Diversity, Abundance and Distribution of O-linked Glycosylation Pathway Enzymes in Prokaryotes-A Comparative Genomics Study

Manjeet Kumar and Peetey V. Balaji*

Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India

Abstract

In prokaryotes, the protein protein N- and O-glycosylation pathways (GlyPW) have been experimentally characterised in some of the organisms. Identifying GlyPWs in other prokaryotes is essential to understand the role of glycosylation. Herein we report a BLASTp and a hidden Markov model (HMM)-profile based comparative genomics approach to identify putative O-glycosylation enzymes in completely sequenced prokaryotic genomes using the experimentally characterized O-GlyPW enzymes as query sequences. Homologs for enzymes of all five categories viz., initiation, modification, extension, flippase and oligosaccharyltransferase are found in 128 organisms and no homolog is found for any of these in 52 organisms. A large number of organisms have homologs for all categories except oligosaccharyltransferases, which show high sequence diversity. Thus, O-GlyPW enzyme homologs are widely prevalent. Most of the 128 organisms are proteobacteria and more than half are pathogenic. The pattern of distribution of homologs indicates species- and strain-specific variations and acquisition of homologs by horizontal gene transfer.

Keywords:

Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| DATDH | 2,4-Diacetamido-2,4,6-trideoxy-hexose |
| diNaCac | N.N’-Diacetamido bacillosamine |
| GaT | Galactosyltransferase |
| GlcT | Glucosyltransferase |
| GT | Glycosyltransferase |
| HMM | Hidden Markov model |
| LPS | Lipopolysaccharide |
| MSA | Multiple Sequence Alignment |
| ORF | Open Reading Frame |
| OT | Oligosaccharyltransferase |
| pgl | Protein glycosylation locus |
| PilO | Pilin glycosylation locus |

Introduction

The pathways for the glycosylation of proteins in prokaryotes have been characterized in some of the organisms and this include. These are the O-glycosylation pathways of Neisseria [1-5], Helicobacter pylori [6], Pseudomonas aeruginosa [7], Bacteroides fragilis [8] and Acinetobacter baumannii [9], and the N-glycosylation pathways of Campylobacter jejuni [10-12], Haloflexa volcanii [13] and Methanococcus voltae [14]. In the genus Neisseria, the O-glycosylation pathway (Figure S1) has been delineated in the species gonorrhoeae [1,5], lactamica [15] and meningitidis [2,3]. The enzymes involved in these pathways have been characterized to various extents [1-4,15-19]. Functional characterization of PilO from Neisseria meningitidis and PilO from Pseudomonas aeruginosa has shown that both these enzymes have relaxed glycan specificity and they require the glycan to be translocated to the periplasm [7]. PilO has preference towards short oligosaccharides whereas the range of glycans that PilO can transfer is structurally more diverse. In Neisseria gonorrhoeae and N. meningitidis, the protein O-glycosylation enzymes are clustered and form the pilin glycosylation locus [20]. Pil polymorphism, phase variability and competition among the enzymes for a common substrate may lead to glycoforms [3,17,20] i.e., variants of a glycoprotein which differ from each other only in the nature of attached glycan [21]. For example, strains which possess NsPgL2 have 4-glyceramido-2-acetamido-2,4,6-trideoxy-hexosamine [22].

Enzymes of the prokaryotic O-glycosylation pathways can be grouped into five categories (Figure S1 and Table S1). Category-I includes the initiation enzymes which catalyse the transfer of a saccharide to a lipid molecule. This forms the first step in the assembly of glycans on a lipid-linked carrier. The N-terminal domain of the enzymes NsPgLB and NsPgLB2 are examples for this category of enzymes. Category-II includes modification enzymes which catalyse the modification of simple saccharides. Examples include the enzymes involved in the biosynthesis of DATDH. These are NsPgLD (dehydratase), NsPgLC (aminotransferase) and the C-terminal domains of NsPgLB and NsPgLB2. Category-II includes extension enzymes. These are glycosyltransferases (GTs) which catalyse the transfer of a saccharide from a nucleotide sugar donor substrate to acceptors in different linkages. These enzymes are responsible for the extension and elaboration of the lipid-linked glycan. The enzymes NsPgLA (α-1,3-GaT), NsPgLH (α-1,3-GlCT) and NsPgLE (β-1,4-GaT) are a few examples. Category-IV includes flippases which flip the pre-assembled glycans from the cytosolic side to the periplasmic side. These enzymes can flip the lipid-linked glycan containing 1, 2 or 3 saccharide moieties (Figure S1). Category-V includes oligosaccharyltransferases (OTs) which transfer the pre-assembled glycan from a lipid-1 in Neisseria, pgl denotes pilin glycosylation locus and contains enzymes of the O-glycosylation pathway. In Campylobacter jejuni, pgl denotes protein glycosylation locus and contains enzymes of the N-glycosylation pathway. The enzymes that constitute these pathways are denoted by the letters of the alphabet e.g., PgL, PgLb, and so on. However,
enzymes sharing the same name have different functions in the two pathways e.g., PglC of Campylobacter jejuni is a galactosyltransferase whereas PglC of Neisseria is an acetyltransferase. Hence, in this study, 2 or 3 letter prefixes denoting the genus and species names of organisms are added to names of proteins Table S1 linked carrier to the acceptor protein. Minimally, an organism requires at least one initiator enzyme (Category-I), a flippase (Category-IV) and an OT (Category-V) for O-glycosylation. Enzymes belonging to Category-II and -III determine the final structure of the glycan.

The identification of enzymes and characterization of their substrate specificities is critical to delineate the glycosylation pathways in various prokaryotes. These also help in understanding the role of glycans in processes such as virulence and pathogenesis. GTs are potential drug targets (see, for example, [9]). In addition, their promiscuous substrate specificity in response to variations in the assay conditions is advantageous for in vitro glycan synthesis [23,24]. Experimental approaches for the identification of new GTs include the use of probes derived from the sequences of hitherto characterized GTs [25] and screening cell lysates for activity [26]. The main disadvantage of such approaches is that they are very time-consuming. Computational approaches can help to reduce the time by narrowing down the possible candidate ORFs. Such an approach has indeed been used to identify putative eukaryotic [27], prokaryotic [28] and archaeal [29] GTs and followed by experimental characterization in a few cases (see, for example, [30]). In view of this, the present study was initiated with the objective of identifying the homologs of the enzymes involved in O-glycosylation pathways using a bioinformatics-based comparative genomics approach. In the present study also, a bioinformatics-based comparative genomics approach has been used for the identification of the homologs of the enzymes involved in O-glycosylation pathways. The amino acid sequence of the ORFs has been used as query for all the database searches.

Methods

Enzymes of the O-glycosylation pathway have been characterized from several organisms (Table S1). The amino acid sequences of these enzymes were used as query for searching their homologs. The proteomes of 865 completely sequenced bacterial genomes constituted the target dataset (Table S2). This dataset is the same as that used for searching the homologs of enzymes that are part of the Campylobacter jejuni N-glycosylation pathway [28]. This dataset was used as such to facilitate comparison of the results from the present study with that obtained on N-glycosylation pathways [23,24]. Experimental approaches for the identification of new GTs include the use of probes derived from the sequences of hitherto characterized GTs [25] and screening cell lysates for activity [26]. The main disadvantage of such approaches is that they are very time-consuming. Computational approaches can help to reduce the time by narrowing down the possible candidate ORFs. Such an approach has indeed been used to identify putative eukaryotic [27], prokaryotic [28] and archaeal [29] GTs and followed by experimental characterization in a few cases (see, for example, [30]). In view of this, the present study was initiated with the objective of identifying the homologs of the enzymes involved in O-glycosylation pathways using a bioinformatics-based comparative genomics approach. In the present study also, a bioinformatics-based comparative genomics approach has been used for the identification of the homologs of the enzymes involved in O-glycosylation pathways. The amino acid sequence of the ORFs has been used as query for all the database searches.

