An improved method for specificity annotation shows a distinct evolutionary divergence among the microbial enzymes of the cholylglycine hydrolase family

Priyabrata Panigrahi, Manas Sule, Ranu Sharma, Sureshkumar Ramasamy and C. G. Suresh

Division of Biochemical Sciences, CSIR-National Chemical Laboratory, Dr Homi Bhabha Road, Pune 411008, India

Received 23 January 2014
Accepted 14 March 2014

Bile salt hydrolases (BSHs) are gut microbial enzymes that play a significant role in the bile acid modification pathway. Penicillin V acylases (PVA) are enzymes produced by environmental microbes, having a possible role in pathogenesis or scavenging of phenolic compounds in their microbial habitats. The correct annotation of such physiologically and industrially important enzymes is thus vital. The current methods relying solely on sequence homology do not always provide accurate annotations for these two members of the cholylglycine hydrolase (CGH) family as BSH/PVA enzymes. Here, we present an improved method [binding site similarity (BSS)-based scoring system] for the correct annotation of the CGH family members as BSH/PVA enzymes, which along with the phylogenetic information incorporates the substrate specificity as well as the binding site information. The BSS scoring system was developed through the analysis of the binding sites and binding modes of the available BSH/PVA structures with substrates glycocholic acid and penicillin V. The 198 sequences in the dataset were then annotated accurately using BSS scores as BSH/PVA enzymes. The dataset presented contained sequences from Gram-positive bacteria, Gram-negative bacteria and archaea. The clustering obtained for the dataset using the method described above showed a clear distinction in annotation of Gram-positive bacteria and Gram-negative bacteria. Based on this clustering and a detailed analysis of the sequences of the CGH family in the dataset, we could infer that the CGH genes might have evolved in accordance with the hypothesis stating the evolution of diderms and archaea from the monoderms.

INTRODUCTION

Bile salt hydrolase (BSH) and penicillin V acylase (PVA) are pharmacologically important classes of enzymes belonging to the cholylglycine hydrolase (CGH) family (Pfam: PF02275). BSH enzymes are widely distributed among gut-inhabiting microbes (Jones et al., 2008), whereas PVA enzymes are found mainly among soil and aquatic microbes (Olsson & Uhlén, 1986; Rathinaswamy et al., 2012; Avinash et al., 2013; Mukherji et al., 2013). Although BSH and PVA enzymes prefer chemically distinct substrates (bile salts and penicillin V, respectively), both cleave a similar amide bond in their substrates.

Bile salts are one of the main components of bile, which is synthesized in the liver from cholesterol, stored in the gallbladder and released into the duodenum for lipid digestion. BSH enzymes protect gut microbes from bile salt toxicity by deconjugating the glycine- and taurine-conjugated bile salts into bile acids and their corresponding amino acid residues, whilst benefiting the host by regulating the delicate balance between cholesterol and bile acid pool. BSH-active bacterial probiotics have been found to be beneficial in lowering serum cholesterol levels (Begley et al., 2006).

PVA enzymes hydrolyse the amide bond in penicillin V yielding phenoxy acetic acid and 6-aminopenicillanic acid as products. 6-Aminopenicillanic acid is used in the pharmaceutical industry for the synthesis of semi-synthetic antibiotics that are more effective against pathogenic bacteria than the naturally occurring penicillins (Rathinaswamy et al., 2012). Although the physiological role of PVA genes is not clear, existing evidence suggests their relation to...
pathogenesis (Kovacikova et al., 2003) as well as aromatic compound degradation pathways (Valle et al., 1991).

Both BSH and PVA enzymes contain the $\alpha/\beta$ Ntn-hydrolase fold (Oinonen & Rouvinen, 2000). They belong to the N-terminal nucleophile hydrolase superfamily with a cysteine as the N-terminal amino acid residue. The catalytic residues Cys2, Arg18, Asp21, Asn175 and Arg228 are conserved, and therefore the mechanisms of hydrolytic reactions in both enzymes are similar (Lodola et al., 2012). Due to a high degree of similarity, they are annotated under a single family in public domain databases like Pfam (family CBAH; Punta et al., 2012), CDD (family Ntn_CGH_like; Marchler-Bauer et al., 2013) and MEROPS (family C59; Rawlings et al., 2012). Sometimes the family members are wrongly annotated, i.e. BSH enzymes annotated as PVA or vice versa. In the CDD database it is observed that the experimentally characterized BSH enzymes from Bifidobacterium longum and Clostridium perfringens have been annotated incorrectly as PVA enzymes (Ntn_PVA family). This issue was addressed previously by Lambert et al. (2008) for Gram-positive BSH/PVA enzymes. Using phylogenetic profiling and molecular modelling, Lambert et al. (2008) correctly annotated the BSH and PVA enzymes from Gram-positive bacteria. However, members from Gram-negative bacteria and archaea were not considered in their analysis.

In the dataset used for the present analysis, we have incorporated experimentally characterized members such as BSH enzymes from Bifidobacterium longum (BBSSH; Kumar et al., 2006), Clostridium perfringens (CpBSH; Coleman & Hudson, 1995) and Bacteroides thetaiotaomicron (BtBSH; Stellwag & Hylemon, 1976), and PVA enzymes from Bacillus subtilis (BspPVA; Ratnawaswamy et al., 2012), Bacillus sphaericus (BspPVA; Olsson & Uhlen, 1986) and Pectobacterium atrosepticum (PaPVA; Avinash et al., 2013) as well as other uncharacterized BSH/PVA enzymes from Gram-positive bacteria, Gram-negative bacteria and archaea. The enzymes BBSSH, CpBSH, BspPVA and BsuPVA belong to Gram-positive bacteria, whereas BtBSH and PaPVA are from Gram-negative bacteria. The entire list of sequences included in the analysis is given in Table S1 (available in the online Supplementary Material).

