Protein Science (1998), 7:1647–1652. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Conserved sequence motifs among bacterial, eukaryotic, and archaeal phosphatases that define a new phosphohydrolase superfamily

MARIA CRISTINA THALLER,[1] SERENA SCHIPPA,[2] AND GIAN MARIA ROSSOLINI[3]

[1]Dipartimento di Biologia, Università di Roma "Tor Vergata," 00133 Rome, Italy
[2]Istituto di Microbiologia, Università di Roma "La Sapienza," 00185, Rome, Italy
[3]Dipartimento di Biologia Molecolare, Sezione di Microbiologia, Università di Siena, 53100, Sienna, Italy

(RECEIVED February 19, 1998; ACCEPTED April 28, 1998)

Abstract: Members of a new molecular family of bacterial nonspecific acid phosphatases (NSAPs), indicated as class C, were found to share significant sequence similarities to bacterial class B NSAPs and to some plant acid phosphatases, representing the first example of a family of bacterial NSAPs that has a relatively close eukaryotic counterpart. Despite the lack of an overall similarity, conserved sequence motifs were also identified among the above enzyme families (class B and class C bacterial NSAPs, and related plant phosphatases) and several other families of phosphohydrolases, including bacterial phosphoglycolate phosphatases, histidinol-phosphatase domains of the bacterial bifunctional enzymes imidazole-glycerolphosphate dehydratases, and bacterial, eukaryotic, and archaeal phosphoserine phosphatases and threalose-6-phosphatases. These conserved motifs are clustered within two domains, separated by a variable spacer region, according to the pattern [FILMAVT]-D-[ILFRMVY]-D-[GSNDE]-[TV]-[ILVAM]-[ATSVILMC]-X-{YFWHKR}-X-{YFWHNQ}-X(102,191)-{KRHNQ}-G-D-{FYWHILVMC}-{QNH}-{FWYGP}-D-{PSNQYW}. The dephosphorylating activity common to all these proteins supports the definition of this phosphatase motif and the inclusion of these enzymes into a superfamily of phosphohydrolases that we propose to indicate as "DDDD" after the presence of the four invariant aspartate residues. Database searches retrieved various hypothetical proteins of unknown function containing this or similar motifs, for which a phosphohydrolase activity could be hypothesized.

Keywords: acid phosphatase; homology; molecular family; molecular superfamily; phosphatase; sequence motif

Phosphohydrolases (phosphatases) are enzymes that catalyze the dephosphorylation of various substrates by hydrolysis of phosphoester or phosphoanhydride bonds (Boyer et al., 1961). Phosphohydrolases are ubiquitous among prokaryotes and eukaryotes. Each cell is normally equipped with several different such enzymes, which play various essential or accessory roles in the cell biology.

Classification of phosphatases was initially based on the functional and biophysical properties of the enzyme, such as pH optimum (acid vs. alkaline), substrate profile (nonspecific vs. specific for certain substrates), and molecular size (high vs. low molecular mass). As molecular sequence data became available, it was evident that also phosphatases, similarly to other proteins, could be grouped into molecular families on the basis of amino acid sequence similarity. The structural criterion has led to the definition of various molecular families of phosphohydrolases for which signature sequence motifs have been defined (Bairoch et al., 1995). The definition of similar conserved sequence motifs is useful for a tentative identification of new hypothetical proteins discovered after large-scale sequencing projects, and may also provide insights into the structure–function relationships of the various enzymes.

Molecular families of bacterial nonspecific acid phosphohydrolases: Concerning bacterial nonspecific acid phosphohydrolases (NSAPs), two different molecular families were previously identified, which we proposed to indicate as molecular classes A and B (Thaller et al., 1994, 1995a, 1995b).

Class A NSAPs are secreted monomeric to oligomeric proteins containing a polypeptide component of approximately 25–27 kDa (Pond et al., 1989; Kasahara et al., 1991; Thaller et al., 1994; Bhargava et al., 1995; Uchiya et al., 1996). This group of enzymes has recently been demonstrated to share some conserved sequence motifs with other bacterial and eukaryotic phosphatases, suggesting that the conserved residues could be essential for catalytic activity and possibly part of the active site of these enzymes (Stukey & Carman, 1997).

Class B NSAPs are secreted homotetrameric metallo-proteins containing a 25-kDa polypeptide component (Uerkvitz, 1988; Thaller et al., 1995b, 1997a) that are completely unrelated to class

Reprint requests to: Maria Cristina Thaller, c/o Istituto di Microbiologia, Università "La Sapienza," Piazzale A. Moro, 5, 00185 Rome, Italy; e-mail: Pezzi@axrma.uniroma1.it.