The search strategy is depicted in Figure S2. Essentially, it involves searching the target database first by BLAST [31]. Hits with E-value <1.0 are selected. This is followed by the identification of hits with high query and subject coverages i.e., the extent to which the alignment covers the query/subject sequences. Hits were combined if the query sequences shared ≥ 75% sequence identity. A Multiple sequence alignment (MSA) of these hits was used to generate a hidden Markov model (HMM) profile using the software HMMER http://hmmer.janelia.org. The dataset of 865 proteomes was re-searched using this HMM profile [32]. Both BLAST and HMMER were installed and run locally. Default values were used for all the parameters except that BLOSUM62 was used as the scoring matrix by setting the composition-based score adjustment to True. E-value cut-off was set to 0.1 for HMMER. Multiple sequence alignment of the chosen BLAST hits was performed using T-Coffee with default values for all the parameters [33].

Results and Discussion

Analysis of BLAST hits. Searching the dataset of 865 proteomes using BLASTp gave a large number of hits for most of the enzymes (Table 1). The number of hits obtained for different enzymes within a category is variable. Hits for Category-I enzymes varied between 764 and 1476; variations in the number of hits is much higher for Category-II, -III and -IV. Very few hits are obtained for Category-V enzymes. Within each category, not all hits are unique. This is because many of the hits share sequence similarity with more than one query enzyme. Query coverage is also important in addition to E-value to establish sequence homology. Hence, query coverages of the hits were plotted against their respective E-values (Figure S3). In addition, cumulative frequencies were plotted to visualize the distribution of E-values. It is seen that nearly 75% of hits in Category-I have E-value <10-10. However, the query coverage is >0.8 for 6,515 hits. This indicates that the sequence homologs of Category-I enzymes have diverged less. In Category-II, 939 hits have E-value <10-10 and query coverage >0.8. Only a small fraction of hits for enzymes of Category-III, -IV and -V have E-value <10-10.

Identification of homologs from HMM profiles. The distribution of E-values and the extent of query coverages of BLAST hits Figure S3 suggest that there can be many false positives vis-à-vis molecular function. It is not possible to ascertain the exact number of false positives without experimental data. Hence, a more stringent strategy was used to identify homologs so that false positives are fewer (Figure S2). Essentially, BLAST hits with very high query coverages and low E-values were chosen to generate HMM profiles. Specifically, the following steps were followed:

(i) Hits with high (>80%) query and subject coverage's were selected. In the case of NmPgIH, NmPgLe, SeWxz, PaWxz and AaWaaL, a lower cut-off for query coverage had to be used since very few hits have higher coverage’s (Table 1). In Category-V, very few hits had high subject coverage’s suggesting that hits are much longer in primary sequence than the query. Hence, only query coverage was used as cut-off criterion in this case. High query and subject coverage criteria led to very few numbers of hits for further analysis in case of NaPgF, PaPlO, PaWaaL.

(ii) HpWaaL, PgWaaL and HpWaaL-G. In these cases BLAST hits having query coverage of ≥ 70% were selected as final hits for analysis.

(iii) Within each category, hits were combined when query sequences shared ≥ 75% sequence identity. For example, in Category-I, hits for the enzymes EcWecA, KpWecA and YeWecA were combined together.

(iv) The dataset of 865 proteomes was re-searched using these HMM profiles and the hits that have HMM profile coverage ≥ 90% were selected for further analysis. These are taken to be the sequence homologs of the query enzymes (Table 1). Setting a high stringency cut-off of at least 90% HMM profile coverage meant a substantial reduction in the number of hits (Table 1). The E-values for the hits satisfying the 90% HMM profile coverage are very low except in nine cases: the highest E-value in these cases lies between 1.0×10^-6 and 0.1 (Table 1). Plots of cumulative frequencies

Citation: Kumar M, Balajia PV (2014) Diversity, Abundance and Distribution of O-linked Glycosylation Pathway Enzymes in Prokaryotes-A Comparative Genomics Study. J Glycomics Lipidomics 4: 117. doi:10.4172/2153-0637.1000117

ISSN: 2153-0637 JGL, an open access journal
of E-values for such hits showed that, even in these cases, most of the hits have E-values <10^-10 (Figure S4). Thus, choosing only hits with low E-value and high alignment coverage ensured that the hits are likely to be functional homologs also. The final hits for further analysis were obtained by combining the hits of all enzymes from that category (Table 2).

**Table 1:** Number of hits obtained from BLAST and HMM searches.

| Protein       | Number of BLASTp hits | Number of BLASTp hits chosen for MSA | QS Coverage threshold | Number of hits pooled for MSAa | Number of HMM hits | Highest E-value (HMM search) |
|---------------|-----------------------|-------------------------------------|-----------------------|---------------------------------|--------------------|-----------------------------|
| **Category I** |                       |                                     |                       |                                 |                    |                             |
| CjPglC        | 1205                  | 60                                  | 95                    | NA                              | 1212               | 150                         | 9.2E-35                   |
| GsWsaP        | 1476                  | 80                                  | 90                    | NA                              | 1684               | 557                         | 3.0E-12                   |
| HpWecA        | 1080                  | 26                                  | 80                    | NA                              | 1498               | 521                         | 1.1E-12                   |
| PaWbpL        | 1381                  | 38                                  | 85                    | NA                              | 1502               | 596                         | 1.1E-15                   |
| PaWsdP        | 1411                  | 38                                  | 90                    | NA                              | 1645               | 502                         | 7.1E-69                   |
| SeWbaP        | 1207                  | 133                                 | 90                    | NA                              | 1211               | 1152                        | 1.1E-20                   |
| SpWchA        | 1206                  | 28                                  | 85                    | NA                              | 1606               | 513                         | 4.6E-62                   |
| YeWbcO        | 1355                  | 44                                  | 85                    | NA                              | 1506               | 583                         | 1.9E-19                   |
| EcWecA        | 976                   | 49                                  | 95                    | NA                              | 118                | 1495                        | 2.5E-13                   |
| KoWecA        | 764                   | 67                                  | 90                    | NA                              | 118                | 1506                        | 2.5E-13                   |
| YeWecA        | 879                   | 117                                 | 90                    | NA                              | 118                | 1506                        | 2.5E-13                   |
| NgPglB        | 1204                  | 54                                  | 95                    | NA                              | 118                | 1506                        | 2.5E-13                   |
| NnPglB        | 1212                  | 51                                  | 95                    | NA                              | 118                | 1506                        | 2.5E-13                   |
| NnPglB2       | 1212                  | 55                                  | 95                    | NA                              | 118                | 1506                        | 2.5E-13                   |
| **Category II** |                       |                                     |                       |                                 |                    |                             |
| NnPglB2       | 129                   | 32                                  | 85                    | NA                              | 1211               | 197                         | 3.3E-34                   |
| NnPglC        | 1867                  | 99                                  | 95                    | NA                              | 4586               | 1411                        | 2.9E-31                   |
| NnPglD        | 1660                  | 71                                  | 95                    | NA                              | 3767               | 357                         | 1.0E-12                   |
| NnPglB        | 2480                  | 54                                  | 85                    | NA                              | 6654               | 231                         | 2.8E-20                   |
| NgPglB        | 523                   | 64                                  | 95                    | NA                              | 105                | 6654                        | 2.8E-20                   |
| **Category III** |                      |                                     |                       |                                 |                    |                             |
| NgPglA        | 2824                  | 77                                  | 95                    | NA                              | 9162               | 2703                        | 0.079                     |
| NnPglH        | 234                   | 26                                  | 70                    | NA                              | 7788               | 1005                        | 4.3E-05                   |
| NnPglG        | 2290                  | 52                                  | 95                    | NA                              | 9652               | 4139                        | 0.1                       |
| NnPglE        | 1253                  | 37                                  | 45                    | NA                              | 7011               | 367                         | 4.4E-05                   |
| **Category IV** |                       |                                     |                       |                                 |                    |                             |
| SeWzx         | 43                    | 19                                  | 70                    | NA                              | 296                | 129                         | 0.083                     |
| EcWzm         | 488                   | 96                                  | 95                    | NA                              | 2150               | 571                         | 5.7E-06                   |
| PaWzx         | 62                    | 25                                  | 65                    | NA                              | 672                | 360                         | 0.1                       |
| BiWzx         | 295                   | 48                                  | 85                    | NA                              | 1561               | 604                         | 0.077                     |
| NsPglT        | 52                    | NA                                  | NA                    | NA                              | NA                 | NA                          | 0.1                       |
| EcWzm         | 278                   | 46                                  | 95                    | NA                              | 1190               | 472                         | 0.071                     |
| PbaWzm        | 348                   | 50                                  | 88                    | NA                              | 1823               | 559                         | 0.005                     |
| PbaWzt        | 49642                 | 82                                  | 85                    | NA                              | 16352              | 160                         | 6.0E-24                   |
| EcWzt_I       | 30738                 | 62                                  | 90                    | 65                              | 16575              | 125                         | 1.8E-24                   |
| EcWzt_II      | 30625                 | 47                                  | 90                    | 65                              | 16575              | 125                         | 1.8E-24                   |
| **Category V** |                       |                                     |                       |                                 |                    |                             |
| PaPilO        | 21                    | NA                                  | NA                    | NA                              | NA                 | NA                          | 3                         |
| PaWaaL        | 55                    | NA                                  | NA                    | NA                              | NA                 | NA                          | 5                         |
| HpWaaL        | 22                    | NA                                  | NA                    | NA                              | NA                 | NA                          | 7                         |
| PgWaaL        | 18                    | NA                                  | NA                    | NA                              | NA                 | NA                          | 2                         |
| HpWaaL-       | 42                    | NA                                  | NA                    | NA                              | NA                 | NA                          | 7                         |
| AaWaaL        | 109                   | 26                                  | 65#                   | 26                              | 670                | 145                         | 0.046                     |
| NnPglL        | 63                    | 63                                  | 65#                   | 24                              | 477                | 49                          | 3.9E-38                   |