The initial phylogenetic analysis of the sequences in the dataset failed to annotate the family members as BSH/PVA enzymes. This inaccuracy thus highlighted the need to develop a better annotation method not based solely on phylogenetic information, but also considering the binding site characteristics as well as substrate specificity information in order to improve the current annotations of the available sequences and to correctly annotate any new members. Using the available structures, a comparison of the binding sites and prediction of substrate-binding modes was carried out by docking and molecular dynamics simulation studies. With the information generated by the above analysis, a binding site similarity (BSS)-based scoring system was developed which helped to annotate correctly CGH family members as BSH or PVA enzymes. The accuracy of annotation of the BSS scoring system was tested against 19 experimentally characterized CGH enzymes as well as the annotation provided previously by Lambert et al. (2008). Lastly, we discuss the evolution of CGH family members and their relationship with the evolution of Gram-positive bacteria, Gram-negative bacteria and archaea.

**METHODS**

**Retrieval of CGH family members and phylogenetic analysis.** Sequences of CGH family members were retrieved from the National Center for Biotechnology Information non-redundant protein database by performing a BLAST search (Altschul et al., 1997) using experimentally characterized BBSSH, CpBSH, BspPVA, BsuPVA, BBSSH and PaPVA protein sequences as queries. These six sequences were considered as 'template sequences' for building the entire dataset. Minimum BLAST score cut-off was kept at 500 and only the best hits from each organism were chosen for further analysis. Sequences lacking the conserved catalytic, nucleophilic Cys residue at their N-terminal were considered to be inactive and were excluded from the analysis. Sequences were clustered at the 60% identity threshold and the non-redundant set containing 198 sequences thus generated was used as the final dataset. The dataset also included the six template enzymes. Multiple sequence alignment was done by using Clustal X (Thompson et al., 1997), whilst MEGA5.2 (Tamura et al., 2011) was used to reconstruct a phylogenetic tree of the CGH family by the neighbour-joining method with a bootstrap value of 1000.

**Prediction of substrate-binding modes using docking analysis.** Prediction of substrate-binding modes among five enzymes of the CGH family with known structures (BBSSH, CpBSH, BtBSH, BspPVA and BsuPVA) was carried out using docking studies. The substrate binding modes in the PaPVA homology model have been reported previously (Avinash et al., 2013). The 3D structures of BBSSH, CpBSH, BtBSH, BspPVA and BsuPVA (PDB IDs: 2HF0, 2RLC, 3HBC, 3PAV and 2OQC) were downloaded from the Protein Data Bank (PDB; Berman et al., 2000). Residues 48–49, 157–162 and 271–273 belonging to the putative substrate-binding site in the BBSSH structure (PDB ID: 3HBC) were found to be missing. These regions were modelled by taking the BBSSH structure as template using Prime (version 3.0; Schrödinger, Ltd). Glycocalic acid (GCA, a bile salt) and penicillin V were used as substrates. The grid-based rigid receptor and flexible ligand-docking program Glide (Friesner et al., 2006) was used to predict accurately the binding modes of ligand in the receptor-binding site. Potential binding sites on each receptor were identified using SiteMap (Halgren, 2007). The topology and parameters for ligands were generated with a general Amber force field using Acype (Sousa da Silva & Vranken, 2012). Dynamics and stability of each receptor–ligand complex were evaluated by conducting explicit solvent molecular dynamics simulations of each complex on a 5 ns timescale using an Amber force field in Gromacs 4.5 (Pronk et al., 2013).

**Estimation of BSS scores for all CGH sequences.** A binding site profile-based scoring system (Fig. S1) was developed to estimate quantitatively the BSS of each CGH family member within the dataset, utilizing the binding site information from each of the six template enzymes (BBSSH, CpBSH, BtBSH, BspPVA, BsuPVA and PaPVA). Each query sequence from the dataset was aligned with the multiple sequence alignment profile of the six template enzymes. Only the binding site positions (corresponding to residues 20–25, 57–67, 79–83, 102 and 127–140 in BBSSH) of the resulting alignment were considered for scoring. The score of the query sequence with the i\*th template (i=1–6) was calculated as:
in Cluster2, the available information was unable to annotate correctly the members of this cluster as BSH/PVA. As mere phylogenetic analysis was not enough to annotate correctly the BSH/PVA sequences, a better annotation method was developed which not only included the phylogenetic information, but also took into consideration the binding site and substrate specificity information of the BSH/PVA enzymes.

**Analysis of the substrate specificity and binding site properties of CGH enzymes**

The enzymes BbBSH, CpBSH, BtBSH, BspPVA, BsuPVA and PaPVA are known to exhibit variations in their substrate specificity, i.e. they vary from being a classic BSH (no PVA activity) to a classic PVA enzyme (no BSH activity). Among the Gram-positive bacteria members, BbBSH is a classic enzyme with only BSH activity (Kumar et al., 2006), whereas BsuPVA is a classic PVA enzyme with only PVA activity (Rathinaswamy et al., 2012). Between these two extremes, CpBSH is a BSH enzyme with low PVA activity and BspPVA is a PVA enzyme with low BSH activity. Quantitatively, the PVA activity of CpBSH is 11% that of BspPVA and the BSH activity of BspPVA is 20% that of BbBSH (Kumar et al., 2006). Among the characterized enzymes from Gram-negative bacteria, BtBSH has BSH activity, whilst its PVA activity has not been verified experimentally (Stellwag & Hylemon, 1976). PaPVA is a Gram-negative classic PVA enzyme (Avinash et al., 2013). Except for PaPVA, the tertiary structure has been determined for the other five enzymes. Except for CpBSH, all determined structures are the apo-form of the enzyme without any bound substrate molecule. The CpBSH structure (2RLC) shows the enzyme bound with its product glycine and cholate (Rossocha et al., 2005) (Fig. 2a).

