Protein Sequence Motif

The PA14 domain, a conserved all-β domain in bacterial toxins, enzymes, adhesins and signaling molecules

Daniel J. Rigden[1], Luciane V. Mello[1,2] and Michael Y. Galperin[3]

[1]School of Biological Sciences, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK
[2]National Center of Genetic Resources and Biotechnology, Cenargen/Embrapa, Brasília, D.F. 70770-900, Brazil
[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Iterative database searches starting from a domain insert sequence in bacterial β-glucosidases reveals the presence of a conserved domain shared by a wide variety of bacterial and eukaryotic proteins. These include other glycosidases, glycosyltransferases, proteases, amidases, adhesins, and bacterial toxins such as anthrax protective antigen (PA). The domain also occurs in the mammalian protein fibrocystin, mutation of which leads to autosomal-recessive polycystic kidney and hepatic disease. The crystal structure of PA shows that this domain (named PA14 after its location in the PA_{20} pro-peptide) has a β-barrel architecture. A PA14 sequence alignment suggests a binding function, rather than a catalytic role, whereas the PA14 domain distribution is compatible with carbohydrate binding.

Sequences of experimentally characterized β-glucosidases from *Kluveromonas fragilis* [1], *Agrobacterium tumefaciens* [2], *Clostridium stercorarium* [3] and *Thermotoga neapolitana* [4] are similar to each other (pairwise identities ~40%) and belong to the same family three of glycosidases (glycoside hydrolases) [5] (see also http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html). The two former sequences, however, are longer than the latter two owing to the presence of a conserved ~150-amino acid insertion in the middle of their C-terminal domains (Pfam entry PF01915 [6]). Here, we report the sequence analysis of this insertion sequence that identified it as a new β-barrel domain found in a variety of bacterial and eukaryotic glycosidases, glycosyl transferases, proteins involved in cell adhesion including medically important surface adhesins of *Candida glabrata* [7,8], and in human polycystic kidney and hepatic disease protein [9–11].

Domain definition

PSI–BLAST search of the NCBI non-redundant protein database (http://www.ncbi.nlm.nih.gov) using the insertion sequence of the *A. tumefaciens* β-glucosidase (residues 361–560 of SWISS-PROT P27034) as the query and inclusion E-value of 0.001 converged after nine iterations retrieving 132 proteins (Figure 1), including a *Bacteroides thetaiotaomicron* genomic [12] sequence BT2948 protein (GenBank accession AAO78054), which contains an insertion of the same domain (hereafter referred to as PA14 domain) into a predicted α-1,2-mannosidase sequence. A PSI–BLAST search using this insert (residues 250–410) as the query converged after 13 iterations, yielding 21 additional proteins for a total of 153 hits. The PA14 domain boundaries were determined from examination of its tandem copies in predicted proteins (Figure 1) – with two copies in proteins from *Pirellula sp.* (GenBank accession CAD78872), *Microbulbifer degradans* (NCBI protein database gi: 23028360), and *Clostridium thermocellum* (gi: 23021222) and three in Npun4489 from *Nostoc punctiforme* (gi: 23128179) and Chlo1407 from *Chloroflexus aurantiacum* (gi: 22971463). The phylogenetic distribution of the PSI–BLAST hits encompassed diverse bacterial and eukaryotic lineages but does not include any archaeal species.

Importantly, PSI–BLAST searches showed that the N-terminal pro-peptide fragment (PA_{20}) of the anthrax protective antigen (PA), a component of the anthrax toxin complex of known 3D structure [13], contains the new domain; PA appeared in the 2nd iteration with E-value 5×10^{-13} starting with the BT2948 domain. In PA_{20}, the domain covers residues 43–179 (SWISS-PROT entry P13423), which corresponds to residues 14–150 of the PDB entry 1ACC [13]). By analogy with the nomenclature of PA fragments, we have named the new domain PA14 because 14 kDa is the theoretical molecular weight of the domain example in PA. After the binding of PA to its cellular receptors, the toxin becomes activated by proteolytic removal of the PA_{20} fragment, which enables oligomerization of the remaining part of the protective antigen (the PA_{63} component) leading to endocytosis of toxin components and eventual intoxication of the target cell [14,15]. Similar processes occur during activation of several closely related toxins from clostridial species [16] and, apparently, of *Bacillus thuringiensis* vegetative insecticidal protein (GenBank accession AY245547). The 167-amino acid PA_{20} fragment, removed during the toxin activation, is not involved in further infection and has attracted much less attention than other components of the anthrax toxin. The PA14 domain forms the core of the PA_{20} fragment and is a β-barrel structure comprising two β-sheets of six and five strands with no significant

