Protein Science (1996), 5:914-922. Cambridge University Press. Printed in the USA.

A novel family of phospholipase D homologues that includes phospholipid synthases and putative endonucleases: Identification of duplicated repeats and potential active site residues

C.P. PONTING[1] AND I.D. KERR[2]

[1] Fibrinolysis Research Unit, The Old Observatory, University of Oxford, South Parks Road, Oxford OX1 3RH, United Kingdom
[2] Laboratory of Molecular Biophysics, Rex Richards Building, University of Oxford, South Parks Road, Oxford, OX1 3QU, United Kingdom

(RECEIVED December 14, 1995; ACCEPTED February 15, 1996)

Abstract

Phosphatidylcholine-specific phospholipase D (PLD) enzymes catalyze hydrolysis of phospholipid phosphodiester bonds, and also transphosphatidylation of phospholipids to acceptor alcohols. Bacterial and plant PLD enzymes have not been shown previously to be homologues or to be homologous to any other protein. Here we show, using sequence analysis methods, that bacterial and plant PLDs show significant sequence similarities both to each other, and to two other classes of phospholipid-specific enzymes, bacterial cardiolipin synthases, and eukaryotic and bacterial phosphatidylserine synthases, indicating that these enzymes form an homologous family. This family is suggested also to include two Poxviridae proteins of unknown function (p37K and protein K4), a bacterial endonuclease (nuc), an *Escherichia coli* putative protein (o338) containing an N-terminal domain showing similarities with helicase motifs V and VI, and a *Synechocystis* sp. putative protein with a C-terminal domain likely to possess a DNA-binding function. Surprisingly, four regions of sequence similarity that occur once in nuc and o338, appear twice in all other homologues, indicating that the latter molecules are bi-lobed, having evolved from an ancestor or ancestors that underwent a gene duplication and fusion event. It is suggested that, for each of these enzymes, conserved histidine, lysine, aspartic acid, and/or asparagine residues may be involved in a two-step ping-pong mechanism involving an enzyme-substrate intermediate.

Keywords: cardiolipin synthase; endonuclease; gene duplication; motif V-VI helicase; phosphatidylserine synthase; phosphodiesterase; phospholipase D; transphosphatidylation

Mammalian PLD enzymes are proposed to mediate intracellular signal transduction, particularly signals from growth factors, via hydrolysis of phosphatidylcholine to phosphatidic acid and choline (reviewed in Nishizuka, 1995). PA is dephosphorylated rapidly by phosphatases to form diacylglycerol, which is a potent activator of protein kinase C. Mammalian PLDs are functionally heterogeneous with varieties of specificities, indicating the presence of distinct PLD gene products. The membrane-associated PC-specific PLD isoenzyme, whose basal activity is stimulated by millimolar concentrations of Ca^{2+} or Mg^{2+} (Okamura & Yamashita, 1994), is a potential mammalian orthologue of plant PC-specific PLDs (Wang et al., 1994; Ueki et al., 1995). The latter enzymes possess an N-terminal domain homologous to PKC C2 domains that may mediate its association with membranes (Ponting & Parker, 1996).

Phospholipase D enzymes catalyze two reactions: hydrolysis of the terminal phosphodiester bond of phospholipids to release PA and transphosphatidylation of specific phospholipids to acceptor alcohols (reviewed in Heller et al., 1976). These reactions are essentially identical, except that, in hydrolysis, water acts as the acceptor, whereas in transphosphatidylation, an appropriate alcohol acts as the polar head-group acceptor. PLD enzymes have been cloned and sequenced from *Streptomyces* species (Iwasaki et al., 1994) and from plants (Wang et al., 1994; Dyer et al., 1995; Ueki et al., 1995). Purification of mammalian PLDs, although problematic in the past, has been more success-

Reprint requests to: C.P. Ponting, University of Oxford, Fibrinolysis Research Unit, The Old Observatory, South Parks Road, Oxford OX1 3RH, United Kingdom; e-mail: chris@biop.ox.ac.uk.

Abbreviations: PLD, phospholipase D; PLC, phospholipase C; PKC, protein kinase C; PC, phosphatidylcholine; PS, phosphatidylserine; PG, phosphatidylglycerol; CL, cardiolipin; PA, phosphatidic acid; DAG, diacylglycerol; ORF, open reading frame; EST, expressed sequence tag; EEV, extracellular enveloped virus.

ful recently (Okamura & Yamashita, 1994; Vinggaard & Hansen, 1995), indicating that sequence information on these shall soon be forthcoming.

The reaction mechanism for cabbage PLD has been proposed to be a ping-pong type reaction on the basis of radiolabel exchange data (Stanacev & Stuhne-Sekalec, 1970). The two-stage reaction pathway is proposed to involve formation of a covalent enzyme-phospholipid intermediate followed either by hydrolysis or by transphosphatidylation (Stanacev & Stuhne-Sekalec, 1970). Phosphodiesterase activity at an equivalent bond to that hydrolyzed by PLD is displayed by several other enzymes involved in lipid metabolism, including cardiolipin and phosphatidylserine synthases, which are both phosphatidyl transferase enzymes. There is evidence to suggest that PLD and PS synthase phosphatidyl transfer reactions occur by similar mechanisms, because both PLD and PS synthase are capable of cardiolipin synthesis (Stanacev & Stuhne-Sekalec, 1970; Nishijima et al., 1988). This relative nonspecificity of PLD and its potential to catalyze both hydrolysis and phosphatidyl transfer reactions (Yang et al., 1967) has elicited comparisons with other enzymes such as alkaline phosphatase (Morton, 1953).

Unlike bacterial and eukaryotic phospholipases C, whose sequence similarities have been demonstrated (Kuppe et al., 1989), PLDs from bacteria and from plants have not previously been shown to be homologues. Furthermore, PLD sequences have not been demonstrated to be homologous to other phosphodiesterases, such as cyclic nucleotide phosphodiesterases (Charbonneau et al., 1986), nucleases (Robins et al., 1994), or to any other proteins. Here we demonstrate a common ancestry for bacterial and eukaryotic PLDs and show that this homologous family includes other enzymes of phospholipid biosynthesis as well as putative endonucleases. Surprisingly, four sequence motifs that appear once in two of these homologues appear twice in the remainder, indicating that the latter have diverged from an ancestor or ancestors that underwent an internal gene duplication and fusion event.