NA denotes not applicable

Hits were grouped together when query sequence identity is ≥ 75%

BLAST hits with query coverage ≥ 70% selected directly for final analysis and HMM profiles were not generated

Only query coverage was taken in these cases

Analysis of HMM hits. Every HMM hit is unique only in the case of Category-II. This is not surprising since enzymes belonging to this category have different molecular functions viz., dehydratase, acetyltransferase and aminotransferase (Table S1). In Category-I, -III and -IV, only a subset of hits are unique indicating that many hits
align with more than one HMM in that category, albeit with different e-values (Table 2). The highest number of hits for a given category is obtained for extension enzymes (Category-III). This may be a reflection of the diversity of the glycan structures. Alternatively, some of these enzymes are part of other glycan biosynthesis pathways e.g., LPS and capsular polysaccharides.

Very few hits are obtained for Category-V and most of them are unique i.e., most of the hits share sequence similarity with only one enzyme in this category (Table 2). Comparison of the amino acid sequences of the experimentally characterized OTs in Category-V Table S1 showed that these enzymes are highly divergent. Statistically significant sequence similarity coupled with adequate query coverage can be observed in only two cases:

(i) Moderate similarity (alignment scores between 29.3 and 47.8 bits; E-values between $3.0 \times 10^{-4}$ and $5.0 \times 10^{-10}$) is shared by a part (residues 185-365) of AaWaaL with the OTs from H. pylori and N. meningitidis.

(ii) The H. pylori enzymes HpWaaL and HpWaaL-G show very high similarity with each other. In contrast, the two P. aeruginosa enzymes PaPilO and PaWaaL have no detectable similarity with each other; so is the case with the two N. meningitidis enzymes NmOTase and NmPglL. A similar observation viz., proteins performing the same function despite the absence of sequence similarity was seen in the case of two OTs involved in the N-glycosylation of prokaryotic proteins. These are Campylobacter jejuni PglB [10] and Pyrococcus furiosus OT [34].

Organisms that have homologs for all the enzyme categories and for none of these pathways have been highlighted. The number of organisms from which the indicated number of unique HMM hits come from.

Categories of glycosylation pathways

| Category          | Name               | Total number of HMM hits | Number of unique hits | Number of organisms |
|-------------------|--------------------|--------------------------|-----------------------|---------------------|
| Category-I        | Initiator enzymes  | 5302                     | 1827                  | 713                 |
| Category-II       | Modification enzymes | 2566                   | 2568                  | 723                 |
| Category-III      | Extension enzymes  | 8214                     | 4914                  | 776                 |
| Category-IV       | Lipases/translocases | 2891                    | 1620                  | 661                 |
| Category-V        | Ligosaccharidyltransferases | 218                  | 204                   | 168                 |

Table 2: Total number of HMM hits, unique hits chosen for further analysis and the number of organisms to which these hits belong.

| Group               | Number of organisms in the dataset | Number of organisms vis-à-vis O-glycosylation pathway | Number of organisms vis-à-vis N-glycosylation pathway |
|---------------------|------------------------------------|-------------------------------------------------------|------------------------------------------------------|
|                     | Homologs for all enzymes           | No homolog for any enzyme                              | Homologs for all enzymes                              | No homolog for any enzyme                              |
| Alphaproteobacteria | 103                                | 10                                                    | 9                                                    | 0                                                    | 3                                                    |
| Bacteroidetes/Chlorobi | 26                                  | 4                                                    | 2                                                    | 0                                                    | 1                                                    |
| Betaproteobacteria  | 62                                  | 51                                                    | 2                                                    | 0                                                    | 0                                                    |
| Chlamydiae/Verrucomicrobia | 13                                | 0                                                    | 3                                                    | 0                                                    | 0                                                    |
| Chloroflexi         | 10                                  | 1                                                    | 0                                                    | 2                                                    | 0                                                    |
| Crenarchaeota       | 36                                  | 4                                                    | 0                                                    | 0                                                    | 0                                                    |
| Cyanobacteria       | 27                                  | 4                                                    | 2                                                    | 1                                                    | 0                                                    |
| Deltaproteobacteria | 24                                  | 1                                                    | 0                                                    | 8                                                    | 0                                                    |
| Epsilonproteobacteria | 179                               | 4                                                    | 17                                                   | 1                                                    | 8                                                    |
| Firmicutes          | 209                                 | 47                                                   | 15                                                   | 0                                                    | 3                                                    |
| Gammaproteobacteria/Spirochaetes | 18                                | 2                                                    | 0                                                    | 0                                                    | 0                                                    |
| Others              | 15                                  | 0                                                    | 1                                                    | 0                                                    | 1                                                    |

Table 3: Number of organisms in each group that have / do not have homologs for enzymes of the O- and N-glycosylation.

The motility status is known for 90 of the 128 organisms; vast majorities (68 out of 90) are motile. It has been shown in some organisms that flagella are O-glycosylated and this has been shown to be important for its assembly [35,36]. In Pseudomonas syringae, it has been suggested that the absence of glycosylation destabilizes the filament structure of flagella and affects the swimming activity of mutants [37]. In addition, in Pseudomonas aeruginosa, it has been suggested that the glycosylation of flagellum and motility can play a crucial role in flagellum-mediated virulence [38]. Nearly half of the organisms (from among the 128) that are motile are also pathogenic...
As mentioned earlier, an organism should minimally have an initiator enzyme, a flippase and an OT to O-glycosylate proteins. A large number of organisms did not have homologs of these three enzymes. In most of the cases, OT is the missing enzyme (Table S3). These organisms probably do have OTs but these have escaped detection in this study because of the high sequence divergence of OTs, as mentioned earlier.

Fifty-two organisms do not have homologs for even a single enzyme of any of the five categories. These organisms also belong to diverse habitats. Their temperature range is mostly mesophilic and they are from different subgroups (Table 5). These organisms have varied morphology. Among different groups, Chlamydiae and Crenarchaeota do not have homologs for any of the five enzyme categories. Out of 52 organisms which do not have homologs for even a single enzyme category, 41 are host-associated and 30 are pathogenic. Comparative genomics studies have shown that large scale genome deletions are

| Tax id | Organism Name | Gram Status | Motile | Habi-tat | Temp. range | Patho-genic |
|--------|---------------|-------------|--------|---------|-------------|-------------|
| 224911 | Bradyrhizobium japonicum USDA 110 | Neg. | Yes | HA | MS | No |
| 288000 | Bradyrhizobium sp. BTA1 | Neg. | Yes | HA | MS | NA |
| 114615 | Bradyrhizobium sp. ORS278 | NA | NA | HA | MS | No |
| 419610 | Methyllobacterium extorquens PA1 | Neg. | Yes | MU | MS | No |
| 323097 | Nitrobacter hamburgensis X14 | Neg. | Yes | TE | MS | No |
| 439375 | Ochrobacterium anthropi ATCC 49188 | NA | NA | TE | MS | Yes |
| 450851 | Phenylobacterium zucineum HLK1 | Neg. | Yes | HA | MS | Yes |
| 347834 | Rhizobium ellii CFN 42 | Neg. | NA | HA | MS | No |
| 491916 | Rhizobium ellii CIAT 652 | Neg. | Yes | HA | MS | No |
| 216596 | Rhizobium leguminosarum bv. viciae 3841 | Neg. | Yes | HA | MS | No |

