Protein Sequence Motif

CASH – a β-helix domain widespread among carbohydrate-binding proteins

Francesca D. Ciccarelli, Richard R. Copley, Tobias Doerks, Robert B. Russell and Peer Bork

In this article, we describe a novel, widespread domain (CASH) that is shared by many carbohydrate-binding proteins and sugar hydrolases. This domain occurs in more than 1000 proteins distributed among all three kingdoms of life. The CASH domain is characterized by internal repetitions of glycines and hydrophobic residues that correspond to the repetitive units of a predicted or observed right-handed β-helix structure of the pectate lyase superfamily.

Surface layer (S-layer) is a simple cell envelope structure present in many bacteria and archaea. It is made of protein or glycoprotein subunits assembled into monomolecular crystalline arrays [1,2]. Despite the general low degree of sequence similarity, S-layer proteins from bacteria share a conserved motif known as the S-layer homology (SLH) domain [3]. This motif has been experimentally shown to bind the external peptidoglycan [4–7]. Because of the absence of this region in archaea, we decided to further investigate S-layer proteins from these organisms to look for a functionally equivalent region. S-layer proteins from thermophilic and hyperthermophilic archaea differ in their domain architecture, some of them containing one or more C-terminal PKD (polycystic kidney disease) domains, and others containing Fn3 (fibronectin type 3) domains (Fig. 1). The PKD domain is thought to mediate interactions with other proteins and with carbohydrates [8,9], whereas the Fn3 domain occurs in many different proteins in eukaryotes, but almost exclusively in glycohydrolases in bacteria [10–12].

Sequence analysis

When studying the regions not covered by these annotated domains, we observed homologous regions among some of the archael S-layer proteins. These regions had counterparts in a variety of other proteins (Figs 1, 2 and supplementary material*). For example, PSI–BLAST searches [13] of the N-terminal part of S-layer protein B from *Archaeoglobus fulgidus* (residues 26–134) against a non-redundant database (NRDB) identified in the first run not only S-layer proteins from other thermophilic and hyperthermophilic archaea, but also the copper-binding component (NosD) of N_2O reductase complexes from different bacteria ($E = 10^{-10}$, violet group in Fig. 2). The second PSI–BLAST iteration revealed significant homologies with three distinct groups of proteins (Fig. 2). One is composed of bacterial alginate epimerases and glycosylated members of the outer shell of the *Ectocarpus silicosus virus* coat [14] ($E = 10^{-05}$, red group). The second cluster contains hypothetical proteins from different species, a serine–threonine kinase from *Leishmania major*, and members of the F-box protein family ($E = 10^{-05}$, pink group). F-box proteins are involved in the degradation of cellular regulatory proteins [15,16]. The third cluster comprises a mouse Shc SH2 domain-binding protein [17], its human homolog, and a human protein expressed by the DNA region encompassing the hereditary prostate cancer (HPC1) and hyperparathyroidism-jaw tumor syndrome (HRPT2) loci [18] ($E = 0.001$, light-blue group). The third PSI–BLAST iteration found the conserved region in a secreted *Streptomyces coelicolor* protein and in a hypothetical *Ectocarpus silicosus virus* protein ($E = 0.001$, orange group). On the fourth iteration, polysaccharidase from *Rhizobium leguminosarum* was retrieved ($E = 10^{-05}$, yellow group), whereas following iterations revealed different hypothetical proteins (blue boxes). Interestingly, just below the PSI–BLAST E-value threshold, members of galacturonase ($E = 0.003$, dark-red group) and pectinesterase ($E = 0.047$, black group) families have been identified. The probable homology of these sequences to the new domain has been verified by significant PSI–BLAST E values using different seed sequences, and by the presence of conserved residues in the alignment (Fig. 2, for the alignment see supplementary material). The homology of all the sequences mentioned in Fig. 2 has been confirmed using Hidden Markov Model (HMM) searches [19].