Results

Eight classes of candidate PLD homologues

Our interest in PLD sequences arose from the identification of the castor bean PLD sequence as possessing an N-terminal C2 domain homologue, which is assumed to mediate an affinity for phospholipids and/or membranes; the remainder of the sequence is assumed to mediate its enzymatic activity (Ponting & Parker, 1996). A BLAST (Altschul et al., 1994) search of sequence databases at the National Center for Biotechnology Information (NCBI) for homologues of the castor bean PLD sequence demonstrated considerable similarity with an hypothetical protein encoded by *Saccharomyces cerevisiae* chromosome XI (SwissProt [sw] accession code sw:YK11_YEAST): the probability that four ungapped alignments were matched by chance was calculated as $p(4) = 1.6 \times 10^{-8}$. Yeast *YK11* is identical to a gene, called *SPO14*, which is required for meiosis and spore formation (Honigberg et al., 1992) and which has been shown recently to encode a phospholipase D (Rose et al., 1995). Additionally, a sequence similar to *S. cerevisiae* YK11/SPO14 is present in *S. pombe* (sw:YA2G_SCHPO), and other eukaryotes also appear to possess YK11/SPO14-orthologous genes as evidenced by YK11/SPO14-similar expressed sequence tag sequences (GenBank [gb] accession codes R34925, R83570, and Z45777 (human) and T76232 and T88610 (*Arabidopsis thaliana*)). The yeast YK11/SPO14 sequences do not appear to contain N-terminal C2 domains and are considerably longer than plant PLD sequences.

The yeast YK11/SPO14 sequences were not alone in their identification as candidate homologues of castor bean PLD. Database searches suggested similarities (lowest p-value $p(5) = 1 \times 10^{-3}$) of PLD and YK11/SPO14 sequences to a bacterial family of homologues. The only member of this family whose function has been characterized is a cardiolipin synthase from *E. coli* (Hiraoka et al., 1991; Tropp et al., 1995). Each of the remaining members (Table 1) is assumed to possess CL synthase activity; however, it is noted that two of these sequences derive from a single organism (*Bacillus subtilis*), which suggests that their functions are distinct.

Further candidate homologues of CL synthases, and thereby also of PLD and YK11/SPO14 sequences, were identified using BLAST searches with CL synthases as query sequences. The CL synthase homologues could be aligned partially with phosphatidylserine synthases, bacterial PLDs, viral proteins (p37K and K4), and a bacterial endonuclease (nuc), with p-values < 0.1 (Fig. 1). To these, two further candidates were added. The first, an open reading frame from *Synechocystis* sp., showed high similarity to nuc in a BLAST search ($p = 2 \times 10^{-7}$). The second, an *E. coli* hypothetical protein (o338), was identified by scanning databases against a multiple alignment of a conserved region (Motifs 3 and 4) of the previously identified sequences, using a local similarity algorithm (Barton, 1993a).

After clustering, each of these sequences could be assigned to one of eight classes (Table 1). Sequences within each class could be multiply aligned and were sufficiently similar to be positively identified as homologues (p-values $< 10^{-8}$). By contrast, sequences taken from different classes could not be identified positively as homologues at this stage, because accurate alignments

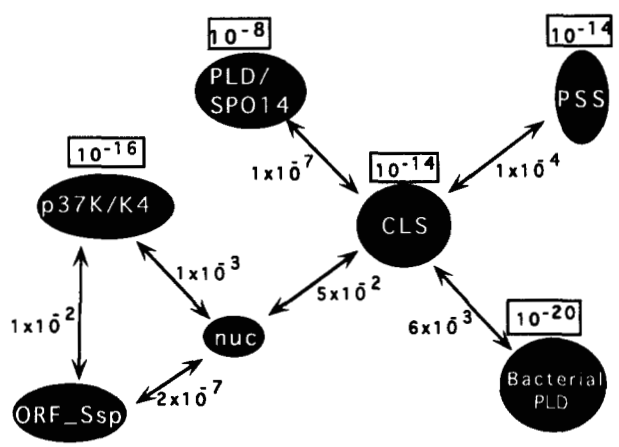

Fig. 1. Representation of the degrees of sequence similarities (p-values) between 7 of the 8 classes of PLD homologues produced by BLAST (Altschul et al., 1994) searches. o338 is not included in this representation because it was not identified using BLAST. Each shaded oval represents a class of homologues; the boxed number above each oval is a typical p-value obtained by pairwise comparisons of sequences within the class. The lowest p-values ($p < 0.1$) obtained by pairwise comparisons of sequences from different classes are shown beside arrows connecting the appropriate classes. For abbreviations, see text and legend to Figure 2.

Table 1. *Eight classes of phospholipase D homologues, their primary activities, and sequences*

Class	Protein/gene	Organism	Primary enzymatic activity[a]	Accession code[b]	References
I	Phospholipase D SPO14	Castor bean	Hydrolysis of PC to PA and choline	pir:A54850	Wang et al. (1994)
		S. cerevisiae		sw:SP14_YEAST	Honigberg et al. (1992)
		S. pombe		sw:YA2G_SCHPO	
		Maize		gb:MZEPHD1	Ueki et al. (1995)
		Rice		gb:RICPHD2	Ueki et al. (1995)
		A. thaliana		gb:ATU36381	Dyer et al. (1995)
II	Phospholipase D toxin	*S. antibioticus*	Hydrolysis of PC to PA and choline	gb:STMPDP	Iwasaki et al. (1994)
		Y. pestis		gb:YPPFRATOX	Cherepanov et al. (1991)
III	CL synthase	*E. coli*	Phosphatidyl transfer from one PG molecule to another.	sw:CLS_ECOLI	Hirschberg and Kennedy (1972)
		S. mutans[c]		sw:YWAP_STRMU	
		P. putida		sw:YLP2_PSEPU	
		B. subtilis		sw:YWJE_BACSU	
		B. subtilis		sw:YWIE_BACSU	
		M. capricolum[c]		gb:Z33226	
IV	PS synthase	*E. coli*	Phosphatidyl transfer from CDP-DAG to serine.	sw:PSS_ECOLI	DeChavigny et al. (1991)
		S. cerevisiae		sw:PEL1_YEAST	
V	p37K	Vaccinia virus	Unknown function	sw:VENV_VACCC	Hirt et al. (1988)
		Fowlpox virus		sw:VENV_FOWPV	
	K4	Vaccinia virus	Unknown function	sw:VK04_VACCC	Boursnell et al. (1988)
		C. elegans		gb:CELE04F6	
		D. discoideum[c]		gb:DD11093A	Giorda et al. (1989)
VI	ORF (*slr0628* gene)	*Synechocystis* sp.	Unknown function	gb:SYCSLRD_69	Kaneko et al. (unpubl.)
VII	nuc	*S. typhimurium*	Endonuclease activity	pir:S41475	Pohlman et al. (1993)
		E. coli[c]		gb:CIBSOG	
VIII	Helicase homologue	*E. coli*	Unknown function	sw:YJHR_ECOLI	Burland et al. (1995)

[a] Activities characterized for the first mentioned sequence for each class.
[b] Abbreviations: sw, swissprot; gb, genbank; pir, protein identification resource.
[c] Incomplete sequences.

throughout their lengths could not be obtained. For each class, multiple alignments of sequences were constructed. The most conserved regions in these alignments corresponded well to the regions identified by BLAST searches as being most similar between sequences taken from different classes of enzyme. This feature is a common occurrence within homologous families as a result of evolutionary pressure to preserve structural regions that are essential for function.