A enzymes at the sequence level. The sequences of four class B bacterial NSAPs are currently available (Table 1). The sequence similarity among these enzymes is overall high (46–89% of identity), and can be followed throughout the protein length (Fig. 1). Two sequence motifs, centered on the most conserved domains, could be proposed as signatures for these enzymes: F-D-I-D-D-T-V-L-F-S-S-P, located in the N-terminal moiety, and Y-G-D-[AS]-D-X-D-[IV] located near the C-terminus [the PROSITE syntax (Bairoch et al., 1995) is used to describe the motifs]. A scan of the Swiss-Prot and Swiss-Trembl databases at the Expasy Server (Appel et al., 1994) using either signature pattern specifically returns the above proteins.

Recently, while screening for production of NSAPs in nonenterobacterial species, we identified an NSAP activity containing a polypeptide component of approximately 30 kDa in *Chryseobacterium* (formerly *Flavobacterium*) *meningosepticum*. The corresponding gene, named *olpA*, was cloned and sequenced (Thaller et al., 1997b; M.C. Thaller, P. Iori, S. Schippa, L. Lauretti, F. Berlutti, & G.M. Rossolini, in prep.; EMBL/GenBank accession #Y12759). The OlpA protein contains an amino-terminal signal sequence typical of precursors of bacterial membrane lipoproteins (Hayashi & Wu, 1990), and does not contain any signature sequence pattern typical of other phosphatases, including those previously proposed for class B NSAPs. Using the amino acid sequence of OlpA in a BLAST search (Altschul et al., 1990) performed at the National Center for Biotechnology Information, two bacterial lipoproteins were identified that showed significant sequence similarity (27 to 45% of identity) throughout the protein length to OlpA (Fig. 1), for neither of which the phosphatase activity had previously been reported: the *e*(P4) outer membrane lipoprotein of *Haemophilus influenzae* (Green et al., 1991) and the LlpC cytoplasmic membrane lipoprotein of *Streptococcus equisimilis* (Gase et al., 1997). Recloning and expression experiments confirmed that both of them were actually endowed with NSAP activity. According to these findings, we proposed to define a third molecular class of bacterial NSAPs, named class C, which includes the above membrane lipoproteins (Thaller et al., 1997b; M.C. Thaller, P. Iori, S. Schippa, L. Lauretti, F. Berlutti, & G.M. Rossolini, in prep.). BLAST searches also showed that bacterial class C NSAPs are similar (32–38% of sequence identity) throughout the protein length

Fig. 1. Alignment of the amino acid sequences of class B and class C bacterial NSAPs, and of plant class C-like phosphatases. Identical amino acid residues are indicated by an asterisk. Conservative amino acid substitutions are indicated by a dot. The name of the proteins are the same as in Table 1.

Table 1. *Members of the "DDDD" superfamily of phosphohydrolases*[a]