Structural analysis

For both galacturonases and pectinesterases, several three-dimensional structures have been determined. Both sequence families belong to the pectin lyase-like structural superfamily (as classified in the SCOP database [20]), members of which adopt a right-handed β-helix structure. Remarkably, the four out of seven families in this group that share significant sequence homology with the new domain are enzymes involved in carbohydrate degradation. The basic structural unit of this family consists of three β strands that form a single turn of the β helix. Each turn contains ~20 amino acids, and is normally repeated between 7 and 11 times to form the elongated helix structure. The repeats show a low degree of sequence identity [21,22] when compared with each other. The region of homology with the new domain corresponds to the core region of the β helix, covering from the second to the sixth repeat (see supplementary material). Secondary structure predictions, using PHD [23], of all families without known structures predict the presence of several β strands, thereby agreeing with the experimentally determined structures of the internal repeats (see supplementary material). This supports the homology of the new

*The supplementary material comprises the multiple sequence alignment of proteins containing CASH domains, and can be accessed at http://www.bork.embl-heidelberg.de/~ciccarel/cash_aln.html, and at http://archive.bmn.com/supp/tibs/cash/ciccarelli.jpg/ The multiple sequence alignment (alignment number ALIGN_000246) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000246.dat).

0968-0004/02/$ – see front matter PII: S0968-0004(01)02046-1

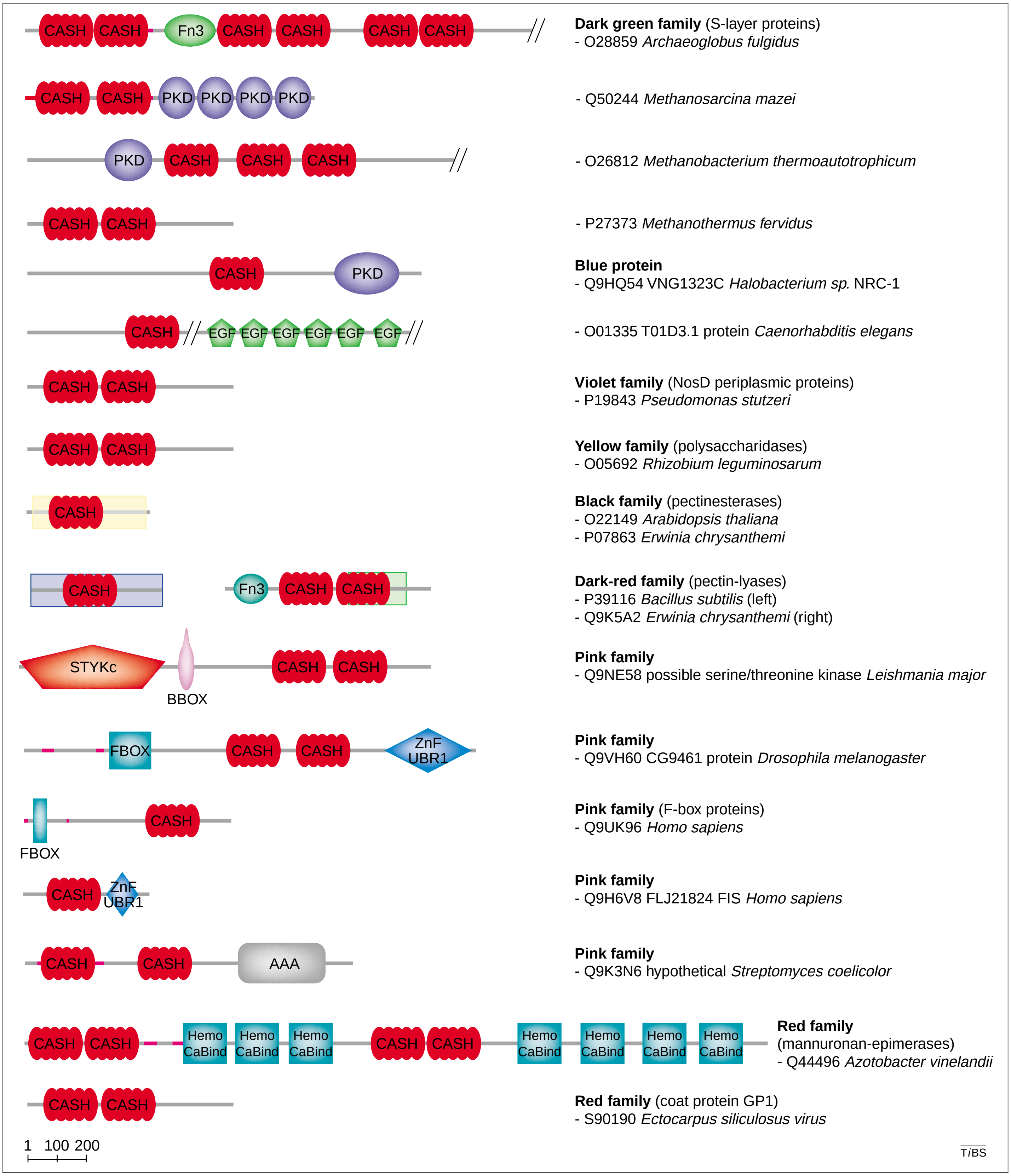
CASH CASH Fn3 CASH CASH CASH CASH
Dark green family (S-layer proteins)
- O28859 Archaeoglobus fulgidus
CASH CASH PKD PKD PKD PKD
- Q50244 Methanosarcina mazei
PKD CASH CASH CASH
- O26812 Methanobacterium thermoautotrophicum
CASH CASH
- P27373 Methanothermus fervidus
CASH PKD
Blue protein
- Q9HQ54 VNG1323C Halobacterium sp. NRC-1
CASH EGF EGF EGF EGF EGF EGF
- O01335 T01D3.1 protein Caenorhabditis elegans