Internal repeats

An unexpected finding was that eukaryotic and bacterial PLDs, PS and CL synthases, p37Ks, K4s, and the ORF from *Synechocystis* sp. each appeared to contain duplicated repeats. Relatively conserved regions appeared to be present twice within each sequence (Motifs 1, 2, 3, and 4 in Fig. 2). Moreover, the order of these motifs within each sequence is invariant. As shall be demonstrated below, the similarities between the duplicated motifs were sufficient to enable their identification as homologous repeats; it is assumed that these enzymes possess bi-lobed structures, with each lobe containing a single copy of each of the four identified motifs. The presence of four motifs in an ordered sequence in these enzymes argues against the possibility that they arose independently, on several occasions, as results of convergent evolution. It is more probable that these enzymes are internally paralogous, i.e., each is a product of internal duplication and fusion within an ancestral gene or genes. The lack of significant sequence similarities in these enzymes away from the four motifs did not allow accurate construction of a phylogenetic tree for these enzymes, therefore, it was not possible to assess whether this duplication/fusion event occurred before or after the emergence of separate gene products with distinct enzymatic activities.

Homology arguments

Multiple alignments and comparisons of sequences using the program MACAW (Schuler et al., 1991) indicated the presence of four motifs, present once in o338 and nuc, and twice in the remaining sequences. MACAW also allowed examination of two propositions: that all or some of the sequences in Table 1 are homologues, and that all or some of the sequences (except o338 and nuc) contain homologous duplicated repeats. For the assessment of the former case, the four motifs from the N-terminal and the C-terminal lobes were aligned separately for two sets of eight sequences drawn from the eight classes of enzyme (Table 1); calculated p-values strongly support an argument that all of these sequences are homologous (lowest p-values for N-terminal

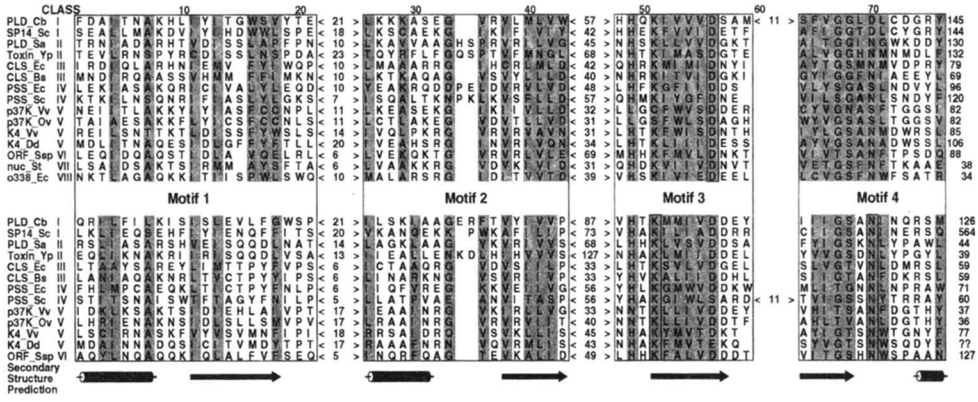

Fig. 2. Multiple sequence alignment of PLD homologues showing internal duplications for classes I–VI, represented using ALSCRIPT (Barton, 1993b). Four sequence blocks ("motifs") of the alignment were defined using the results of BLAST (Altschul et al., 1994) and MACAW (Schuler et al., 1991) algorithms. Positions are shaded if the chemical character of residues (defined as hydrophobic residues [A, C, I, L, V, M, F, Y, and W], small residues [G, A, and S], and acidic, and amides of acidic, residues [D, E, N, and Q]) is conserved in >66% of sequences within either subunit. Positions are boxed if residues are absolutely conserved in all sequences of either subunit. Numbers in angled brackets represent the numbers of amino acids that intervene between alignment blocks. Numbers at the end of the alignment represent the distance in amino acids to the C-terminal ends, with the exception of the numbers for Classes I–VI (top lines only) that represent the distance to Motif 1 of the second subunit. Secondary structure predictions were undertaken using PHD (Rost & Sander, 1993); α-helices are represented by filled cylinders, and β-strands by arrows. Accession codes and sequences represented are: PLD_Cb, pir:A54850 (residues 215–682); SP14_Sc, sw:YK11_YEAST (698–1119); PLD_Sa, gb:STMPDP (120–512); Toxin_Yp, gb:YPPFRATOX (2–491); CLS_Ec, sw:CLS_ECOLI (135–427); CLS_Bs, sw:YWJE_BACSU (59–339); PSS_Ec, sw:PSS_ECOLI (41–381); PSS_Sc, sw:PEL1_YEAST (49–429); p37K_Vv, sw:VENV_VACCC (37–335); p37K_Ov, gb:OVU06671 (38–342); K4_Vv, VK04_VACCC (32–347); K4_Dd, gb:CELE04F6; ORF_Ssp, gb:SYCSLRD_69 (87–426); nuc_St, pir:S41475 (42–139); o338_Ec, sw:YJHR_ECOLI (194–304). Abbreviations: Cb, castor bean; Sc, *S. cerevisiae*; Sa, *S. antibioticus*; Yp, *Y. pestis*; Ec, *E. coli*; Bs, *B. subtilis*; Vv, vaccinia virus; Ov, orf virus; Dd, *D. discoideum*; Ssp, *Synechocystis* sp.; St, *S. typhimurium*; SP14, *SPO14* gene product.

motifs 2, 3, and 4, $p = 5.4 \times 10^{-2}$, $p = 4.0 \times 10^{-16}$, and $p = 0.93$, respectively, and for C-terminal motifs, $p = 4.9 \times 10^{-3}$, $p = 8.0 \times 10^{-14}$, and $p = 5.4 \times 10^{-7}$, respectively). For assessment of the second argument, motifs from both N- and C-terminal lobes were aligned for two sets of six sequences, taken from classes I–VI; calculated p-values support the proposition that class I–VI sequences are internally paralogous (lowest p-values for motifs 2,3, and 4, $p = 3.8 \times 10^{-5}$, $p = 9.2 \times 10^{-19}$, and $p = 2.6 \times 10^{-5}$). Calculated p-values for motif 1 did not provide additional evidence for homology. However, it is noted that hydrophobic residues (ILVMFYWCA) are conserved at several positions in this motif (Fig. 2).