Corresponding author: Daniel J. Rigden (drigden@liverpool.ac.uk).
Available online 7 June 2004

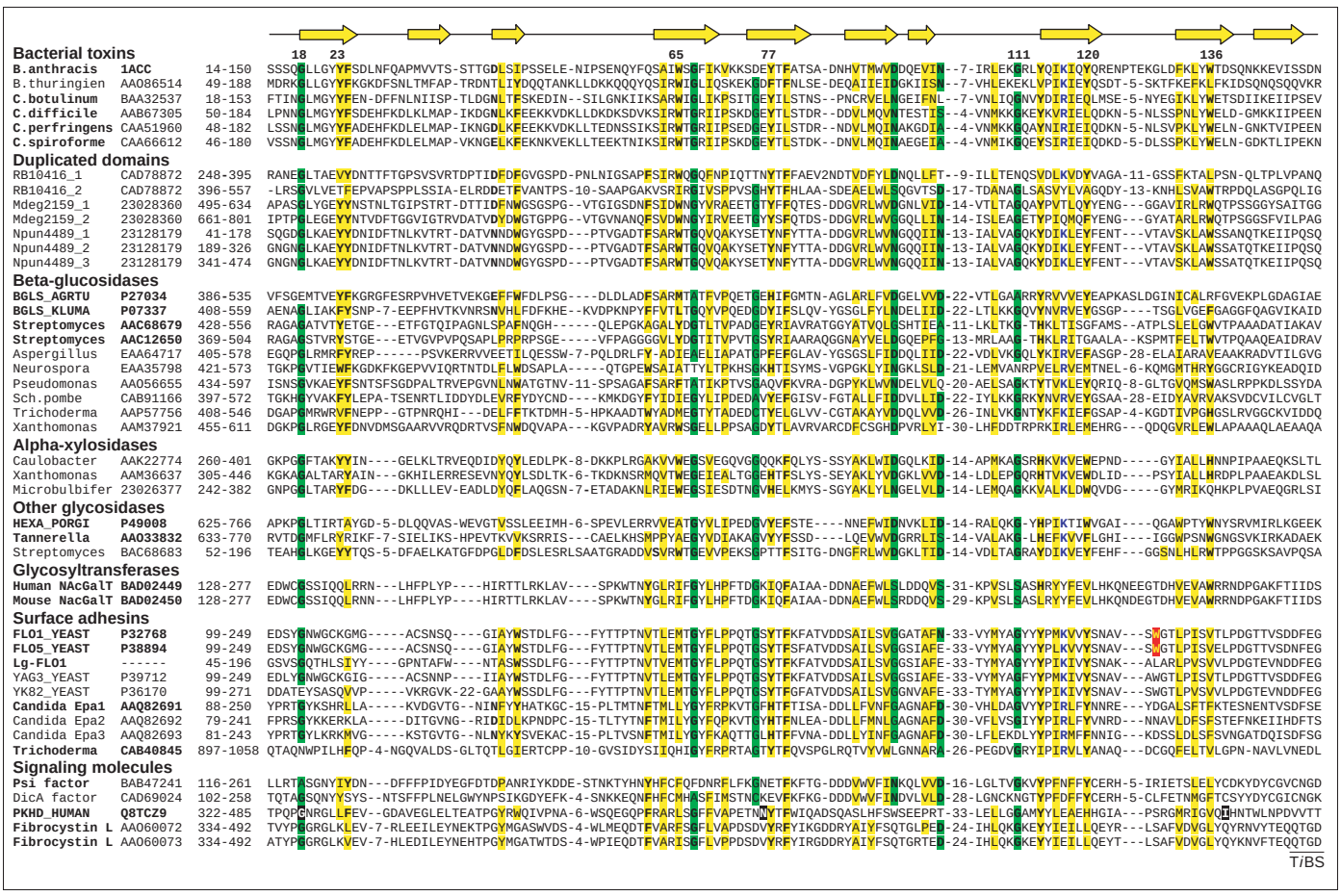

Figure 1. Multiple sequence alignment of the PA14 (named PA14 after its location in the PA_{20} pro-peptide) domains. The alignment was constructed on the basis of PSI–BLAST search results, followed by Smith–Waterman alignment of selected sequences and minimal manual editing. The proteins are listed by the genus name of the source organism, followed by SWISS-PROT or GenBank accession numbers (where available) or NCBI gi numbers. The names of experimentally characterized proteins are in bold; names of those shown in Figure 2 are in blue. The secondary structure of PA14 domain from anthrax protective antigen itself (PDB code: 1ACC) is shown above the alignment and its most conserved residues are numbered. Conserved residues conforming to 80% consensus, as determined using the Consensus program by Brown and Lai (http://www.bork.embl-heidelberg.de/Alignment/consensus.html), are shown in bold and/or colored as follows: Arg and Lys, blue; hydrophobic, yellow background; small (Gly, Ser, Ala, Asp, Asn), green background. The numbers between the aligned blocks indicate the lengths of variable inserts in the respective protein sequences. Trp228 in the yeast flocculin, which is involved in sugar recognition [21], is indicated by red background. The sequence of mutant flocculin Lg-FLO1 was taken from [23]. The residues of human fibrocystin whose mutations cause autosomal recessive polycystic kidney disease with pre- or post-natal death [29], are indicated as white letters on black background. The PA14 domain has been deposited in Pfam with the accession number PF07691 (http://www.sanger.ac.uk/Software/Pfam/).

structural similarity (in the SCOP database [17] or by independent searches) with any other domain of known structure. The N and C termini of the domain are close together in space, presumably, thereby facilitating the insertion of the PA14 domain into other recognized domains without structural disruption.

