)	41.6 (12.3)
N	AAC.....	22.6 (8.0)	25.8 (9.6)	22.0 (7.1)	18.2 (4.8)
P	CCT.....	10.2 (4.1)	8.1 (4.2)	2.7 (3.2)	10.4 (3.9)
P	CCC.....	3.8 (3.3)	7.5 (4.5)	5.5 (3.9)	3.5 (3.2)
P	CCG.....	21.0 (7.1)	25.2 (9.4)	38.5 (16.0)	7.3 (6.5)
P	CCA.....	8.4 (4.2)	9.7 (4.4)	0.7 (1.4)	10.3 (6.6)
Q	CAG.....	18.9 (7.0)	29.4 (7.8)	39.0 (10.3)	10.2 (5.0)
Q	CAA.....	16.5 (5.8)	15.2 (8.7)	2.8 (3.5)	26.3 (3.7)
R	CGT.....	6.9 (3.8)	20.7 (7.8)	8.0 (6.1)	6.8 (6.4)
R	CGC.....	10.4 (5.6)	24.3 (7.6)	33.4 (11.2)	3.7 (4.2)
R	CGG.....	7.4 (3.8)	5.6 (3.0)	11.3 (6.8)	2.6 (2.3)
R	CGA.....	4.3 (3.3)	2.9 (3.1)	0.0 (0.0)	3.7 (3.5)
R	AGG.....	4.5 (3.8)	1.0 (1.7)	0.7 (2.0)	4.6 (5.3)
R	AGA.....	9.7 (4.7)	1.5 (3.1)	0.6 (1.1)	16.3 (9.3)
S	TCT.....	9.4 (5.5)	6.5 (4.2)	2.9 (3.0)	18.2 (8.0)
S	TCC.....	8.0 (3.8)	6.7 (4.2)	13.4 (4.8)	5.7 (6.1)
S	TCG.....	7.4 (4.9)	7.6 (2.7)	16.4 (6.3)	4.5 (3.5)
S	TCA.....	12.4 (4.2)	5.0 (3.7)	3.5 (5.6)	15.6 (8.3)
S	AGT.....	4.9 (3.2)	8.7 (4.4)	1.2 (1.6)	9.0 (3.9)
S	AGC.....	13.7 (5.4)	16.6 (5.9)	20.7 (6.4)	9.5 (4.0)
T	ACT.....	5.5 (3.7)	6.3 (3.4)	4.0 (3.8)	12.4 (5.1)
T	ACC.....	8.6 (5.0)	21.7 (6.6)	42.7 (7.0)	7.4 (6.0)
T	ACG.....	17.7 (7.3)	14.0 (6.0)	18.0 (4.8)	9.6 (7.3)
T	ACA.....	23.1 (7.8)	5.7 (3.9)	0.7 (1.2)	19.5 (8.3)
V	GTT.....	15.4 (5.4)	12.5 (4.6)	4.8 (3.5)	20.1 (7.8)
V	GTC.....	18.3 (9.0)	15.1 (5.8)	14.9 (7.3)	8.2 (4.6)
V	GTG.....	17.5 (5.2)	25.1 (9.5)	41.5 (11.8)	11.2 (8.5)
V	GTA.....	9.8 (5.3)	8.6 (3.5)	2.9 (2.4)	13.6 (5.5)
W	TGG.....	21.9 (9.4)	27.9 (10.1)	18.3 (11.4)	15.8 (9.3)
Y	TAT.....	29.6 (8.7)	21.5 (6.0)	10.7 (6.3)	33.8 (12.0)
Y	TAC.....	16.3 (6.4)	18.2 (6.4)	20.1 (18.0)	14.9 (6.0)
·	TGA.....	0.6 (0.9)	0.4 (0.7)	1.5 (1.8)	0.4 (0.9)
·	TAG.....	0.3 (0.8)	0.1 (0.4)	0.0 (0.0)	0.3 (0.7)
·	TAA.....	1.1 (1.1)	1.3 (1.0)	1.2 (1.1)	1.6 (1.3)

NOTE.—Values are mean frequencies per thousand (SD). The data for *B. subtilis* and *E. coli* groups consist of genes described in table 1. The ''outlier'' genes (see fig. 2) *xynA, bglS,* and *xynB* from *B. subtilis* and *sgcX, yegX,* and *yagH* from *E. coli* were not included. The *E. coli* and *B. subtilis* prophage groups consist of 7 and 11 genes (described in fig. 3) and include the *yagH* and *xynB* genes, respectively.