Secondary structure predictions

Secondary structure predictions were undertaken using PHD (Rost & Sander, 1993) for multiply aligned sequences of classes I–V and single sequences of classes VI–VIII. Within the expected levels of accuracy of this algorithm, these predictions were consistent with the alignment of Figure 2, including the observation of duplicated motifs. A secondary structure prediction is shown in Figure 2 for positions where the type of predicted secondary structure was conserved in ≥6 of the 10 multiple alignments in PLD classes I–V.

Discussion

Compared with the considerable knowledge gained recently of the domain organization and functions of phospholipases C and A_2, little is known of phospholipases D, due to considerable difficulties in their purification and sequencing. However, several plant PLDs have been sequenced recently (Wang et al., 1994; Ueki et al., 1995), which has led to identification of PKC-like C2 domains within their N-terminal regions (Ponting & Parker, 1996). It is shown here that the remaining portions of their sequences are homologous to a variety of phosphodiesterases that can be classified into eight groups. These include CL and PS synthases, which are enzymes with phospholipase D-like activities, and bacterial PLDs. Other members of this homologous family are an *E. coli* endonuclease (nuc), and Poxviridae viral proteins p37K and K4, with unknown functions. The methods used here did not indicate that eukaryotic phosphatidylinositol-glycan-specific PLDs (Scallon et al., 1991) are homologous to the enzymes discussed here.

Whereas the enzymes of classes I–VI are bi-lobed, it is apparent that classes VII and VIII molecules, nuc and o338, are monolobed, because they contain single copies of four motifs. However, because one of these (nuc) appears to be multimeric in solution (Lackey et al., 1977; Pohlman et al., 1993), a requirement for two or more domains, either from a single polypeptide chain (classes I–VI), or from several (classes VII and VIII), for activity, remains a possibility. This proposition is analogous to the demonstrated homology between viral mono-lobed aspartic proteinases, which are active as dimers, and mammalian aspartic proteinases, which are bi-lobed monomers (Miller et al., 1989). The presence of internal duplications within the catalytic domains of enzymes is not unusual. Rhodanese, aspartic proteinases, methionyl aminopeptidases, creatine amidinohydrolases, and actin-like ATPases, among others, are thought to contain such duplications (Tang et al., 1978; Keim et al., 1981; Bork et al., 1992; Murzin, 1993). The active sites of these en-

zymes are usually situated at the interface between the two lobes, to which catalytic residues from both lobes contribute.

Identification of PLD homologues was undertaken using sequence information only, without recourse to molecular functions. Importantly, however, the known functions of these homologues are either identical or are similar to the phosphodiesterase and transphosphatidylation functions of eukaryotic PLDs and add, therefore, substantially to evidence for their common evolutionary ancestry. These homologues are likely to be individual PLD isoenzymes that differ in their specificities.

Phospholipases D (classes I and II)

One of the most striking identifications of sequence similarity demonstrated here was the identification of bacterial extracellular PLDs (class II sequences; Iwasaki et al., 1994) as homologues of plant intracellular PC-specific PLDs (class I), indicating an origin for this homologous family prior to the divergence of eukaryotes and bacteria. These PLDs catalyze identical reactions: cabbage PLD (which is an orthologue of castor bean PLD [Wang et al., 1994]) and *Streptomyces* PLDs demonstrate phosphodiesterase activities on PS substrates, and also transphosphatidylation activities converting PC to PS in the presence of L-serine (Juneja et al., 1989). Unlike plant PLDs, however, bacterial PLDs do not possess an N-terminal C2 domain. Significant similarities ($p = 4 \times 10^{-21}$) between the sequences of *Streptomyces* PLDs and a toxin encoded by the *Yersinia pestis* plasmid pFra (Cherepanov et al., 1991) suggests that these proteins are orthologues.

Cardiolipin synthases (class III)

E. coli CL synthase catalyzes the formation of CL and glycerol from two molecules of phosphatidylglycerol. This occurs via a transfer mechanism, similar to that of PLD, whereby a phosphatidyl group is transferred from one PG molecule to another (Hirschberg & Kennedy, 1972; Tunaitus & Cronan, 1973). A second indication that the reaction mechanisms of plant PLDs and bacterial CL synthases are similar is that, not only is the product CL an inhibitor of CL synthesis, but phosphatidate, the product of PLD hydrolysis of PC, is also (Ragolia & Tropp, 1994). Furthermore, although hydrolysis reactions catalyzed by CL synthase have not been reported, cabbage PLD has been shown to possess a low CL synthase activity (Stanacev & Stuhne-Sekalec, 1970).

In eukaryotes, CL synthesis occurs via an alternative pathway involving the reaction between CDP-DAG and PG to form CL and CMP. No enzyme that catalyzes this reaction in eukaryotes has yet been sequenced, and hence any evolutionary relationship of eukaryotic CL synthases to the enzymes discussed here remains to be determined.

Phosphatidylserine synthases (class IV)

Synthesis of PS in *S. cerevisiae* occurs via two independent mechanisms. A PS synthase encoded by the *CHO1* gene, catalyzes the displacement of CMP from CDP-DAG by serine, and is a member of the CDP-alcohol phosphatidyltransferase gene family (Nikawa et al., 1987). A second PS synthase encoded by the *PEL1* gene, is homologous to an *E. coli* PS synthase that catalyzes hydrolysis of CDP-DAG to PA and CMP, and also catalyzes transphosphatidylation of CDP-DAG to serine (DeChavigny et al., 1991). It is this second enzyme that is homologous to PLDs. An enzyme homologous to the yeast *PEL1* gene product is present in the human genome as indicated by ESTs (GenBank codes: M77859, R55725 and T12593). *E. coli* PS synthase has also been suggested to catalyze cardiolipin synthesis via transphosphatidylation (Nishijima et al., 1988).

p37K and K4 homologues (class V)

p37K and K4 proteins are close homologues in the Poxviridae viral family that have not been suggested previously to possess enzymatic activities. Vaccinia virus p37K (gene *F13L*) is a palmitoylated protein that is a major antigenic component of the extracellular enveloped virus form (Payne, 1978). A prediction that p37K is a transmembrane protein (Hirt et al., 1986) was not substantiated using the prediction methods of Jones et al. (1994) and Persson and Argos (1994). The EEV form differs from its preceding form, the intracellular mature or naked virus (IMV or INV) form, in containing an additional membrane that is acquired from the *trans*-Golgi network prior to its release from the cell. Deletion of the p37K gene (*F13L*) in vaccinia has been shown to prevent EEV formation and cell–cell viral transmission (Blasco & Moss, 1991), whereas a mutation in the vaccinia gene resulting in a single amino acid substitution of tyrosine for aspartic acid 279 confers resistance to an inhibitor of virus envelopment and release (Schmutz et al., 1991). This has led to the suggestion (Blasco & Moss, 1991) that p37K is involved in the process of viral envelopment via the acquisition of cellular membranes. By contrast, the finding that p37K is a PLD homologue suggests that it may serve to process the phospholipids of the EEV envelope acquired from the host's *trans*-Golgi network.

