Article No. jmbi.1999.2827 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 289, 729–745

JMB

Eukaryotic Signalling Domain Homologues in Archaea and Bacteria. Ancient Ancestry and Horizontal Gene Transfer

C. P. Ponting[1]*, L. Aravind[1], J. Schultz[2], P. Bork[2] and E. V. Koonin[1]

[1]*National Center for Biotechnology Information National Library of Medicine National Institutes of Health Bldg. 38A, Bethesda MD 20894, USA*

[2]*EMBL, Meyerhofstr. 1 69012 Heidelberg, Germany*

**Corresponding author*

Phyletic distributions of eukaryotic signalling domains were studied using recently developed sensitive methods for protein sequence analysis, with an emphasis on the detection and accurate enumeration of homologues in bacteria and archaea. A major difference was found between the distributions of enzyme families that are typically found in all three divisions of cellular life and non-enzymatic domain families that are usually eukaryote-specific. Previously undetected bacterial homologues were identified for# plant pathogenesis-related proteins, Pad1, von Willebrand factor type A, src homology 3 and YWTD repeat-containing domains. Comparisons of the domain distributions in eukaryotes and prokaryotes enabled distinctions to be made between the domains originating prior to the last common ancestor of all known life forms and those apparently originating as consequences of horizontal gene transfer events. A number of transfers of signalling domains from eukaryotes to bacteria were confidently identified, in contrast to only a single case of apparent transfer from eukaryotes to archaea.

Keywords: horizontal gene transfer; signalling domains; homology; genome comparison; sequence profiles

Introduction

Recent genome sequencing of organisms representing each of the three divisions of cellular life (archaea, bacteria and eukaryota) provides opportunities to infer the genetic heritage of the entire set of their genes and portions of genes encoding individual protein domains. Domains are spatially compact units of three-dimensional protein structure. Those domains that show significant sequence similarity to each other or are similar in fold, with common functions and active or binding sites, are considered homologous, that is are thought to have evolved from a common ancestor. Homologous proteins and domains may arise either due to speciation, in which case they are products of orthologous genes, or else as a consequence of an intra-genome gene duplication, resulting in paralogous gene products (Fitch, 1970, 1995;

Abbreviations used: AP-ATPase, APAF-1-like ATPase; BRCT, breast cancer C-terminal domain; C1, protein kinase C constant region 1 domain; C2, protein kinase C constant region 2 domain; CBS, domain present twice in cystathionine β-synthase; cNMP, cyclic nucleotide monophosphate binding domain; EGF, epidermal growth factor-like domain; FHA, forkhead associated domain; GAP, GTPase activator protein; GEF, guanine nucleotide exchange factor; HECT, domain homologous to E6-AP carboxyl terminus; IG, immunoglobulin domain; LRR, leucine-rich repeat; LysM, lysin-like motif; MATH, Meprin and TRAF homology domain; MPN, Mpr1p and Pad1p N-terminal domain; NHL, NCL-1, HT2A and LIN-41 repeats; NIβ, Na^{+}-Ca^{2+} exchanger/integrin subunit β4 domain; Pad1, domain homologous to *Schizosaccharomyces pombe* Pad1p (also called MPN or JAB domains); PDZ, PSD-95, Dlg, ZO-1/2 domain; PH, pleckstrin homology domain; PKD, domain in polycystic kidney disease 1 protein; PR-1, plant pathogenesis-related proteins of group 1-like domains; PTB/PI, phosphotyrosine binding/interaction domain; PX, phox homology domain; SAM, sterile alpha motif domain; SET, Suvar3-9, enhancer-of-zeste, trithorax domain; SH2, src homology 2 domain; SH3, src homology 3 domain; SWIB, SWI complex, BAF60b domain; TGFβ, transforming growth factor β-like domain; TIR, toll-interleukin-1-resistance domain; TPR, tetratrico peptide repeat; vWFA, von Willebrand factor A domain; WD40, repeat containing conserved Trp and Asp residues; WW, domain containing conserved Trp residues.

E-mail address of the corresponding author: ponting@ncbi.nlm.nih.gov

0022-2836/99/240729–17 $30.00/0

Doolittle, 1995; Henikoff *et al.*, 1997; Tatusov *et al.*, 1997).

The issue of orthology is, however, less than straightforward due to the existence of multidomain proteins and the prominence of domain rearrangements in evolution. Numerous multidomain proteins are present in all organisms but complex domain architectures are particularly common among certain classes of eukaryotic-specific proteins. Primarily, these are proteins involved in different forms of signal transduction, such as membrane receptors, protein kinases and phosphatases, adapter proteins, phospholipases, and chromatin-associated proteins. Domain accretion, that is step-by-step addition of new domains to pre-existing cores, seems to have played a major role in the evolution of these functional classes of proteins. Domain architectures of large, complex regulatory proteins are quite variable among phylogenetically distant eukaryotes. In particular, there are relatively few yeast orthologues of multidomain proteins from multicellular eukaryotes such as the nematode *Caenorhabditis elegans* (e.g. see Mushegian *et al.*, 1998; Chervitz *et al.*, 1998). Few, if any, archaeal and bacterial orthologues with identical collinear domain arrangements are detectable for eukaryotic multidomain proteins.

Eukaryotic signalling domains are often discussed in the literature as if their progenitors and their functions arose early in eukaryotic evolution. However, prokaryotic homologues have been identified for many enzymes involved in eukaryotic signal transduction, such as phospholipase C, phospholipase D, protein kinase, diacylglycerol kinase, protein phosphatases and 3′-5′ cyclic nucleotide phosphodiesterases (Heinz *et al.*, 1995; Ponting & Kerr, 1996; Koonin, 1996; Smith & King, 1995; Kennelly & Potts, 1996; Leonard *et al.*, 1998; Schultz *et al.*, 1998; Bork *et al.*, 1996a; Aravind & Koonin, 1998). Similarly, typical extracellular proteolytic enzymes of eukaryotes, such as trypsin-like serine proteases and matrix metalloproteases, have readily detectable homologues among prokaryotic enzymes (Delbaere *et al.*, 1975; McKerrow, 1987).

Prokaryotic homologues have been detected also for some non-enzymatic domains that are essential to eukaryotic signal transduction and are typically involved in mediating protein-protein interactions. These domains, however, are typically much less conserved, at least at the sequence level, than enzymes, and the discovery of bacterial and archaeal counterparts of eukaryotic domains required careful application of sensitive sequence analysis methods (e.g. see Gibson *et al.*, 1993; Hofmann & Bucher, 1995; Koonin *et al.*, 1996; Ponting, 1997a,b; Bateman, 1997; Bork *et al.*, 1997a; Ponting & Aravind, 1997; Aravind & Ponting, 1997).

Recent improvements in computational methods of sequence analysis (e.g. see Altschul *et al.*, 1997; Altschul & Koonin, 1998) and the rapid increase in available sequence information now allow the reliable identification of even subtly conserved domains in completed archaeal, bacterial and eukaryotic genomes. This provides insights into the origins and evolution of these domains from the analysis of their phyletic distributions. Furthermore, these advances have provided considerable evidence that gene-transfer between species, and indeed between divisions, has been a major force for change during prokaryotic evolution (Koonin *et al.*, 1997; Aravind *et al.*, 1998; Woese, 1998; Doolittle, 1998). Horizontal gene transfer from bacteria to eukaryotes, particularly in the context of organellar endosymbiosis but probably also under other circumstances, also is a major evolutionary phenomenon (Doolittle, 1998). By contrast, instances of apparent horizontal transfer in the opposite direction, from eukaryotes to bacteria, have been reported only anecdotally (Bork, 1993; Aravind *et al.*, 1999a; Baumgartner *et al.*, 1998; Little *et al.*, 1994).

Here we investigate the phyletic distribution of domains commonly found in eukaryotic signalling proteins (Bork *et al.*, 1997b). Our intention was to use the latest tools of sequence analysis to search for prokaryotic homologues of these domains, to infer whether these homologues are related to their eukaryotic counterparts by vertical descent or by horizontal transfer, and to provide insight into the preservation, or otherwise, of function among diverse homologues.

Results and Discussion

Occurrences of 213 domain families in 18 complete genomes were investigated using the PSI-BLAST and HMMER2 sequence analysis methods. The hits were evaluated in terms of their statistical significance as indicated in Methods. Additionally, all alignments were examined for the conservation of signature motifs that are typical of specific domains, in order to eliminate possible false positives. These families represent domains that participate in eukaryotic intracellular and extracellular signal transduction pathways. Approximately one-quarter of these domains (55 in total) were found to be represented both in eukaryotic proteins and in prokaryotic proteins. Table 1 contains the numbers of proteins found in complete genomes that contain "eukaryotic" signalling domains; only those domains that are found at least once in complete prokaryotic genomes are included. A dramatic difference in the phyletic distributions of the enzymatic and non-enzymatic signalling domains was immediately apparent, most of the former were readily detectable in all three divisions of life, whereas the majority of the latter are eukaryote-specific (Figure 1). A Table representing the GenBank identifier codes (GIs) of proteins from the complete genomes that contain these domains is available from: ftp://ncbi.nlm.nih.gov/pub/koonin/SIGNALLING. Alignments of domain families, in general, may be acquired from the SMART Web site

Table 1. Distributions of enzymatic and non-enzymatic domains in completely sequenced genomes