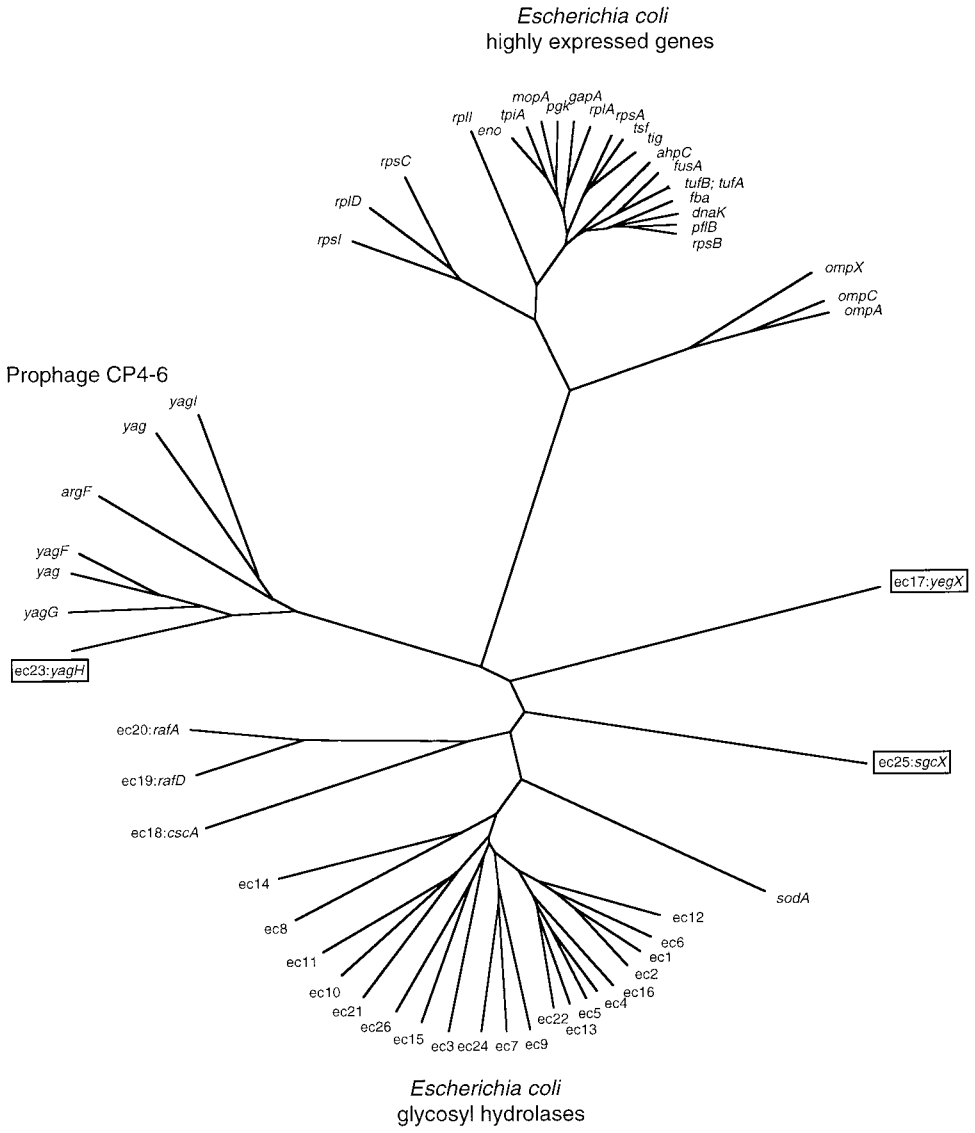

FIG. 3.—Dendrogram of the relationships between the codon usages of *E. coli* GH genes, prophage CP4-6 genes, and the set of highly expressed genes. The dendrogram was constructed from the matrix of correlation distances by the program GROWTREE from the GCG sequence analysis software package, using the UPGMA method (see *Materials and Methods*). Genes that cluster anomalously are called by their names and boxed.