Protein K4 is a p37K homologue, with which it shares approximately 25% amino acid identity (Boursnell et al., 1988). Its function appears to be inessential in vaccinia virus replication because it has been shown that deletion of the K4 gene (*K4L*) does not inhibit either the formation of EEVs or viral transmission (Blasco & Moss, 1991). K4 orthologues are not confined to the Poxviridae family because it would appear that homologous molecules are present in eukaryotes. Although not recognized as such previously, a fragment of an ORF 3′ to gene 3 of a developmentally regulated gene family in *Dictyostelium discoideum* (Giorda et al., 1989) encodes a portion of a K4 orthologue. Furthermore, a similar sequence is encoded by *Caenorhabditis elegans* cosmid E04F6 (GenBank code: CELE04F6_1). Humans also possess a K4-similar gene, as evidenced by overlapping ESTs (GenBank: R11447, F05415, R50605, H15300, R72685, T97936, and R69902, allowing for frame shifts), whose translations may be aligned against almost the entire vaccinia K4 sequence.

Synechocystis sp. ORF (class VI)

This predicted ORF from the cyanobacterium *Synechocystis* sp. (strain PCC6803) (Kaneko et al., 1995) encodes a protein of unknown function. A BLAST search with this sequence (553 amino acids) yielded strong similarity in its N-terminal region to *Salmonella typhimurium* nuc. Furthermore, as shall be demonstrated elsewhere (C.P. Ponting, L.C. Serpell, & A.J. Doherty, in prep.), the C-terminal region of this sequence is homologous

to several DNA-binding proteins and includes a "helix-hairpin-helix" DNA-binding motif, similar to that proposed to be present in endonuclease III (Thayer et al., 1995). The putative DNA-binding function of the C-terminal extension of the *Synechocystis* PLD homologue, and its similarity to the endonuclease nuc, suggests that this ORF may also encode an endonuclease.

nuc, an endonuclease (class VII)

The only member of the seventh class of homologues whose entire sequence is known is an endonuclease, termed nuc, encoded by the bacterial drug resistance IncN plasmid pKM101 (Pohlman et al., 1993). This enzyme hydrolyzes single-stranded and duplex DNA, substrates very different to those of other PLD homologues. However, the PLD homologous enzymes discussed here, including nuc, catalyze essentially identical reactions as each is a phosphodiesterase. It is plausible, therefore, that the catalytic mechanisms of PLDs, CL synthases, and PS synthases are similar to that of nuc. The finding of significant sequence similarities between enzymes of different specificities is not a rare occurrence, as shown elsewhere (Murzin, 1993; Mushegian & Koonin, 1994; and references therein).

o338, an unusual homologue of helicases (class VIII)

The only member of the eighth class of PLD homologues is a hypothetical protein sequenced as part of the *E. coli* genome project (Burland et al., 1995). ORF *o338* encodes a putative protein whose C-terminal region contains single copies of the four PLD-characteristic motifs (Fig. 2). Its N-terminal region has been shown elsewhere to be homologous to SEN1-like helicases (Koonin & Rudd, 1996). Although without endonuclease activity itself, SEN1 is known to be required for the activity of an endonuclease essential for tRNA splicing (DeMarini et al., 1992). Considering this, and the endonuclease activity of nuc, it is possible that *o338* encodes an N-terminal helicase homologue with a C-terminal endonuclease domain.

The o338 sequence is unusual in that it contains only two motifs (V and VI) of six motifs characteristic of helicases (Gorbalenya & Koonin, 1993). This is not the only example of a "Motif V–VI" SEN1-like helicase homologue: we have identified an ORF, encoded near cytotoxin-associated gene A in *Helicobacter pylori* (Covacci et al., 1993), which is homologous to helicases, and also only contains motifs V and VI (Fig. 3).

Catalytic mechanisms and active sites

The finding that these proteins are members of a homologous family suggests that they share similar catalytic mechanisms mediated by common active sites. Of all these enzymes, only the reaction mechanism of *E. coli* PS synthase has been well characterized (Raetz et al., 1987). This study demonstrated that the cleavage of the P-O bond between the β-phosphorus and the α-β oxygen of CPD-DAG occurs with retention of configuration at the phosphorus atom, indicative of a two-step ping-pong reaction mechanism involving a phosphatidylated enzyme intermediate. This is similar to the reaction mechanism suggested for cabbage PLD (Stanacev & Stuhne-Sekalec, 1970).

Candidate active site residues of a family of homologous enzymes may usually be identified as those that are absolutely conserved. However, this is complicated here by the presence of two homologous subunits within the majority of PLD-homologous enzymes: it is unknown whether these molecules harbor two active sites, or whether a single active site is formed by residues from within either one or both subunits. The former case, although rare, is not unknown (e.g., a yeast thiolase [Mathieu et al., 1994]), whereas the latter is the more usual (e.g., the hexokinase/actin/hsp70 family [Bork et al., 1992]).

It is noted that residues at several positions in motifs 3 and 4 are present either in one or in both subunits: these are His 49,