Protein family	Protein (organism, accession #)[b]	Domain A[c]	Domain B[c]	References[d]
Bacterial class B NSAPs	AphA (*Escherichia coli*, P32697)	67-FDIDDTVLFSSP-	110-YGDSDNDI-40	Thaller et al. (1997a)
	AphA (*Salmonella enterica*, O08430)	66-FDIDDTVLFSSP-	110-YGDSDNDI-40	
	NapA (*Morganella morganii*, Q59554)	66-FDIDDTVLFSSP-	110-YGDADADI-40	Thaller et al. (1995b)
	NapA (*Haemophilus influenzae*, Y07615)	65-FDIDDTVLFSSP-	110-YGDSDDDV-40	
Bacterial class C NSAPs	OlpA (*Chryseobacterium meningosepticum*, O08351)	73-LDIDETVLDNSP-	102-FGDNLSDF-72	Thaller et al. (1997b)
	e(P4)(*H. influenzae*, P26093)	82-ADLDETMLDNSP-	104-VGDNLDDF-68	Green et al. (1991)
	LlpC (*Streptococcus equisimilis*, O05471)	97-LDIDETVLDNSP-	103-FGDNLVDF-65	Gase et al. (1997)
	HP1285 (*Helicobacter pylori*, AE000633)	56-LDLDETVLNTFD-	102-VGDTLHDF-52	Tomb et al. (1997)
Plant class C-like	APS1 (*Lycopersicon esculentum*, P27061)	107-FDVDETLLSNLP-	108-SGDQWSDL-20	
	AP (*Glycine max*, AJ223074)	102-FDIDETTLSNLP-	111-IGDQWSDL-20	
	VSP (*Arabidopsis thaliana*, Q39259)	120-FDLDDTLLSSIP-	108-IGDQWADL-20	
Bacterial GPHs	CbbZ (*E. coli*, P32662)	11-FDLDGTLVDSAP-	166-VGDSRNDI-55	Lyngstadaas et al. (1995)
	CbbZ (*H. influenzae*, P44755)	9-FDLDGTLVNSLP-	153-VGDSKNDI-42	Fleischmann et al. (1995)
	CbbZ (*Rhodobacter sphaeroides*, P95680)	5-FDLDGTLVHSAP-	147-VGDSEVDA-46	Gibson et al. (1991)
	CbbZ (*Rhodobacter capsulatus*, U23145)	6-FDLDGTLIDSAP-	146-VGDSEIDA-47	
	CbbZ (*Synechococcus* sp., Q55039)	5-FDFDGTLVDSLP-	139-DGDETRDV-48	
	CbbZ (*Synechocystis* sp., P73525)	7-FDFDGTIADTHD-	139-VGDETRDI-55	Kaneko et al. (1996)
	CbbZ (*Alcaligenes eutrophus*, P40852)	12-IDLDGTLVDSAP-	148-VGDSAVDV-51	Schaeferjohann et al. (1993)
	YvoE (*Bacillus subtilis*, AF017113)	10-FDLDGTLINTNE-	141-VGDNYHDV-45	
	(*Borrelia burgdorferi*, AE001168)	7-FDMDGTLVNSIM-	150-IGDSDVDM-43	Fraser et al. (1997)
	RifM (*Amycolatopsis mediterranei*, AF040570)[e]	24-FDLDGVVVDSFA-	139-IGDAPTDL-49	
Bacterial HPPases	His7 (*E. coli*, P06987)	7-IDRDGTLISEPP-	108-IGDRATDI-60	Chiariotti et al. (1986)
	His7 (*H. influenzae*, P44327)	6-IDRDGTLIDEPK-	107-IGDRETDV-70	Fleischmann et al. (1995)
	His7 (*S. enterica*, P10368)	7-IDRDGTLISEPP-	108-IGDRATDI-60	Carlomagno et al. (1988)
	LMBK (*Streptomyces lincolnensis*, Q54364)[f]	16-FDRDGVLIEATV-	108-VGDRWRDV-46	Peschke et al. (1995)
PSPases	SerB (*E. coli*, P06862)	114-MDMDSTAIQIEC-	143-IGDGANDL-45	Neuwald and Stauffer (1985)
	SerB (*H. influenzae*, P44997)	107-MDMDSTAIQIEC-	143-IGDGANDL-44	Fleischmann et al. (1995)
	SerB (*H. pylori*, AE000578)	6-FDFDSTLVNAET-	143-VGDGANDL-38	Tomb et al. (1997)
	(*Mycobacterium tuberculosis*, AL021287)	183-FDVDSTLVQGEV-	143-VGDGANDI-63	
	SerB (*Archaeoglobus fulgidus*, AE000956)	133-FDMDSTLVEAEI-	143-VGDGANDR-48	Klenk et al. (1997)
	(*Methanococcus jannaschii*, Q58989)	9-FDFDSTLVNNET-	143-VGDGANDI-39	Bult et al. (1996)
	(*Methanobacterium thermoautotrophicum*, AE000921)	10-FDLDNVIIDGEA-	142-VGDGANDI-323	Smith et al. (1997)
	SerB (*Saccharomyces cerevisiae*, P42941)	95-FDMDSTLIYQEV-	145-VGDGGNDL-49	Guerreiro et al. (1996)
	(*Schyzosaccharomyces pombe*, P78910)	80-FDMDSTLIQQEC-	145-VGDGANDL-40	
	(*Schistosoma mansoni*, Q26545)	12-LDVDSTVCEDEG-	145-IGDGMTDA-46	Davis et al. (1995)
	(*Homo sapiens*, P78330)	18-FDVDSTVIREEG-	146-IGDGATDM-41	Collet et al. (1997)
T6Pases	OtsB (*E. coli*, P31678)	18-FDLDGTLAEIKP-	165-LGDDLTDE-63	Kaasen et al. (1994)
	OtsB (*Rhizobium* sp., P55611)	33-LDIDGTLLDLAT-	165-IGDDVTDE-47	Freiberg et al. (1997)
	(*M. thermoautotrophicum*, AE000931)	24-TDIDGTISEIAP-	166-LGDDITDA-54	Smith et al. (1997)
	OtsP (*M. leprae*, Q49734)	179-FDFDGTLSDIVD-	174-LGDDITDE-56	
	(*M. tuberculosis*, AL009198)	145-FDFDGTLSDIVE-	170-LGDDITDE-56	
	YW11 (*M. tuberculosis*, Q10850)[g]	285-LDFDGTLSDIVE-	172-IGDDLTDE-850	
	TPS2 (*S. cerevisiae*, P31688)	575-FDYDGTLTPIVK-	191-LGDDFTDE-110	De Virgilio et al. (1993)
	(*S. pombe*, Z97209)	568-MDYDGTLTPIVR-	176-AGDDRTDE-53	
	(*S. pombe*, Z99167)	579-LDYDGTLIESAR-	172-AGDDKTDE-78	
	(*Emericella nidulans*, Q00786)	1-FDYDGTLTPIVK-	176-SGDDFTDE-30	Borgia et al. (1996)
	(*A. thaliana*, Z97344)[h]	598-LDFDGTMVQPGS-	174-VGDDRSDE-73	
Unassigned showing the "DDDD" motif	(*Sulfolobus sulfataricus*, P95067)	6-VDLDGTLTEDRE-	151-IGDSSTDI-51	Sensen et al. (1996)
	(*S. cerevisiae* Q12486)	17-FDMDGTLCLPQP-	130-VGDSFDDM-51	
	YNBO (*S. cerevisiae*, P53981)	7-TDFDGTVTLEDS-	157-CGDGVSDL-57	
YAED	YAED (*E. coli*, P31546)	9-LDRDGTINVDHG-	112-VGDKLEDM-50	Blattner et al. (1997)
	YAED (*H. influenzae*, P46452)	6-LDRDGTLNIDYG-	111-VGDKVEDL-47	Fleischmann et al. (1995)
COF	YUPP (*Mycoplasma hominis*, P43051)	15-IDLDGTLLADSA-	204-MGDSYNDL-43	
	YBHA (*E. coli*, P21829)	7-LDLDGTLLTPKK-	201-FGDNFNDI-44	Walkenhorst et al. (1995)
	COF (*E. coli*, P46891)	6-FDMDGTLLMPDH-	191-FGDAMNDR-55	
	YIGL (*E. coli*, P27848)	6-SDLDGTLLSPDH-	192-LXDGMNDA-47	Daniels et al. (1992)
	YIGL (*H. influenzae*, P44771)	9-SDLDGTLLTPEH-	194-FGDGMNDV-49	Fleischmann et al. (1995)