Species:	Eubacteria												Archaea				Eukaryota	
	Aa	Bs	Bb	Ct	Ec	Hi	Hp	Mt	Mg	Mp	Sy	Tp	Af	Mj	Mth	Ph	Ce	Sc
A. *Enzymes*																		
Acid P-ase	1	3	-	-	3	1	3	2	-	-	1	1	-	3	1	1	7	7
Adenoviral-type Cys-protease	-	-	-	2	1	-	-	-	-	-	-	-	-	-	-	-	7	2
AP-ATPase	-	-	1	-	-	-	-	6	-	-	-	-	-	-	-	1	1	-
Calcineurin-like P-ase	6	12	1	3	10	6	3	5	1	1	5	2	12	8	10	10	66	23
Caspase	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	1
cAMP/cGMP cyclases	-	-	-	-	-	-	-	16	-	-	3	1	-	-	-	-	36	1
Collagenase-like metalloprotease	1	-	-	-	-	-	-	-	-	-	1	-	1	1	-	2	43	-
DAG kinase	-	3	-	-	1	-	-	2	-	-	1	-	1	-	-	-	10	2
DAG kinase-like	1	2	1	-	1	1	1	1	1	1	2	1	1	1	1	1	1	2
IPPc-like	-	1	1	-	3	1	1	2	-	-	1	2	1	-	2	-	33	11
Guanylate kinase	3	3	-	1	2	1	1	1	1	1	1	-	-	1	-	1	10	1
Lysozyme	3	6	1	-	8	3	2	5	-	-	1	2	-	-	-	-	-	-
Papain-like Cys protease	-	-	-	-	-	-	-	-	-	-	-	1	1	-	-	1	30	1
Phospholipase C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6	1
Phospholipase D	1	3	-	7	6	1	4	1	-	-	2	-	1	-	-	-	6	4
PP2A P-ase	-	1	-	-	3	1	-	-	-	-	1	-	1	-	-	-	49	13
PP2C P-ase	2	3	-	2	-	-	1	1	1	1	3	1	-	-	-	-	10	9
S2P-like metalloprotease	2	4	1	1	1	1	2	3	-	-	4	1	4	3	3	2	1	-
Ser/Thr/Tyr kinase	2	4	-	3	3	1	1	13	1	1	12	-	4	4	4	4	435	116
Ser/Thr/Tyr and Tyr P-ase	-	-	-	-	1	-	-	1	-	-	-	-	-	2	-	1	112	14
Small GTPase	3	-	-	-	-	-	-	1	-	-	-	-	1	1	3	-	62	28
Transglutaminase	2	1	-	-	-	1	-	5	-	-	4	-	4	1	5	3	1	2
Trypsin-like Ser protease	1	5	1	1	5	2	1	6	-	-	3	3	-	-	1	-	25	1
B. *Non-enzymatic domains*																		
Ankyrin	-	-	-	-	3	-	-	-	-	-	1	2	-	-	-	-	86	19
BRCT	1	1	1	1	1	1	1	2	1	2	1	1	-	-	-	-	26	10
Bulb-type lectin	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-
Cadherin	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	17	-
CBS	7	15	3	5	9	5	3	12	1	1	8	5	11	15	18	8	13	10
cNMP-binding	2	2	-	1	3	3	-	10	-	-	13	4	1	-	-	-	17	2
Discoidin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-
EF-hands	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	65	18
Fasicilin-like	-	-	-	-	-	-	-	2	-	-	2	-	-	-	-	-	1	1
FHA	-	-	-	1	-	-	-	7	-	-	10	-	-	-	-	-	12	14
Fibronectin III	1	3	1	-	-	-	1	-	-	-	2	2	1	-	-	2	55	2
Integrin α domain	-	-	-	-	-	-	-	-	-	-	3	-	-	-	-	-	2	1
Kelch	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	15	6
LRR	-	-	-	-	1	-	-	-	-	-	-	1	-	-	-	-	69	12
LysM	1	17	4	2	8	2	1	2	-	-	2	7	-	-	-	-	5	-
NIβ	-	-	-	-	-	-	-	-	-	-	4	-	-	-	-	-	3	-
Pad1-like (MPN)	1	-	-	-	-	-	-	1	-	-	1	-	1	-	1	2	7	4
PDZ	4	9	3	2	7	4	3	5	-	-	9	5	1	1	2	1	64	2
PKD	-	-	-	-	-	-	-	-	-	-	1	-	2	1	4	1	-	-
PR-1-like	-	2	1	-	-	-	-	-	-	-	-	-	1	-	-	-	33	3
PTX	-	-	1	-	-	-	-	-	-	-	2	1	-	-	-	-	1	-
Sel-1-like	-	-	-	-	3	1	9	-	-	-	-	-	-	-	-	-	2	4
SET	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	3	2
Src homology 3	-	6	-	1	1	1	1	-	-	-	3	-	-	-	-	-	58	24
SWIB	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	3	2
TIR	-	1	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1	-
TPR	15	16	20	9	13	4	5	1	-	-	24	23	3	9	8	6	41	29
VWA	1	5	5	-	4	-	1	4	-	-	4	3	2	3	2	1	65	3
WD40	-	-	-	-	-	-	-	-	-	-	6	-	-	-	-	-	127	110
YWTD	-	-	1	-	-	-	-	4	-	-	2	1	-	-	-	-	15	-
Zf_AN1	-	-	-	-	-	-	-	-	-	-	-	-	3	-	-	-	3	2
Number of genes	1522	4099	1215	893	4289	1709	1565	3918	467	677	3169	1031	2407	1869	1770	2064	19,099	6526

However, three domains are listed here due to their identification in incompletely sequenced prokaryotic genomes: homologues of caspases (e.g. in *S. coelicolor* pk3 (gi 625077)), phospholipase C (e.g. *B. cereus* PLC (gi 130297)) and discoidin-like (e.g. in *C. septicum* sialidase (gi 266601)) domains. A putative enzymatic domain family that contains motifs similar to those in DAG kinases is listed as DAG kinase-like. S2P-like metalloproteases are homologues of human S2P (Rawson *et al.*, 1997). Species abbreviations: Aa, *Aquifex aeolicus*; Bs, *Bacillus subtilis*; Bb, *Borrelia burgdorferi*; Ct, *Chlamydia trachomatis*; Ec, *Escherichia coli*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori*; Mg, *Mycoplasma genitalium*; Mp, *Mycoplasma pneumoniae*; Mt, *Mycobacterium tuberculosis*; Sy, *Synechocystis* PCC6803; Tp, *Treponema pallidum*; Af, *Archaeoglobus fulgidus*; Mth, *Methanobacterium thermoautotrophicum*; Mj, *Methanococcus jannaschii*; Ph, *Pyrococcus horikoshii*; Ce, *Caenorhabditis elegans*; and Sc, *Saccharomyces cerevisiae*. Other abbreviations: P-ase, phosphatase; IPPc, inositol polyphosphate phosphatase; DAGK, diacylglycerol kinase; PTX, pentraxin-like domain; other abbreviations are in the text and in abbreviations used. The number of genes in complete genomes is shown at the foot of the Table.

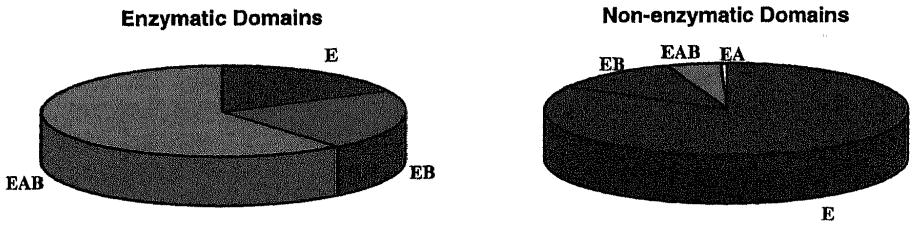

Figure 1. Phyletic distribution of enzymatic and non-enzymatic eukaryotic signalling domains. Abbreviations: E, number of domain families observed only in eukaryota; EA, number of domain families observed only in eukaryota and archaea; EB, number of domain families observed only in eukaryota and bacteria; EAB, number of domain families observed in all three divisions, eukaryota, archaea and bacteria.

(http://coot.embl-heidelberg.de/SMART/) or else directly from the authors. In addition, the SMART Web site provides information relating to the phyletic distributions, structures and functions of the domains discussed here.

Phyletic distributions of enzyme families

Notably, 17 of the 28 analysed families of enzymes that are involved in various forms of eukaryotic signalling were identified in the genomes from each of the three divisions of cellular life, archaea, bacteria and eukarya (Figure 1). These included homologues of protein and lipid phosphatases, protein and lipid kinases, zinc-dependent proteases, cysteine and serine proteases, phospholipases D, small GTPases, ATPases, and transglutaminases. As has been suggested for "eukaryotic-type" protein kinases (Leonard *et al.*, 1998), these enzymes are likely to have an ancient provenance, having originated in the progenitor of the three divisions of life, although their subsequent history appears to have included horizontal gene transfers as well as lineage-specific family expansions (see below).

The cellular functions performed by homologues from within an enzyme family, however, are usually diverse and, at least in some cases, appear to differ in eukaryotes and prokaryotes. For example, inositol polyphosphate phosphatases that hydrolyse the second messenger inositol 1,4,5-trisphosphate in eukaryotic signalling pathways are homologues of prokaryotic neutral sphingomyelinases, exodeoxyribonucleases III and DNases I (Matsuo *et al.*, 1996; L. A., unpublished results). Similarly, an archaeal homologue of animal transglutaminases has been characterized as a protease, leading to the suggestion that many, if not all, prokaryotic members of the large superfamily of transglutaminase-like enzymes may be proteases (Makarova *et al.*, 1999a). Other enzymes, however, most notably S/T protein kinases and PP2C, and calcineurin-like phosphatases are likely to perform signalling roles in both prokaryotes and eukaryotes (Zhang *et al.*, 1998; Leonard *et al.*, 1998).

The eukaryotic cyclic nucleotide phosphodiesterases, a major class of signal-transducing enzymes, represent a special case. While no classic cyclic nucleotide phosphodiesterases are detectable in archaea and bacteria, distantly related enzymes of the HD superfamily of hydrolases have been identified (Aravind & Koonin, 1998). The combination of some of these proteins with other signalling modules (Aravind & Koonin, 1998) suggests that they are likely to function as cNMP phosphodiesterases.

The phyletic distribution of another group of classical eukaryotic signalling enzymes, the small, Ras-like GTPases, suggests an unusual case of horizontal gene transfer from archaea to bacteria. These enzymes are found in multiple copies in all eukaryotes, whereas the majority of the archaea have between one and three representatives of a distinct subfamily. Bacterial Ras-like GTPases similar to archaeal homologues were detected only in *A. aeolicus*, *M. tuberculosis* and in *M. xanthus*. Experimental evidence from *Myxococcus* suggests that the eukaryotic Ras-like GTPase SAR1 can function in the place of its bacterial homologue (Hartzell, 1997). At least in the case of *Aquifex* an archaeal origin of these genes is particularly credible, given observations of apparently massive gene flux from archaea to this hyperthermophilic bacterium (Aravind *et al.*, 1998).

Homologues of six families of eukaryotic enzymes, namely lysozymes, adenoviral-type cysteine proteases, caspases (Chen *et al.*, 1998; Aravind *et al.* 1999b), PP2C-type phosphatases, phospholipases C and adenylate/guanylate cyclases, were identified in bacteria but not among currently known archaeal sequences. Bacterial phospholipases C are detectable only in the *Bacillacaea* family whereas eukaryotic versions are widespread in metazoa, plants and fungi. As suggested elsewhere (Heinz *et al.*, 1998), bacterial homologues are likely to have arisen from horizontal gene transfer from a eukaryotic source. A counter argument that phospholipases C homologues in bacteria lineages other than the *Bacillacaea* and in archaea have been lost during evolution appears to be less parsimonious, that is, would invoke a greater number of distinct evolutionary events.

The evolutionary heritage of adenylate/guanylate cyclases is less clear, since the bacterial versions are distributed among several diverse lineages of bacteria. However, it is notable that some of the 16 *M. tuberculosis* cyclases, such as Rv1625c, demonstrate significantly greater

sequence similarity to vertebrate cyclases than they do with bacterial cyclases, and also group with them in neighbour joining phylogenetic trees with greater than 80% bootstrap support (data not shown). This suggests that at least one "eukaryote type" cyclase entered the prokaryotes from a eukaryotic source early in evolution. Other *M. tuberculosis* cyclases are more similar to other cyclases scattered among bacterial lineages than they are to eukaryotic homologues (e.g. Rv1318c, which is more similar to 22 bacterial cyclases ($2 \times 10^{-80} \leqslant E \leqslant 1 \times 10^{-8}$) than it is to eukaryotic ones ($E \geqslant 1 \times 10^{-8}$)). The latter molecules may have originated in an earlier horizontal gene transfer of a eukaryotic cyclase and then disseminated in the prokaryotic world through a series of further horizontal transfer events. This evolutionary scenario is analogous to that proposed previously for the PKN2-type protein kinases (Leonard *et al.*, 1998).

An important, recurrent feature of the evolution of enzymes involved in signalling functions is the lineage-specific addition of adapter domains to the termini of core enzymatic domains. Well-known examples of this include the addition of SH2, SH3, PH, C1, C2 and FHA domains to protein kinases in eukaryotes. A similar phenomenon is seen in bacteria: SH3 (see below), WD40, penicillin-binding, cyclase and YWTD domains have been fused to PKN2-type kinase domains.