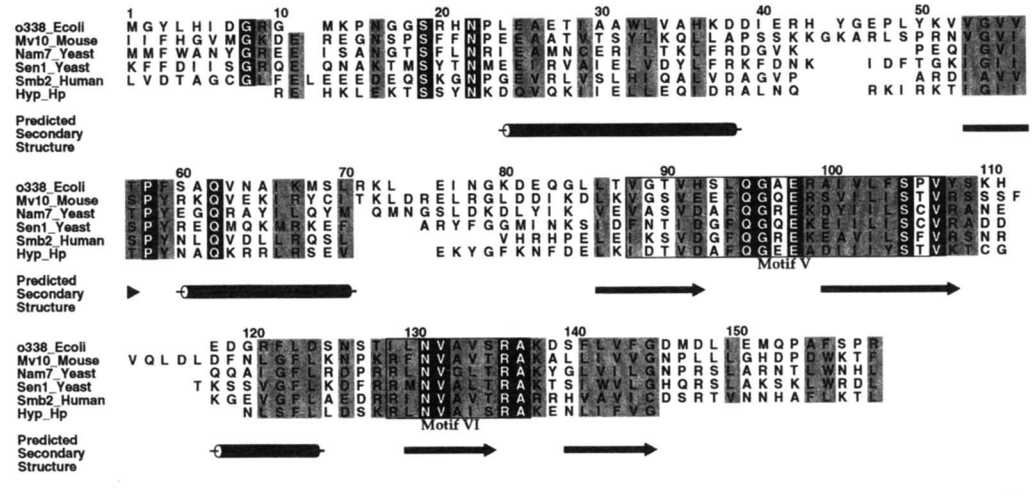

Fig. 3. AMPS (Barton, 1990) multiple alignment of the *E. coli o338* hypothetical gene product N-terminal sequence with regions of SEN1-like helicases and an *H. pylori* (Hp) hypothetical protein containing Motifs V and VI, represented using ALSCRIPT (Barton, 1993b). Residues are shown in white on black if they are absolutely conserved across all sequences; positions are shaded if the chemical character of residues (as Fig. 2) is conserved in >66% of sequences. Helicase motifs V and VI (Gorbalenya & Koonin, 1993) are boxed. Secondary structure prediction was undertaken using PHD (Rost & Sander, 1993). Sequences are represented by their SwissProt accession codes (except o338 [sw:YJHR_ECOLI] and the *H. pylori* ORF [gb:HPCAI 4464–4817]).

Lys 51, Asp 56, and Asn 70 (numbering as Fig. 2), with the exception that His 49 is not present in viral p37K proteins. It is significant, perhaps, that these residues are similar to the active site residues that catalyze hydrolysis of phosphodiester bonds by nucleases (Saenger, 1991). In these enzymes, an active site water molecule, activated by a metal or by a histidine residue, acts as a nucleophile by attacking the phosphate, which causes cleavage of the phosphodiesterase bond. In most nucleases, residues that are involved in this process, either directly or otherwise, are His, Asp, Glu, and for pancreatic ribonucleases A, Lys (reviewed in Saenger, 1991).

His 49 or Lys 51 may be involved in the formation of covalently bonded enzyme–substrate intermediates by analogy with the histidine–phosphorous intermediates during catalyses by, inter alia, nucleoside diphosphate kinase (NDP kinase) (Moréra et al., 1995), succinyl CoA synthetase (Wolodko et al., 1994), phosphoglycerate mutase (Campbell et al., 1974), and acid phosphatase (Schneider et al., 1993), and the lysine-phosphorous intermediates that occur for mRNA capping enzymes (Cong & Shuman, 1993) and DNA ligases (Engler & Richardson, 1982). Phosphorylation of Asp 56 in the intermediate state remains a possibility, by analogy with that observed elsewhere (Sanders et al., 1989).

Of these examples, the residues of the NDP-kinase active site are most similar to the conserved residues of the PLD homologues. In NDP-kinase, the phosphohistidine intermediate is stabilized by two arginine residues, and single lysine, asparagine, glutamate, and tyrosine residues (Moréra et al., 1995). Thus, a possible reaction scheme for the PLD homologues involves activation of the acceptor group, water for hydrolysis or an appropriate polar-head group for transphosphatidylation, by one or both histidine residues at position 49, and formation of histidine- or lysine- or aspartyl-phosphorous intermediates, stabilized by the remaining conserved residues. Hydrolysis or transphosphatidylation of this intermediate would then lead to product formation. This mechanism may be modified for nuc, which alone of all PLD homologues, absolutely requires divalent cations for activity (Lackey et al., 1977; cf. Raetz et al., 1987; Hiraoka et al., 1991; Vinggaard & Hansen, 1995; and references therein), and for p37K, which lacks His 49. These hypotheses are appropriate for examination by site-directed mutagenesis of the conserved His 49, Lys 51, Asp 56, and Asn 70 residues.

Materials and methods

Sequence comparisons were undertaken using amino acid and nucleotide sequence databases held at the NCBI, USA, and a subset of these at the University of Oxford, UK. Candidate homologues were detected using two algorithms: BLAST (Altschul et al., 1994) at the NCBI, and the local similarity algorithm (scanps) of Barton (1993a). The former enabled pairwise sequence comparisons and estimations of the probabilities (or p (m)-values) that sequence similarities within m ungapped blocks arose by chance. The latter allowed comparisons of gapped multiple alignments with amino acid sequence databases, resulting in hierarchical similarity score lists.

Estimation of p-values for ungapped blocks within multiple alignments was provided by the program MACAW (Schuler et al., 1991). Care was taken not to underestimate p-values by using only one member from each class of homologous enzymes (Table 1) in their calculation. Separate comparisons were undertaken for motifs in each lobe using two representative sets of domain sequences. These were: (set A) castor bean PLD, *Streptomyces antibioticus* phospholipase D, *E. coli* PS synthase, vaccinia p37K, ORF from *Synechocystis* sp., *S. typhimurium* nuc, and *E. coli* helicase homologue, and (set B) *S. cerevisiae* YK11/SPO14, *Streptomyces* sp. phospholipase D, *S. cerevisiae* PS synthase, vaccinia K4, *B. subtilis* CL synthase, ORF from *Synechocystis* sp., *S. typhimurium* nuc, and *E. coli* helicase homologue. The same sequence sets were used for the comparisons of motifs in both lobes, except for the omission of the single-lobed molecules (nuc and o338). The search space N (Schuler et al., 1991) in the comparison of n sequences was set as the product sequence lengths l_n, which encompassed the four sequence motifs (Fig. 2). Sequences, judged using MACAW as homologous, were aligned using AMPS (Barton, 1990).

For each of these computational methods, the BLOSUM62 amino acid substitution matrix (Henikoff & Henikoff, 1992) was used. Secondary structure predictions for all eight classes were provided by the neural network method (PHD) of Rost and Sander (1993). BLAST (Altschul et al., 1994) input query sequences were preprocessed using SEG (Wootton & Federhen, 1993) in order to mask low-complexity sequences.

Predictions of transmembrane helices were performed using MEMSAT (Jones et al., 1994) and TMS-PRED (Persson & Argos, 1994). TMS-PRED identifies potential membrane-spanning helices from multiple sequence alignments, whereas MEMSAT analyzes individual sequences. For the analysis by TMS-PRED of viral p37K and K4 proteins, sequences were aligned using AMPS (Barton, 1990).

Note added in proof

Following completion of this work a sequence of the human orthologue of yeast (class 1) PLDs was reported (Hammond SM, Altshuller YM, Sung TC, Rudge SA, Rose K, Engebrecht JA, Morris AJ, Frohman MA. 1995. Human ADP-ribosylation factor-activated phosphatidylcholine-specific phospholipase D defines a new and highly conserved gene family. *J Biol Chem 270*:29640–29643).