(*continued*)

Table 1. *Continued*

Protein family	Protein (organism, accession #)[b]	Domain A[c]	Domain B[c]	References[d]
COF	YIDA (*E. coli*, P09997)	7-IDMDGTLLLPDH-198-IGDQENDI-45		Burland et al. (1993)
	YWPJ (*B. subtilis*, P94592)	5-IDLDGTLLNSKH-217-VGDSLNDK-43		
	Y003 (*H. influenzae*, P44447)	7-SDFNGTLLTSQH-190-FGDNFNDL-53		Fleischmann et al. (1995)
	Y125 (*Mycoplasma capricolum*, P53661)	6-IDIDGTVYSRKH-203-FGDGENDL-38		Bork et al. (1995)
	YXHE (*B. subtilis*, P54947)	6-IDMDGTLLNDHH-198-IGDNGNDL-46		Yoshida et al. (1995)
	Y265 (*Mycoplasma pneumoniae*, P75399)	9-FDLDGTLLSWGH-206-FGDGDNDV-47		Himmelreich et al. (1996)
	Y265 (*Mycoplasma genitalium*, P47507)	7-FDLDGTLLSSNQ-205-FGDADNDV-46		Fraser et al. (1995)
	YCSE (*B. subtilis*, P42962)	14-IDMDGTLLNDEQ-171-MGDSLNDI-44		Yamane et al. (1996)
	Y125 (*M. genitalium*, P47371)	6-LDLDGTLLSKTK-207-IGDSWNDY-52		Fraser et al. (1995)
	Y125 (*M. pneumoniae*, P75511)	6-LDLDGTLLSRTR-207-IGDSLNDR-48		Himmelreich et al. (1996)
	Y263 (*M. genitalium*, P47505)	9-SDLDGTIVSWNP-217-CGDGDNDI-45		Fraser et al. (1995)
	Y263 (*M. pneumoniae*, P75401)	9-SDLDGTIVSWNP-218-FGDGDNDI-60		Himmelreich et al. (1996)
	YA90 (*M. pneumoniae*, P75360)	14-CDLDGTLLRYQN-210-LGDSYNDL-46		Himmelreich et al. (1996)
	(*B. burgdorferi*, AE001120)	8-SDLDGTLLLSKS-205-FGDGFNDT-50		Fraser et al. (1997)

[a]Hypothetical proteins of unknown function containing motifs either typical of "DDDD" phosphohydrolases or closely related to them are also included in the list.

[b]Accession numbers for Swiss-Prot or Swiss-Trembl entries are provided when available. Other accession numbers are for GenBank-EMBL entries.

[c]Protein sequences were aligned relative to their phosphatase motifs; numbers preceding domain A indicate the distance from the N-terminus to domain A; numbers between the two domains indicate the length of the spacer between the two domains; numbers following domain B indicate the distance from domain B to the C-terminus.

[d]All sequences listed without references were deposited directly into databases.

[e]The RifM phosphatase of *A. mediterranei* was included into the GPH family owing to the its overall sequence similarity with other members of this family.

[f]The LMBK protein of *S. lincolnensis*, although being described as an IGPD, contains only the HPPase domain typical of the bacterial bifunctional IGPDs.

[g]The large YW11 protein of *M. tuberculosis*, which is classified as a glycosyl hydrolase, was included in this family because it exhibits significant similarity to the *E. coli* T6Pase over a 255 amino acid overlap, being most probably a multifunctional protein with a T6Pase domain located at residues 241–527 and a trehalase domain located at residues 528–1327.