A structural comparison of the available structures showed similar positional preference of their catalytic residues in the active site region. The catalytic framework observed in the CpBSH structure is shown in Fig. 2(b). However, the enzymes show significant variation in terms of size and properties of their binding site pockets. This variation is due to differential folding and conformations of the loops near the binding site (Fig. 2c). The substrate binding sites in these enzymes consist mainly of four loops, i.e. loop1–loop4, comprising residues 22–27, 58–65, 129–139 and 261–269, respectively, in BbBSH. In the BbBSH structure, coordinates for residues 48–49 of loop1, 157–162 of loop3 and 271–273 of loop4 were found to be missing, presumably due to the disorder of these highly dynamic loops. Therefore, these loop regions were modelled using BbBSH as the template.

Out of the four binding site loops in these enzymes, loop2–loop4 show significant differences in terms of their folding and conformation (Fig. 2c). Loop3 in PVA-type enzymes (BspPVA and BsuPVA) is oriented more inside the cavity compared with BSH-type enzymes (BbBSH, CpBSH and BtBSH), reducing the size of the binding site pockets. Loop2 in BtBSH is oriented more into the cavity as compared with the others, thereby shifting the binding site

**RESULTS AND DISCUSSION**

**Dataset generation and phylogenetic analysis of the BSH/PVA sequences**

A dendrogram prepared based on the phylogenetic analysis of the 198 sequences in the dataset (Table S1) resulted in the formation of two distinct clusters (Fig. 1a): Cluster1 (75 sequences) and Cluster2 (123 sequences). The majority of the sequences in Cluster1 belonged to Gram-positive bacteria (phyllum Firmicutes: 60; Actinobacteria: 12), whilst three were from archaea (Methanobacterium formicicum, Methanobrevibacter smithii and Methanosphaera stadtmanae). Cluster2 included a majority of the sequences from Gram-negative bacteria (phyllum Proteobacteria: 65; Bacteroidetes: 24; Cyanobacteria: 12; Planctomycetes: 6; and three from other phyla), whilst 10 were from Gram-positive Actinobacteria and three were archaeal sequences (Natratella aquyptia, Natrinema gari and Methanoplanus petrolearius). Thus, Gram-positive and archaeal members were distributed across both clusters with the majority of Gram-positive sequences grouped in Cluster1, whereas Gram-negative sequences were found only in Cluster2. Among the 10 Gram-positive Actinobacteria of Cluster2, eight actually belong to the order Corynebacterineae (Fig. 1a), which are the intermediates between true Gram-positive and Gram-negative bacteria (Gupta, 2011). These intermediates show positive Gram staining but also have an additional highly ordered layer of mycolic acid resembling the outer membrane of true Gram-negative bacteria (Gupta, 2011). The other two Gram-positive Actinobacteria in Cluster2 are Kitasatospora setae and Streptomyces sp. MgI belonging to the family Streptomycetaceae.

The experimentally characterized BSH (BbBSH and CpBSH) and PVA (BspPVA and BsuPVA) enzymes from Gram-positive bacteria belonged to Cluster1, whereas the BSH (BtBSH) and PVA (PaPVA) enzymes from Gram-negative bacteria belonged to Cluster2 (Fig. 1a). In Cluster1, the BSH enzymes CpBSH and BSSH were distributed in two different branches; CpBSH was observed to be grouped along with PVA enzymes (BsuPVA and BspPVA). Similarly,
pocket and increasing its solvent accessibility. These results are summarized quantitatively in Table 1, which describes the binding site volume, solvent accessibility, and hydrophobic and hydrophilic properties in these enzymes. As compared with PVA-type enzymes (\textit{BspPVA} and \textit{BsuPVA}), the BSH-type enzymes (\textit{BlBSH}, \textit{CpBSH} and \textit{BtBSH}) were observed to have a larger, more exposed and hydrophilic binding site (Table 1), in order to accommodate the bile salt molecule, which is larger than penicillin V.