Acknowledgments

This is a contribution from the Oxford Centre for Molecular Sciences (OCMS), which is supported by the UK Engineering and Physical Sciences Research Council, the Biotechnology and Biological Science Research Council, and the Medical Research Council (MRC). C.P.P. is a MRC Training Fellow. We thank Drs. C.M. Dobson and M.S.P. Sansom (University of Oxford) for providing stimulating working environments, and Drs. P.J. Parker (Imperial Cancer Research Fund, London) and G.L. Smith (University of Oxford) for helpful discussions.

References

Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nature Genet 6*:119–129.

Barton GJ. 1990. Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol 183*:403–428.

Barton GJ. 1993a. An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Comput Applic Biosci 9*:729–734.

Barton GJ. 1993b. ALSCRIPT: A tool to format multiple sequence alignments. *Protein Eng 6*:37–40.

Blasco R, Moss B. 1991. Extracellular vaccinia virus formation and cell-to-cell virus transmission are prevented by deletion of the gene encoding the 37,000-Dalton outer envelope protein. *J Virol 65*:5910–5920.

Boursnell MEG, Foulds IJ, Campbell JI, Binns MM. 1988. Non-essential genes in the vaccinia virus HindIII fragment: A gene related to serine protease inhibitors and a gene related to the 37K vaccinia virus major envelope antigen. *J Gen Virol 69*:2995–3003.

Bork P, Sander C, Valencia A. 1992. An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc Natl Acad Sci USA 89*:7290–7294.

Burland V, Plunkett G III, Sofia HJ, Daniels DL, Blattner FR. 1995. Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucleic Acids Res 23*:2105–2119.

Campbell JW, Watson HC, Hodgson GI. 1974. Structure of yeast phosphoglycerate mutase. *Nature 250*:301–303.

Charbonneau H, Beier N, Walsh KA, Beavo JA. 1986. Identification of a conserved domain among cyclic nucleotide phosphodiesterases from diverse species. *Proc Natl Acad Sci USA 83*:9308–9312.

Cherepanov PA, Mikhailova TG, Karimova GA, Zakharova NM, Ershov YV, Volkovoi KI. 1991. Cloning and detailed mapping of the Fra-ymt region of the *Yersinia pestis* plasmid pFra. *Mol Gen Mikrobiol Virusol 12*:19–26.

Cong P, Shuman S. 1993. Covalent catalysis in nucleotidyl transfer: A KTDG motif essential for enzyme-GMP complex formation by mRNA capping enzyme is conserved at the active sites of RNA and DNA ligases. *J Biol Chem 268*:7256–7260.

Covacci A, Censini S, Bugnoli M, Petracca R, Burroni D, Macchia G, Massone A, Papini E, Xiang Z, Figura N, Rappuoli R. 1993. Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. *Proc Natl Acad Sci USA 90*:5791–5795.

DeChavigny A, Heacock PN, Dowhan W. 1991. Sequence and inactivation of the *pss* gene of *Escherichia coli*. Phosphatidylethanolamine may not be essential for cell viability. *J Biol Chem 266*:5323–5332.

DeMarini DJ, Winey M, Ursic D, Webb F, Culbertson MR. 1992. SEN1, a positive effector of tRNA-splicing endonuclease in *Saccharomyces cerevisiae*. *Mol Cell Biol 12*:2154–2164.

Dyer JD, Zheng L, Wang X. 1995. Cloning and nucleotide sequence of a cDNA (accession no. U36381) encoding phospholipase D from *Arabidopsis* (PGR95-096). *Plant Physiol*. Forthcoming.

Engler MJ, Richardson CC. 1982. DNA ligases. In: Boyer PD, ed. *The enzymes, vol XV*. New York: Academic Press. pp 3–29.

Giorda R, Ohmachi T, Ennis HL. 1989. Organization of a gene family developmentally regulated during *Dictyostelium discoideum* spore germination. *J Mol Biol 205*:63–69.

Gorbalenya AE, Koonin EV. 1993. Helicases: Amino acid sequence comparisons and structure-function relationships. *Curr Opin Struct Biol 3*:419–429.

Heller M. 1976. Phospholipase D. *Adv Lipid Res 16*:267–326.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*:10915–10919.

Hiraoka S, Nukui K, Uetake N, Ohta A, Shibuya I. 1991. Amplification and substantial purification of cardiolipin synthase of *Escherichia coli*. *J Biochem 110*:443–449.

Hirschberg CB, Kennedy EP. 1972. Mechanism of the enzymatic synthesis of cardiolipin in *Escherichia coli*. *Proc Natl Acad Sci 69*:648–651.

Hirt P, Hiller G, Wittek R. 1986. Localisation and fine structure of a vaccinia virus gene encoding an envelope antigen. *J Virol 58*:757–764.

Honigberg SM, Conicella C, Esposito RE. 1992. Commitment to meiosis in *Saccharomyces cerevisiae*: Involvement of the *SPO14* gene. *Genetics 130*:703–716.

Iwasaki Y, Nakano H, Yamane T. 1994. Phospholipase D from *Streptomyces antibioticus*: Cloning, sequencing, expression, and relationship to other phospholipases. *Appl Microbiol Biotechnol 42*:290–299.

Jones DT, Taylor WR, Thornton JM. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry 33*:3038–3049.

Juneja LR, Kazuoka T, Goto N, Yamane T, Shimizu S. 1989. Conversion of phosphatidylcholine to phosphatidylserine by various phospholipases D in the presence of L- or D-serine. *Biochim Biophys Acta 1003*:277–283.

Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S. 1995. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis spp. strain PCC6803: I. Sequence features in the 1Mb region from map positions 64% to 92% of the genome. *DNA Res 2*:153–166.

Keim P, Heinrikson RL, Fitch WM. 1981. An examination of the expected degree of sequence similarity that might arise in proteins that have conserved similar conformational states. *J Mol Biol 151*:179–197.

Koonin EV, Rudd KE. 1996. Two domains of superfamily I helicases may exist as separate proteins. *Protein Sci 5*:178–180.

Kuppe A, Evans LM, McMillen DA, Griffith OH. 1989. Phosphatidylinositol-specific phospholipase C of *Bacillus cereus*: Cloning, sequencing, and relationship to other phospholipases. *J Bacteriol 171*:6077–6083.

Lackey D, Walker GC, Keng T, Linn S. 1977. Characterization of an endonuclease associated with the drug resistance plasmid pKM101. *J Bacteriol 131*:583–588.