[h]The *Arabidopsis thaliana* 98-kDa protein, which is classified as a trehalose-6-phosphate synthase homologue, was included in this family because it also contains a T6Pase domain, its structure being overall similar to the yeast T6Pases, which are bifunctional proteins with an N-terminal trehalose-6-P synthase domain and a C-terminal T6Pase domain.

to a hypothetical protein of *Helicobacter pylori* (HP1285, Tomb et al., 1997) that could represent another member of this protein family, although not containing any N-terminal motif typical of bacterial lipoprotein signal sequences (Fig. 1). Two sequence motifs, centered on the most conserved domains, could be proposed as signatures for bacterial class C NSAPs: [IV]-[VAL]-D-[IL]-D-E-T-[VM]-L-X-[NT]-X(2)-Y, located in the N-terminal moiety, and [IV]-[LM]-X(2)-G-D-[NT]-L-X-D-F, located near the C-terminus. A scan of the Swiss-Prot and Swiss-Trembl databases at the Expasy Server (Appel et al., 1994) using either signature pattern specifically returns the above proteins.

Further sequence analysis revealed that, although more distantly, class C NSAPs are also related to class B enzymes. The similarity (14–22% of sequence identity) can be followed throughout the entire protein length, being stronger within two domains that correspond to the most conserved ones within each enzyme family (Fig. 1).

Significant sequence similarity (12–22% of identity) throughout the protein length was also observed between bacterial class C NSAPs and a family of plant proteins including the APS1 tomato acid phosphatase, a soybean acid phosphatase, and an *Arabidopsis thaliana* protein for which a phosphatase activity has not been reported but whose primary structure is similar to the tomato and soybean enzymes (Fig. 1). This represents the first example of a family of bacterial NSAPs that has a relatively close eukaryotic counterpart. Also, in this case, the similarity is stronger within two domains, which correspond to the most conserved ones within class C NSAPs and between class C and class B NSAPs (Fig. 1). Bacterial class B NSAPs appear to be more distantly related than class C enzymes to the plant class C-like phosphatases (Fig. 1).

Definition of a new molecular superfamily of phosphohydrolases: Additional analyses, performed using the sequences of the two above domains as probes for homology searches, revealed that similar conserved domains are also present in members of other protein families that, although not showing an overall similarity with either bacterial class B or class C NSAPs, or plant class C-like phosphatases, are endowed with phosphohydrolase activity. These include: (a) bacterial phosphoglycolate phosphatases (GPHs, EC 3.1.3.18); (b) histidinol–phosphatase (HPPases, EC 3.1.3.15) domains of the bacterial bifunctional enzymes imidazole–glycerolphosphate dehydratase (IGPD); (c) bacterial, eukaryotic, and archaeal phosphoserine phosphatases (PSPases, EC 3.1.3.3); (d) bacterial, eukaryotic, and archaeal threalose-6-phosphatases (T6Pases, EC 3.1.3.12) (Table 1).

The sequence motifs shared by all these enzymes are: [FILMAVT]-D-[ILFRMVY]-D-[GSNDE]-[TV]-[ILVAM]-[ATSVILMC]-X-{YFWHKR}-X-{YFWHNQ} (domain A), and {KRHNQ}-G-D-{FYWHILVMC}-{QNH}-{FWYGP}-D-{PSNQYW} (domain B), separated by a number of amino acid residues ranging from 102 to 191 (Table 1). These structural similarities, together with the dephosphorylating activity exhibited by the members of each family, support the definition of this phosphatase motif and the inclusion of all these enzymes into a molecular superfamily of phosphohydrolases, which we propose to

indicate as "DDDD" due to the couple of invariant aspartate residues present in each domain (Table 1). The invariant residues could be essential for the phosphohydrolase catalytic activity of these enzymes and part of the active site. In addition, other residues located within the domains could be critical to the phosphohydrolase activity, because the more degenerated sequence motif D-X-D-X-[TV]-X(109,198)-G-D-X(3)-D is also found in members of other protein families that do not possess phosphohydrolase activity and in which this sequence pattern does not appear to be a typical family signature. Within each conserved domain a preference for certain residues is evident, at some positions, depending on the protein family (Table 1). In particular, at position 3 of domain A, arginine appears to be peculiar of HPPases, while members of other families carry a hydrophobic residue; at position 5 of domain A, a negatively charged residue (aspartate or glutamate) is peculiar of class B and class C bacterial NSAPs and of plant class C-like phosphatases, while members of other families usually carry either glycine (GPHs, HPPases, and T6Pases) or serine (PSPases); at position 11 of domain A, glutamate characterizes the PSPases, while members of the other families usually carry noncharged residues; at position 4 of domain B, aspartate characterizes the T6Pases, glycine the PSPases, arginine the HPPases, and glutamine the plant class C-like phosphohydrolases.