ure 2 cluster anomalously in this figure and can be distinguished from *B. subtilis* GH genes and the set of highly expressed genes. The prophage 5 *xynB* gene, which clusters with the viral genes, encodes a β-xylosidase from GH family 43 (the similarity of this encoded protein and the encoded *E. coli yagH* xylosidase was 53.8%). However, the endo-1,4-β-D-xylan xylanohydrolase (EC 3.2.1.8) *xynA* gene, which belongs to prophage 6, clusters anomalously. We observed that the *B. subtilis xynA* gene is highly similar to the *B. circulans xlnA* gene. The Pearson correlation coefficient between these genes was found to be 0.996, and the index of similarity was 99.5%, which corresponds to one single amino acid change in 213 residues. The above observation was particularly interesting and, consequently, in the *B. subtilis* codon usage correlation experiments, we included a group of *B. circulans* GH genes. The *B. subtilis xynA* and *B. circulans xlnA* genes cluster both together and anomalously in figure 4. G+C content data of both the *xynA* and the *xlnA* genes agree with this observation. The G+C content of the *xynA* gene differs from the mean G+C content of the *B. subtilis* GH genes as shown in figure 1, while the *xlnA* gene G+C content (total: 42.4%; first codon position: 38.3%; second codon position: 52.8%; third codon position: 35.9%) differ from the mean G+C content of the *B. circulans* GH genes (total, 50.3 ± 4.4; first codon position, 48.0 ± 4.8%; second codon position, 44.8 ± 4.1%; third codon position, 58.1 ± 12.2%).

To investigate the close relationship of the DNA sequences of both the *B. subtilis xynA* and the *B. circulans xlnA* genes, we went on to study the flanking regions of the *B. circulans xlnA* xylanase gene. We matched the 1,349-bp DNA sequence (EMBL file