Mathieu M, Zeelen JP, Pauptit RA, Erdmann R, Kunau WH, Wierenga RK. 1994. The 2.8 Å crystal structure of peroxisomal 3-ketoacyl-CoA thiolase of *Saccharomyces cerevisiae*: A five-layered $\alpha\beta\alpha\beta\alpha$ structure constructed from two core domains of identical topology. *Structure 2*: 797–808.

Miller M, Jaskólski M, Rao JKM, Leis J, Wlodawer A. 1989. Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature 337*:576–579.

Moréra S, Chiadmi M, LeBras G, Lascu I, Janin J. 1995. Mechanism of phosphate transfer by nucleoside diphosphate kinase: X-ray structures of the phosphohistidine intermediate of the enzymes from *Drosophila* and *Dictyostelium*. *Biochemistry 34*:11062–11070.

Morton RK. 1953. Transferase activity of hydrolytic enzymes. *Nature 172*: 65–68.

Murzin AG. 1993. Can homologous proteins evolve different enzymatic activities? *Trends Biochem Sci 18*:403–405.

Mushegian AR, Koonin EV. 1994. Unexpected sequence similarity between nucleosidases and phosphoribosyltransferases of different specificity. *Protein Sci 3*:1081–1088.

Nikawa JI, Kodaki T, Yamashita S. 1987. Primary structure and disruption of the phosphatidylinositol synthase gene of *Saccharomyces cerevisiae*. *J Biol Chem 262*:4876–4881.

Nishijima S, Asami Y, Uetake N, Yamagoe S, Ohta A, Shibuya I. 1988. Disruption of the *Escherichia coli cls* gene responsible for cardiolipin synthesis. *J Bacteriol 170*:775–780.

Nishizuka Y. 1995. Protein kinase C and lipid signaling for sustained cellular responses. *FASEB J 9*:484–496.

Okamura S, Yamashita S. 1994. Purification and characterization of phosphatidylcholine phospholipase D from pig lung. *J Biol Chem 269*: 31207–31213.

Payne L. 1978. Polypeptide composition of extracellular enveloped vaccinia virus. *J Virol 27*:28–37.

Persson B, Argos P. 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J Mol Biol 237*:182–192.

Pohlman RF, Liu F, Wang L, Moré MI, Winans SC. 1993. Genetic and biochemical analysis of an endonuclease encoded by the IncN plasmid pKM101. *Nucleic Acids Res 21*:4867–4872.

Ponting CP, Parker PJ. 1996. Extending the C2 domain family: C2s in PKCs δ, ϵ, η, θ, phospholipases, GAPs, and perforin. *Protein Sci 5*:162–166.

Raetz CRH, Carman GM, Dowhan W, Jiang RT, Waszkuc W, Loffredo W, Tsai MD. 1987. Phospholipids chiral at phosphorus. Steric course of the reactions catalyzed by phosphatidylserine synthase from *Escherichia coli* and yeast. *Biochemistry 26*:4022–4027.

Ragolia L, Tropp BE. 1994. The effects of phosphoglycerides on *Escherichia coli* cardiolipin synthase. *Biochim Biophys Acta 1214*:323–332.

Robins P, Pappin DJC, Wood RD, Lindahl T. 1994. Structural and functional homology between mammalian DNase IV and the 5′-nuclease domain of *Escherichia coli* DNA polymerase I. *J Biol Chem 269*:28535–28538.

Rose K, Rudge SA, Frohman MA, Morris AJ, Engebrecht JA. 1995. Phospholipase D signaling is essential for meiosis. *Proc Natl Acad Sci USA 92*:12151–12155.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Saenger W. 1991. Structure and catalytic function of nucleases. *Curr Opin Struct Biol 1*:130–138.

Sanders DA, Gillececastro BL, Stock AM, Burlingame AL, Koshland DE. 1989. Identification of the site of phosphorylation of the chemotaxis response regulator protein, CheY. *J Biol Chem 264*:21770–21778.

Scallon BJ, Fung WJC, Tsang TC, Li S, Kado-Fong H, Huang KS, Kochan JP. 1991. Primary structure and functional activity of phosphatidylinositol-glycan-specific phospholipase D. *Science 252*:446–448.

Schmutz C, Payne LG, Gubser J, Wittek R. 1991. A mutation in the gene encoding the vaccinia virus 37,000-M_r protein confers resistance to an inhibitor of virus envelopment and release. *J Virol 65*:3435–3442.

Schneider G, Lindqvist Y, Vihko P. 1993. Three-dimensional structure of rat acid phosphatase. *EMBO J 12*:2609–2615.

Schuler GD, Altschul SF, Lipman DJ. 1991. A workbench for multiple alignment construction and analysis. *Proteins Struct Funct Genet 9*:180–190.

Stanacev NZ, Stuhne-Sekalec L. 1970. On the mechanism of enzymatic phos-

phatidylation. Biosynthesis of cardiolipin catalysed by phospholipase D. *Biochim Biophys Acta 210*:350–352.

Tang J, James MNG, Hsu IN, Jenkins JA, Blundell TL. 1978. Structural evidence for gene duplication in the evolution of the acid proteases. *Nature 271*:618–621.

Thayer MM, Ahern H, Xing D, Cunningham RP, Tainer JA. 1995. Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *EMBO J 14*:4108–4120.

Tropp BE, Ragolia L, Xia W, Dowhan W, Milkman R, Rudd KE, Ivanisevic R, Savic DJ. 1995. Identity of the *Escherichia coli cls* and *nov* genes. *J Bacteriol 177*:5155–5157.

Tunaitus E, Cronan JE Jr. 1973. Characterization of the cardiolipin synthetase activity of *Escherichia coli* envelopes. *Arch Biochem Biophys 155*:420–427.

Ueki J, Morioka S, Komari T, Kumashiro T. 1995. Purification and characterization of phospholipaseD (PLD) from rice (*Oryza sativa* L) and cloning of cDNA for PLD from rice and maize (*Zea mays* L). *Plant Cell Physiol 36*:903–914.

Vinggaard AM, Hansen HS. 1995. Characterization and partial purification of phospholipase D from human placenta. *Biochim Biophys Acta 1258*: 169–176.

Wang X, Xu L, Zheng L. 1994. Cloning and expression of phosphatidylcholine-hydrolyzing phospholipase D from *Ricinus communis* L. *J Biol Chem 269*:20312–20317.

Wolodko WT, Fraser ME, James MNG, Bridger WA. 1994. The crystal structure of succinyl-CoA synthetase from *Escherichia coli* at 2.5 Å resolution. *J Biol Chem 269*:10883–10890.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem 17*:149–163.

Yang SF, Freer S, Benson AA. 1967. Transphosphatidylation by phospholipase D. *J Biol Chem 242*:477–484.