Considering the above patterns and also the sequences immediately surrounding the two conserved domains, it was possible to define peculiar motifs, centered around either one or both domains, specific for each group of "DDDD" phosphohydrolases. In particular, apart from the patterns already discussed for the bacterial class B and C NSAPs (see above), the plant Class C-like phosphatases and the T6Pases are specifically recognized by motifs centered on domain B: I-V-G-X(2)-G-D-Q-W-X-D-L and [ILVSA]-G-D-D-[RKFILV]-[ST]-D-[AE]-[ASGND]-[AGM], respectively. The GPHs are specifically recognized by a motif centered on domain A: [IV]-X-[IF]-D-[FLM]-D-G-[TV]-[LIV]-[AIV]-[NDH]-[ST]-X(3)-[FILV]. The HPPases are specifically recognized by a motif centered on domain A, [FI]-D-R-D-G-[TV]-L-I, and by a motif centered on domain B, S-[FY]-V-[IV]-G-D-R-X(2)-D-[IV]-X(2)-A. The PSPases are specifically recognized by a motif centered on domain A, [IV]-G-D-G-[AMG]-[NT]-D-X(2)-[MA]-X(3)-[AS]-X(3)-[IV]-[AG]-[FWY], and by a motif centered on domain B, [LFM]-D-[VLFM]-D-[SN]-T-[ILVA]-[CIV]-X(2)-E-X-[IL]-[DE].

Hypothetical proteins of unknown function possibly related to the "DDDD" phosphohydrolases: The proposed sequence motif specific for members of the "DDDD" phosphohydrolase superfamily was also found in three hypothetical proteins of unknown function (two of *Saccharomyces cerevisiae* and one of *Sulfolobus sulfataricus*; Table 1) that do not exhibit overall sequence similarity to any of the seven families of phosphohydrolases included in the "DDDD" superfamily. It would be interesting to verify whether also these proteins are endowed with phosphatase activity.

Moreover, sequence motifs very similar, although not identical, to that proposed for members of the "DDDD" phosphohydrolase superfamily were found to be represented in the hypothetical YAED proteins and in members of the recently described COF family of hypothetical proteins (Bairoch et al., 1995; PROSITE PDOC00944) (Table 1). The YAED proteins exhibit significant sequence similarity (24–30% of identity) with the HPPase domain of the bacterial IGPDs, but are not recognized by the proposed "DDDD" signature for the presence of an asparagine residue at position 8 of domain A (Table 1). Members of the COF family do not exhibit overall sequence similarity to any of the seven families of phosphohydrolase included in the "DDDD" superfamily, but are characterized by the presence of two sequence motifs that partially overlap the "DDDD" conserved domains and usually retain the same pattern of invariant residues. However, peculiar patterns of noninvariant residues and/or the length of the spacer region between the two domains always prevent their recognition by the proposed "DDDD" motif (Table 1). The hypothesis that at least some of these proteins may be endowed with phosphohydrolase activity appears suggestive and should deserve further investigation.

Acknowledgments: The support of grant 96.03391.CT04 from the Italian National Research Council (C.N.R.), and of a grant from Università di Siena, Quota 60% to G.M.R., are gratefully acknowledged.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic alignment search tool. *J Mol Biol 215*:403–410.

Appel RD, Bairoch A, Hochstrasser DF. 1994. A new generation of information retrieval tools for biologists: The example of the ExPASy WWW server. *Trends Biochem Sci 19*:258–260.

Bairoch A, Bucher P, Hofmann K. 1995. The PROSITE database, its status in 1995. *Nucleic Acids Res 24*:189–196.

Bhargava T, Datta S, Ramachandran V, Ramakrishnan R, Roy RK, Sankaran K, Subrahmanyam YVBK. 1995. Virulent *Shigella* codes for a soluble apyrase: Identification, characterization and cloning of the gene. *Curr Sci 68*:293–300.

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science 277*:1453–1462.

Borgia PT, Miao Y, Dodge CL. 1996. The *orlA* gene from *Aspergillus nidulans* encodes a trehalose-6-phosphate phosphatase necessary to normal growth and chitin synthesis at elevated temperatures. *Mol Microbiol 20*:1287–1296.

Bork P, Ouzounis C, Casari G, Schneider R, Sander C, Dolan M, Gilbert W, Gillevet PM. 1995. Exploring the *Mycoplasma capricolum* genome: A minimal cell reveals its physiology. *Mol Microbiol 16*:955–967.

Boyer PD, Lardy H, Mayback K, eds. 1961. *The enzymes*, vol. V. New York: Academic Press.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, Fitzgerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Weidman JF, Fuhrmann JL, Presley EA, Nguyen D, Utterback TR, Kelley JM, Peterson JD, Sadow PW, Hanna MC, Cotton MD, Hurst MA, Roberts KM, Kaine BP, Borodovsky M, Klenk HP, Fraser CM, Smith HO, Woese CR, Venter JC. 1996. Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*. *Science 273*:1058–1073.

