Eur. J. Biochem. **271**, 2705–2715 (2004) © FEBS 2004 doi:10.1111/j.1432-1033.2004.04199.x

Evolutionary relationships of the prolyl oligopeptidase family enzymes

Jarkko I. Venäläinen, Risto O. Juvonen and Pekka T. Männistö

Department of Pharmacology and Toxicology, University of Kuopio, Finland

The prolyl oligopeptidase (POP) family of serine proteases includes prolyl oligopeptidase, dipeptidyl peptidase IV, acylaminoacyl peptidase and oligopeptidase B. The enzymes of this family specifically hydrolyze oligopeptides with less than 30 amino acids. Many of the POP family enzymes have evoked pharmaceutical interest as they have roles in the regulation of peptide hormones and are involved in a variety of diseases such as dementia, trypanosomiasis and type 2 diabetes. In this study we have clarified the evolutionary relationships of these four POP family enzymes and analyzed POP sequences from different sources. The phylogenetic trees indicate that the four enzymes were present in the last common ancestor of all life forms and that the β-propeller domain has been part of the family for billions of years. There are striking differences in the mutation rates between the enzymes and POP was found to be the most conserved enzyme of this family. However, the localization of this enzyme has changed throughout evolution, as three archaeal POPs seem to be membrane bound and one third of the bacterial as well as two eukaryotic POPs were found to be secreted out of the cell. There are also considerable distinctions between the mutation rates of the different substrate binding subsites of POP. This information may help in the development of species-specific POP inhibitors.

Keywords: acylaminoacyl peptidase; dipeptidyl peptidase IV; evolution; oligopeptidase B; prolyl oligopeptidase.

The prolyl oligopeptidase family of serine proteases (clan SC, family S9) includes a number of peptidases, from which prolyl oligopeptidase (POP, EC 3.4.21.26), dipeptidyl peptidase IV (DPP IV, EC 3.4.14.5), oligopeptidase B (OB, EC 3.4.21.83) and acylaminoacyl peptidase (ACPH, EC 3.4.19.1) have been the enzymes under the most intense study [1–3]. This enzyme family is different from the classical serine protease families, trypsin and subtilisin, in that they cleave only peptide substrates while excluding large proteins. The mechanism of preventing the digestion of bigger proteins was recently clarified when the 3D structure of POP was solved [4]. The enzyme consists of a peptidase and seven-bladed β-propeller domains. The narrow entrance of β-propeller prevents larger proteins from entering into the enzyme active site. A similar β-propeller consisting of eight instead of seven blades was recently identified in DPP IV when its crystal structure was solved [5].

Correspondence to J. I. Venäläinen, Department of Pharmacology and Toxicology, University of Kuopio, P.O. Box 1627, FIN-70211 Kuopio, Finland. Fax: + 358 17 162424, Tel.: + 358 17 163774, E-mail: Jarkko.Venalainen@uku.fi

Abbreviations: ACPH, acylaminoacyl peptidase; DPPII, dipeptidyl peptidase II; DPP IV, dipeptidyl peptidase IV; OB, oligopeptiadase B; POP, prolyl oligopeptidase; GPI, glycosylphosphatidylinositol; LUCA, last universal common ancestor.

Enzymes: prolyl oligopeptidase (EC 3.4.21.26); dipeptidyl peptidase IV (EC 3.4.14.5); oligopeptidase B (EC 3.4.21.83); acylaminoacyl peptidase (EC 3.4.19.1).

Note: The departmental website is available at http://www.uku.fi/farmasia/fato/indexe.htm

(Received 28 March 2004, revised 28 April 2004, accepted 4 May 2004)

The enzymes of the POP family have different substrate specificities: POP hydrolyzes peptides at the carboxyl side of the proline residue, DPP IV liberates dipeptides where the penultimate amino acid is proline, OB cleaves peptides at lysine and arginine residues and ACPH removes N-acetylated amino acids from blocked peptides. DPP IV is a membrane bound enzyme, and in this way different from the rest of the POP family members that are cytoplasmic proteins [3]. However, a membrane bound form of POP has also been characterized from bovine brain but the sequence of this protein is not available at the present time [6].

Many of the POP family enzymes have become targets of the pharmaceutical industry, e.g. POP degrades many neuropeptides involved in learning and memory, such as substance P, thyrotropin releasing hormone and arginine-vasopressin. Indeed, POP inhibitors have been shown to reverse scopolamine-induced amnesia in rats and to improve cognition in old rats and 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-treated Parkinsonism model monkeys [7–9]. A number of the antitrypanosomal drugs in widespread use are OB inhibitors [10]. In addition, inhibition of DPP IV has been proposed as a therapeutic approach to the treatment of type 2 diabetes as this enzyme is involved in the metabolic inactivation of a glucagon-like peptide 1 that stimulates insulin secretion [11]. Recently, DPP IV knockout mice were found to be protected against obesity and insulin resistance [12].

In this study, based on public databanks and a number of computer programs, we have clarified the evolutionary relationships of these four POP family enzymes by generating phylogenetic trees including POP family enzymes from different species. First, important amino acids for the enzyme function were sought by analyzing multiple alignments of 72 aligned POP family sequences. Secondly, we analyzed POP sequences from different species because POP

can be considered as a model enzyme of this family, as its crystal structure is available and many details about its catalytic mechanism are known. In this analysis we created a conservation profile of POP to study the mutation rates of amino acids involved in substrate binding and to find other essential amino acids. Finally, we pinpointed signal sequences, and transmembrane and lipid anchor sequences from POP enzymes of different sources to study if the localization of the enzyme has changed during evolution.