CASH CASH
Violet family (NosD periplasmic proteins)
- P19843 Pseudomonas stutzeri
CASH CASH
Yellow family (polysaccharidases)
- O05692 Rhizobium leguminosarum
CASH
Black family (pectinesterases)
- O22149 Arabidopsis thaliana
- P07863 Erwinia chrysanthemi
CASH Fn3 CASH CASH
Dark-red family (pectin-lyases)
- P39116 Bacillus subtilis (left)
- Q9K5A2 Erwinia chrysanthemi (right)
STYKc BBOX CASH CASH
Pink family
- Q9NE58 possible serine/threonine kinase Leishmania major
FBOX CASH CASH ZnF UBR1
Pink family
- Q9VH60 CG9461 protein Drosophila melanogaster
FBOX CASH
Pink family (F-box proteins)
- Q9UK96 Homo sapiens
CASH ZnF UBR1
Pink family
- Q9H6V8 FLJ21824 FIS Homo sapiens
CASH CASH AAA
Pink family
- Q9K3N6 hypothetical Streptomyces coelicolor
CASH CASH Hemo CaBind Hemo CaBind Hemo CaBind CASH CASH Hemo CaBind Hemo CaBind Hemo CaBind Hemo CaBind
Red family
(mannuronan-epimerases)
- Q44496 Azotobacter vinelandii
CASH CASH
Red family (coat protein GP1)
- S90190 Ectocarpus siliculosus virus
1 100 200
TiBS

Fig. 1. Domain architectures of selected proteins containing one or more CASH domain. Only proteins with different organization and at least another domain are shown. The domains are named according to the SMART database (http://smart.embl-heidelberg.de). The semi-transparent boxes indicate the borders of the family-specific PFAM annotations [27] for pectinesterases (in yellow), pectin lyases (in blue) and pectinase (in green). Note that some of the different domain architectures are only supported by a single protein and might have been caused by erroneous gene predictions. Abbreviations: AAA, ATPase domain associated with different cellular activities; BBOX, B-box-type zinc finger domain; CASH, carbohydrate-binding proteins and sugar hydrolases; EGF, epidermal growth factor domain; FBOX, domain used as link to ubiquitination target; Fn3, fibronectin type 3 domain; HemoCaBind, hemolysin-type calcium-binding domain; PKD, polycystic kidney disease domain; STYKc, serine–threonine–tyrosine protein kinase domain; ZnF_UBR1, zinc finger domain involved in recognition of N-end rule substrates in yeast UBR1.