Group: Bacteroidetes/Chlorobi

| Tax id | Organism Name | Gram Status | Motile | Habi-tat | Temp. range | Patho-genic |
|--------|---------------|-------------|--------|---------|-------------|-------------|
| 331678 | Chlorobium phaeobacteroides BS1 | Neg. | No | AQ | MS | No |
| 319225 | Pelodictyon luteolum DSM 273 | Neg. | No | MU | MS | No |
| 431947 | Porphyromonas gingivalis ATCC 33277 | Neg. | No | HA | MS | Yes |
| 242819 | Porphyromonas gingivalis W83 | Neg. | No | HA | MS | Yes |

Group: Betaproteobacteria

| Tax id | Organism Name | Gram Status | Motile | Habi-tat | Temp. range | Patho-genic |
|--------|---------------|-------------|--------|---------|-------------|-------------|
| 397945 | Acidovorax citrulli AAC00-1 | Neg. | Yes | MU | MS | Yes |
| 232721 | Acidovorax sp. JS42 | Neg. | Yes | TE | MS | No |
| 62928 | Atzoracis sp. BHT2 | Neg. | Yes | HA | MS | No |
| 360910 | Bordetella avium 197N | Neg. | Yes | HA | NA | Yes |
| 257310 | Bordetella bronchiseptica RB50 | Neg. | Yes | HA | MS | Yes |
| 257311 | Bordetella parapertussis 12822 | Neg. | NA | HA | MS | Yes |
| 257313 | Bordetella pertussis Tohama I | Neg. | NA | HA | MS | Yes |
| 340100 | Bordetella petrii DSM 12804 | Neg. | No | AQ | MS | No |
| 339670 | Burkholderia ambifaria AMMD | Neg. | Yes | MU | MS | NA |
| 398577 | Burkholderia ambifaria MC40-6 | Neg. | NA | MU | MS | Yes |
| 331271 | Burkholderia cenocepacia AU 1054 | NA | NA | NA | NA | Yes |
| 331272 | Burkholderia cenocepacia HI2424 | NA | NA | NA | NA | NA |
| 216591 | Burkholderia cenocepacia J2315 | Neg. | Yes | MU | NA | Yes |
| 406425 | Burkholderia cenocepacia MCO-3 | Neg. | Yes | MU | MS | Yes |
| 482957 | Burkholderia lata | NA | NA | NA | NA | NA |
| 243160 | Burkholderia maltophilia ATCC 23344 | Neg. | No | HA | MS | Yes |
| 412022 | Burkholderia maltophilia NCTC 10229 | Neg. | No | HA | MS | Yes |
| 320389 | Burkholderia maltophilia NCTC 10247 | Neg. | No | HA | MS | Yes |
| 320388 | Burkholderia maltophilia SAVP1 | Neg. | No | HA | MS | Yes |
| 395019 | Burkholderia multivorans ATCC 17816 | Neg. | NA | HA | MS | Yes |
| 391038 | Burkholderia phytofirmans PsJN | Neg. | Yes | HA | NA | No |
| 398527 | Burkholderia pseudomallei 1105a | Neg. | Yes | TE | MS | Yes |
| 357348 | Burkholderia pseudomallei 1716b | Neg. | Yes | TE | MS | Yes |
| 320372 | Burkholderia pseudomallei 668 | Neg. | Yes | TE | MS | Yes |
| 272560 | Burkholderia pseudomallei K96243 | Neg. | Yes | TE | MS | Yes |
| 271848 | Burkholderia thailandensis E264 | Neg. | Yes | TE | MS | NA |
| Tax id  | Organism Name                               | Gram Status | Motile | Habi-tat | Temp. range† | Patho- genic |
|--------|---------------------------------------------|-------------|--------|---------|--------------|--------------|
| 269482 | Burkholderia vietnamiensis G4               | Neg.        | Yes    | MU      | NA           | Yes          |
| 266265 | Burkholderia xenovorans LB400               | Neg.        | Yes    | MU      | MS           | Yes          |
| 243365 | Chromobacterium violaceum ATCC 12472        | Neg.        | Yes    | MU      | MS           | Yes          |
| 977880 | Cupriavidus taiwanensis LMG 19424           | NA          | NA     | NA      | NA           | NA           |
| 398578 | Deftia acidovorans SPH-1                    | Neg.        | NA     | MU      | MS           | NA           |
| 535289 | Diaphorobacter sp. TPSY                     | Neg.        | Yes    | AQ      | MS           | No           |
| 204773 | Hemminmonas arsenicoxydans                  | NA          | NA     | AQ      | MS           | No           |
| 375286 | Janthinobacterium sp. Marseille             | NA          | Yes    | AQ      | MS           | NA           |
| 557598 | Laribacter hongkongensis HLMK9              | Neg.        | Yes    | HA      | MS           | Yes          |
| 395495 | Legothrix cholodni SP-6                     | NA          | No     | AQ      | MS           | No           |
| 242231 | Neisseria gonorrhoeae FA 1090               | Neg.        | NA     | HA      | MS           | Yes          |
| 521006 | Neisseria gonorrhoeae NCCP11945             | Neg.        | NA     | HA      | MS           | Yes          |
| 374833 | Neisseria meningitidis 053442              | Neg.        | NA     | HA      | MS           | Yes          |
| 272831 | Neisseria meningitidis FAM18                | Neg.        | NA     | HA      | MS           | Yes          |
| 122586 | Neisseria meningitidis MC58                 | Neg.        | NA     | HA      | MS           | Yes          |
| 122587 | Neisseria meningitidis Z2491                | Neg.        | NA     | HA      | MS           | Yes          |
| 335283 | Nitrosomonas europa C91                    | Neg.        | Yes    | MU      | NA           | NA           |
| 323848 | Nitrosospira multiformis ATCC 25196         | Neg.        | Yes    | TE      | MS           | NA           |
| 296591 | Polaromonas sp. JS666                      | Neg.        | No     | MU      | MS           | No           |
| 381666 | Ralstoniaeutropha H16                      | Neg.        | Yes    | SP      | MS           | NA           |
| 264198 | Ralstoniaeutropha JMP134                   | NA          | Yes    | MU      | MS           | NA           |
| 266264 | Ralstonia metallidurans CH34               | Neg.        | NA     | SP      | MS           | No           |
| 402626 | Ralstonia pickettii 12J                    | Neg.        | NA     | MU      | MS           | No           |
| 338969 | Rhodotherax femireducens T118              | Neg.        | Yes    | MU      | MS           | No           |
| 383372 | Roseiflexus castenholzii DSM                | NA          | Yes    | AQ      | TH           | No           |
| 43989  | Cyanobacteriaceae sp. ATCC 51142            | NA          | NA     | AQ      | MS           | No           |
| 65393  | Cyanobacteriaceae sp. PCC 7424             | NA          | No     | AQ      | MS           | No           |
| 84588  | Synechococcus sp. WH 8102                  | NA          | Yes    | AQ      | MS           | No           |
| 1148   | Synechocystis sp. PCC 6803                 | NA          | NA     | AQ      | MS           | No           |
| 177437 | Desulfobacterium autotrophicum HRM2         | Neg.        | Yes    | MU      | MS           | No           |
| 525146 | Desulfovibrio desulficans subsp. desulfuricans str. ATCC 27774 | Neg. | Yes | MS | No |
| 883    | Desulfovibrio vulgaris str. 'Miyazaki F'   | Neg.        | NA     | MU      | MS           | No           |
| 351605 | Geobacter uraniireducens R4                 | Neg.        | NA     | MU      | MS           | No           |
| 387093 | Sulfurovum sp. NBC37-1                     | Neg.        | No     | SP      | MS           | No           |
| 272562 | Clostridium acetobutylicum ATCC 824         | Pos.        | Yes    | MU      | MS           | No           |
| 212717 | Clostridium tetani E88                     | Pos.        | Yes    | MU      | MS           | Yes          |
| 203119 | Clostridium thermoceluum ATCC 27405        | Pos.        | Yes    | MU      | TH           | No           |
| 373903 | Halothermothrix crenul H 168               | Neg.        | NA     | AQ      | TH           | No           |
| 480119 | Acinetobacter baumannii AB0057             | Neg.        | No     | MU      | MS           | Yes          |
| 509170 | Acinetobacter baumannii SDF                | Neg.        | NA     | AQ      | MS           | Yes          |
| 62977  | Acinetobacter sp. ADP1                     | Neg.        | No     | MU      | MS           | Yes          |
| 434271 | Actinobacillus pleuropneumoniae serovar 3 str. JL03 | Neg. | NA | HA | NA | Yes |
| 537457 | Actinobacillus pleuropneumoniae serovar 7 str. AP76 | Neg. | NA | HA | MS | Yes |
| 380703 | Aeromonas hydrophila subsp. hydrophila ATCC 7966 | Neg. | Yes | MU | MS | Yes |
| 382245 | Aeromonas salmonicida subsp. salmonicida A449 | Neg. | Yes | AQ | MS | Yes |
| 316275 | Aliivibrio salmonicida LF11238              | Neg.        | Yes    | AQ      | PS           | Yes          |
| 465817 | Erwinia tasmaniensis E1199                 | Neg.        | Yes    | HA      | MS           | No           |
| 458234 | Francisella tularensis subsp. holocerca FTNF002-00 | Neg. | No | MU | MS | Yes |
| 376619 | Francisella tularensis subsp. holocerca LVS | NA | NA | NA | NA | NA |
| 441952 | Francisella tularensis subsp. mediastatica FSC147 | Neg. | No | HA | MS | Yes |
| 393115 | Francisella tularensis subsp. tularensis FSC198 | Neg. | No | AQ | MS | Yes |
| 177416 | Francisella tularensis subsp. tularensis SCHU S4 | Neg. | No | AQ | NA | Yes |
| 481316 | Francisella tularensis subsp. tularensis WY96-3418 | Neg. | No | HA | MS | Yes |
suggestive of the inherent sequence divergence of the glycosylation pathway enzymes are from Neisseria, a Betaproteobacteria. This is similar to that observed for O-glycosylation pathway i.e., all Category-II and V enzymes are from C. jejuni, an Epsilonproteobacteria. This scenario is similar to the group Epsilonproteobacteria (Table 3) and the query enzymes have homologs for all enzymes of the N-glycosylation pathway belong to this group. It is seen that the maximum number of organisms that have homologs of the enzymes of the N-glycosylation pathway are from Desulfovibrio desulfuricans, a Deltaproteobacteria, and in none of the species Escherichia and Thermotoga. This probably can be a possible role for glycosylation of proteins in these two organisms can be good model systems to study the effects of glycosylation and exploitation as microbial factory for glycosylating heterologous proteins. Species and strain-specific variations in the presence of homologs. Analysis of the presence of homologs for enzymes of different categories in different species of a genus did not show much variation, especially in the case of the presence of homologs for enzymes of different categories in different species of a genus. This scenario is similar to that observed for O-glycosylation pathway enzymes, i.e., all Category-II and -III query enzymes are from Neisseria, a Betaproteobacteria. This is suggestive of the inherent sequence divergence of the glycosylation pathway enzymes.