Five families of enzymes were identified only in eukaryotes. These were cytosolic phospholipases A_2, pancreatic-type phospholipases A_2, ubiquitin-conjugating enzymes E2, HECT E3 ubiquitin-protein ligases and ubiquitin C-terminal hydrolases. The apparent absence in prokaryotes of the enzymes involved in the ubiquitin pathway might have been anticipated, since this major eukaryotic regulatory system appears to have no counterpart in either archaea or bacteria.

Phyletic distributions of non-enzymatic signalling domains: domains specific to eukaryotes

The majority (83%) of the non-enzymatic domain families analysed were found only in eukaryotes. The extreme sequence divergence of some eukaryotic domain families, such as PH, PTB/PI, Ca^{2+}-binding C2, SAM, RA, PX and WW domains among others, suggests that prokaryotic homologues of these, if they exist, may not be detectable by sequence similarity alone (at least using current search methods and databases). On the other hand, the fact that certain domain families, such as, for example GAL4 homologues in fungi (Bork & Gibson, 1996), are lineage-specific, argues that some domains are indeed represented only among eukaryotes. Those in this category are likely to include guanine nucleotide exchange factors (GEFs) and GTPase activator proteins (GAPs) for the small GTPases Ras and Rho, since GTPases of these classes are not found in prokaryotes. Interestingly, however, the *Rickettsia prowazekii* genome is unique among sequenced prokaryotic genomes in encoding a protein (gi 3860933) that is a homologue of mammalian Sec7-like GEFs that are specific for ARFs (Wolf *et al.*, 1999b). It is most likely that the gene coding for this protein has entered the rickettsial genome *via* horizontal transfer from the eukaryotic host. Its presence in an organism lacking small GTPases of the ARF subclass is intriguing and suggests a regulation of host GTPases as part of the pathogenic mechanism of the bacterium.

Other domains that were only detectable in eukaryotes were those that contained several disulphide bridges and also lacked extensive secondary structure, such as epidermal growth-factor-like (EGF), kringle and TGFβ-like cystine knot domains. Disulphide bridges are thought to increase stability and resistance to proteolysis in these domains (Doolittle, 1995; Bork *et al.*, 1996b). Consequently such domains are only fully folded outside of the reducing environment of the cytosol. It has been argued that these domains evolved only after the Earth's atmosphere became oxidising (Doolittle, 1995) lending credence to the notion that disulphide bridge-dependent domains emerged in the eukaryotic lineage following its divergence from bacteria and archaea. It is notable that prokaryotic homologues of immunoglobulin-like, discoidin-like, von Willebrand factor A, lysin motif (LysM) and PR-1 domains (see below), in contrast to their eukaryotic counterparts, possess no disulphide bridges. These domains, however, contain extensive secondary structures and hydrogen bond networks, and consequently, disulphide bridges are non-essential for their structure and functions.

Phyletic distributions of non-enzymatic signalling domains: domains present in all three divisions

The phyletic distributions of non-enzymatic eukaryotic signalling domains are in stark contrast to those of signalling enzymes (Figure 1). Only nine out of the 185 non-enzymatic domain families considered were identified in genomes from each of the three divisions of cellular life: cNMP, CBS, kelch, Pad1, PDZ, PKD, PR-1, TPR and vWFA domains (Table 1). Of these, the PDZ domain family is perhaps the best-documented (Koonin *et al.*, 1996; Ponting, 1997a) and representatives among prokaryotic proteins includes HtrA-like serine proteases and S2P-like metalloproteases. Experimental evidence (reviewed by Ponting, 1997a) indicates that C-terminal polypeptide-binding function of these domains may be common to all three divisions of life and, by inference, might have been established already in their last common ancestor. The participation of PDZ domains in regulating signalling pathways, however, appears to be a metazoan invention, since domain architectures of PDZ-containing proteins

would suggest that these functions are absent in *S. cerevisiae* and among the available plant sequences.

The cNMP-binding domain is widespread among bacteria and eukaryotes while present only in *Archaeoglobus* amidst the archaea. Consequently, it appears most likely that the archaeal protein has entered the *Archaeoglobus* genome *via* a horizontal transfer event from bacteria.

CBS (cystathionine-β synthase) domains are a unique case of a (predicted) signalling domain (Bateman, 1997; Ponting, 1997b) that is present in all currently available complete genome sequences. TPR repeats (Sikorski *et al.*, 1990) are nearly as common, being apparently absent only in *M. pneumoniae* and *M. genitalium*. Currently the functions of these two domain families in these diverse organisms are unclear. Significant, and probably independent, expansions of these domains are observed in several species, e. g. CBS in the archaeon *M. jannaschii* and TPR in the cyanobacterium *Synechocystis* sp.; these are likely to have adaptive importance, and elucidating the processes they are involved in will be of great interest.

Another domain family whose functions are unknown, even though a representative's tertiary structure has been determined (Fernández *et al.*, 1997), includes homologues of the anti-fungal plant pathogenesis-related proteins of group 1 (PR-1: Kitajima & Sato, 1999). Prokaryotic homologues of this family had not been observed previously, but were detected during this study in the archaeon *A. fulgidus*, and two bacteria, *B. subtilis* and *B. burgdorferi* (Table 1 and Figure 2(a)). As indicated in Methods, the findings made using the PSI-BLAST profile for the PR-1 domain were verified by reciprocal BLAST searches. For example, a search with *B. subtilis* YkwD (gi 2632218) as query revealed significant similarities with a *S. cerevisiae* PR-1 homologue, YKR013w (gi 539306), with $E = 4 \times 10^{-4}$. Conservation of structurally clustered charged groups in PR-1 homologues has led to the suggestion that these are enzymes (Szyperski *et al.*, 1998). However, none of these residues is absolutely conserved in either eukaryotic or prokaryotic homologues, and only one of these is conserved among all four bacterial homologues. More recent studies indicate that PR-1 homologues are serine protease inhibitors (Yamakawa *et al.*, 1998).

Similarly, a functionally uncharacterised eukaryotic family of Pad1-like domains (Hershey *et al.*, 1996; Aravind & Ponting, 1998) (also called MPN domains (Hofmann & Bucher, 1998)) has additional archaeal and bacterial representatives (Table 1 and Figure 2(b)). These domains have been found as subunits of eukaryotic 26 S proteasome, translation initiation factor-3 and transcriptional activation complexes. A His-(Ser/Thr)-His tripeptide that is conserved in approximately 50% prokaryotic and 50% eukaryotic Pad1-like domains might indicate that this subset of homologues possesses catalytic activity. All of these domains in complete prokaryotic genomes occur as single domain proteins. The only prokaryotic Pad1-like domain-containing protein with more than a single domain is bacteriophage tail assembly protein K (Sanger *et al.*, 1982), which is involved in the assembly of the initiator complex for tail polymerisation. This protein contains a C-terminal peptidase family U-20 domain, common to several prokaryotic cell lysins including p60 (Bubert *et al.*, 1992). This suggests that the Pad1-like domain

(a)

```
T19C9.5 Ce    AEAIDNIVFIHNKLRNAAS(14)MQLLSWNESLVAEAENEKY----YCEPADNK(4)KLGDNIYQYDV---NTYDDIDGVGAMGSINKDTH--
AG5     Di    PTERKNIVTQINKYRSRLI(17)MLRMRWDCKLEKSAQNWAN----MCVFGHSP(5)GIGENVYAYW(8)-KKTAGTDAGRLWWSELEKYYSD
AG5     Vv    EAEKQEILKVHNDFRQKVA(17)MNNLVWNDELANIAQVWAS----QCNYGHDT(7)PVGQNIAKRS(4)-LFDSPGKLVKMWENEVKDFN--
Sc7     Scc   QSEIDQWLKAHNNERAQHG----AVALVWNQTLSDKAADWAS----QCIWEHSNSGQNLAAWFSPQAN---KPMNISQGVGGWNAEEPDYN--
SCP1    Mm    MSVQEEIVSKHNQLRRMVS(4)DLLKMEWNYDAQVNAQQWAD----KCTFSHSP(8)RCGENLFMSS----YLASWSSAIQGWYNEYKDLT--
Glip    Hs    EDFIKDCVRIHNKFRSEVK(4)DMLYMTWDPALAQIAKAWAS----NCQFSHN(14)SLGENIWTGSV---PIFSVSSAITNWYDEIQDYN--
YlbC    Bs    TTSEKQILDLTNVIRVKHG----LAKLEWDQPTAEVAFGHSE--------------DMKENNYFSHVSK-KYGSLKDRLEEGHVDFQQAG--
YkwD    Bs    SAYEKKVVELTNAERQKQG----LKPLQIDETLSKSARAKSQ--------------DMKDKNYFDHQSP-TYGSPFDMMKSFGISYKTAG--
AF0003  Af    EESSKAAIEYLNQLRAQNG----LPPVKWNKTLYEFALERLE--------------DMHERGYYSHYDPVTHETLIYRYVEGYVGECILNGV
BB0689  Bb    KEDMKILYSEIAELRKKLN----LNHLEIDDTLEKVAKEYAI--------------KLGENRTITHTL--FGTTPMQRIHKYDQSFNLTR--
1CFE    Le    QNSPQDYLAVHNDARAQVG----VGPMSWDANLASRAQNYANSRAGDCNLIHS----GAGENLAKGG----GDFTGRAAVQLWVSERPSYN--
2-structure      HHHHHHHHHHHHHH          EE  HHHHHHHHHHHHH                  EE          HHHHHHHHHHHHHEE
consensus     .p.bcphlp.hNpbR.bhs....b..bpWs.sL.p.Ab.bup.............h.-p.h.....hs..p.h..b.spbpph...

T19C9.5 Ce    --DALKSEAK---AAKNRLRQMLYSKSKSIGCIYESCD(7)NYNTRLLICKYSPPLENIDEKLF 3879974/28-185
AG5     Di    -NPSNNLTSEVAMENILHFTQMAWGETYKLGSGVDHNIV-MVARTLVFICHYFPGGNMVKDLIY 2245508/26-195
AG5     Vv    -PNIEWSKNNL--KKTGHYTQMVWAKTKEIGCGSVKYVK-DEWYTHYLVCNYGPSGNFRNEKLY 549194/40-203
Sc7     Scc   --TTTYSGA-------GHWTQVVWKSTTSVGCAAYSCP(12)KTLWYYVCNYYRPGNVSPRDKY 548902/55-197
SCP1    Mm    -YDVGPKQPD---SVVGHYTQVVWNSTFQVACGVAECPK--NPLRYYYVCHYCPVGNYQGRLYT 417753/36-182
Glip    Hs    -FKTRICK-----KVCGHYTQVVWADSYKVGCAVQFCPK(7)SNGAHFICNYGPGGNYPIW-PY 2507371/31-186
YlbC    Bs    --ENIAYNY-----VDGPAAVEGWLNSEGHRKALLNSD------YTHLGVGVDR--KYYTQNFI 2339999/224-443
YkwD    Bs    --ENIAKGQ-----KTPEEVVKAWMNSEGHRKNILNPN------FTHIGVGYVESGSIWTQQFI 2632218/134-255
AF0003  Af    RGTNLLSNGLQSLFGYEEEAIDIWSKSTMHKLILTDKR------FTDAAVACKYDMCVLIMTGG 2650663/33-164
BB0689  Bb    --EILASG------IELNRVVNAWLNSPSHKEALINTD-----TDKIGGYRLKTTDNIDIFVVL 2688622/27-147
1CFE    Le    --YATNQCVGG--KKCRHYTQVVWRNSVRLGCGRARCNN----GWWFISCNYDPVGNWIGQRPY 2624502/1-135
2-structure     HHH EE          HHHHH     EEEEEEEE        EEEEEEE
consensus     ..ps..p.....cbs..sW.po..h.psh.pss.....hhhbh..b..shsh..b..b
```