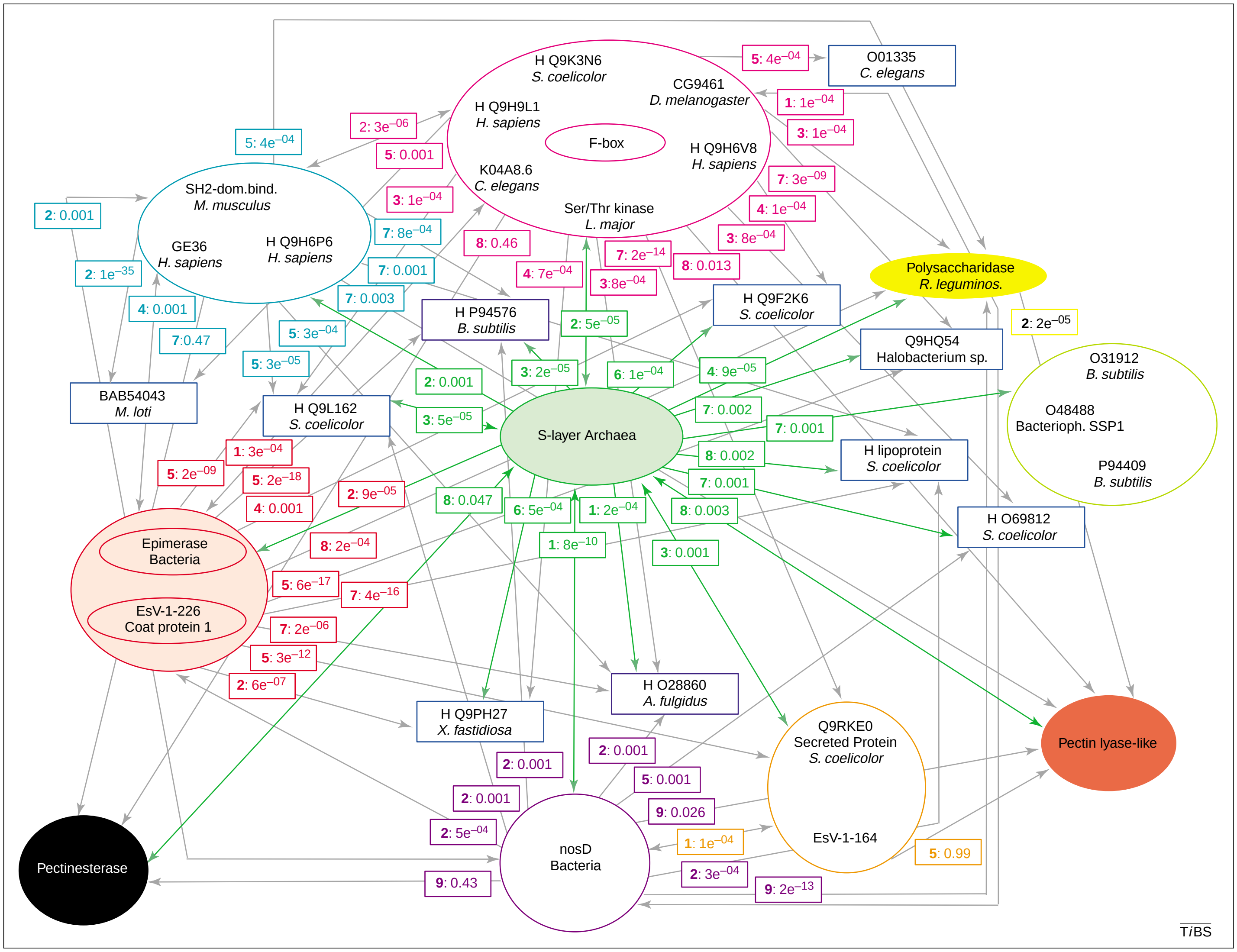

Fig. 2. Grouping of homologous sequences using PSI–BLAST [25,26]. The starting seed sequence was the N-terminal part of S-layer protein B from *Archaeoglobus fulgidus* (residues 26–134). All the sequences retrieved at the zero iteration of PSI–BLAST were assumed to form a family. Different colors are associated with different families. Circles indicate the presence of more than one sequence, whereas boxes denote individual sequences. Filled circles indicate families for which interactions with carbohydrates have been described. The connections between S-layer proteins and all other families are shown by green arrows. The boxes associated with arrows indicate the iteration number (in bold) and the E value at which the first sequence of the family was found. Reciprocal PSI–BLAST searches were performed using the first retrieved member of each family. The connections between all other families with significant E values are indicated by grey arrows. Double-headed arrows indicate bidirectional PSI–BLAST detections. In these cases, the most significant E values are displayed. Note that, as a consequence of the widespread nature of the domain, this procedure might not be able to identify all the members of the family. For example, sequence similarity between some of the S-layer and NosD protein CASH domains and extracellular acid proteases from *Thermoplasma acidophilum* has been reported [28].

region to pectate lyases and the predicted right-handed β helix. Although, in principle, there are few limitations on the number of turns per β helix, the homologous regions detected are usually ~150 residues long, with clearly defined domain borders (Fig. 1). Even in cases with more than one homologous region in the same sequence, it seems that they have arisen by duplications of domains, rather than by the successive addition of turns. For example, in Q9VH60, the two successive regions share >40% sequence identity. Thus, despite the fact that the pectate lyase superfamily contains a few additional turns of the β helix, we believe that we have identified a true domain.

Functional implications

Because of its occurrence in many carbohydrate-interacting proteins (Fig. 2), we named the region CASH domain (carbohydrate-binding proteins and sugar hydrolases).