It is found that Roseflexus castenholzii and Desulfovibrio desulfuricans have homologs for both N- and O-glycosylation pathway enzymes. R. castenholzii belongs to the group Chloroflexi whereas D. desulfuricans is a Deltaproteobacteria. These two organisms differ from each other in their habitat, oxygen requirements and temperature range. Despite these differences, they both seem to have N- and O-linked glycosylation pathway enzymes. Glycosylation is known to play a role in the stabilization of the folded form of proteins [41] and this can be a possible role for glycosylation of proteins in R. castenholzii, a thermophile.

Desulfovibrio desulfuricans species shows the potential of being pathogenic since it has been found that it can cause bacteremia in immunocompetent man [42]. These two organisms can be good model systems to study the effects of glycosylation and exploitation as microbial factory for glycosylating heterologous proteins. Species and strain-specific variations in the presence of homologs. Analysis of the presence of homologs for enzymes of different categories in different species of a genus did not show much variation, especially when Category-V is excluded (Table 6). Homologs of OTs are present in only a few species in genera such as Pseudomonas and Leptospira when Category-V is excluded (Table 6). Homologs of OTs are present in only a few species in genera such as Pseudomonas and Leptospira. This is indicative of the inherent sequence divergence of the glycosylation pathway enzymes.

| Organism                        | Category | Presence | Homolog | Nature | Manner |
|---------------------------------|----------|----------|---------|--------|--------|
| Marinobacter aquaeolei VT8      |          | Neg.     | Yes     | AQ     | MS     | No     |
| Marimonas sp. MWY1L             |          | Neg.     | Yes     | AQ     | MS     | No     |
| Pectobacterium atrosepticum SCR1043 |        | Neg.     | Yes     | MU     | MS     | Yes    |
| Pseudomonas aeruginosa LESB58   |          | Neg.     | Yes     | MU     | MS     | Yes    |
| Pseudomonas aeruginosa PA7      |          | Neg.     | Yes     | MU     | MS     | Yes    |
| Pseudomonas aeruginosa PAO1     |          | Neg.     | Yes     | MU     | MS     | Yes    |
| Pseudomonas aeruginosa UCBPP-PA14 |       | Neg.     | Yes     | MU     | MS     | Yes    |
| Pseudomonas putida WB19         |          | Neg.     | Yes     | MU     | MS     | No     |
| Psychromonas ingrahami 37       |          | Neg.     | No      | AQ     | PS     | No     |
| Serratia proteamaculans 568     |          | NA      | Yes     | MU     | MS     | Yes    |
| Shewanella sp. MR-4             |          | Neg.     | Yes     | MU     | MS     | NA     |
| Sodalis glossinisidus 'morsitans' |       | Neg.     | No      | HA     | MS     | No     |
| Stenotrophomonas maltophilia K279a |        | Neg.     | NA      | MU     | MS     | Yes    |
| Stenotrophomonas maltophilia R551-3 |      | Neg.     | NA      | MU     | MS     | NA     |
| Vibrio cholerae M66-2           |          | Neg.     | Yes     | MU     | MS     | Yes    |
| Vibrio cholerae O1 biovar El Tor str. N16961 | | Neg. | Yes | AQ | MS | Yes |
| Vibrio cholerae O395            |          | Neg.     | Yes     | AQ     | MS     | Yes    |
| Vibrio fetus ES114              |          | Neg.     | Yes     | MU     | MS     | No     |
| Vibrio vulnificus CMCP6         |          | Neg.     | Yes     | AQ     | MS     | Yes    |
| Vibrio vulnificus YJ016         |          | Neg.     | Yes     | AQ     | MS     | Yes    |
| Xanthomonas axonopodis pv. citi str. 306 | | Neg. | Yes | HA | MS | Yes |
| Xanthomonas campestris pv. campestris str. 8004 | | Neg. | Yes | HA | MS | Yes |
| Xanthomonas campestris pv. campestris str. ATCC 33913 | | Neg. | Yes | HA | MS | Yes |
| Xanthomonas campestris pv. campestris str. B100 | | NA | NA | NA | NA | NA |
| Xanthomonas campestris pv. vesicatoria str. B5-10 | | Neg. | Yes | HA | MS | Yes |
| Xanthomonas oryzae pv. oryzae KACC10331 | | Neg. | NA | HA | MS | Yes |
| Xanthomonas oryzae pv. oryzae MAFF 311018 | | Neg. | Yes | HA | MS | Yes |
| Xanthomonas oryzae pv. oryzae PX099A | | Neg. | Yes | HA | MS | Yes |
| Xylella fastidiosa 9a5c         |          | Neg.     | NA      | HA     | MS     | Yes    |
| Xylella fastidiosa M12          |          | Neg.     | NA      | HA     | MS     | Yes    |
| Xylella fastidiosa M23          |          | Neg.     | NA      | HA     | MS     | NA     |
| Xylella fastidiosa Temecula1    |          | NA      | NA      | HA     | MS     | Yes    |