Figure 2. (*Legend shown on page 737*)

(b)

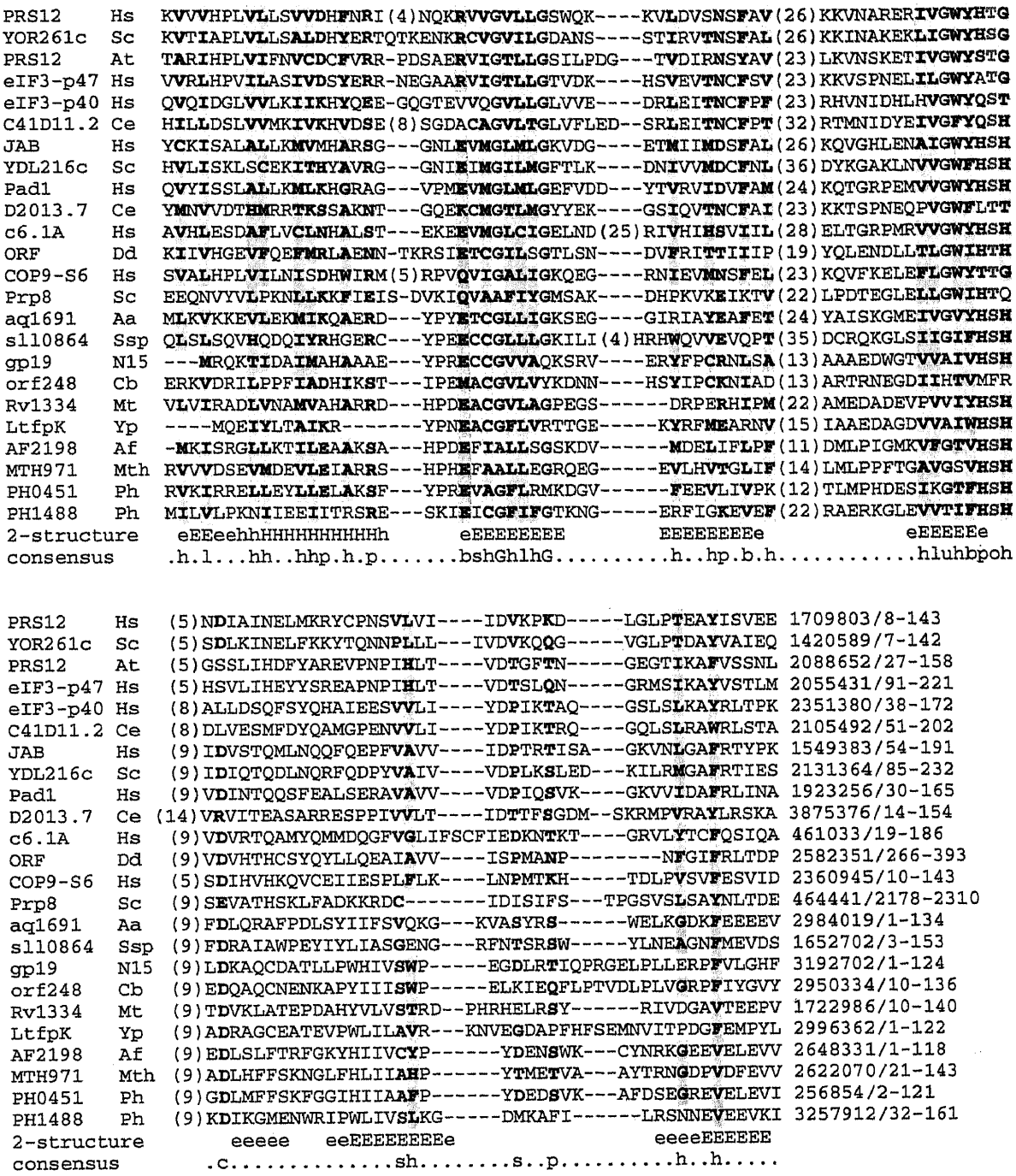

Figure 2. (*Legend shown on page 737*)

may mediate attachment of the lysin to the target cell wall.

PKD domains (The International Polycystic Kidney Disease Consortium, 1995; Sandford *et al.*, 1997) are found in proteins from all four completely sequenced archaeal genomes, but were not detected in complete bacterial genomes (Bycroft *et al.*, 1999). However, the presence of these domains in proteases and cellulases from bacteria whose genomes are currently incomplete has led to suggestions that these domains may have resulted from a past horizontal transfer from animals (Bycroft *et al.*, 1999). In eukaryotes, PKD domains are currently restricted to mammals and birds. This suggests that while it has been disseminated by horizontal transfers in prokaryotes its origin in eukaryotes is unclear.

An unexpected finding was that homologues of von Willebrand factor type A (vWFA) domains are widespread among prokaryotes, being present in all four archaeal and eight of the 11 bacterial genomes studied (Table 1 and Figure 2(c)). Consequently, it is proposed that the vWFA domain was already present in the last common ancestor of the

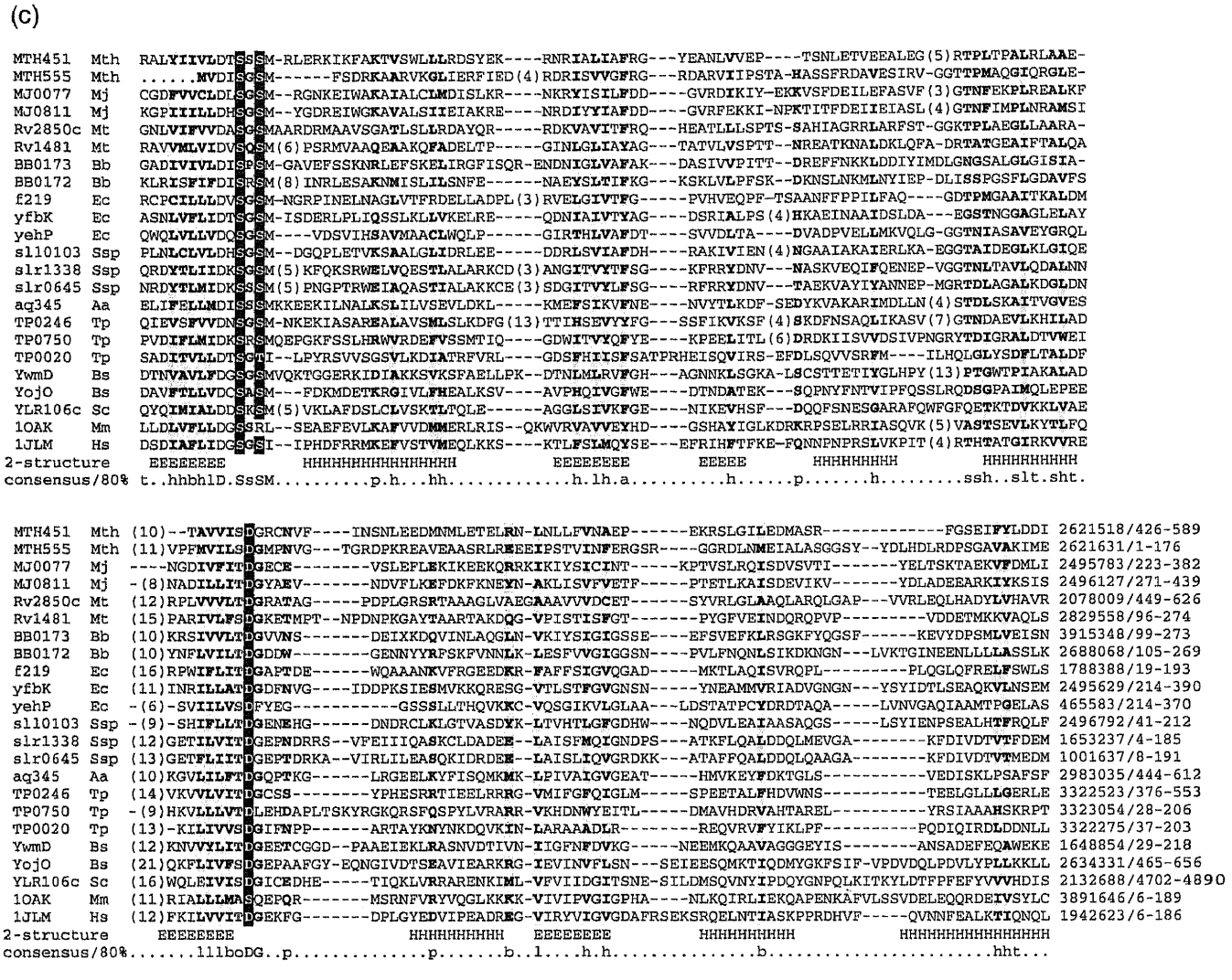

Figure 2. (*Legend shown on page 737*)

(e)

Figure 2. Multiple sequence alignments of previously undetected prokaryotic homologues of eukaryotic signalling domains: (a) PR-1-domains; (b) PAD1-domains; (c) vWFA-domains; (d) SH3-domains; and (e) AN1-like Zn-finger domains. Amino acid residues are coloured according to a 80% consensus (calculated using http://www.bork.embl-heidelberg.de/Alignment/consensus.html; N. Brown & J. Lai, unpublished results): + indicates positively charged residues (H, K and R, coloured in green), −indicates negatively charged residues (D and E, coloured in green), a indicates aromatic residues (F, H, W and Y, highlighted in yellow), b indicates big residues (E, F, I, K, L, M, Q, R, W and Y, indicated in grey or yellow), c indicates charged residues (D, E, H, K and R, coloured in green), h indicates hydrophobic residues (A, C, F, H, I, L, M, V, W and Y, highlighted in yellow), l indicates aliphatic residues (I, L and V, highlighted in yellow), o indicates alcohol residues (S and T, coloured in pink), p indicates polar residues (D, E, H, K, N, Q, R, S and T, coloured in dark blue), s indicates small residues (A, C, S, T, D, N, V, G and P, coloured in light blue), u indicates tiny residues (A, G and S, coloured in light blue). Residues that are predicted to form disulphide bridges (a), that co-ordinate Mn^{2+} in vWFA domains (c) and that are predicted to bind Zn^{2+} (e) are shown as white-on-black. Experimental (Fernandez *et al.*, 1997; Celikel *et al.*, 1998; Wittekind *et al.*, 1994) or predicted (Rost & Sander, 1993) secondary structures are indicated below the alignments, and GenBank identifiers and residue numbers are shown following each alignment. Abbreviations: species: Aa, *Aquifex aeolicus*; Af, *Archaeoglobus fulgidus*; At, *Arabidopsis thaliana*; Bb, *Borrelia burgdorferi*; Bs, *Bacillus subtilis*; Cb, *Coxiella burnetii*; Ce, *Caenorhabditis elegans*; Cp, *Clostridium perfringens*; Cs, *Ciona savignyi*; Ct, *Chlamydia trachomatis*; Dd, *Dictyostelium discoideum*; Di, *Dirofilaria immitis*; Ec, *Escherichia coli*; Gg, *Gallus gallus*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori*; Hs, *Homo sapiens*; Le, *Lycopersicon esculentum* (tomato); Ls, *Listeria welshimeri*; Mj, *Methanococcus jannaschii*; Mm, *Mus musculus*; Mt, *Mycobacterium tuberculosis*; Mth, *Methanobacterium thermoautotrophicum*; N15, bacteriophage N15; PbCV1, *Paramecium bursaria* chlorella virus 1; Ph, *Pyrococcus horikoshii*; phTw, *Staphylococcus* phage Twort; Pv, *Phaseolus vulgaris*; Sa, *Staphylococcus aureus*; Sc, *Saccharomyces cerevisiae*; Scc, *Schizophyllum commune*; Sm, *Streptococcus mutans*; Sp, *Schizosaccharomyces pombe*; Ss, *Staphylococcus simulans*; Ssp, *Synechocystis* sp.; Tp, *Treponema pallidum*; Vv, *Vespula vidua*; Xl, *Xenopus laevis*; and, Yp, *Yersinia pestis*. Other abbreviations: AG5, venom allergen antigen Ves vi 5; Alys, autolysin; eIF3, eukaryotic initiation factor 3; Glip, Glioma pathogenesis-related protein; JAB, Jun activation domain binding protein; LtfpK, lambda tail fiber protein K; Lys, lysin; Lysosta, lysostaphin; LytH, *N*-acetylmuramoyl-L-alanine amidase; Prp8, Pre-mRNA splicing factor; PRS12, 26 S proteasome regulatory subunit S12.