**Mode of substrate binding among CGH enzymes**

**Modes of GCA binding.** GCA is a bile salt molecule synthesized in the liver, formed by the conjugation of a

---

**Fig. 1.** (a) Dendrogram prepared based on the phylogenetic analysis of CGH family sequences. Two distinct clusters are labelled as Cluster1 and Cluster2. Members are coloured according to their source (red, Gram-positive bacteria; blue, Gram-negative bacteria; pink, \textit{Corynebacterineae}; green, archaea). Experimentally characterized BSH and PVA enzymes of each cluster are labelled. (b) BSS-based annotation of CGH family members into BSH or PVA enzymes. Green, BSH; brown, PVA; black, BSH/PVA. (c) Line plots depicting the estimated BSS scores of each CGH family member with six template enzymes (\textit{Bl}, \textit{BlBSH}; \textit{Cp}, \textit{CpBSH}; \textit{Bsp}, \textit{BspPVA}; \textit{Bsu}, \textit{BsuPVA}; \textit{Bt}, \textit{BtBSH}; \textit{Pa}, \textit{PaPVA}). Separate plots are drawn for each subgroup. The \(y\)-axes correspond to the BSS scores.
cholic acid moiety to the amino acid glycine through an amide bond (Fig. 3a). The N-terminal Cys residue of CGH enzymes carries out a nucleophilic attack on this scissile amide bond to release glycine (leaving group) while the cholate moiety remains bound to the enzyme as the acyl-enzyme adduct (Lodola et al., 2012). In the crystal structure of CpBSH (PDB ID: 2RLC), the adduct group cholate occupies the binding site (siteA; formed by four loops: red). (b) Geometrical arrangement of the catalytic residues in the active site of CpBSH. During catalysis, the N-terminal Cys2 acts as both a nucleophile and base. Arg18 and Asp21 form hydrogen-bonding interactions with Cys2, whereas Asn82 and Asn175 form the putative oxyanion hole. Arg228 helps in transition-state stabilization (Lodola et al., 2012). This arrangement of catalytic residues remains conserved in all CGH enzymes. (c) Superposition of the four substrate-binding site loops (loop1–loop4) from BlBSH (red), CpBSH (magenta), BtBSH (yellow), BspPVA (blue) and BsuPVA (green). The differential folding and conformations of the above-defined loops in these enzymes can be seen, resulting in variance in the size of their binding site pockets. The active site Cys residue is shown and labelled.

In all BSH-active enzymes (BlBSH, CpBSH, BtBSH and BspPVA), the modes of GCA binding were in agreement with the binding mode seen in 2RLC wherein the adduct group cholate occupies siteA whilst the leaving group glycine occupies siteL. The BlBSH–GCA complex is shown in Fig. 3(c) and other complex structures are illustrated in Fig. S2. It was observed that GCA binding shows a directional preference for the amide bond orientation (CO–N) in the direction from siteA to siteL, with reference

### Table 1. SiteMap quantitative estimation of binding site properties of CGH enzymes

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Volume (Å³)*</th>
<th>Exposure†</th>
<th>Hydrophobic</th>
<th>Hydrophilic</th>
<th>Hydrophobic/hydrophilic‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>BspPVA</td>
<td>153.32</td>
<td>0.31</td>
<td>3.98</td>
<td>0.46</td>
<td>8.61</td>
</tr>
<tr>
<td>BsuPVA</td>
<td>344.37</td>
<td>0.46</td>
<td>2.55</td>
<td>0.59</td>
<td>4.34</td>
</tr>
<tr>
<td>CpBSH</td>
<td>485.35</td>
<td>0.57</td>
<td>0.91</td>
<td>1.04</td>
<td>0.87</td>
</tr>
<tr>
<td>BlBSH</td>
<td>553.77</td>
<td>0.45</td>
<td>0.95</td>
<td>1.07</td>
<td>0.88</td>
</tr>
<tr>
<td>BtBSH</td>
<td>718.58</td>
<td>0.60</td>
<td>0.38</td>
<td>1.04</td>
<td>0.37</td>
</tr>
</tbody>
</table>

*Binding site volume (Å³) measured using the shrink-wrap approach of SiteMap.
†Exposure is a measure of how open the site is to solvent; the higher the value, the more exposed it is. It is calculated as a ratio and does not carry any units.
‡Hydrophilic and hydrophobic terms are a measure of the hydrophilic and hydrophobic nature of the site; the higher the ratio of hydrophobic to hydrophilic values, the more hydrophobic the site is. It is calculated as a ratio and does not carry any units.
Favourable values of free energy of binding along with shorter and stable nucleophilic attack distance values during dynamics suggest the suitability of these binding modes for BSH activity in these enzymes (Table 2, Fig. S2f).

Interestingly, even in BsuPVA, a BSH-inactive enzyme, the predicted mode of GCA binding was found to be similar to that of all BSH-active enzymes (Fig. S2e). However, in the modelled structure of the BSH-inactive PaPVA enzyme, the GCA molecule was predicted to bind in a reversed amide bond orientation, from siteL to siteA (Avinash et al., 2013).

Polar complementarity: probable basis for GCA specificity. GCA is a planar amphipathic molecule with its hydrophobic surface consisting of the methyl groups of the steroid ring, whilst the hydrophilic surface is formed by the three hydroxyl groups (3α-, 7α- and 12α-OH; Fig. 3a). Structural analysis of all predicted enzyme–GCA complex structures revealed an important correlation (Fig. 4a) between GCA specificity and the degree of hydrophilic complementarity of the three hydroxyl groups. Using the radial distribution function, the maximum probability density of receptor polar atoms within 5 Å of each hydroxyl groups was estimated (Fig. 4a, Table 3). Furthermore, the hydrogen-bonding interaction of these hydroxyl groups with nearest receptor polar groups during dynamics was also assessed quantitatively (Table 3). The radial distribution function $g(r)$ for the polar complementarity was fixed at 0.5 for the analysis. Values $>0.5$ were considered to have polar complementarity at the particular hydroxyl group.