Domain architectures

Comparisons of the retrieved sequences against Pfam [6], SMART [18] and CDD [19] domain databases revealed many diverse domain combinations involving the PA14 domain (Figure 2). Most of the experimentally characterized PA14-containing proteins are involved in carbohydrate binding and/or metabolism (Figure 2), including glycoside hydrolase domains of families 2, 3, 10, 20 and 31 in the classification of Coutinho and Henrissat (http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html) and the recently characterized mammalian chondroitin β-1,4-*N*-acetylgalactosaminyltransferases [20]. Among sequences lacking obvious catalytic domains, a carbohydrate-binding function has been shown for *Saccharomyces cerevisiae* flocculation proteins [21] and their distant homologs in the pathogenic yeast *Candida glabrata* that mediate adherence to human cells [7]. In the *C. glabrata* adhesin Epa1p [8] and *S. cerevisiae* flocculins [21–22], carbohydrate binding is associated with the N-terminal third of the protein, which has therefore been assigned as a new domain in Pfam (PF06660), covering residues 29–240 of FLO1. This region largely overlaps the PA14 domain (residues 90–255 of FLO1_YEAST; Figure 1). The flocculin N-terminal domain might, therefore, be considered as one of the many PA14 domain variants.

Several PA14-containing proteins are involved in adhesion and/or signaling (Figure 2), which is consistent with their ability to bind carbohydrate-containing ligands. In a putative adhesin from *M. degradans* (gi: 23028360), twin PA14 domains follow a LamGL jellyroll domain, implicated in diverse cellular functions, including adhesion [23]. In the *Streptomyces coelicolor* protein SCO0505, a PA14 domain is combined with a putative glucose dehydrogenase domain and a region that matches the family 6 carbohydrate-binding module (CBM) in PFAM [6] and the CAZY (http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html) family CBM 35. In a hypothetical *Plasmodium*