three divisions of cellular life. Archaeal and bacterial vWFA homologs were detected using the PSI-BLAST profile and investigated in additional detail. For example, a search with the C-terminal region (residues 371-676) of *Synechocystis* PCC6803 ChlD (gi 3913253) as query revealed significant similarity with the vWFA domain of porcine inter-α-trypsin inhibitor heavy chain (gi 3024051), with $E = 1 \times 10^{-3}$. The majority of prokaryotic vWFA domains show conservation of aspartic acid and serine residues known from structures of eukaryotic vWFA domains to interact with divalent cations (e.g. Asp156, Ser158, Ser160 and Asp258 of human CD11b; Figure 2(c)). This suggests that these domains also bind divalent cations.

von Willebrand factor is an adhesive multidomain glycoprotein, present in mammals, that mediates platelet adhesion to the injured vascular subendothelium *via* the first and third of its four vWFA domains (reviewed by Sadler, 1989). vWFA domains have also been recognized in a variety of other metazoan proteins, including collagens and integrin α-subunits, and much has been made of metal ion-dependent adhesion sites (MIDAS) present in a subset of vWFA domains that mediate ligand-binding events (Baldwin *et al.*, 1998).

The functions of the majority of prokaryotic vWFA homologues remain unknown. However, the data on two functionally characterized bacterial vWFA domain homologues suggest that the metal ion and/or protein-binding functions of prokaryotic vWFA domains may be similar to those of their eukaryotic homologues. Specifically, the terY protein of IncHI2 plasmid R478 assists in protecting *E. coli* from the toxic effects of heavy metals (Whelan *et al.*, 1997) and hence might bind metal ions. The vWFA domain-containing D subunit of the enzyme magnesium-protoporphyrin IX chelatase interacts with the enzyme's I subunit in a Mg^{2+}-dependent manner (Jensen *et al.*, 1998). How-

ever, the diverse cellular functions of eukaryotic vWFA domains, including membrane trafficking regulation (Creutz *et al.*, 1998), intracellular proteolysis and transcription regulation (it is present in the S5a and p44 subunits of the 26 S proteasome and TFIIH complexes, respectively; Aravind & Ponting, 1998), suggest that the functions of the vWFA domains in prokaryotes are likely also to be variable. This notion is supported by contrasting domain architectures and cellular localisations of prokaryotic vWFA domain-containing proteins. For example, *E. coli* b2073 and *Synechocystis* sp. slr0645 and slr1338, and *M. thermoautotrophicum* MTH555 contain only vWFA domains; *B. subtilis* ywmC and ywmD, *E. coli* yfbK and *T. pallidum* TP0750 are predicted to be secreted vWFA domains; *M. tuberculosis* Rv1481, and *B. burgdorferi* BB0172 and BB0173 are predicted to contain three transmembrane regions and a vWFA domain; and, ChlI/BchI magnesium chelatase subunits contain a AAA ATPase domain (Neuwald *et al.*, 1999) and a vWFA domain.

Phyletic distributions of non-enzymatic signalling domains: domains specific to eukaryota and bacteria, but not archaea

Detection of past horizontal gene transfer events relies on the examination of the phyletic distribution of a given family of orthologous genes (domains) and/or on observation of significant differences between the topology of the phylogenetic tree for the family and that of the organismal tree. Whenever a given domain is common and present in a variety of different contexts in one division (in the context of the present work, that of the eukaryotes) but shows a sporadic distribution without a significant diversity in another division (e.g. bacteria), a case for horizontal transfer may exist. Counter-arguments might be advanced that such cases arise primarily from multiple gene deletions in diverse organisms. As the volume of sequence data from diverse organisms increases, it is becoming increasingly possible to construct most parsimonious evolutionary scenarios for individual gene (domain) families that account for the contributions of both horizontal gene transfer and lineage-specific gene loss (Koonin *et al.*, 1997; Aravind *et al.*, 1998, 1999a; Wolf *et al.*, 1999b; Doolittle & Logsdon, 1998; Doolittle, 1998; Woese, 1998; Lawrence & Ochman, 1998).

The observed phyletic distribution of signalling domains (Table 1) strongly suggests horizontal transfer of genes from eukaryotic genome donors to bacterial genome acceptors as the most likely evolutionary route for a number of domains. Together with the aforementioned arguments for this phenomenon for members of the phospholipase C, cyclase and Sec7 domain families, the following were detected and are proposed here to be results of horizontal gene transfers: leucine-rich repeat proteins in *E. coli* (b1471) and in *T. pallidum* (TpLRR), a SET domain (Suvar3-9, enhancer-of-zeste, trithorax (Tschiersch *et al.*, 1994)) protein in *C. trachomatis* (CT737), two SWIB domain (SWI complex, BAF60b domains (Wang *et al.*, 1996)) proteins in *C. trachomatis* (topA and CT460), and toll-interleukin-1-resistance (TIR) domain (Whitham *et al.*, 1994) proteins in *B. subtilis* (yddK), *Synechocystis* sp. (slr0658), *Rhizobium* sp. NGR234 (Y4LF) and *Streptomyces coelicor* (SC6A9.25, SC6A9.38 and SC7H1.23).

We also observed that tandem repeats present in bacterial NHL (NCL-1, HT2A, LIN-41) domains (Slack & Ruvkun, 1998) are homologues of the YWTD repeat family of proposed β-propeller domains that are widespread in eukaryotic extracellular proteins (Springer, 1998). Among bacteria, these repeats are restricted to human pathogens (Slack & Ruvkun, 1998), which suggests that they have been acquired *via* horizontal gene transfer from their eukaryotic host. It is particularly striking that none of these domain families that have apparently been transferred from eukaryotes to bacteria was detected in known archaeal sequences.

The lineage of *Synechocystis* PCC6803 appears to be particularly acquisitive of eukaryotic-type domains, since this cyanobacterial genome (Kaneko *et al.*, 1996) was found to contain considerably more proteins containing such domains than any other prokaryote. The present analysis added several new domains to the list of likely horizontal transfers from eukaryotes in *Synechocystis*. The domain repertoire of this organism include a TIR domain (as above), EF-hands (talB), cadherin repeats (slr2046), ankyrin repeats (slr1109), fasciclin-like domains (sll1483 and sll1735), a bulb-type lectin domain (slr1028), integrin α and β4-subunit domains (slr1403, slr0408 and slr1028; May & Ponting, 1999) and WD40 repeat-containing β-propellers (sll0163, slr0143, sll1491, slr1410, slr0439 and slr1409; Dagnall & Saier, 1997; Bedu & Joset, 1997; Zhang *et al.*, 1998).). Sequence similarities of these domains to eukaryotic homologues are considerable, indicating that these are relatively recent acquisitions to this cyanobacterium's lineage. The predominance of eukaryotic signalling domains in its proteome suggests an intimate symbiotic association with a eukaryotic host some time in the evolutionary history of this cyanobacterium.

The proposition that horizontal gene transfer, distinct from the influx of mitochondrial and chloroplast genes into the nuclear genome, has occurred between the bacterial and eukaryotic lineages is not new. Of the non-enzymatic domains or repeats considered here bacterial ankyrin repeats (Bork, 1993), discoidin-like domains (Baumgartner *et al.*, 1998), and fibronectin type 3 domains (Little *et al.*, 1994) had been proposed as having arisen from horizontal transfer events from eukaryotic genomes. Conversely, eukaryotic AP-ATPases have been proposed as having been transferred from bacteria but not from the organelles (Aravind *et al.*, 1999b). Although previously proposed as having arisen from vertical transfer only (Bateman

et al., 1996), at least some of the bacterial immunoglobulin (IG)-like domains are also likely to have arisen from horizontal gene transfer from eukarya, since they are considerably more similar to eukaryotic IG domains than they are to any other prokaryotic sequence. For example, two rounds of PSI-BLAST (Altschul *et al.*, 1997) with the two IG domains of *Cellulomonas fimi* cellulase C (amino acid residues 918-1101) as query, retrieved 499 eukaryotic IG homologues and no other bacterial homologues with an *E*-value threshold of 0.001. No IG domains were identified in complete prokaryotic genomes.

It is evident from this study that considerable numbers of bacterial genes have been acquired from eukaryotic sources *via* horizontal gene transfer. Given that the examples discussed here are only those supported by clean-cut sequence-based evidence, it is likely that many more such instances have occurred that are less evident from sequence information.

Non-enzymatic signalling domains specific to eukaryota and bacteria, but not archaea: the case of bacterial lysis proteins and SH3 domains

Src homology 3 (SH3) domains are widespread among metazoan intracellular signalling proteins and typically bind proline-rich polypeptides (Mayer & Eck, 1995). The search protocol using PSI-BLAST-constructed profiles for eukaryotic SH3 domains (see Methods) detected several bacterial proteins containing homologues of these domains (Figures 2(d) and 3). A reverse PSI-BLAST search of the non-redundant sequence database with *B. subtilis* YfhK (gi 2804541) as query revealed significant similarities with eukaryotic SH3 domains using an *E*-value inclusion threshold of 0.01. Although this threshold was greater than that applied elsewhere for this project, a HMMER2 search with a HMM calculated from a multiple alignment of bacterial SH3 domain sequences also detected significant similarity to a eukaryotic SH3 domain of known structure (mouse Grb2, N-terminal domain; Wittekind *et al.*, 1994), among others, with significance ($E = 8.3 \times 10^{-3}$). After submission of the manuscript, SH3 domains were predicted independently in a subset of prokaryotic proteins presented here (Whisstock & Lesk, 1999).