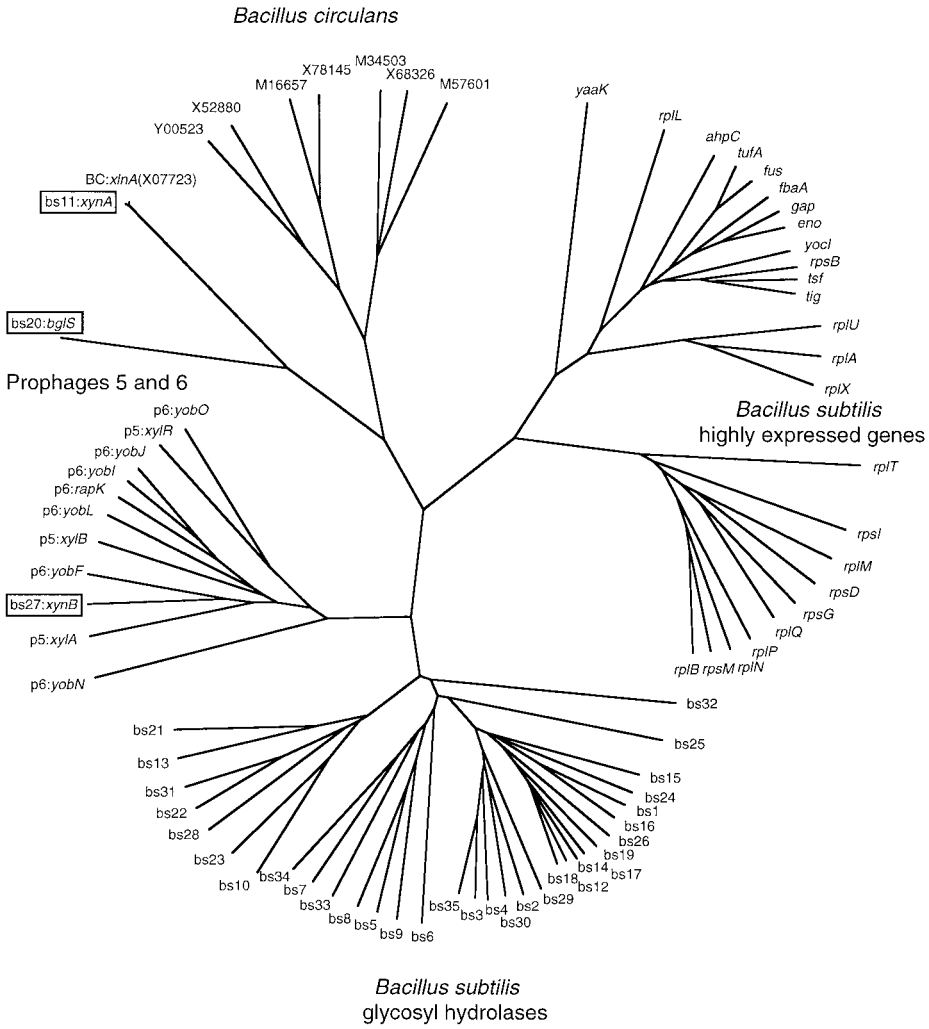

FIG. 4.—Dendrogram of the relationships between the codon usages of *B. subtilis* and *B. circulans* GH genes, prophage 5 and 6 genes, and the set of highly expressed genes. The dendrogram was constructed from the matrix of correlation distances by the program GROWTREE from the GCG sequence analysis software package, using the UPGMA method (see *Materials and Methods*). Genes that cluster anomalously are called by their names and boxed. The EMBL accession numbers of *B. circulans* genes are shown.

X07723) with the complete genome of *B. subtilis*. The *B. circulans xlnA* region is similar to two widely separated fragments of the *B. subtilis* genome. The *B. circulans* DNA fragment between positions 1 and 279 matches the complementary DNA chain of the *B. subtilis ykrQ* gene (which codifies for an unknown-function protein, similar to the two-component sensor histidine kinase) (identity: 99.6%) which is between positions 1420112 and 1420391 of the *B. subtilis* chromosome. Furthermore, the *B. circulans* DNA fragment, between positions 273 and 1346, matches between positions 2053353 and 2054613 of the *B. subtilis* chromosome (identity: 98.2%). This *B. subtilis* DNA region includes the TATA (-10) 5′-flanking region of the *xynA* coding region and a considerable part of the 3′-flanking region. The first 6 bp sequence and the DNA fragment between positions 273 and 279 of the *B. circulans xlnA* sequence are restriction sites of the *Eco*RI restriction enzyme, which was used in the nucleotide analysis of the gene (Yang, Mackenzie, and Narang 1988).

Figure 5 illustrates the phylogenetic congruency test for investigating the possible horizontal transfer of the *B. subtilis xynA* gene. It should be emphasized that the protein tree correlates well with the standard species tree. However, *B. subtilis* and *B. circulans* xylanases (bootstrap: 1,000) cluster anomalously together with two more xylanases (*xynA* from *Bacillus stearothermophilus* and *xynA* from *Aeromonas caviae*) that, respectively, show 75% and 77% similarity at the protein level with both *B. subtilis* and *B. circulans* genes. According to taxonomy, the *Aeromonas* xylanase belongs to the proteobacteria phylogenetic branch. In addition, the branch on which these genes cluster is closer to actinomycetes than the costridial firmicute branch itself.

Discussion

Codon usage of genes which are highly expressed in *E. coli* (Ikemura 1981), *B. subtilis* (Shields and Sharp 1987), and *Saccharomyces cerevisiae* (Sharp and Cowe