In B/BSH, polar complementarity was observed for all three hydroxyl groups along with good hydrogen-bonding

![Fig. 3.](image-url) (a) GCA (a bile salt) and (b) penicillin V. The scissile amide bonds that are hydrolysed by the CGH enzymes are marked with a line. The leaving and adduct groups in each substrate along with the three polar hydroxyl groups (3α-, 7α- and 12α-OH) of GCA are labelled. PAA, phenoxy acetic acid; 6-APA, 6-aminopenicillanic acid. (c) Mode of binding of GCA in B/BSH. (d) Mode of binding of penicillin V in BsuPVA. In both complexes, the adduct groups occupy siteA, whilst directing the leaving groups towards the active site (siteL), positioning the scissile amide bond just inside the cleft of the enzyme, close to the N-terminal Cys residue, in an orientation favourable for the nucleophilic attack. The four substrate-binding loops are shown in cartoon representation. The modes of binding of GCA and penicillin V among other CGH enzymes are shown in Figs S2 and S3, respectively.
interactions (Table 3). The values for both these factors at the hydroxyl positions 7α- and 12α-OH in CpBSH might support the fact that the activity of this enzyme is lower than BbBSH (Kumar et al., 2006). In the case of BtBSH, although no comparative experimental evidence is available, the calculated values at 3α- and 12α-OH suggest that the activity in this case could also be lower than that of BbBSH. In BspPVA, polar complementarity values at all positions are >0.5, but hydrogen-bonding interactions only at the 3α-OH advocate the available experimental evidence showing that it has low levels of BSH activity (Kumar et al., 2006). However, in BsuPVA, only the 3α-OH group shows complementarity with a low percentage of hydrogen-bonding interactions at the same site, which could be the reason for being BSH-inactive. The significance of these hydroxyl groups in contributing to binding

**Table 2. Summary of free energy of binding (GlideScore) of all predicted protein–ligand complex structures**

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>GCA</th>
<th>Penicillin V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GlideScore (kcal mol⁻¹)</td>
<td>Nucleophilic attack distance (Å) (mean ± sd)</td>
</tr>
<tr>
<td>BbBSH</td>
<td>−12.19</td>
<td>4.4 ± 0.3</td>
</tr>
<tr>
<td>CpBSH</td>
<td>−9.58</td>
<td>4.4 ± 0.2</td>
</tr>
<tr>
<td>BtBSH</td>
<td>−7.41</td>
<td>4.4 ± 0.1</td>
</tr>
<tr>
<td>BspPVA</td>
<td>−10.96</td>
<td>4.5 ± 0.2</td>
</tr>
<tr>
<td>BsuPVA</td>
<td>−8.85</td>
<td>3.9 ± 0.2</td>
</tr>
</tbody>
</table>

**Fig. 4.** (a) Radial distribution of receptor polar atoms around three hydroxyl groups of GCA (3α-, 7α- and 12α-OH) in BbBSH (black), CpBSH (red), BtBSH (magenta), BspPVA (green) and BsuPVA (blue). The y-axes correspond to the probability density of receptor polar atoms [g(r)] and the x-axes correspond to distance (in Å). Polar complementarity at the respective hydroxyl group has been estimated at a 5 Å distance threshold as the maximum probability of finding receptor polar atoms [max. g(r)] within 5 Å. Except for BsuPVA, a BSH-inactive enzyme, polar complementarity is observed at more than one hydroxyl group in all BSH-active enzymes. In BsuPVA, the polar complementarity is observed around only 3α-OH. (b) Box plot depicting the distribution of angles between the phenyl ring planes of substrate penicillin V and aromatic residues in its vicinity.
affinity is further supported by Batta et al. (1984). Hence, polar complementarity constitutes an important factor influencing the GCA specificity among CGH enzymes. Indel mutations of polar residues in the loop3 region in PVA enzymes can be considered as an engineering strategy for producing GCA specificity in these enzymes.

**Modes of penicillin V binding among CGH enzymes.** In all CGH enzymes, whether PVA-active (BspPVA, BsuPVA, CpBSH) or PVA-inactive (BtBSH), and also in BtBSH, the mode of penicillin V binding was observed to be similar. The BsuPVA–penicillin V complex structure is shown in Fig. 3(d) and other penicillin V complexes are depicted in Fig. S3. The adduct group phenoxy acetic acid occupies siteA and the leaving group 6-aminopenicillanic acid occupies siteL. Consequently, establishing a directional preference for the amide bond (CO–N) with respect to the nucleophilic residue Cys1. All complexes remained stable during simulation with shorter and stable nucleophilic attack distance values (Table 2). During the simulation of the BsuPVA–penicillin V complex, penicillin V was observed to form a hydrogen bond with Arg228, Asn175 and Cys1 residues (~98, 45 and 14 %, respectively). This observation supports the proposed reaction mechanism of Cys–Ntn hydrolases, in which the Arg228 residue is shown to have a direct role in transition-state stabilization and Asn175 is shown to have a role for substrate recognition through hydrogen bonding (Lodola et al., 2012).

**Aromatic interactions in the active site might influence penicillin V binding.** The presence of aromatic–aromatic interactions between the phenyl ring of penicillin V and aromatic residues in its vicinity is deemed important in all enzyme–penicillin V complexes (Fig. S4). In BspPVA and BsuPVA, Phe21 and Tyr82 residues interact with the phenyl ring of penicillin V in a strict geometry. This Phe21, Tyr82 aromatic pair is substituted with (Trp21, Asn81), (Ile22, Asn82) and (Trp46, Phe105) pairs in BtBSH, CpBSH and BtBSH, respectively. In BtBSH, the possible involvement of Trp residues in penicillin V binding is evident from experimental data, which show quenching of tryptophan fluorescence as a result of penicillin V binding (Kumar et al., 2006). In CpBSH, although Phe21 is substituted by Ile22, Tyr24 is involved in aromatic interaction with penicillin V, which compensates for the loss of Phe21. As in PVA enzymes, in BtBSH, both Trp46 and Phe105 interact with the phenyl ring of penicillin V in a strict geometrical arrangement, suggesting that BtBSH might show a low degree of PVA activity.