A subset of bacterial proteins that contain the newly detected SH3 domains also contains cell lytic enzymatic domains, such as a lysozyme homologous domain in *B. subtilis* LytD (Figure 3). Consequently these are predicted to be cell lysins that function during pathogen-host interaction or at completion of cell division. One of such SH3 domain-containing lysins is *Staphylococcus simulans* lysostaphin that cleaves peptidoglycans of target *S. aureus* cells. A 92 residue C-terminal portion of *S. simulans* lysostaphin that entirely encompasses its predicted SH3 domain is known to mediate the binding of lysostaphin to the *S. aureus* cell wall

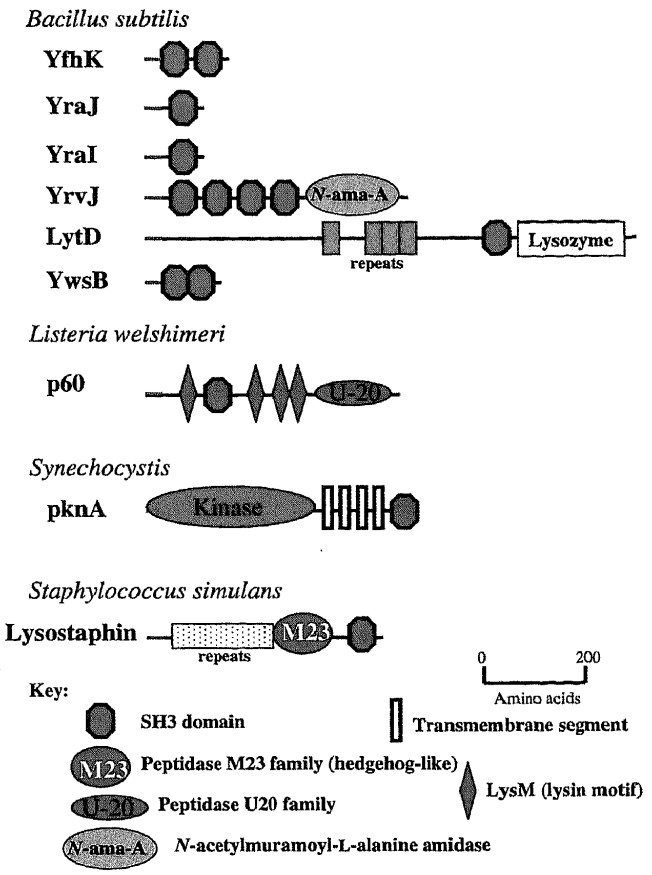

Figure 3. The domain architectures of representative bacterial proteins that contain previously undetected prokaryotic homologues of eukaryotic SH3 domains.

(Baba & Schneewind, 1996). Consequently it is predicted that other bacterial SH3 domains possess similar functions. Whether the binding ligand for these domains is similar to the proline-rich ligands of eukaryotic SH3s deserves further experimental investigation. The presence of a well-conserved arginine residue (highlighted in Figure 2(e)) in bacterial SH3 domains, that is less well conserved among eukaryotic SH3s, argues for a function of bacterial homologues that differs from their eukaryotic counterparts. The known sequences of neither archaea nor plants possess detectable SH3 domain homologues. This distribution suggests an early horizontal gene transfer between eukaryota and bacteria, although the direction of transfer is unclear.

Synechocystis PCC6803 protein sll0776 (gi 1006577) contains a C-terminal SH3 domain preceded by a Pkn2-type (Leonard *et al.*, 1998) protein kinase (Figure 3). Since bacterial Pkn2-type kinases seem to have arisen from horizontal gene transfer from eukaryota to bacteria (Leonard *et al.*, 1998), this protein appears to contain two distinct domains that have each undergone horizontal gene transfer between bacterial and eukaryotic divisions. However, since there are no known examples of eukaryotic proteins with Pkn2-type

kinase domains followed by C-terminal SH3 domains, it is unlikely that these horizontal transfer events were concurrent.

Several of the bacterial SH3 domain-containing proteins also contain a repeat motif noted to be present in bacterial lysins (Birkeland, 1994); these are denoted LysM in Table 1 and Figure 3. HMMER2 searches of current databases with a multiple alignment of LysM repeats demonstrated significant similarities to repeats in *C. elegans* and *Kluyveromyces lactis* chitinases. The *K. lactis* protein is known to contain subunits of a killer toxin that inhibits growth of sensitive yeast cells (Stark & Boyd, 1986). In common with many other extracellular eukaryotic homologues of intracellular prokaryotic domains (see above), the yeast and worm repeats contain conserved cysteine residues, indicative of disulphide bridges where their bacterial counterparts contain none.

Phyletic distributions of non-enzymatic signalling domains: domains specific to eukaryota and archaea, but not bacteria

Since eukaryotic to bacterial horizontal gene transfer has apparently been so widespread, it was unexpected that only a single possible example of horizontal transfer from eukarya to archaea was detected in this study (Figure 1). This example relates to a putative zinc-chelating domain, distinct from other Zn fingers, contained in the ubiquitin-like fusion protein AN1 (Linnen *et al.*, 1993) in metazoa, plants and fungi. We found that this domain is also present in three proteins (AF1011, AF1187 and AF1454) of the archaeon *A. fulgidus* but not in any other prokaryote (Figure 2(e)). The reasons for the lack of such cases may relate to the fact that all currently available complete archaeal genomes are from extreme thermophiles that have limited contact with eukaryotes in the environs where they thrive. A recent comprehensive analysis of the evolutionary patterns for these archaeal species revealed very few instances of apparent horizontal transfer from eukaryotes (Makarova *et al.*, 1999b). Sequencing of genomes from mesophylic archaea, particularly symbiotic ones, such as *Cenarchaeum symbiosum*, is eagerly awaited. Analysis of these genomes is expected to resolve whether the current near absence of detectable cases of horizontal transfer from eukaryotes to archaea is due to the extreme lifestyle of the currently studied archaea, or whether there are fundamental reasons precluding such evolutionary events.

Concluding remarks: evolutionary interpretation of the phyletic distribution of eukaryotic-type signalling domains

In terms of their phyletic distribution and likely evolutionary routes, eukaryotic signalling domains can be classified into three groups: (i) those of universal provenance; (ii) those found only in two of the divisions of life, typically eukaryotes and bacteria; and (ii) those detectable only in eukaryotes (Table 1 and Figure 1). Figure 4 summarises, in a schematic form, our current ideas of the history of these domains. Among the universally distributed domains, some, in particular the PDZ, CBS, vWFA,

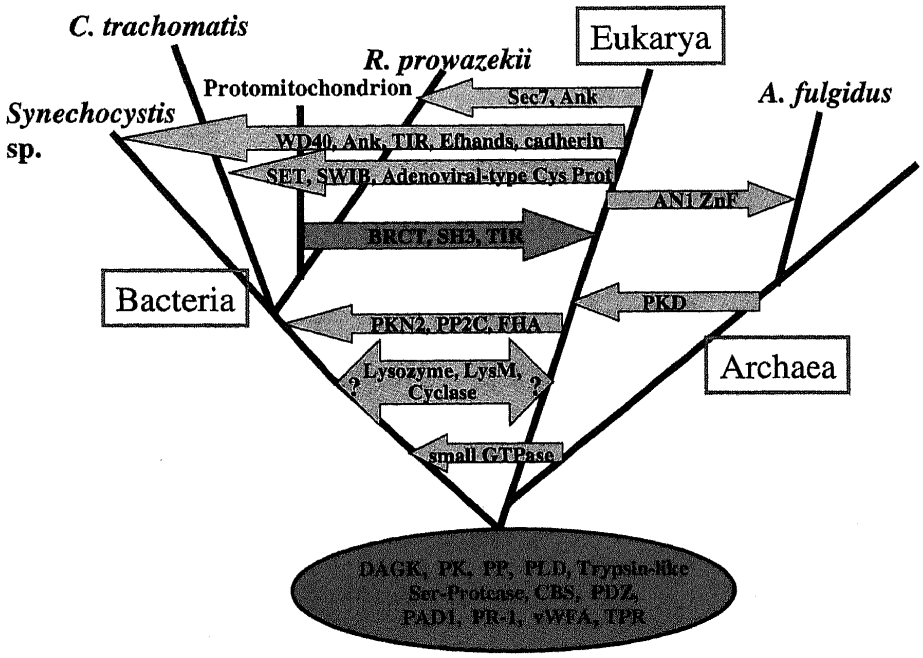

Figure 4. Schematic representation of proposed evolutionary histories of selected signalling domains. Blue arrows indicate proposed horizontal gene transfer events, and the red arrow represents gene acquisition from mitochondrial endosymbiosis. Domains represented within the green oval are suggested to have been present in the last common ancestor of archaea, eukaryota and bacteria. The directions of proposed horizontal transfers of lysozyme, LysM, LRR and cyclase domains between eukaryotes and bacteria are not apparent from our analyses.

Pad1-like and PR-1 domains and the TPR motifs, are likely to have emerged in the common ancestor of all life forms. Notably, although conservation of domain architectures of signalling proteins across divisions is not typical, some distinct arrangements containing these universal domains, such as the combination of PDZ domains with the S2P protease domain, are represented in all three divisions.

The domains that are present in only two of the divisions, typically eukaryotes and bacteria, are likely to have evolved by lateral dissemination (Figure 4) unless there has been a very early gene loss in the third division (archaea). Comparisons of the phyletic distributions of the given domain in eukaryotes and in bacteria enables one to determine which of these two evolutionary scenarios is the more likely for each particular case. In a number of cases, such as, for example, those of the bulb lectin, integrin domains, SET, SWIB, Sec-7, WD40 and YWTD repeat-containing domains, the presence of the signalling domain in bacteria appears to be a result of a direct horizontal transfer from a eukaryote. This conclusion is supported by the finding that these domains are found in a single bacterial group, being otherwise uniquely eukaryotic. Furthermore, the bacteria, in which these domains have been detected, such as *Chlamydiae, Rickettsiae* and *Synechocystis*, are in an intimate contact with the host eukaryote as a part of their parasitic life cycle or as the result of symbiosis, thereby providing ample opportunity for horizontal gene transfer to occur. Other domains, such as the classical signalling enzymes, PKN2-type kinases and adenylyl cyclases, may be cases of ancient horizontal transfer from the eukaryotes which have undergone further dispersal by means of inter-prokaryotic transfers. Interestingly, FHA domains that bind phosphorylated proteins occur only in those genomes which also encode eukaryote-type PKN-2 kinases. This suggests that these two domains have been disseminated in a co-ordinated manner by horizontal transfer during bacterial evolution.

The most puzzling group of signalling domains are those that occur only in two of the divisions of life, namely bacteria and eukaryotes, and are ubiquitous or at least very common in each division. These include the SH3, TIR and BRCT domain families. The most parsimonious explanation in these cases may be a single ancient horizontal transfer between the two divisions. In particular, it seems possible that these domains originally have evolved in bacteria, have been transferred to eukaryotes as a result of the mitochondrial endosymbiosis and subsequently have undergone major expansions during the evolution of the eukaryotes. However, the alternative model of an origin in the last common ancestor, with a loss at the base of the archaeal lineage, cannot be ruled out for these domains. One interesting feature of the distribution of "eukaryotic signalling domains" in bacteria is their preponderance in organisms with a multi-stage developmental cycle, such as the mycobacteria, the cyanobacteria and chlamydia, as against those with simple developmental cycles such as *E. coli*. This suggests that just as with eukaryotes these proteins have been recruited to participate in complex signalling events that are characteristic of development.

Analysis of "eukaryotic" signalling domains in prokaryotes is of major interest for understanding their evolution and their role in different aspects of bacterial physiology, including pathogenicity. It appears that the currently available 20 or so complete prokaryotic genomes contain sufficient information to define the principal conserved protein families involved in essential cellular functions. The situation with signalling domains is quite different in that further increase in the diversity of the collection of sequenced genomes is expected to significantly extend our understanding of their functions and fates in evolution.