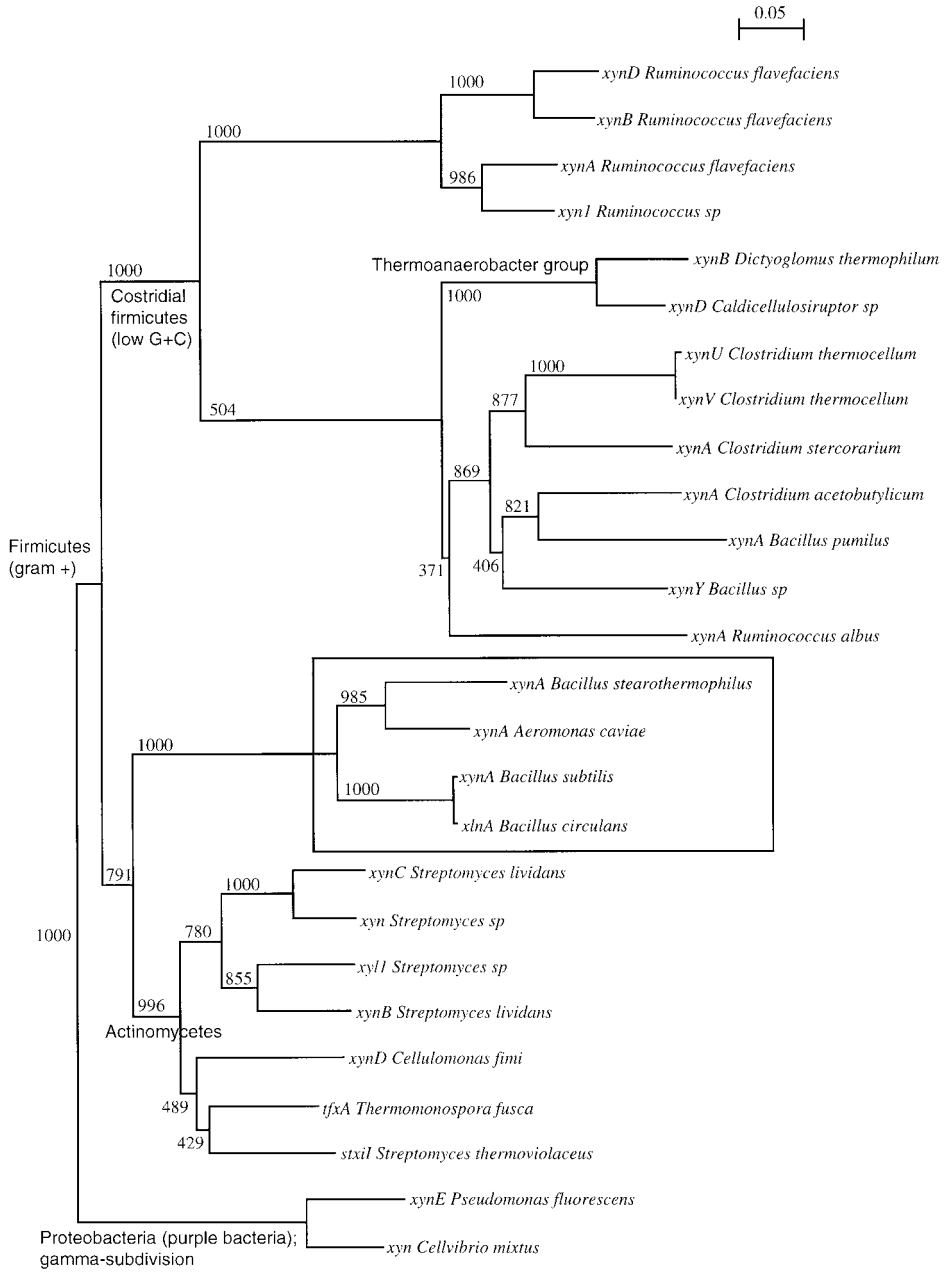

FIG. 5.—Phylogenetic tree of bacterial xylanase family 11, constructed with the CLUSTAL W software package. There were 1,000 bootstrap replicates, and the bootstrap values are indicated above each branch. Operational taxonomic units are referred to by the gene names and species, and the taxonomic classification is shown. The *B. subtilis* and *B. circulans* xylanase genes are shown in a box.

1991) correlate very strongly with the known abundance of iso-accepting tRNAs. The advantage of this translational selection system consists in the use of a codon for which there is a corresponding tRNA that can speed up the process of mRNA translation (Sharp et al. 1993). It has been shown for *E. coli* that genes expressed at high levels have a strong preference for certain codons. However, for genes expressed at low levels, codon usage is more uniform (Ikemura 1981). In this sense, the CAI value measures the extent to which codon usage agrees with a reference set from highly expressed genes (Sharp and Li 1987). The CAI values of the set of *E. coli* and *B. subtilis* GH genes ranged from 0.35 to 0.41 and from 0.35 to 0.49, respectively (individual data not shown), and the codon usage correlation dendrograms clearly distinguish GH genes from the set of highly expressed genes. This suggests that the encoded GH genes are expressed at intermediate levels. Syvanen (1994) concluded that codon usage comparison is a poor procedure for identifying horizontal gene transfers because Médigue et al. (1991) showed clearly that codon usage patterns are influenced by the level of gene expression. In this study, we demonstrate that the proposed horizontally transferred genes can be distinguished from highly expressed genes, and therefore our Pearson correlation routine, which can reveal an anomalous codon usage

pattern in a set of genes, could, in conjunction with G+C content analysis and the congruency phylogenetic test, become a useful tool for detecting horizontal gene transfer events.