**Table 4:** Some characteristics of organisms that have at least one homolog for each category of O- glycosylation pathway enzymes

| Organism                        | Category | Presence | Homolog | Nature | Manner |
|---------------------------------|----------|----------|---------|--------|--------|
| Leptospira borgopersenii serovar Hardjo-bovis JB197 | | Neg. | Yes | HA | MS | Yes |
| Leptospira borgopersenii serovar Hardjo-bovis L550 | | Neg. | Yes | HA | MS | Yes |

Data are not available in some cases; these are denoted as NA (not available)

AQ: Aquatic; HA: Host-Associated; MU: Multiple; SP: Specialized; TE: Terrestrial

MS: Mesophilic; PS: Psychrophilic; TH: Thermophilic

Characteristics of host-associated organisms/symbionts [39,40].
| Tax id | Organism Name | Gram Status | Motility | Habi-ta§ | Temp range† | Patho-genic |
|--------|---------------|-------------|----------|----------|------------|-------------|
| 320483 | Anaplasma marginale str. Florida | NA | NA | HA | MS | Yes |
| 234826 | Anaplasma marginale str. St. Maries | NA | NA | HA | NA | Yes |
| 212042 | Anaplasma phagocytophilum H2 | Neg. | NA | HA | MS | Yes |
| 269484 | Ehrlichia caris str. Jake | Neg. | NA | HA | NA | Yes |
| 205920 | Ehrlichia chaffeensis str. Arkansas | NA | NA | HA | NA | Yes |
| 302409 | Ehrlichia ruminantium str. Gardel | Neg. | NA | HA | MS | Yes |
| 254945 | Ehrlichia ruminantium str. Welgevonden | Neg. | NA | HA | MS | Yes |
| 570417 | Wolbachia endosymbiont of Culex quinquefasciatus Piel | Neg. | NA | HA | MS | No |
| 292805 | Wolbachia endosymbiont strain TRS of Brugia malayi | Neg. | No | HA | MS | NA |
| 511995 | Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2 | NA | NA | SP | MS | No |
| 444179 | Candidatus Sulcia muelleri GWSS | NA | NA | NA | NA | NA |
| 269483 | Burkholderia sp. | Neg. | Yes | MU | NA | Yes |
| 164546 | Cupriavidus taiwanensis | Neg. | Yes | HA | MS | No |
| 218497 | Chlamydophila abortus S26/3 | Neg. | NA | HA | MS | Yes |
| 227941 | Chlamydophila caviae GPIC | Neg. | NA | HA | MS | Yes |
| 264202 | Chlamydophila felis Fe/C-56 | Neg. | NA | HA | MS | Yes |
| 453591 | Ignicoccus hospitalis KIN4/I | Neg. | Yes | AQ | HT | NA |
| 269799 | Geobacter metallireducens GS-15 | Neg. | Yes | AQ | MS | No |
| 338963 | Pelobacter carbinolicus DSM 2380 | Neg. | NA | AQ | MS | No |
| 322098 | Aster yellows witches’-broom phytoplasma YWB | NA | NA | HA | MS | Yes |
| 59748 | Candidatus Phytoplasma australense | NA | NA | HA | MS | Yes |
| 220668 | Lactobacillus plantarum WCFS1 | Pos. | NA | HA | MS | No |
| 265311 | Mesoplasma florula L1 | Neg. | No | HA | MS | Yes |
| 347257 | Mycoplasma agalactiae PG2 | Neg. | No | HA | PS | Yes |
| 243272 | Mycoplasma arthritidis 15BL3-1 | Neg. | No | HA | MS | Yes |
| 340047 | Mycoplasma capricolum subsp. capricolum ATCC 27343 | Neg. | No | HA | MS | Yes |
| 233150 | Mycoplasma gallisepticum R | Neg. | Yes | HA | MS | Yes |
| 243273 | Mycoplasma genitalium G37 | Neg. | Yes | HA | MS | Yes |
| 295358 | Mycoplasma hyopneumoniae 232 | Neg. | No | HA | MS | Yes |
| 262722 | Mycoplasma hyopneumoniae 7448 | Neg. | No | HA | MS | Yes |
| 262719 | Mycoplasma hyopneumoniae J | Neg. | No | HA | MS | Yes |
| 267748 | Mycoplasma mobile 163K | Neg. | Yes | HA | MS | Yes |
| 272633 | Mycoplasma penetrans HF-2 | Neg. | No | HA | MS | Yes |
| 272634 | Mycoplasma pneumoniae M129 | Neg. | No | HA | MS | Yes |
| 272635 | Mycoplasma pulmonis UAB CTIP | Neg. | No | HA | MS | Yes |
| 262723 | Mycoplasma synoviae 53 | Neg. | No | HA | MS | Yes |
| 314275 | Alteromonas macleodi ‘Deep ecotye’ | Neg. | Yes | AQ | MS | No |
| 374463 | Baumannia cicadellinicola str. Hc (Homalodisca coagulata) | NA | NA | HA | NA | No |
| 563178 | Buchnera aphidicola str. 5A (Acyrthosiphion pismum) | Neg. | NA | HA | MS | No |
| 617005 | Buchnera aphidicola str. APS (Acrithosiphion pismum) | Neg. | NA | HA | MS | No |
| 372461 | Buchnera aphidicola str. Cc (Cinara cedri) | Neg. | NA | HA | MS | No |
| 198804 | Buchnera aphidicola str. Sg (Schizaphis graminum) | Neg. | NA | HA | MS | No |
| 561501 | Buchnera aphidicola str. TuC7 (Acrithosiphion pismum) | Neg. | NA | HA | MS | No |
| 291272 | Candidatus Blochmannia pennsylvanicus str. BPEN | NA | NA | NA | NA | No |
| 203907 | Candidatus Blochmannia floridanus | Neg. | NA | SP | MS | No |
| 387662 | Candidatus Carsonella ruddii PV | NA | NA | SP | NA | No |
| 412965 | Candidatus Vescomycocysus okutaniii HA | NA | NA | HA | MS | No |
| 316407 | Escherichia coli str. K-12 substr. W3110 | Neg. | Yes | HA | MS | NA |
| 119857 | Francisella tularense subsp. holarctica | Neg. | No | HA | MS | Yes |
| 41514 | Salmonella enterica subsp. arizonae serovar 62:z4,z23:-- | Neg. | Yes | HA | MS | Yes |
The non-uniform occurrence of homologs for different categories may be due to the local needs/habitat of that particular organism. Homologs for few categories were absent in other genera (Table 6). In Acinetobacter baumannii, one strain has homologs for all five enzyme categories whereas a few others have homologs for only three or four category enzymes. This variability is suggestive of strain-specific variability as observed in Neisseria. In Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7, the variation of homologs in different organisms is not surprising since, even among Neisseria, species- and strain-specific polymorphisms have been reported [20, 45].

Overall, few genera had homologs for all enzyme categories whereas homologs for few categories were absent in other genera (Table 6). This may be due to the local needs/habitat of that particular organism [45]. The non-uniform occurrence of homologs for different categories across different genera as well as within the same genus hints at heterogeneity in the glycans synthesized by these organisms. Such kind of heterogeneity is likely to be present in different organisms of a species also. In one related study, it was established that glycan structures with different chain length are present in the genus Campylobacter when grouped on the basis of thermotolerance [44]. The variation of homologs in different organisms is not surprising since, even among Neisseria, species- and strain-specific polymorphisms have been reported [20, 45].