Despite differences in enzymatic functions, structure-based sequence alignments [24] and analysis of published accounts of the active sites for the pectate-lyase-like superfamily show that these sites tend to occur in a common location: roughly in the middle of the β helix on a concave surface that is thought to be formed owing to the twist of the β sheets formed by the helical structure [21,22] (green bars below the alignment, see supplementary material). It is tempting to suggest that other sugar-binding functions could be localized in an equivalent region of the structure. Furthermore, there is even the possibility that some, if not all, of the domains detected here (including those of S-layer proteins) might harbor sugar hydrolase activity. In any case, CASH domains seem to have important carbohydrate-binding functions as they have been detected in more than 1000 proteins from all three kingdoms of life.

Acknowledgements

We are grateful to A. Lupas for his helpful comments on the manuscript. The work is supported by NATO and DFG. F.D. C. is supported by an Istituto Mario Negri-Milano fellowship.

References

1 Sleytr, U.B. (1978) Regular arrays of macromolecules on bacterial cell walls: structure, chemistry, assembly, and function. *Int. Rev. Cytol.* 53, 1–62

2 Sleytr, U.B. *et al.* (1993) Crystalline bacterial cell surface layers. *Mol. Microbiol.* 10, 911–916

3 Lupas, A. *et al.* (1994) Domain structure of the *Acetogenium kivui* surface layer revealed by electron crystallography and sequence analysis. *J. Bacteriol.* 176, 1224–1233

4 Egelseer, E.M. *et al.* (1998) The S-layer proteins of two *Bacillus stearothermophilus* wild-type strains are bound via their N-terminal region to a secondary cell wall polymer of identical chemical composition. *J. Bacteriol.* 180, 1488–1495

5 Olabarria, G. *et al.* (1996) A conserved motif in S-layer proteins is involved in peptidoglycan binding in *Thermus thermophilus*. *J. Bacteriol.* 178, 4765–4772

6 Lemaire, M. *et al.* (1995) OlpB, a new outer layer protein of *Clostridium thermocellum*, and binding of its S-layer-like domains to components of the cell envelope. *J. Bacteriol.* 177, 2451–2459

7 Mesnage, S. *et al.* (1997) Bacterial SLH domain proteins are non-covalently anchored to the cell surface via a conserved mechanism involving wall polysaccharide pyruvylation. *EMBO J.* 19, 4473–4484