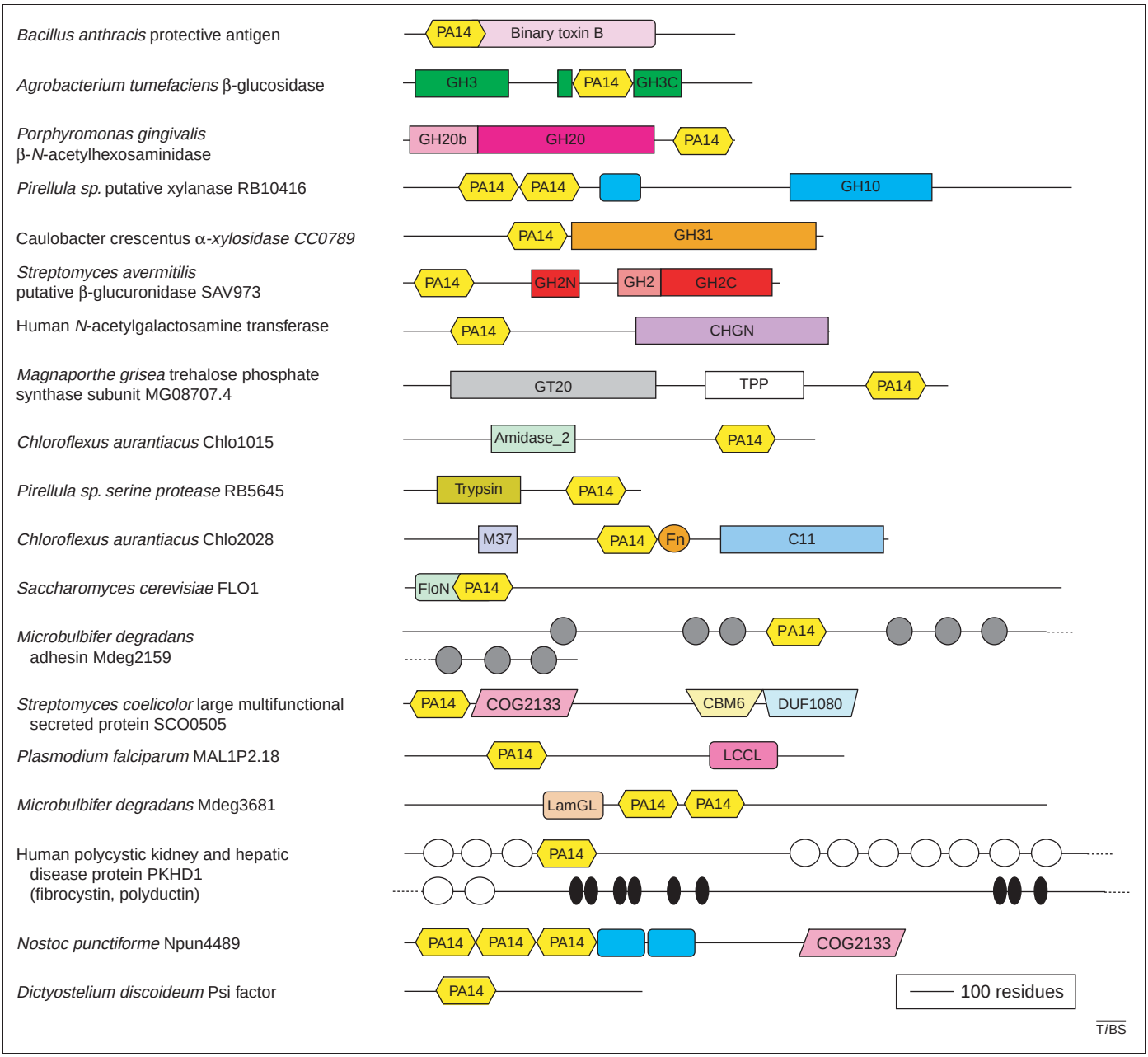

Figure 2. Domain organization of proteins containing the PA14 domain (named PA14 after its location in the PA_{20} pro-peptide). Full names and accession numbers of the proteins are as in Figure 1. Domain composition of individual proteins was deduced by comparing them with Pfam [6], SMART [18] and CDD [19] databases and drawn approximately to scale. The 50 C-terminal residues of the PA14 domain overlap with the beginning of the Binary toxin B domain (Pfam entry PF03495). The flocculin N-terminal domain (FloN; PF06660) overlaps with the PA14 domain. Other glycoside hydrolase domain abbreviations and Pfam database entries are as follows: GH2N (PF02837), GH2 (PF00703), GH2C (PF02836), GH3 (PF00933), GH3C (PF01915), GH10 (PF00331), GH20 (PF00728), GH20b (PF02838) and GH31 (PF01055). Glycosyl transferase catalytic domains are GT20 (PF00982) and CHGN (chondroitin N-acetylgalactosaminyltransferase; PF05679), the former additionally in combination with a trehalose phosphate phosphatase domain (PF02358). Other presumed catalytic domains are Amidase_2 (PF01510), Trypsin (PF00089), M37 (PF01551) and C11 (PF03415). Carbohydrate-binding module family 6 (CBM6; PF03422), LCCL (PF03815) and LamGL (SMART entry: SM00560) domain are also found in combination with PA14 and are labeled, as is a domain of unknown function (DUF1080; PF06439). Smaller repeated domains are drawn as gray circles (CADG; SM00736), white ellipses (TIG; PF01833) or black ellipses (PbH1; SM00710). Cyan rounded boxes represent CALX-β domains (PF03160) and the orange ellipse is a Fibronectin type III domain (PF00041).