Participation of aromatic residues in penicillin V binding has also been shown experimentally in PaPVA, where the Trp23 and Trp87 aromatic pair is known to interact with the phenyl ring of penicillin V (Avinash et al., 2013). Fig. 4(b) shows the distribution of the planar angle between the phenyl ring of penicillin V and the aromatic planes in its vicinity. The PVA enzymes, BsuPVA and BspPVA, show less deviation in these planar angles compared with the BSH enzymes, BtBSH and CpBSH, resulting in firm binding of the penicillin V molecule and thus showing higher PVA activity. In BtBSH, similar to BspPVA, only a slight deviation was observed, suggesting possible penicillin V binding affinity. It is probable that these aromatic residues might interact with the incoming substrate through stacking interactions and help it to initially orient favourably in the binding site, influencing the binding affinity.

In summary, it was observed that among CGH enzymes, substrate binding (bile salt or penicillin V) involves a directional and orientational preference of adduct and leaving groups, and therefore the scissile amide bond direction, with respect to the position of the nucleophile Cys residue. The polar complementarities of the three hydroxyl groups of the GCA molecule might also influence its binding affinity with these enzymes. Similarly, the presence of aromatic residues in the active site and their arrangement with respect to the phenyl ring of bound penicillin V might play a decisive role in penicillin V binding and affinity.

**Substrate specificity annotation of family members**

Enzymes possessing similar residues in their binding site pocket may have similar mechanisms of action and substrate preferences (Haupt et al., 2013). Based on this assumption, each of the 198 sequences was assigned a total of six numeric scores (BSS scores) corresponding to their BSS with BtBSH, CpBSH, BspPVA, BsuPVA, BtBSH and PaPVA. Based on initial phylogenetic clustering and the

---

### Table 3. Quantitative estimation of polar complementarities for the three hydroxyl groups of the GCA molecule and percentage of times their involvement in hydrogen-bonding interactions (%Hbond) during molecular dynamics simulation of each enzyme–GCA complex

<table>
<thead>
<tr>
<th>Property</th>
<th>Hydroxyl group</th>
<th>BtBSH</th>
<th>CpBSH</th>
<th>BtBSH</th>
<th>BspPVA</th>
<th>BsuPVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. prob. density within 5 Å (estimated by radial distribution function) 12x</td>
<td>0.73</td>
<td>0.77</td>
<td>0.60</td>
<td>1.05</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>7x</td>
<td>0.55</td>
<td>1.80</td>
<td>0.30</td>
<td>0.51</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>3x</td>
<td>1.02</td>
<td>1.28</td>
<td>0.96</td>
<td>3.18</td>
<td>3.18</td>
<td></td>
</tr>
<tr>
<td>%Hbond</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12x</td>
<td>21.67</td>
<td>4.63</td>
<td>16.39</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7x</td>
<td>4.83</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3x</td>
<td>4.51</td>
<td>6.95</td>
<td>72.07</td>
<td>100</td>
<td>8.47</td>
<td></td>
</tr>
</tbody>
</table>
calculated BSS scores, the 198 CGH enzyme sequences were annotated into five subgroups using a cut-off of score difference set at 30 between the highest scores with BSH and PVA enzymes. During the analysis, sequences belonging to Cluster1 could be annotated easily as BSH/PVA (Fig. 1b) on the basis of the score difference of >30. The subgroups in Cluster1 were defined as Cluster1-BSH (members scoring highest with Cluster1 BSH enzymes; either BtBSH or CpBSH, or both) and Cluster1-PVA (scoring highest with Cluster1 PVA enzymes; either BspPVA or BsuPVA, or both) (Fig. 1c). During the annotation of Cluster2 sequences into BSH/PVA, a few sequences showed similar scores for both the Cluster2 BSH enzyme BtBSH as well as the Cluster2-PVA enzyme PaPVA. Sequences in Cluster2 were annotated into three subgroups. The first two subgroups showing a score difference of >30 were defined as Cluster2-BSH (scoring highest with Cluster2 BSH enzyme BtBSH) and Cluster2-PVA (scoring highest with Cluster2 PVA enzyme PaPVA), whilst the third was defined as Cluster2-BSH/PVA (scoring highest with both BtBSH and PaPVA, a score difference of <30) (Fig. 1b, c).

Amongst the 75 Cluster1 sequences, 59 were annotated as BSH and 16 as PVA enzymes, whilst in the 123 Cluster2 sequences, 21 were annotated as BSH, 49 as PVA and 53 as BSH/PVA enzymes. The detailed BSS scores and the resulting annotations are given in Table S1.

The BSS-based scoring system was validated using experimentally verified BSH/PVA enzymes from Gram-positive bacteria, Gram-negative bacteria and archaea to check the accuracy of the annotations predicted (Table S2). Amongst the experimentally characterized Gram-positive BSH enzymes, those from the Firmicutes and Actinobacteria were predicted correctly as Cluster1-BSH enzymes. Similarly, the PVA enzyme from Listeria monocytogenes EGDe was annotated correctly as a Cluster1-PVA enzyme. In the case of Gram-negative bacteria, the BSH enzymes from Brucella abortus and Bacteroides vulgatus were annotated correctly as Cluster2-BSH enzymes. Similarly, the known BSH archaeal enzymes from Methanosphaera stadtmanae and Methanobrevibacter smithii were also predicted correctly as Cluster1-BSH enzymes. The BSS scoring system was further validated against the Gram-positive CGH enzymes annotated previously as BSH/PVA by Lambert et al. (2008). The BSS-based functional assignment was found to be in agreement with the earlier annotations (Table S3).