We have shown that *E. coli yagH* and *B. subtilis xynA* and *xynB* genes have G+C contents and a codon usage patterns that differ greatly from those of the other GHs of these organisms. Such observations, and the proof of the phylogenetic congruency test, suggest that these genes have been acquired by a horizontal gene transfer. The precise localization of these genes within the *E. coli* and *B. subtilis* complete genomes (Blattner et al. 1997; Kunst et al. 1997) clearly confirms that they have been acquired through a transduction/integration mechanism, since the three genes belong to cryptic prophages. The origins of the *yagH* and *xynB* genes are unknown. More ortholog sequences will be needed to resolve this question, whereas the *xynA* gene could have its origin in an actinomycetes-related organism. This methodology can be applied to other systems to detect horizontal gene transfer events, even if the complete genome of an organism is not available. In this sense, experiments for detecting other cases of horizontal gene transfers in the GH superfamily are in progress in our biocomputing group, especially for organisms living in ecosystems such as the rumen, where cellulose and plant hemicellulose constitute the main raw nutritive substrate.

The strong similarity between the *B. subtilis xynA* and *B. circulans xlnA* genes needs to be mentioned. Both genes are 99.5% similar, both genes differ in G+C content and codon usage from their respective genomes, and *xynA* is integrated into a prophage. Furthermore, both genes are clustered together in the same phylogenetic branch near the actinomycetes group (see fig. 5). The strong similarity between these genes and the fact that the 5′-flanking region of *xlnA* matches two widely separated regions of the *B. subtilis* chromosome may suggest that contamination could have occurred when Yang, Mackenzie, and Narang (1988) sequenced the *xlnA* gene from *B. circulans*. In fact, these authors may have sequenced the *xynA* gene from *B. subtilis.* The boundaries of the matching regions are both *Eco*RI sites for the enzyme used in the nucleotide analysis of the gene (Yang, Mackenzie, and Narang 1988). This suggests that a cloning artifact could have been produced when the *xlnA* gene was sequenced. Further analysis of the *B. circulans* genome will resolve this question.

Acknowledgments

S.G.-V. was the recipient of a fellowship (FI/96-7.030) from the Catalan Governmental Agency CIRIT (Generalitat de Catalunya). We thank the reviewers for valuable suggestions and John Bates and Kevin Costello (from the Language Service of our university) for their help during the writing of this paper. This work has not been awarded grants by any research-supporting institution.

LITERATURE CITED

BAUER, M. W., L. E. DRISKILL, and R. M. KELLY. 1998. Glycosyl hydrolases from hyperthermophilic microorganisms. Curr. Opin. Biotechnol. **9**:141–145.

BLATTNER, F. R., G. PLUNKETT III, C. A. BLOCH et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1462.

DAVIES, G., and B. HENRISSAT. 1995. Structures and mechanisms of glycosyl hydrolases. Structure **3**:853–859.

GILBERT, H. J., G. P. HAZLEWOOD, J. I. LAURIE, C. G. ORPIN, and G. P. XUE. 1992. Homologous catalytic domains in a rumen fungal xylanase: evidence for gene duplication and prokaryotic origin. Mol. Microbiol. **6**:2065–2072.

GRABNITZ, F., K. P. RUCKNAGEL, M. SEISS, and W. L. STAUDENBAUER. 1989. Nucleotide sequence of the *Clostridium thermocellum bglB* gene encoding thermoestable β-glucosidase B: homology to fungal β-glucosidases. Mol. Gen. Genet. **217**:70–76.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE, and R. MERCIER. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9**: r43–r74.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER, and A. PAVÉ. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8**:r49–r62.

HEINEMANN, U., J. AY, O. GAISER, J. J. MULLER, and M. N. PONNUSWAMY. 1996. Enzymology and folding of natural and engineered bacterial beta-glucanases studied by X-ray crystallography. Biol. Chem. **377**:447–454.

HENRISSAT, B., and A. BAIROCH. 1996. Updating the sequence based classification of glycosyl hydrolases. Biochem. J. **316**:695–696.

HORI, H., and S. OSAWA. 1987. Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Mol. Biol. Evol. **4**:445–472.

IKEMURA, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein sequence: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

JEANMOUGIN, F., J. D. THOMPSON, M. GOUY, D. G. HIGGINS, and T. J. GIBSON. 1998. Multiple sequence alignment with Clustal X. Trends Biochem. Sci. **23**:403–405.

KOONIN, E. V., and M. Y. GALPERIN. 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr. Opin. Genet. Dev. **7**:757–763.

KUNST, F., N. OGASAWARA, I. MOSZER et al. (151 co-authors). 1997. The complete genome sequence of the Gram positive bacterium *Bacillus subtilis.* Nature **390**:249–256.

MÉDIGUE, C., T. ROUXEL, P. VIGIER, A. HENAUT, and A. DANCHIN. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222**:851–856.