falciparum protein, MAL1P2.18, the PA14 domain is combined with an LCCL domain, the naming and implication in lipopolysaccharide binding of which are discussed in Ref. [24]. In the *M. degradans* protein Mdeg3681, the PA14 domain is placed after three of a total of nine cadherin-like domains (CADG, SM00576), which are found in dystroglycans and sarcoglycans [25], as well as in hemagglutinins and neuraminidases [26]. A signaling role in the transition from amoeba to prespore cells has been demonstrated for *Dictyostelium discoideum* prespore-cell-inducing Psi factor [27], in which PA14 is the only recognizable globular domain, and can be suggested in several other proteins. In the extracellular signaling molecule DicA (published only in the database, GenBank accession CAD69024), the PA14 domain is followed by seven copies of the *Dictyostelium* (slime mold) repeat (PF00526) and a transmembrane anchor (not shown). Finally, the largest protein containing the PA14 domain, fibrocystin (also known as polyductin or tigmin; Figure 2), is also lacking known molecular function, although both

this 4074-amino acid transmembrane protein and its variant, referred to as fibrocystin L, appear to function as receptors in cellular differentiation [10,28]. At the phenotypic level, mutations in the gene encoding fibrocystin are responsible for autosomal recessive polycystic kidney and hepatic disease [9,10,29]. Several known lethal mutations in fibrocystin map to the PA14 domain (Figure 1), whereas the most common disease-causing nonsense mutation Arg496 → Xaa, typically resulting in perinatal death [29], causes formation of a protein truncated shortly after its PA14 domain.

Domain function

Taken together, several lines of evidence suggest that the PA14 domain could be a novel carbohydrate-binding module:

- The PA14 domain is combined in a mosaic manner with various catalytic or non-catalytic domains directly or indirectly implicated in binding carbohydrate or peptidoglycan.
- For the yeast flocculins, carbohydrate binding has been demonstrated for a region overlapping the PA14 domain [21].
- The notion that the PA14 β-sandwich domain might have a carbohydrate-binding function is also consistent with the fact that all CBMs of known structure are composed principally of β-strands with half of these similarly described as β-sandwiches in the SCOP database.
- The alignment of PA14 sequences (Figure 1) has few conserved hydrophilic residues, in agreement with a passive binding role rather than a catalytic function [30].
- Based on structural interpretation of sequence conservation, a putative carbohydrate binding site [31] on the PA14 domain could be located in the vicinity of highly conserved aromatic residues 77 and 136 although its less than complete conservation might suggest that carbohydrate binding is not maintained in all PA14 domains. Interestingly, Ile473 of human fibrocystin (Figure 1), mutation of which causes moderate polycystic kidney disease [29], borders this putative site.

Future perspectives

The hypothesis that the PA14 domain binds carbohydrates raises interesting possibilities about its participation in anthrax toxin activation. Simple hydrolysis of the PA_{20}–PA_{63} bond in intact PA is insufficient in itself to give subunit separation [14], which raises the question of what drives separation of the PA_{20} and PA_{63} *in vivo*. Attachment of the PA14 domain of the PA_{20} fragment to an extracellular matrix component could provide an explanation for this apparent inconsistency and offer an insight into the activation of this family of toxins. We hope that the recognition of the PA14 domain in the polycystic kidney disease protein will encourage the identification of its ligand(s) and, thereby, enable a better understanding of its biological function and role in disease.