In Cluster1, most of the enzymes annotated as BSH enzymes were found to belong to the gut-inhabiting bacteria (Firmicutes and Actinobacteria) and archaea (Methanobacterium fumaricum, Methanobrevibacter smithii and Methanosphaera stadtmanae). The enzyme from Planococcus antarcticus, an environmental bacterium isolated from cyanobacterial mat samples in the lakes of Antarctica (Reddy et al., 2002), was also classified as a BSH enzyme. Interestingly, the enzymes annotated in Cluster2 as BSH enzymes were distributed widely among both gut-inhabiting bacteria and environmental microbes (e.g. Burkholderia sp. Y123, Blastopirellula marina, Desulfovibrio fructosovorans and Rhodococcus vannielii). In addition, the Cluster2 enzyme from Rickettsia felis, a pathogen causing flea-borne spotted fever in cats and which is also known to infect humans, was annotated as a BSH enzyme. The presence of BSH enzymes among pathogenic bacteria such as L. monocytogenes and Enterococcus faecalis (an opportunistic pathogen) was reported previously (Begley et al., 2006). The presence of BSH genes among the gut-inhabiting micro-organisms can be attributed to their role in bile acid resistance, thereby protecting these organisms in the host gastrointestinal tract (Jones et al., 2008). However, the physiological roles of the BSH genes among pathogens such as Rickettsia felis, environmental bacteria such as Burkholderia sp. and others still need to be explored.

The enzymes annotated as PVA in the dataset were found to be distributed widely among both pathogen and environment degrading organisms. Pathogenic bacteria include Pectobacterium, Agrobacterium, Brevundimonas, Bordetella, Acinetobacter, Yersinia, Proteus, Providencia and others. The involvement of PVA genes in the virulence of the pathogen Vibrio cholerae was reported previously (Kovacikova et al., 2003). Reports on the involvement of acyl-homoserine lactone acylase enzymes (Ser-Ntn-hydrolases, distant homologues of PVA) in quorum quenching amongst opportunistic pathogens such as Pseudomonas are also known (Bokhove et al., 2010). However, more experimental evidence may be required to ascertain the possible role of PVA genes in the pathogenesis in these organisms. PVA genes were also found to be distributed among organisms of soil and aquatic ecosystems capable of degrading compounds containing aromatic rings. Examples include Achromobacter xylosoxidans (haloaromatic), Rhodococcus qingshengii (isolated from carbendazim-contaminated soil), Mycobacterium vanbaalenii (polycyclic aromatic hydrocarbon-metabolizing bacteria isolated from petroleum-contaminated estuarine sediments), Delftia acidovorans (able to grow on chlorophenyl herbicides), etc. It has been postulated that the penicillin acylase genes are related to pathways involved in the assimilation of aromatic compounds as a carbon source by scavenging for phenylacetylated compounds in the non-parasitic environment (Valle et al., 1991). However, more experimental evidence may be required to ascertain these roles of PVA genes.

**Evolutionary basis for the divergence of CGH family members into two clusters**

The detailed sequence analysis of the members of the dataset revealed a crucial 13–19 aa indel, which resulted in the separation of the members into these two distinct clusters. This indel corresponds to the presence or absence of the ‘assembly motif’ in the sequences. Irrespective of their source (Gram-positive bacteria, Gram-negative bacteria or archaea) and function (BSH or PVA), all members of Cluster1 possessed this motif, whereas those belonging to Cluster2 lacked the motif (Fig. 5b, c). Most CGH family members form homotetramers as quaternary structures. In BBBSH, this assembly motif is about 26 Å long (Fig. 5a) and
comprises residues 188–220 (Kumar et al., 2006). The long assembly motif of each monomer extends into its neighbouring monomer (diagonally opposite) helping in tetramer assembly and stabilization (Fig. 5d). However, in the case of BtBSH belonging to Cluster2, the absence of this assembly motif (Fig. 5a) results in the quaternary association (Fig. 5e) being less thermodynamically stable than that of BlBSH. Theoretical estimation of thermodynamic stabilities of the tetramers by PISA (Krissinel & Henrick, 2007) showed that the absence of this motif in the BtBSH structure reduces the thermodynamic stability of its quaternary association compared with BlBSH. The values of $\Delta C^{\text{diss}}$ (free energy of assembly dissociation) corresponding to BlBSH and BtBSH assembly were estimated as 48.5 and 4.9 kcal mol$^{-1}$, respectively (positive values indicating thermodynamically stable assembly and an external driving force would be required to dissociate it). The two tetramers also differed in terms of the extent of their subunit interface area, and number of non-bonded, hydrogen bonding and salt bridge interactions between the subunits (Table 4). A major difference observed was in the AD and BC interface (Fig. 5d, e), where the absence of the assembly motif reduced significantly the interface area in the BtBSH structure (Table 4).

Another difference observed between the members of the two clusters was the length of pre-peptide sequence preceding the N-terminal catalytic cysteine residue. The CGH enzymes undergo a post-translational modification to autocatalytically remove this pre-peptide sequence to obtain a mature enzyme (Chandra et al., 2005). Most members of Cluster2 possess a comparatively longer pre-peptide sequence than Cluster1 members (Fig. 5c). Among the Cluster2 members (mostly Gram-negative), in most cases