Burland VD, Plunkett G, Daniels DL, Blattner FR. 1993. DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: Organizational symmetry around the origin of replication. *Genomics 16*:551–561.

Carlomagno MS, Chiariotti L, Alifano P, Nappo AG, Bruni CB. 1988. Structure and function of the *Salmonella typhimurium* and *Escherichia coli* K-12 histidine operons. *J Mol Biol 203*:585–606.

Chiariotti L, Nappo AG, Carlomagno MS, Bruni CB. 1986. Gene structure in the histidine operon of *Escherichia coli*. Identification and nucleotide sequence of the *hisB* gene. *Mol Gen Genet 202*:42–47.

Collet JF, Gerin I, Rider MH, Veiga-da-Cunha M, Van Schaftingen E. 1997. Human phosphoserine phosphatase: Sequence, expression and evidence for a phosphoenzyme intermediate. *FEBS Lett 408*:281–284.

Daniels DL, Plunkett G III, Burland V, Blattner FR. 1992. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science 257*:771–778.

Davis RE, Hardwick C, Tavernier P, Hodgson S, Singh H. 1995. RNA trans-splicing in flatworms. Analysis of trans-spliced mRNAs and genes in the human parasite, *Schistosoma mansoni*. *J Biol Chem 270*:21813–21819.

De Virgilio C, Burckert N, Bell W, Jenoe P, Boller T, Wiemken A. 1993. Disruption of *TPS2*, the gene encoding the 100 kDa subunit of the trehalose-6-phosphate synthase/phosphatase complex in *Saccharomyces cerevisiae*

causes accumulation of trehalose-6-phosphate and loss of trehalose-6-phosphate phosphatase activity. *Eur J Biochem 212*:315–323.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields CA, Gocayne JD, Scott JD, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T,Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science 269*:496–512.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton RA, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van-Vugt R, Palmer N, Adams MD, Gocayne JD, Weidman J, Utterback T, Watthey L, Mcdonald L, Artiach P, Bowman C, Garland S, Fujii C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO, Venter JC. 1997. Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*. *Nature 390*:580–586.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann JL, Nguyen DT, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu Pc, Lucier TS, Peterson SN, Smith HO, Hutchison CA III, Venter JC. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science 270*:397–403.

Freiberg C, Fellay R, Bairoch A, Broughton WJ, Rosenthal A, Perret X. 1997. Molecular basis of symbiosis between *Rhizobium* and legumes. *Nature 387*:394–401.

Gase K, Liu G, Bruckmann A, Steiner K, Ozegowski J, Malke H. 1997. The *lppC* gene of *Streptococcus equisimilis* encodes a lipoprotein that is homologous to the outer membrane protein *e* (P4) from *Haemophilus influenzae*. *Med Microbiol Immunol 186*:63–73.

Gibson JL, Falcone DL, Tabita FR. 1991. Nucleotide sequence, transcriptional analysis, and expression of genes encoded within the form I CO_2 fixation operon of *Rhodobacter sphaeroides*. *J Biol Chem 266*:14646–14653.

Green BA, Farley JE, Quinn-Dey T, Deich RA, Zlotnick GW. 1991. The *e*(P4) outer membrane protein of *Haemophilus influenzae*: The structural gene. *Infect Immun 59*:3191–3198.

Guerreiro P, Barreiros T, Soares H, Cyrne L, Maia E, Silva A, Rodrigues-Pousada C. 1996. Sequencing of a 176 kb segment on the right arm of yeast chromosome VII reveals 12 ORFs, including *CCT*, *ADE3* and *TR-1* genes, homologues of the yeast *PMT* and *EF1G* genes, of the human and bacterial electron-transferring flavoproteins (beta-chain) and of the *Escherichia coli* phosphoserine phosphohydrolase, and five new ORFs. *Yeast 12*:273–280.

Hayashi S, Wu HC. 1990. Lipoproteins in bacteria. *J Bioenerg Biomembr 22*:451–471.

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res 24*:4420–4449.

Kaasen I, McDougall J, Strom AR. 1994. Analysis of the *otsBA* operon for osmoregulatory trehalose synthesis in *Escherichia coli* and homology of the OtsA and OtsB proteins to the yeast trehalose-6-phospatase. *Gene 145*:9–15.

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S. 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.

Kasahara M, Nakata A, Shinagawa H. 1991. Molecular analysis of the *Salmonella typhimurium phoN* gene, which encodes nonspecific acid phosphatase. *J Bacteriol 173*:6770–6775.

Klenk HP, Clayton RA, Tomb J, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, Mckenney K, Adams MD, Loftus B, Peterson S, Reich CI, Mcneil LK, Badger JH, Glodek A, Zhou L, Overbeek R, Gocayne JD, Weidman JF, Mcdonald L, Utterback T, Cotton MD, Spriggs T, Artiach P, Kaine BP, Sykes SM, Sadow PW, D'Andrea KP, Bowman C, Fujii C, Garland SA, Mason TM, Olsen GJ, Fraser CM, Smith HO, Woese CR, Venter JC. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature 390*:364–370

Lyngstadaas A, Lobner-Olesen A, Boye E. 1995. Characterization of three genes in the *dam*-containing operon of *Escherichia coli*. *Mol Gen Genet 247*:546–554.