References

1 Raynal, A. *et al.* (1987) Sequence and transcription of the β-glucosidase gene of *Kluyveromyces fragilis* cloned in *Saccharomyces cerevisiae*. *Curr. Genet.* 12, 175–184

2 Castle, L.A. *et al.* (1992) Cloning and sequencing of an *Agrobacterium tumefaciens* β-glucosidase gene involved in modifying a *vir*-inducing plant signal molecule. *J. Bacteriol.* 174, 1478–1486

3 Schwarz, W. *et al.* (1989) Cloning and expression of *Clostridium stercorarium* cellulase genes in *Escherichia coli*. *Biotechnol. Lett.* 11, 461–466

4 Zverlov, V. *et al.* (1997) *Thermotoga neapolitana bglB* gene, upstream of *lamA,* encodes a highly thermostable β-glucosidase that is a laminaribiase. *Microbiol.* 143, 3537–3542

5 Henrissat, B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* 280, 309–316

6 Bateman, A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.* 32 (Database issue), D138–D141

7 Cormack, B.P. *et al.* (1999) An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells. *Science* 285, 578–582

8 Frieman, M.B. *et al.* (2002) Modular domain structure in the *Candida glabrata* adhesin Epa1p, a β-1,6 glucan-cross-linked cell wall protein. *Mol. Microbiol.* 46, 479–492

9 Onuchic, L.F. *et al.* (2002) PKHD1, the polycystic kidney and hepatic disease 1 gene, encodes a novel large protein containing multiple immunoglobulin-like plexin-transcription-factor domains and parallel β-helix 1 repeats. *Am. J. Hum. Genet.* 70, 1305–1317

10 Ward, C.J. *et al.* (2002) The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. *Nat. Genet.* 30, 259–269

11 Xiong, H. *et al.* (2002) A novel gene encoding a TIG multiple domain protein is a positional candidate for autosomal recessive polycystic kidney disease. *Genomics* 80, 96–104

12 Xu, J. *et al.* (2003) A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science* 299, 2074–2076

13 Petosa, C. *et al.* (1997) Crystal structure of the anthrax toxin protective antigen. *Nature* 385, 833–838

14 Leppla, S.H. (1991) The anthrax toxin complex. In *Sourcebook of Bacterial Protein Toxins* (Alouf, J.E. and Freer, J.H., eds), pp. 277–302, Academic Press

15 Collier, R.J. and Young, J.A. (2003) Anthrax toxin. *Annu. Rev. Cell Dev. Biol.* 19, 45–70

16 Perelle, S. *et al.* (1997) Production of a complete binary toxin (actin-specific ADP-ribosyltransferase) by *Clostridium difficile* CD196. *Infect. Immun.* 65, 1402–1407

17 Andreeva, A. *et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32, D226–D229

18 Letunic, I. *et al.* (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* 32 (database issue), D142–D144

19 Marchler-Bauer, A. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 31, 383–387

20 Sato, T. *et al.* (2003) Molecular cloning and characterization of a novel human β 1,4-*N*-acetylgalactosaminyltransferase, β 4GalNAc-T3, responsible for the synthesis of *N*,*N*′-diacetyllactosediamine, galNAc β 1-4GlcNAc. *J. Biol. Chem.* 278, 47534–47544

21 Kobayashi, O. *et al.* (1998) Region of FLO1 proteins responsible for sugar recognition. *J. Bacteriol.* 180, 6503–6510

22 Groes, M. *et al.* (2002) Purification, crystallization and preliminary X-ray diffraction analysis of the carbohydrate-binding domain of flocculin, a cell-adhesion molecule from *Saccharomyces carlsbergensis*. *Acta Crystallogr. D Biol. Crystallogr.* 58, 2135–2137

23 Beckmann, G. *et al.* (1998) Merging extracellular domains: fold prediction for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins. *J. Mol. Biol.* 275, 725–730

24 Trexler, M. *et al.* (2000) The LCCL module. *Eur. J. Biochem.* 267, 5751–5757

25 Dickens, N.J. *et al.* (2002) Cadherin-like domains in α-dystroglycan, α/ε-sarcoglycan and yeast and bacterial proteins. *Curr. Biol.* 12, R197–R199

26 Jost, B.H. *et al.* (2001) Cloning, expression, and characterization of a neuraminidase gene from *Arcanobacterium pyogenes*. *Infect. Immun.* 69, 4430–4437