MORACCI, M., M. CIARAMELLA, R. NUCCI, L. H. PEARL, I. SANDERSON, A. TRINCONE, and M. ROSSI. 1994. Thermostable beta glycosidase from *Sulfolobus solfataricus.* Biocatalysis **11**:89–103.

MOSZER, I. 1998. The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. FEBS Lett. **430**:28–36.

MOSZER, I., P. GLASER, and A. DANCHIN. 1995. SubtiList: a relational database for the *Bacillus subtilis* genome. Microbiology **141**:261–268.

NESTI, G., G. POLI, M. CHICCA, P. AMBROSINO, C. SCAPOLI, and I. BARRAI. 1995. Phylogeny inferred from codon usage pattern in 31 organisms. CABIOS **11**:167–171.

OSAWA, S., T. H. JUKES, K. WATANABE, and A. MUTO. 1992. Recent evidence for evolution of the genetic code. Microbiol. Rev. **56**:229–264.

PERRIERE, G., M. GOUY, and T. GOJOBORI. 1998. The nonredundant *Bacillus subtilis* (NRSub) database: update 1998. Nucleic Acids Res. **26**:60–62.

RILEY, M. 1993. Functions of the gene products of *Escherichia coli.* Microbiol. Rev. **57**:862–952.

ROGGENTIN, P., R. SCHAUER, L. L. HOYER, and E. R. VIMIR. 1993. The sialidase superfamily and its spread horizontal gene transfer. Mol. Microbiol. **9**:915–921.

SHARP, P. M., and E. COWE. 1991. Synonymous codon usage in *Saccharomyces cerevisiae.* Yeast **7**:657–678.

SHARP, P. M., E. COWE, D. G. HIGGINS, D. C. SHIELDS, K. H. WOLFE, and F. WRIGHT. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity. Nucleic Acids Res. **16**:8207–8221.

SHARP, P. M., and W. H. LI. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res. **14**:7737–7749.

———. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

SHARP, P. M., M. STENICO, J. F. PEDEN, and A. T. LLOYD. 1993. Codon usage: mutational bias, translation selection, or both? Biochem. Soc. Trans. **21**:835–841.

SHIELDS, D. C., and P. M. SHARP. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res. **15**:8023–8040.

SKORY, C. D., S. N. FREER, and R. J. BOTHAST. 1996. Properties of an intracellular beta glucosidase purified from the cellobiose fermenting yeast *Candida wickerhamii.* Appl. Microbiol. Biotechnol. **46**:353–359.

SMITH, M. W., D. F. FENG, and R. F. DOOLITTLE. 1992. Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem. Sci. **17**:489–493.

SNEDECOR, G. W., and W. G. COCHRAN. 1995. Statistical Methods. 8th edition. Iowa State University Press, Ames, Iowa.

STALBRAND, H., A. SALOHEIMO, J. VEHMAANPERA, B. HENRISSAT, and M. PENTTILA. 1995. Cloning and expression in *Saccharomyces cerevisiae* of a *Trichoderma reesei* beta mannanase gene containing a cellulose binding domain. Appl. Environ. Microbiol. **61**:1090–1097.

SUEOKA, N. 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc. Natl. Acad. Sci. USA **47**:1141–1149.

SYVANEN, M. 1994. Horizontal gene transfer: evidence and possible consequences. Annu. Rev. Genet. **28**:237–261.

TEERI, T. T. 1997. Crystalline cellulose degradation: new insight into the function of cellobiohydrolases. Trends Biotechnol. **15**:160–167.

VAN VLIET, F., A. BOYEN, and N. GLANSDORFF. 1988. On interspecies gene transfer: the case of the *argF* gene of *Escherichia coli.* Ann. Inst. Pasteur Microbiol. **139**:493–496.

YANG, R. C. A., C. R. MACKENZIE, and S. NARANG. 1988. Nucleotide sequence of a *Bacillus circulans* xylanase gene. Nucleic Acids Res. **16**:7187.

ZHOU, L., G.-P. XUE, C. G. ORPIN, G. W. BLACK, H. J. GILBERT, and G. P. HAZLEWOOD. 1994. Intronless *celB* from the anaerobic fungus *Neocallimastix patriciarum* encodes a modular family A endoglucanase. Biochem. J. **297**:359–364.

MANOLO GOUY, reviewing editor

Accepted May 11, 1999