Methods

Databases

Twelve complete bacterial genomes (namely, those of *Aquifex aeolicus, Bacillus subtilis, Borrelia burgdorferi, Chlamydia trachomatis, Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Mycoplasma genitalium, Mycoplasma pneumoniae, Mycobacterium tuberculosis, Synechocystis* PCC6803 and *Treponema pallidum*), four complete archaeal genomes (*Archaeoglobus fulgidus, Methanobacterium thermoautotrophicum, Methanococcus jannaschii* and *Pyrococcus horikoshii*) and two complete eukaryotic genomes (*Caenorhabditis elegans* and *Saccharomyces cerevisiae*) were investigated.

Sequence data for all genomes were acquired from the GenBank database *via* Entrez (http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html). A total of 205 previously prepared multiple alignments of intracellular and extracellular signalling domains were acquired from the SMART tool and database (http://coot.embl-heidelberg.de/SMART/: Schultz *et al.*, 1998; Ponting *et al.*, 1999). Information concerning acronyms, phyletic distributions, known structures and links to literature information is available from the SMART Web site. Other multiple alignments, namely those of papain and adenoviral-type cysteine proteases, transglutaminase, kelch, TIR, NIβ, MATH, SET, AP-ATPase, BRCT, Pad1, LysM, fasciclin-like, SWIB, Sel1-like and metalloprotease domains, were prepared *de novo* (C.P.P., L.A., J.S., P.B. & E.V.K., unpublished results), or were brought up-to-date from previously published alignments. Multiple alignments were "seeded" to remove one of each pair of sequences that were greater than either 67% or approximately 80% identical (for details see Schultz *et al.* (1998)). Searches against all existing sequence data employed the non-redundant database available from the National Center for Biotechnology Information (ftp://ncbi.nlm.nih.gov/blast/db).

Sequence analysis

Sequences were considered homologues if one or both of the following two methods provided evidence for significant similarities in sequence. The position-specific iterative BLAST (PSI-BLAST) method (Altschul *et al.*, 1997) was used primarily, whereas the Hidden Markov model (HMM) database searching algorithm of HMMER2 (http://hmmer.wustl.edu/) was of particular use in identifying homologues containing multiple repeats. All sequences in SMART alignments were extracted and used as queries in PSI-BLAST searches. Typically, a maximum of eight iterations and an *E*-value inclusion threshold of 0.001 were used, and no sequence filters (such as those for regions of biased composition (Wootton & Federhen, 1996) or coiled-coil domains (Lupas *et al.*, 1991)) were applied. These parameters are similar to those used elsewhere that were found to achieve optimal low rates of false positive predictions (Park *et al.*, 1998). The resulting PSI-BLAST output was inspected for the inclusion of false positive sequences and for the detection of true positive sequences from prokaryotic organisms. Such assessments were assisted by the use of the taxonomy analysis procedures of SEALS (Walker & Koonin, 1997). Prediction of true positive *versus* false positive sequences made use of reciprocal PSI-BLAST and HMMER2 searches, and visual inspection for the conservation of signature motifs, characteristic of individual domains.

The identities and numbers of proteins containing domains or repeats from a particular homologous family were acquired using PSI-BLAST and HMMER2 database searches. PSI-BLAST searches of complete genome sequences employed profiles, derived from verified multiple alignments of sequences identified from the non-redundant sequence database with *E*-values <0.001 within eight iterations. These profiles were saved and used to search the protein sets from complete genomes (Wolf *et al.*, 1999a). HMMER2 searches of complete genomes used a HMM derived from verified multiple alignments and calculated using default parameters. Sequences identified with *E*-values $<10^{-3}$ (PSI-BLAST) or *E*-values $<10^{-2}$ (HMMER2) were considered to represent true homologues. Sequences identified with $10^{-3} < E < 10^{-1}$ (PSI-BLAST) or $10^{-2} < E < 10$ (HMMER2) were further investigated using reciprocal PSI-BLAST searches. Of these, only sequences that were predicted with significance (PSI-BLAST $E < 10^{-3}$) were deemed to be homologues. Detection of short repeats, such as tetratricopeptide repeats (TPRs) or leucine-rich repeats (LRRs), was straight-forward using HMMER2, since this method employs a heuristic procedure that estimates a single *E*-value from one or more alignments within the same sequence (S.R. Eddy, unpublished; http://hmmer.wustl.edu/). Homologues were predicted from sequence analysis only, without recourse to analyses of structural data.

The necessity of using *E*-value thresholds in this study to reduce the numbers of false positives has the consequence that the numbers of proteins containing a particular domain or repeat represented in complete genomes are likely to be underestimates. Indeed, sequence-based methods are expected to identify only the minority of true homologues (Park *et al.*, 1998). Additionally, individual domain families may be identified as divergent representatives of a larger set of families (e.g. see Beckmann *et al.*, 1998). It is anticipated that improved analytical methods (Huynen *et al.*, 1998; Wolf *et al.*, 1999a,b) and the results of structural genomics projects (Rost, 1998) will continue to identify divergent homologues with the consequence that the number of distinct domain families will diminish (Aravind *et al.*, 1996b; Bork *et al.*, 1997b).

Acknowledgements

We apologise to the many investigators whose important contributions to the understanding of signalling domains, their evolution, function and structure, are not cited here because of space restrictions. This work has been funded, in part, by a National Research Council (USA) Senior Associateship (to C.P.P.).

References

Altschul, S. F. & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444-447.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Aravind, L. & Koonin, E. V. (1998). The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **23**, 469-472.

Aravind, L. & Ponting, C. P. (1997). The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem. Sci.* **22**, 458-459.

Aravind, L. & Ponting, C. P. (1998). Homologues of 26 S proteasome subunits are regulators of transcription and translation. *Protein Sci.* **7**, 1250-1254.

Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. (1998). Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14**, 442-444.

Aravind, L., Walker, D. R. & Koonin, E. V. (1999a). Conserved domains in DNA repair proteins and evolution of repair systems. *Nucl. Acids Res.* **27**, 1223-1242.

Aravind, L., Dixit, V. M. & Koonin, E. V. (1999b). The domains of death: evolution of the apoptosis machinery. *Trends Biochem. Sci.* **24**, 47-53.

Baba, T. & Schneewind, O. (1996). Target cell specificity of a bacteriocin molecule: a C-terminal signal directs lysostaphin to the cell wall of *Staphylococcus aureus*. *EMBO J.* **15**, 4789-4797.

Baldwin, E. T., Sarver, R. W., Bryant, G. L., Jr., Curry, K. A., Fairbanks, M. B., Finzel, B. C., Garlick, R. L., Heinrikson, R. L., Horton, N. C., Kelley, L. L., Mildner, A. M., Moon, J. B., Mott, J. E., Mutchler, V. T., Tomich, C. S., Watenpaugh, K. D. & Wiley, V. H. (1998). Cation binding to the integrin CD11b I domain and activation model assessment. *Structure*, **6**, 923-935.

Bateman, A. (1997). The structure of a domain common to archaebacteria and the homocystinuria disease protein. *Trends Biochem. Sci.* **22**, 12-13.

Bateman, A., Eddy, S. R. & Chothia, C. (1996). Members of the immunoglobulin superfamily in bacteria. *Protein Sci.* **5**, 1939-1941.

Baumgartner, S., Hofmann, K., Chiquet-Ehrismann, R. & Bucher, P. (1998). The discoidin domain family revisited: new members from prokaryotes and a

homology-based fold prediction. *Protein Sci.* **7**, 1626-1631.

Beckmann, G., Hanke, J., Bork, P. & Reich, J. G. (1998). Merging extracellular domains: fold predictions for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins. *J. Mol. Biol.* **275**, 725-730.

Bedu, S. & Joset, F. (1997). HatA and HatR, implicated in the regulation of inorganic carbon uptake process in *Synechocystis* PCC6803, may contain WD domains. *Mol. Microbiol.* **24**, 230.

Birkeland, N. K. (1994). Cloning, molecular characterization, and expression of the genes encoding the lytic functions of lactococcal bacteriophage phi LC3: a dual lysis system of modular design. *Canad. J. Microbiol.* **40**, 658-665.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally-diverse proteins: mobile modules that cross phyla horizontally. *Proteins: Struct. Funct. Genet.* **17**, 363-374.

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

Bork, P., Brown, N. P., Hegyi, H. & Schultz, J. (1996a). The protein phosphatase 2C (PP2C) superfamily: detection of bacterial homologues. *Protein Sci.* **5**, 1421-1425.

Bork, P., Downing, A. K., Kieffer, B. & Campbell, I. D. (1996b). Structure and distribution of modules in extracellular proteins. *Quart. Rev. Biophys.* **29**, 119-167.

Bork, P., Hofmann, K., Bucher, P., Neuwald, A. F., Altschul, S. F. & Koonin, E. V. (1997a). A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J.* **11**, 68-76.

Bork, P., Schultz, J. & Ponting, C. P. (1997b). Cytoplasmic signalling domains: the next generation. *Trends Biochem. Sci.* **22**, 296-298.

Bubert, A., Kuhn, M., Goebel, W. & Kohler, S. (1992). Structural and functional properties of the p60 proteins from different Listeria species. *J. Bacteriol.* **174**, 8166-8171.

Bycroft, M., Bateman, A., Clarke, J., Hamill, S. J., Sandford, R., Thomas, R. L. & Chothia, C. (1999). The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease. *EMBO J.* **18**, 297-305.

Celikel, R., Varughese, K. I., Madhusudan, , Yoshioka, A., Ware, J. & Ruggeri, Z. M. (1998). Crystal structure of the von Willebrand factor A1 domain in complex with the function blocking NMC-4Fab. *Nature Struct. Biol.* **5**, 189-194.

Chen, J. M., Rawlings, N. D., Stevens, R. A. & Barrett, A. J. (1998). Identification of the active site of legumain links it to caspases, clostripain and gingipains in a new clan of cysteine endopeptidases. *FEBS Letters*, **441**, 361-365.

Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, J. M. & Botstein, D. (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022-2028.

Creutz, C. E., Tomsig, J. L., Snyder, S. L., Gautier, M. C., Skouri, F., Beisson, J. & Cohen, J. (1998). The copines, a novel class of C2 domain-containing, calcium-dependent, phospholipid-binding proteins conserved from Paramecium to humans. *J. Biol. Chem.* **273**, 1393-1402.

Dagnall, B. H. & Saier, M. H., Jr (1997). HatA and HatR, implicated in the uptake of inorganic carbon in *Synechocystis* PCC6803, contain WD40 domains. *Mol. Microbiol.* **24**, 229-230.

Delbaere, L. T., Hutcheon, W. L., James, M. N. & Thiessen, W. E. (1975). Tertiary structural differences between microbial serine proteases and pancreatic serine enzymes. *Nature*, **257**, 758-763.

Doolittle, R. F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287-314.

Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307-311.

Doolittle, W. F. & Logsdon, J. M., Jr (1998). Archaeal genomics: do archaea have a mixed heritage? *Curr. Biol.* **8**, R609-R611.

Fernández, C., Szyperski, T., Bruyère, T., Ramage, P., Mösinger, E. & Wüttrich, K. (1997). NMR solution structure of the pathogenesis-related protein P14a. *J. Mol. Biol.* **266**, 576-593.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99-113.