27 Nakagawa, M. *et al.* (1999) A prespore-cell-inducing factor in

Dictyostelium discoideum: its purification and characterization. *Biochem. J.* 343, 265–271
28 Hogan, M.C. *et al.* (2003) PKHDL1, a homolog of the autosomal recessive polycystic kidney disease gene, encodes a receptor with inducible T lymphocyte expression. *Hum. Mol. Genet.* 12, 685–698
29 Bergmann, C. *et al.* (2003) Spectrum of mutations in the gene for autosomal recessive polycystic kidney disease (ARPKD/PKHD1). *J. Am. Soc. Nephrol.* 14, 76–89
30 Koonin, E.V. and Galperin, M.Y. (2002) Sequence – Evolution – Function. *Computational Approaches in Comparative Genomics*, Kluwer Academic Publishers, Boston, MA, USA
31 Quiocho, F.A. and Vyas, N.K. (1999) Atomic interactions between proteins/enzymes and carbohydrates. In *Bioinorganic Chemistry: Carbohydrates* (Hecht, S.M., ed.), pp. 441–457, Oxford University Press

0968-0004/$ - see front matter
doi:10.1016/j.tibs.2004.05.002

DCC proteins: a novel family of thiol-disulfide oxidoreductases

Krzysztof Ginalski[1], Lisa Kinch[2], Leszek Rychlewski[3] and Nick V. Grishin[1,2]

[1]Department of Biochemistry, University of Texas, Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9038, USA
[2]Howard Hughes Medical Institute, University of Texas, Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA
[3]BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznań, Poland

Using top-of-the-range fold-recognition methods, we have assigned a thioredoxin-like structure to a family of previously uncharacterized hypothetical proteins of bacterial origin. The DCC family, named after the conserved N-terminal DxxCxxC motif, encompasses proteins of unknown function from DUF393 (in Pfam database) and COG3011. The presence of two invariant potentially catalytic cysteine residues indicates that DCC proteins function as thiol-disulfide oxidoreductases.

The thioredoxin-like fold represents a prototype for several different protein families including a large family of thiol-disulfide oxidoreductases. These enzymes – which encompass, among others, thioredoxin (Trx), glutaredoxin (Grx) and disulfide bond isomerase (DsbC) – are found in all kingdoms of life and typically use an active-site CxxC motif to mediate target protein activity by dithiol–disulfide exchange. In addition, thioredoxin-like proteins, such as Grx or glutathione *S*-transferase (GST), use the small thiol-containing molecule glutathione (GSH) in their various reactions. Through redox regulation of different target proteins, thiol-disulfide oxidoreductases control diverse cellular functions including apoptosis, cell proliferation, protein folding, oxidative stress and signal transduction. For instance, the genomes of most organisms possess several identified variants of Trx (at least seven in *Arabidopsis thaliana* [1]). This abundance makes identification of specific physiologic roles difficult considering potential functional redundancies. In this and other cases, precise knowledge of all existing thioredoxin-like molecules is thus imperative to determining their physiology.

Identification and structural assignment for DCC proteins

We have assigned a thioredoxin-like fold to a family of previously uncharacterized hypothetical proteins, which we call DCC after a conserved characteristic N-terminal DxxCxxC motif. This finding is a result of a large-scale structure–function annotation performed for all PfamA protein families [2] of unknown function (DUF) with a newly developed meta profile [3] alignment method Meta-BASIC (http://basic.bioinfo.pl) [4]. This fold-recognition approach uses comparison of sequence profiles combined with predicted secondary structure (meta profiles) enabling detection of distant similarity between proteins. Specifically, Meta-BASIC mapped the consensus sequence of DUF393, which contains several DCC proteins, with an above-threshold (>12) Z score to both a glutaredoxin family (PF00462) and the structure of *Escherichia coli* glutaredoxin 3 (Grx3) [5]. Our benchmarking results show that predictions with Z scores of >12 have $<5\%$ probability of being incorrect. Both PSI-Blast [6] and RPS-Blast were unable to find any reliable matches (with E-value < 0.01) to other PfamA families or to known protein structures.

The Meta-BASIC assignment was further confirmed by 3D-Jury [7] (http://bioinfo.pl/meta), the consensus method of fold-recognition servers that has proven to be one of the best performing approaches in CASP5 [8]. 3D-Jury assigned reliable scores of >50 (correspond to correct fold predictions in $>90\%$ of the cases [9]) to the thioredoxin-like superfamily for both the consensus sequence of DUF399 and its family member, the *yugD* gene product from *Bacillus subtilis* (gi 16080202). In particular, the highest scoring 3D-Jury predictions pointed to several thioltransferases (mainly from Grx family) and to GST proteins as the closest structural templates. Additional indicators of the correct fold assignment for DUF399

Corresponding author: Krzysztof Ginalski (kginal@chop.swmed.edu).
Available online 27 April 2004