Distribution of different enzyme categories among the organisms. OT is critical for glycosylation and the existence of its homologs in an organism strengthens the prediction that O-glycosylation occurs in these organisms. Homologs for OT were found in 168 organisms (Table 2). Few of these have more than one homolog. Most of these 168 organisms are proteobacteria; others include Actinobacteria, Bacteroidetes/Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes and Spirochaetes. Twenty-one organisms have at least one homolog for all category enzymes except Category-IV (Table S4). It can be surmised that a divergent class of flippases are involved in these cases for transferring the oligosaccharide across the membrane in these organisms. Some organisms belonging to Betaproteobacteria and Gammaproteobacteria groups are missing homologs for extension enzymes (Category-III). This type of heterogeneity is likely to be present in different organisms (Category-IV). Some organisms belonging to Bacteroidetes/Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes and Spirochaetes. Twenty-one organisms have at least one homolog for all category enzymes except Category-IV (Table S4). It can be surmised that a divergent class of flippases are involved in these cases for transferring the oligosaccharide across the membrane in these organisms. Some organisms belonging to Betaproteobacteria and Gammaproteobacteria groups are missing homologs for extension enzymes (Category-III). This type of heterogeneity is likely to be present in different organisms (Category-IV). Some organisms belonging to Bacteroidetes/Chlorobi and Chloroflexi groups are missing homologs for extension enzymes (Category-III). This type of heterogeneity is likely to be present in different organisms. Some organisms belonging to Betaproteobacteria and Gammaproteobacteria groups are missing homologs for extension enzymes (Category-III). This type of heterogeneity is likely to be present in different organisms.
categories. There are 85 organisms in the dataset that are tagged as antibiotic resistant by the Center for Disease Control and Prevention, Atlanta (www.cdc.gov/drugresistance DiseasesConnectedAR.html#1). Nine of these have homologs for all five enzyme categories and these hints at the existence of O-glycosylation pathway. The genomes of all of these organisms are >2 Mb with 39-57% GC content. The habitat is either host-associated or multiple. All are mesophiles and live in aerobic environment. Recently, the antibiotic-resistant Acinetobacter baumannii ATCC 17978 has been reported to have the O-glycosylation pathway [9]. Even the present study shows that this organism has homologs for all five enzyme categories and hence can potentially glycosylate the proteins.

Distribution of organisms in the phylogenetic tree. 16S rRNA based phylogenetic analysis shows that the organisms which have homologs for all five enzyme categories are scattered in the phylogenetic tree and so do those that do not have homologs for any of the five enzyme categories (Figure 1). Organisms having homologs for all categories and for none of the categories are clustered in only a few branches. Variations in the occurrence of homologs belonging to different categories are observed among closely related organisms in certain subtrees (Figure 2). For example, in the Bradyrhizobium subtree, except two organisms, the other three have homologs for all enzymes categories (Figure 2A). These two organisms viz. Rhodopseudomonas palustris and Oligotropha carboxidivorans have homologs for all enzyme categories except Category-V and Categories-IV and -V, respectively. In the subtree containing some Betaproteobacteria,Ralstonia metallidurans and Ralstonia eutropha have homologs for all enzyme categories but their immediate neighbour Cupriavidus taiwanensis does not have homolog for any enzyme category Figure 2B. In this subtree, Ralstonia solanacearum has homologs for all enzymes except flippases. Polynucleobacter necessarius lacks homologs for Category-III and -V. The presence/absence of homologs and variations in the number of homologs represent significant diversity among the members of the subtree. All except two organisms in the subtree containing Diaphorobacter sp. and Leptothrix cholodni have homologs for all enzyme categories (Figure 2C). These two organisms are Methylibium petroleiphilum and Polaromonas napthalenivorans which lack homologs for Category-V enzymes. As discussed earlier, even these organisms may have OTs and the reason for not finding the homologs may be because of the sequence divergence.

The organisms which lack homologs for any of the five categories were also mapped in the phylogenetic tree. In one of the subtrees, most of the members are from Mycoplasma (Figure 2D). Homologs are absent in all organisms except Mycoplasma mycoides. It is intriguing that many of these organisms also lack homologs for enzymes involved in N-linked glycosylation as reported earlier [28]. The absence of both N- and O-linked glycosylation in these parasitic organisms suggests that these organisms have very different pathways for glycosylation or have evolved other, as yet, unknown mechanisms to serve the role played by glycosylation.

In some subtrees, one organism has homologs for enzymes of all categories whereas its neighbour does not have homolog for enzymes of any category. For example, Geobacter uraniireducens (a Deltaproteobacteria) has homologs from all five enzyme categories but its neighbour lacks homologs for only Category-V (Figure 2E and 2G). Uraniireducens has the largest genome (5.1 Mb) size among all the Geobacter which are part of this study. It is tempting to speculate that the high genome size of this organism is the reason for it having homologs for all five enzyme categories. Interestingly, it is the only Geobacter in the dataset which is microaerophilic; all others are anaerobic. Additionally, the homolog of Category-V enzyme in G. uraniireducens has significantly low GC content (40%) than the GC content of this organism in whole (54%). This suggests the presence of horizontally transferred genes in this organism. Also, other members in this subtree viz., Geobacter metallireducens and Pelobacter carbinolicus do not have homologs even for a single enzyme category.

The variation in the number of homologs belonging to different categories in case of many organisms reflects the diversity of the O-glycosylation pathway as has been demonstrated in Neisseria gonorrhoea [5,15]. These variations can be attributed to the horizontal gene transfer and selective loss of genetic material [46-48]. Moreover, a gene may exist in a phase variable form in few strains but not in others [16]. This gene might give benefit to one organism in the form of constitutive gene whereas another strain of the same species may get advantage from it as a contingency gene [49]. One such example is from Haemophilus influenzae which uses mechanisms such as homologous recombination and slipped-strand mispairing to generate high-frequency changes in expression of genes belonging to polysaccharide (LPS, CPS) and fimbrial category [49]. The understanding of the O-linked glycosylation system and its effects are likely to be more complicated since a dynamic interplay between O-glycosylation and other post-translational modifications such as the addition of phosphoethanolamine / phosphocholine has been reported [50].

In summary, homologs for all five enzymes categories are found in 128 organisms. The number is likely to be even more since a significant number of organisms have homologs for all five categories except OTs, which are known to be highly divergent in their sequences. Besides, the criteria used to identify homologs were kept very stringent to avoid false positives. Overall, this study clearly shows that the O-glycosylation pathway enzyme homologs are widely prevalent. Analyses of the pattern of distribution of homologs indicate species- and strain-specific variations in glycan structures and acquisition of O-glycosylation pathway enzyme homologs by horizontal gene transfer in certain clades.

There are several examples of proteins which share sequence
similarity but varying levels of functional similarity. In view of this, it is not possible to ascertain exactly the nature of donor and acceptor substrates used by the homologs of different enzyme categories which are identified in this study. Further bioinformatics analyses, combined with experimental data, and are essential to ascertain the specific functions of these enzymes. The experimental characterization of the substrate specificities, combined with the spatiotemporal pattern of expression of these genes, will lead to a better understanding of their involvement in various biological processes. The homologs identified are a good starting point for experimental characterization of their molecular functions.

Acknowledgement

Manjeet Kumar is grateful to the Council of Scientific and Industrial Research, India for research fellowship.