Neuwald AF, Stauffer GV. 1985. DNA sequence and characterization of the *Escherichia coli* serB gene. *Nucleic Acids Res 13*:7025–7039.

Peschke U, Schmidt H, Zhang HZ, Piepersberg W. 1995. Molecular characterization of the lincomycin-production gene cluster of *Streptomyces lincolnensis* 78-11. *Mol Microbiol 16*:1137–1156.

Pond JL, Eddy CK, Mackenzie KF, Conway T, Borecky DJ, Ingram LO. 1989. Cloning, sequencing, and characterization of the principal acid phosphatase, the *phoC*$^+$ product, from *Zymomonas mobilis*. *J Bacteriol 171*:767–774.

Schaeferjohann J, Yoo JG, Kusian B, Bowien B. 1993. The *cbb* operons of the facultative chemoautotroph *Alcaligenes eutrophus* encode phosphoglycolate phosphatase. *J Bacteriol 175*:7329–7340.

Sensen CW, Klenk HP, Singh RK, Allard G, Chan CC, Liu QY, Penny SL, Young F, Schenk ME, Gaasterland T, Doolittle WF, Ragan MA, Charlebois RL. 1996. Organizational characteristics and information content of an archaeal genome: 156 kb of sequence from *Sulfolobus solfataricus* P2. *Mol Microbiol 22*:175–191.

Smith DR, Doucette-Stamm LA, Deloughery C, Lee HM, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu D, Spadafora R, Vicare R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush D, Safer H, Patwell D, Prabhakar S, McDougall S, Shimer G, Goyal A, Pietrovski S, Church GM, Daniels CJ, Mao JI, Rice P, Nolling J, Reeve JN. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. *J Bacteriol 179*:7135–7155.

Stukey J, Carman GM. 1997. Identification of a novel phosphatase sequence motif. *Protein Sci 6*:469–472.

Thaller MC, Berlutti F, Schippa S, Iori P, Passariello C, Rossolini GM. 1995a. Heterogeneous patterns of acid phosphatases containing low molecular mass polypeptides in members of the family *Enterobacteriaceae*. *Int J Syst Bacteriol 45*:255–261.

Thaller MC, Berlutti F, Schippa S, Lombardi G, Rossolini GM. 1994. Characterization and sequence of PhoC, the principal phosphate-irrepressible acid phosphatase of *Morganella morganii*. *Microbiology 140*:1341–1350.

Thaller MC, Lombardi G, Berlutti F, Schippa S, Rossolini GM. 1995b. Cloning and characterization of the NapA acid phosphatase/phosphotransferase of *Morganella morganii*: Identification of a new family of bacterial acid phosphatase-encoding genes. *Microbiology 141*:147–154.

Thaller MC, Schippa S, Bonci A, Cresti S, Rossolini GM. 1997a. Identification of the gene (*aphA*) encoding the class B acid phosphatase/phosphotransferase of *Escherichia coli* MG1655 and characterization of its product. *FEMS Microbiol Lett 146*:191–198.

Thaller MC, Schippa S, Iori P, Berlutti F, Rossolini GM. 1997b. Cloning of *Chryseobacterium meningosepticum* acid phosphatase-encoding gene: Identification of a family of outer membrane bacterial phosphatases. Abs 97th General Meeting of the American Society for Microbiology, Miami Beach, Florida, 4–8 May 1997. p 286.

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature 388*:539–547.

Uchiya K, Tohsuji M, Nikai T, Sugihara H, Sasakawa C. 1996. Identification and characterization of phoN-Sf, a gene on the large plasmid of *Shigella flexneri* 2a encoding a nonspecific phosphatase. *J Bacteriol 178*:4548–4554.

Uerkvitz W. 1988. Periplasmic nonspecific phosphatase II from *Salmonella typhimurium* LT2. *J Biol Chem 263*:15823–15830.

Walkenhorst HM, Hemschemeier SK, Eichenlaub R. 1995. Molecular analysis of the molybdate uptake operon, *modABCD*, of *Escherichia coli* and *modR*, a regulatory gene. *Microbiol Res 150*:347–361.

Yamane K, Kumano M, Kurita K. 1996. The 25 degrees–36 degrees region of the *Bacillus subtilis* chromosome: Determination of the sequence of a 146 kb segment and identification of 113 genes. *Microbiology 142*:3047–3056.

Yoshida K, Fujimura M, Yanai N, Fujita Y. 1995. Cloning and sequencing of a 23-kb region of the *Bacillus subtilis* genome between the *iol* and *hut* operons. *DNA Res 2*:295–301.