---

**Fig. 5.** (a) Superposition of the structures of BtBSH (cyan) and BlBSH (brown) showing the missing ‘assembly motif’ in the BtBSH structure, marked by an arrow. (b) Multiple sequence alignment of five enzymes to show the absence of the ‘assembly motif’ in the BtBSH (PDB ID: 3HBC) sequence. (c) Distribution of sequence length of the assembly motif and pre-peptide sequence in Cluster1 and Cluster2 enzymes. The x-axes correspond to sequence length and the y-axes correspond to the frequency of enzymes. (d, e) The quaternary structures of the enzymes BtBSH (d) and BlBSH (e). The individual subunits of the homotetramers are labelled A–D. The 26 A˚ loop extensions of subunits A and D of BtBSH, each interacting with the neighbouring subunit, are highlighted by circles.
the pre-peptide sequence was also found to act as a signal sequence, hinting at the possible translocation of these enzymes into the periplasmic space. It is known in Gram-negative bacteria that many enzymes participating in the inactivation of antibiotics are located in the periplasmic space (Gupta, 2011). The localization of penicillin G acylases and cephalosporin acylases in the periplasm involved in the inactivation of antibiotics is already known. As the above enzymes belong to the Ntn-hydrolase superfamily to which the CGH family belongs, it might be assumed that a similar translocation could also occur in the Gram-negative CGH family members. Of the three archaea belonging to Cluster2, only *Natrialba aegyptia* possesses a four-residue pre-peptide sequence. A majority of members of Cluster1 (mostly Gram-positive) were found to have a single methionine residue, preceding the N-terminal cysteine, which in most cases is proteolytically removed by methionyl aminopeptidase enzyme (Ben-Bassat et al., 1987). Only 16 members of Cluster1 contained a pre-peptide of more than one residue (Fig. 5c).

The above observations could thus help to understand the evolution of CGH family members along with the evolution of Gram-positive bacteria, Gram-negative bacteria and archaea. One hypothesis states that antibiotic selection pressure could be a major evolutionary force behind the evolution of diderms (true Gram-negative bacteria) from monoderms (true Gram-positive bacteria) (Gupta, 2011). As many CGH enzymes are known to inactivate antibiotics such as penicillin V, it could be hypothesized that during the evolution of diderms from monoderms the assembly motif could have been deleted in the case of the members of the CGH family belonging to Cluster2 (Gram-negative). The existence of eight CGH enzymes belonging to Cluster2 of the order *Corynebacterineae* (Gram-positive bacteria) shows this transition wherein the enzymes lack the assembly motif similar to that of Gram-negative bacteria (Table S1).

In the case of archaeal CGH enzymes, the presence of the tetramer assembly motif among the three archaeal members of Cluster1 (Table S1) indicates their close relation to Gram-positive members, which is also supported by a recent study proposing the emergence of archaea from Gram-positive bacteria in response to antibiotic selection pressure (Valas & Bourne, 2011). Interestingly, the three archaeal members of Cluster2 also lacked the assembly motif and were grouped phylogenetically along with the *Corynebacterineae* members, the intermediates between Gram-positive and Gram-negative (Fig. 1a).

The possibility of CGH genes being transferred by horizontal gene transfer was also analysed. The comparison of GC content of the available genome sequences in the dataset, their CGH genes and the region flanking the CGH genes ruled out the possibility of horizontal gene transfer as these were found to be almost similar.

**CONCLUSION**

Extending from the annotation system developed by Lambert et al. (2008), we present an improved method for the annotation of members of the CGH family based on the phylogenetic, substrate specificity and binding site information of the enzymes. The method presented here could annotate correctly the BSH/PVA sequences in the dataset into five distinct groups, based on the BSS scores. Based solely on sequence information, this method could thus be used to annotate correctly any putative CGH family members.

The presence of BSH genes among gut-inhabiting microbes supports the role of BSH in the protection of microbes in the host gastrointestinal tract. The occurrence of PVA genes in pathogens and organisms degrading molecules with aromatic rings suggests the need for further exploration of the physiological roles of PVA enzymes.

The emergence of diderms from monoderms represents a crucial juncture in the evolutionary history of microbes. Using the method described in this paper, we have identified two distinct subfamilies within the CGH family showing divergent evolution. Most members of Cluster1 are Gram-positive bacteria, whereas Cluster2 is rich in Gram-negative members and archaeal members are distributed across both subfamilies. The detailed sequence analysis of the CGH family members reveals that the members of two subfamilies

---

**Table 4. Quantitative estimation of interface area and the number of non-bonded interactions between individual subunits of Bt/BSH and Bl/BSH in their quaternary structures**

<table>
<thead>
<tr>
<th>Property</th>
<th>Interface</th>
<th>BtBSH</th>
<th>BlBSH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AB/CD AC/BD AD/BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interface area (Å²)</td>
<td>1954.8 302.6</td>
<td>1506.7</td>
<td>1588.4</td>
</tr>
<tr>
<td>ΔG (kcal mol⁻¹)</td>
<td>-26.4 -1.0</td>
<td>-23.9</td>
<td>-14.1</td>
</tr>
<tr>
<td>No. interface residues</td>
<td>40 6</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td>No. hydrogen bonds</td>
<td>22 4</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>No. salt bridges</td>
<td>6 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No. non-bonded contacts</td>
<td>280 32</td>
<td>178</td>
<td>144</td>
</tr>
</tbody>
</table>

---

**Downloaded from www.microbiologyresearch.org by**

IP: 54.70.40.11

Microbiology 160
differ not only by a 19–23 aa indel signature, which alters the thermodynamic stabilities of their quaternary structure, but also in terms of the length of their pre-peptide sequence.

The above analysis thus provides a supporting explanation for the antibiotic selection pressure theory, whilst also opening new dimensions for exploration of the true significance of tetramer assembly loops.

ACKNOWLEDGEMENTS

P. P., M. S. and R. S. thank the Council of Scientific and Industrial Research, New Delhi for Senior Research Fellowships. S. K. is D. S. T. Ramanujan Fellow. This research was carried out under CSIR-NCL’s Centre of Excellence in Scientific Computation (CoESC).

REFERENCES


---

Edited by: R. Lan