References

1. Aas FE, Vik A, Vedde J, Koomey M, Egge-Jacobsen W (2007) Neisseria gonorrhoeae O-linked glycosylation: functional analyses define both the biosynthetic pathway and glycan structure. Mol Microbiol 65: 607-624.
2. Power PM, Roddam LF, Dieckmann M, Srikantha YN, Tan YC, et al. (2000) Genetic characterization of pilin glycosylation in Neisseria meningitidis. Microbiology 146 : 967-979.
3. Power PM, Roddam LF, Rutter K, Fitzpatrick SZ, Srikantha YN, et al. (2003) Genetic characterization of pilin glycosylation and phase variation in Neisseria meningitidis. Mol Microbiol 49: 833-847.
4. Stimson E, Virji M, Makepeace K, Dell A, Morris HR, et al. (1995) Meningococcal pilin: a glycoprotein substituted with digalactosyl 2,4-diacyetamido-2,4-dideoxyhexose. Mol Microbiol 17: 1201-1214.
5. Hartley MD, Morrison MJ, Aas FE, Berud B, Koomey M, et al. (2011) Biochemical characterization of the O-glycosylation pathway in Neisseria gonorrhoeae responsible for biosynthesis of protein glycans containing N,N'-diacyetamidoacetamido. Biochemistry 50: 4936-4948.
6. Schirm M, Soo EC, Aubry AJ, Austin J, Thibault P, et al. (2003) Structural, genetic and functional characterization of the flagella glycosylation process in Helicobacter pylori. Mol Microbiol 48: 1579-1592.
7. Faridmoayer A, Fentabili MA, Mills DC, Klassen JS, Feldman MF (2007) Functional characterization of bacterial oligosaccharyltransferases involved in O-linked protein glycosylation. J Bacteriol 189: 8088-8098.
8. Flétche CM, Coyne-MJ, Villa OF, Chatzidaki-Livanis M, Comstock LE (2009) A general O-glycosylation system important to the physiology of a major human intestinal symbiont. Cell 137: 321-331.
9. Iwashkiew JA, Seper A, Weber BS, Scott NE, Vinogradov E, et al. (2012) Identification of a general O-linked protein glycosylation system in Acinetobacter baumannii and its role in virulence and biofilm formation. PLoS Pathog 8: e1002756.
10. Linton D, Dorrell N, Hitchen PG, Amber S, Karfyshew AV, et al. (2005) Functional analysis of the Campylobacter jejuni N-linked protein glycosylation pathway. Mol Microbiol 55: 1695-1703.
11. Szymanski CM, Yao R, Ewing CP, Trust TJ, Guerry P (1999) Evidence for a system of general protein glycosylation in Campylobacter jejuni. Mol Microbiol 32: 1022-1030.
12. Wacker M, Linton D, Hitchen PG, Nita-Lazar M, Haslam SM, et al. (2002) N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. Science 298: 1790-1793.
13. Yurist-Doutsch S, Eichler J (2009) Manual annotation, transcriptional analysis, and protein expression studies reveal novel genes in the agl cluster responsible for N-glycosylation in the halophilic archaeon Haloferax volcanii. J Bacteriol 191: 3068-3075.
14. Larkin A, Chang MM, Whitworth GE, Imperiali, B (2013) Biochemical evidence for an alternate pathway in N-linked glycoprotein biosynthesis. Nature Nat Chemical Biology 9: 367–373.
15. Berud B, Aas FE, Vik A, Wnther-Larsen HC, Egge-Jacobsen W, et al. (2010) Genetic, structural, and antigenic analyses of glycan diversity in the O-linked protein glycosylation systems of human Neisseria species. J Bacteriol 192: 2816-2826.
16. Banerjee A, Wang R, Supernavage SL, Ghosh SK, Parker J, et al. (2002) Implications of phase variation of a gene (pgA) encoding a pilin galactosyl transferase in gonococcal pathogenesis. J Exp Med 196: 147-162.
17. Berud B, Vituriene R, Hartley MD, Paulsen BS, Egge-Jacobsen W, et al. (2011) Genetic and molecular analyses reveal an evolutionary trajectory for glycansynthesis in a bacterial protein glycosylation system. Proc Natl Acad Sci USA 108: 9643-9648.
18. Power PM, Seib KL, Jennings MP (2006) Pilin glycosylation in Neisseria meningitidis occurs by a similar pathway to wzy-dependent O-linked biosynthesis in Escherichia coli. Biochem Biophys Res Commun 347: 904-908.
19. Vik A, Aas FE, Anonsen JH, Bilsborough S, Schneider A, et al. (2009) Broad spectrum O-linked protein glycosylation in the human pathogen Neisseria gonorrhoeae. Proc Natl Acad Sci U S A 106: 4447-4452.
20. Kahler CM, Martin LE, Tzeng YL, Miller YK, Sharkey K, et al. (2001) Polymorphisms in pilin glycosylation Locus of Neisseria meningitidis expressing class II pil. Infect Immun 69: 3597-3604.
21. Rudd PM, Joao HC, Coghil E, Fstellen P, Saunders MR, et al. (1994) Glycophobolates modify the dynamic stability and functional activity of an enzyme. Biochemistry 33: 17-22.
22. Charnot-Rooke J, Rousseau B, Lantermier F, Mikiyq G, Mairtry E, et al. (2007) Alternative Neisseria spp. type IV pilin glycosylation with a glucosamine acetamido trideoxyhexose residue. Proc Natl Acad Sci U S A 104: 14763-14789.
23. Faridmoayer A, Fentabili MA, Haurat MF, Yi W, Woodward R, et al. (2008) Extreme substrate promiscuity of the Neisseria oligosaccharyltransferase involved in protein O-glycosylation. J Biol Chem 283: 34596-34604.
24. Yamamoto T (2010) Marine bacterial sialytransferases. Mar Drugs 7: 2781-2794.
25. Mine T, Kataya S, Kajiwara H, Tsunashima M, Tsukamotsu H, et al. (2010) An alpha2,6-sialytransferase cloned from Photobacterium leiognathi strain JT-SHIZ-119 shows both sialytransferase and neuraminidase activity. Glycobiology 20: 158-165.
26. Tsukamoto H, Takakura Y, Yamamoto T (2007) Purification, cloning, and expression of an alpha/beta-galactoside alpha-2,3-sialytransferase from a luminous marine bacterium, Photobacterium phosphoreum. J Biol Chem 282: 29794-29802.
27. Hashimoto K, Tomitotsu T, Kawanoo S, Yoshizawa SC, Okuda S, et al. (2009) Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans. Carbohydr Res 344: 881-887.
28. Kumar M, Balaji PV (2011) Comparative genomics analysis of completely sequenced microbial genomes reveals the ubiquity of N-linked glycosylation in prokaryotes. Mol Biosyst 7: 1629-1645.
29. Magidovich H, Eichler J (2009) Glycosyltransferases and oligosaccharyltransferases in Archaea: putative components of the N-glycosylation pathway in the third domain of life. FEMS Microbiol Lett 300: 122-130.
30. Morrison MJ, Imperiali B (2013) Biosynthesis of UDP-N,N'-diacyetamidoacetamido in Acinetobacter baumannii: Biochemical characterization and correlation to existing pathways. Arch Biochem Biophys 536: 72-80.
31. Allschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1999) Basic local alignment search tool. J Mol Biol 215: 403-410.
32. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.
33. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 215-217.
34. Iigura M, Maita N, Kamishikiryo J, Yamada M, Obika T, et al. (2008) Structure-guided identification of a new cataytic motif of oligosaccharyltransferase. EMBO J 27: 234-243.
35. Logan SM (2006) Flagellar glycosylation - a new component of the motility repertoire? Microbiology 152: 1249-1262.
36. Josenhans C, Vossebeen L, Friedrich S, Suerbaum S (2002) The neuA/pod gene cluster of Helicobacter pylori is involved in flagellar biosynthesis and flagellin glycosylation. FEMS Microbiol Lett 210: 165-172.
37. Takagi F, Hishita S, Suzuki T, Ogawa Y, Aizawa S, et al. (2008) Effects of glycosylation on swimming ability and flagellar polymorphic transformation in Pseudomonas syringae pv. tabaci 6605. J Bacteriol 190: 764-768.
38. Arora SK, Neely AN, Blair B, Lory S, Ramphal R (2005) Role of motility and flagellin glycosylation in the pathogenesis of Pseudomonas aeruginosa burn wound infections. Infect Immun 73: 4395-4398.
39. Moran NA (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. Curr Opin Microbiol 6: 512-518.
40. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct 4: 13.

41. Solá RJ, Griebenow K (2009) Effects of glycosylation on the stability of protein pharmaceuticals. J Pharm Sci 98: 1223-1245.

42. Goldstein EJ, Citron DM, Peraino VA, Cross SA (2003) Desulfovibrio desulfuricans bacteremia and review of human Desulfovibrio infections. J Clin Microbiol 41: 2752-2754.

43. Hug I, Feldman MF (2011) Analogies and homologies in lipopolysaccharide and glycoprotein biosynthesis in bacteria. Glycobiology 21: 138-151.

44. Nothaft H, Szymanski CM (2013) Bacterial protein N-glycosylation: new perspectives and applications. J Biol Chem 288: 6912-6920.

45. Virji M (2009) Pathogenic neisseriae: surface modulation, pathogenesis and infection control. Nat Rev Microbiol 7: 274-286.

46. Isambert H, Stein RR (2009) On the need for widespread horizontal gene transfers under genome size constraint. Biol Direct 4: 28.

47. Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36: 6688-6719.

48. Yanai I, Camacho C, DeLisi C (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys Rev Lett 85: 2641-2644.

49. Moxon ER, Rainey PB, Nowak MA, Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol 4: 24-33.

50. Anonsen JH, Egge-Jacobsen W, Aas FE, Barud B, Koomey M (2012) Novel protein substrates of the phospho-form modification system in Neisseria gonorrhoeae and their connection to O-linked protein glycosylation. Infection and Immunity 80:22–30.