8 The European Polycystic Kidney Disease Consortium (1994) The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. *Cell* 77, 881–894

9 Hughes, J. *et al.* (1995) The polycystic kidney disease 1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. *Nat. Genet.* 10, 151–160

10 Little, E. *et al.* (1994) Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *J. Mol. Evol.* 39, 631–643

11 Schultz, J. *et al.* (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5857–5864

12 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234

13 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

14 Klein, M. *et al.* (1995) Coat protein of the *Ectocarpus siliculosus* virus. *Virology* 206, 520–526

15 Bai, C. *et al.* (1996) SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. *Cell* 86, 263–274

16 Patton, E.E. *et al.* (1998) Combinatorial control in ubiquitin-dependent proteolysis: don't Skp the F-box hypothesis. *Trends Genet.* 14, 6–14

17 Schmandt, R. *et al.* (1999) Cloning and characterization of mPAL, a novel Shc SH2 domain-binding protein expressed in proliferating cells. *Oncogene* 18, 1867–1879

18 Sood, R. *et al.* (2001) Cloning and characterization of 13 novel transcripts and the human rgs8 gene from the 1q25 region encompassing the hereditary prostate cancer (hpc1) locus. *Genomics* 73, 211–222

19 Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* 14, 755–763

20 Murzin, A.G. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540

21 Chothia, C. and Murzin, A.G. (1993) New folds for all-β proteins. *Structure* 1, 217–222

22 Yoder, M.D. (1995) Protein motifs. 3. The parallel β helix and other coiled folds. *FASEB J.* 9, 335–342

23 Rost, B. *et al.* (1994) PHD an automatic mail server for protein secondary structure prediction. *CABIOS* 10, 53–60

24 Russell, R. and Barton G. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14, 309–323

25 Bork, P. and Gibson, T.J. (1996) Applying motif and profile searches. *Methods Enzymol.* 266, 162–184

26 Bork, P. *et al.* (1995) Divergent evolution of a β/α-barrel subclass: detection of numerous phosphate-binding site, by motif search. *Protein Sci.* 4, 268–274

27 Sonnhammer, E.L. *et al.* (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28, 405–420

28 Ruepp, A. *et al.* (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* 407, 508–513

Francesca D. Ciccarelli*
Tobias Doerks
Peer Bork
Max-Delbrueck-Centrum, PO Box 740238,
D-13092 Berlin, Germany; and European
Molecular Biology Laboratory,

Richard R. Copley
Robert B. Russell
European Molecular Biology Laboratory,
69012 Heidelberg, Meyerhofstr. 1, Germany.
*e-mail: ciccarel@embl-heidelberg.de

TiBS Editorial Policy

As the leading review journal in biochemistry and molecular biology, *TiBS* enables researchers, teachers and their students to keep up with new and recent developments across this broad field.

Review forms the foundation of each monthly issue. These articles, invited from leading researchers in a specific field, summarize and assess recent and important developments. Similarly commissioned, Opinion articles highlight new ideas and challenge existing models. Meeting reports, research news and protein sequence motifs are discussed in the Research Update section; whereas News & Comment houses short articles highlighting recent research papers of particular interest, as well as a biochemical news section and the letters column of the journal. Computer Corner reports on new software applications, and information resources available on the Internet. Also in our Forum section are Q & A, which aims to outline essential points of a research topic and is made accessible by a question and answer format, and the Historical Perspective column, which provides a historical overview of a particular subject.

Articles for *TiBS* are generally invited by the Editors, but ideas for Reviews and, in particular, Opinion features are welcome. Prospective authors should send a brief summary, citing key references, to the Staff Editor in London, who will supply guidelines on manuscript preparation if the proposal is accepted.

The submission of completed articles without prior consultation is strongly discouraged. As much of the journal content is planned in advance, such manuscripts might be rejected primarily because of lack of space.

Authors should note that all Review and Feature articles for *TiBS* are peer reviewed before acceptance and publication cannot be guaranteed.