Fitch, W. M. (1995). Uses for evolutionary trees. *Phil. Trans. Roy. Soc. ser. B*, **349**, 93-102.

Gibson, T. J., Thompson, J. D. & Heringa, J. (1993). The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acid. *FEBS Letters*, **324**, 361-366.

Hartzell, P. L. (1997). Complementation of sporulation and motility defects in a prokaryote by a eukaryotic GTPase. *Proc. Natl Acad. Sci. USA*, **94**, 9881-9886.

Heinz, D. W., Ryan, M., Bullock, T. L. & Griffith, O. H. (1995). Crystal structure of the phosphatidylinositol-specific phospholipase C from *Bacillus cereus* in complex with myo-inositol. *EMBO J.* **14**, 3855-3863.

Heinz, D. W., Essen, L.-O. & Williams, R. L. (1998). Structural and mechanistic comparison of prokaryotic and eukaryotic phosphoinositide-specific phospholipases C. *J. Mol. Biol.* **275**, 635-650.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K. & Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609-614.

Hershey, J. W., Asano, K., Naranda, T., Vornlocher, H. P., Hanachi, P. & Merrick, W. C. (1996). Conservation and diversity in the structure of translation initiation factor eIF3 from humans and yeast. *Biochimie*, **78**, 903-907.

Hofmann, K. & Bucher, P. (1995). The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem. Sci.* **20**, 347-349.

Hofmann, K, & Bucher, P. (1998). The PCI domain: a common theme in three multiprotein complexes. *Trends Biochem. Sci.* **23**, 204-205.

Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S. & Bork, P. (1998). Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323-326.

Jensen, P. E., Gibson, L. C. D. & Hunter, C. N. (1998). Determinants of catalytic activity with the use of purified I, D and H subunits of the magnesium protoporphyrin IX chelatase from *Synechocystis* PCC6803. *Biochem. J.* **334**, 335-344.

Kaneko, T., Sato, S., Kotani, H., Tamaka, A., Asamiza, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T.,

Matsuno, A., Muraki, A. & Nakazaki, N. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109-136.

Kennelly, P. J. & Potts, M. (1996). Fancy meeting you here! A fresh look at "prokaryotic" protein phosphorylation. *J. Bacteriol.* **178**, 4759-4764.

Kitajima, S. & Sato, F. (1999). Plant pathogenesis-related proteins: molecular mechanisms of gene expression and protein function. *J. Biochem.* **125**, 1-8.

Koonin, E. V. (1996). A duplicated catalytic motif in a new superfamily of phosphohydrolases and phospholipid synthases that includes poxvirus envelope proteins. *Trends Biochem. Sci.* **21**, 242-243.

Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996). *Escherichia coli*-Functional and evolutionary implications of genome scale computer-aided protein sequence analysis. In *Genomes of Plants and Animals* (Gustafson, J. P. & Flavell, R. B., eds), pp. 177-210, Plenum Press, New York.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997). Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* **25**, 619-637.

Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA*, **95**, 9413-9417.

Leonard, C. J., Aravind, L. & Koonin, E. V. (1998). Novel families of putative protein kinases in bacteria and archaea: evolution of the "Eukaryotic" protein kinase superfamily. *Genome Res.* **8**, 1038-47.

Linnen, J. M., Bailey, C. P. & Weeks, D. L. (1993). Two related localized mRNAs from *Xenopus laevis* encode ubiquitin-like fusion proteins. *Gene*, **128**, 181-188.

Little, E., Bork, P. & Doolittle, R. F. (1994). Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *J. Mol. Evol.* **39**, 631-643.

Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, **252**, 1162-1164.

Makarova, K. S., Aravind, L. & Koonin, E. V. (1999a). A superfamily of archaeal, bacterial and eukaryotic proteins homologous to animal transglutaminases. *Protein Sci.* **in the press.**

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I. & Koonin, E. V. (1999b). Comparative genomics of the archaea: evolution of conserved protein families, the stable core and the variable shell. *Genome Res.* **in the press.**

Matsuo, Y., Yamada, A., Tsukamoto, K., Tamura, H.-O., Ikezawa, H., Nakamura, H. & Nishikawa, K. (1996). A distant evolutionary relationship between bacterial sphingomyelinase and mammalian DNase I. *Protein Sci.* **5**, 2459-2467.

May, A. P. & Ponting, C. P. (1999). Integrin α- and β4-subunit-domain homologues in cyanobacterial proteins. *Trends Biochem. Sci.* **24**, 12-13.

Mayer, B. J. & Eck, M. J. (1995). SH3 domains. Minding your p's and q's. *Curr. Biol.* **5**, 364-367.

McKerrow, J. H. (1987). Human fibroblast collagenase contains an amino acid sequence homologous to the zinc-binding site of Serratia protease. *J. Biol. Chem.* **262**, 5943.

Mushegian, A. R., Garey, J. R., Martin, J. & Lin, L. X. (1998). Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.* **8**, 590-598.

Neuwald, A. F., Aravind, L., Spouge, J. L. & Koonin, E. V. (1999). AAA+: a class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.* **9**, 27-43.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.

Ponting, C. P. (1997a). Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Sci.* **6**, 464-468.

Ponting, C. P. (1997b). CBS domains in ClC chloride channels implicated in myotonia and nephrolithiasis (kidney stones). *J. Mol. Med.* **75**, 160-163.

Ponting, C. P. & Aravind, L. (1997). PAS: a multifunctional domain family comes to light. *Curr. Biol.* **7**, R674-R677.

Ponting, C. P. & Kerr, I. D. (1996). A novel family of phospholipase D homologues that includes phospholipid synthases and putative endonucleases: identification of duplicated repeats and potential active site residues. *Protein Sci.* **5**, 914-922.

Ponting, C. P., Schultz, J., Milpetz, F. & Bork, P. (1999). Identification and annotation of domains from signalling and extracellular protein sequences. *Nucl. Acids Res.* **27**, 229-232.

Rawson, R. B., Zelenski, N. G., Nijhawan, D., Ye, J., Sakai, J., Hasan, M. T., Chang, T. Y., Brown, M. S. & Goldstein, J. L. (1997). Complementation cloning of S2P, a gene encoding a putative metalloprotease required for intramembrane cleavage of SREBPs. *Mol. Cell*, **1**, 47-57.

Rost, B. (1998). Marrying structure and genomics. *Structure*, **6**, 259-263.

Rost, B. & Sander, C. (1993). Prediction of protein structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599.

Sadler, J. E. (1989). The molecular biology of human von Willebrand factor. In *Coagulation and Bleeding Disorders: The Role of Factor VIII and von Willebrand Factor* (Zimmerman, T. S. & Ruggeri, Z. M., eds), vol. 9, pp. 117-136, Dekker, New York/Basel.

Sandford, R., Sgotto, B., Aparicio, S., Brenner, S., Vaudin, M., Wilson, R. K., Chissoe, S., Pepin, K., Bateman, A., Chothia, C., Hughes, J. & Harris, P. (1997). Comparative analysis of the polycystic kidney disease 1 (PKD1) gene reveals an integral membrane glycoprotein with multiple evolutionary conserved domains. *Hum. Mol. Genet.* **6**, 1483-1489.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. (1982). Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162**, 729-773.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857-5864.

Sikorski, R. S., Boguski, M. S., Goebl, M. & Hieter, P. (1990). A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell*, **60**, 307-317.

Slack, F. J. & Ruvkun, G. (1998). A novel repeat domain that is often associated with RING finger and B-box motifs. *Trends Biochem. Sci.* **23**, 474-475.

Smith, R. F. & King, K. Y. (1995). Identification of a eukaryotic-like protein kinase gene in archaebacteria. *Protein Sci.* **4**, 126-129.

Springer, T. A. (1998). An extracellular β-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components. *J. Mol. Biol.* **283**, 837-862.

Stark, M. J. & Boyd, A. (1986). The killer toxin of Kluyveromyces lactis: characterization of the toxin subunits and identification of the genes which encode them. *EMBO J.* **5**, 1995-2002.

Szyperski, T., Fernández, C., Mumenthaler, C. & Wüttrich, K. (1998). Structure comparison of human glioma pathogenesis-related protein GliPR and the plant pathogenesis-related protein P14a indicates a functional link between the human immune system and a plant defense system. *Proc. Natl Acad. Sci. USA*, **95**, 2262-2266.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631-637.

The International Polycystic Kidney Disease Consortium (1995). Polycystic kidney disease: the complete structure of the *PKD1* gene and its protein. *Cell*, **81**, 289-298.

Tschiersch, B., Hofmann, A., Krauss, V., Dorn, R., Korge, G. & Reuter, G. (1994). The protein encoded by the *Drosophila* position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J.* **13**, 3822-3831.

Walker, D. R. & Koonin, E. V. (1997). SEALS: a system for easy analysis of lots of sequences. *Ismb*, **5**, 333-339.

Wang, W., Xue, Y., Zhou, S., Kuo, A., Cairns, B. R. & Crabtree, G. R. (1996). Diversity and specialization of mammalian SWI/SNF complexes. *Genes Dev.* **10**, 2117-2130.

Whelan, K. F., Sherburne, R. K. & Taylor, D. E. (1997). Characterization of a region of the IncHI2 plasmid R478 which protects *Escherichia coli* from toxic effects specified by components of the telluric, phage and colicin resistance cluster. *J. Bacteriol.* **179**, 63-71.

Whisstock, J. C. & Lesk, A. M. (1999). SH3 domains in prokaryotes. *Trends Biochem. Sci.* **24**, 132-133.

Whitham, S., Dinesh-Kumar, S. P., Choi, D., Hehl, R., Corr, C. & Baker, B. (1994). The product of the tobacco mosaic virus resistance gene *N*: similarity to toll and the interleukin-1 receptor. *Cell*, **78**, 1101-1115.

Wittekind, M., Mapelli, C., Farmer, B. T., II, Suen, K. L., Goldfarb, V., Tsao, J., Lavoie, T., Barbacid, M., Meyers, C. A. & Mueller, L. (1994). Orientation of peptide fragments from Sos proteins bound to the N-terminal SH3 domain of Grb2 determined by NMR spectroscopy. *Biochemistry*, **33**, 13531-13539.

Woese, C. (1998). The universal ancestor. *Proc. Natl Acad. Sci. USA*, **95**, 6854-6859.

Wolf, Y., Brenner, S. E., Bash, P. A. & Koonin, E. V. (1999a). Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9**, 17-26.

Wolf, Y. I., Aravind, L. & Koonin, E. V. (1999b). *Rickettsiae* and *Chlamydiae*-distinct patterns of horizontal transfer in two groups of obligate intracellular parasites and evidence of gene exchange between them. Evidence of horizontal gene transfer and gene exchange. *Trends Genet.* **15**, 173-175.

Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.

Yamakawa, T., Miyata, S., Ogawa, N., Koshikawa, N., Yasumitsu, H., Kanamori, T. & Miyazaki, K. (1998). cDNA cloning of a novel trypsin inhibitor with similarity to pathogenesis-related proteins, and its frequent expression in human brain cancer cells. *Biochim. Biophys. Acta*, **1395**, 202-208.

Zhang, C.-C., Gonzalez, L. & Phalip, V. (1998). Survey, analysis and genetic organization of genes encoding eukaryotic-like signaling proteins on a cyanobacterial genome. *Nucl. Acids Res.* **26**, 3619-3625.

Edited by J. M. Thornton

(Received 1 March 1999; received in revised form 13 April 1999; accepted 26 April 1999)