Materials and methods

Multiple sequence alignment and construction of phylogenetic trees of the POP family

The POP family enzymes from different sources were identified by BLASTP searches from the NCBI nr database against human POP (NP_002717), human DPP IV (CDHU26), human ACPH (P13798), *Escherichia coli* OB (E64946) and rat DPP II (JC7668) sequences. To be identified as a POP family member, the sequence had to have the catalytic triad topology of Ser-Asp-His which is different from the classical serine proteases [13]. The iterative PSI-BLAST feature was not applied in these searches. The aim of the searches was to obtain a large enough number of sequences for the analysis, not to find all the existing POP family sequences. As a result, 28 POP, 10 ACPH, 14 DPP IV, 20 OB and seven DPP II sequences from different species were manually selected for the analysis. The selected sequences and their accession codes are presented in Table 1.

A multiple sequence alignment of the 79 selected sequences was constructed by a combination of T-COFFEE and CLUSTALX programs [14,15]. A structure based sequence alignment of pig POP (1QFS) and human DPP IV (IJ2E) was created using the T-COFFEE program and other proteins were subsequently added to this alignment using the CLUSTALX program until the multiple sequence alignment of 79 sequences was obtained. The alignment was manually edited based on the initial 3D alignment. The neighbor-joining tree was constructed for the peptidase domains of the enzymes (corresponding to the pig POP residues 1–72 and 428–710) and for the complete sequences using CLUSTALX. Bootstrap values were calculated with 1000 resamplings. DPP II sequences were used as the outgroup in this analysis, as this enzyme is a close neighbor to the POP family and a member of the serine protease family S28. The NJPLOT program was used to display the constructed phylogenetic tree. The phylogenetic trees were also constructed using the maximum likelihood method with the program TREE-PUZZLE [16]. The TREEVIEW program (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) was used to view the maximum likelihood tree.

Conservation profile of POP sequences

To study the conservation profile of POP, 28 POP sequences alone were aligned using T-COFFEE. Multiple sequence alignments were visualized and analyzed using GENEDOC program (http://www.psc.edu/biomed/genedoc/) alongside the pig POP sequence. The conservation rates of each of the 710 amino acids were divided into four groups: 1st, ≤49%; 2nd, between 50 and 74%; 3rd, between 75 and 99% and 4th, 100% similarity at an alignment position. The similarities of amino acids were based on BLOSUM62 substitution matrix.

Prediction of transmembrane regions, lipid anchors and signal peptides in POP sequences

All of the 28 POP sequences from different sources were analyzed with TMHMM program [17] to decide whether the enzymes contain transmembrane sequences. Lipid anchor sites were searched with program BIG PI [18]. The presence of signal sequences in the POP enzymes and their possible cleavage sites were predicted with the SIGNALP V2.0 program using hidden Markov model method [19].

Results and Discussion

Multiple sequence alignment of the POP family enzymes

As can be seen from Table 1, POP and ACPH are distributed in archaeal, bacterial and eukaryotic species whereas DPP IV and OB were not found from archaeal sources. Although POP and ACPH are present in all three forms of organisms (Bacteria, Archaea, Eucaryota), there are several organism groups in which these enzymes were not found. For example, POP was not found in Fungi. Table 2 lists some identity and similarity percentages within POP family enzymes when the whole sequences or just the catalytic domains of the enzymes are taken into account. In general, the sequence identity percentages between the four enzymes are low, below 20%. The peptidase domain is slightly more conserved, as shown by the higher identity/similarity percentages. However, despite the low sequence homology and distinct substrate specificities, the multiple sequence alignment revealed 10 invariant residues between the 72 aligned enzymes of the POP family: Arg505, Gly506, Gly511, Asp529, Gly552, Ser554, Gly556, Gly557, Asp641 and His680 (numbering according to the pig POP sequence, the residues are shown with downward arrows in Fig. 1). All of these amino acids are located at the active site of the enzyme. This was expected, as it has been reported previously that the greatest similarities between the amino acid sequences of POP family members are located in the C-terminal third of the alignment [20]. Of these conserved residues, Ser554, Asp641 and His680 form the catalytic triad of POP and the small residues Gly552, Gly556 and Gly557 are clustered around the catalytic serine. The three glycine residues have been proposed to improve the binding of substrate by preventing steric hindrance [4]. Arg505 and Gly506 are situated in a loop between the β4-strand and the αB′-helix at the active site, and Gly511 is the first residue of that α-helix. The high degree of conservation of these residues suggests that this turn between the secondary structure elements is crucial for the POP family enzyme function or for its structural stability.

Figure 2 represents some amino acid similarity percentages of whole sequences and catalytic domains between human and some eukaryotic, bacterial and archaeal sequences of POP, DPP IV and ACPH. The similarities between human and rat sequences are very high for POP (98/98%; whole sequences and catalytic domains, respectively)

Table 1. Prolyl oligopeptidase family and DPP II enzymes from different species used in this analysis.

Enzyme	Species	Domain of life	Accession number
POP	Human	Eukarya	NP_002717
	Pig	Eukarya	P23687
	Bovine	Eukarya	Q9XTA2
	Mouse	Eukarya	NP_035286.1
	Rat	Eukarya	NP_112614.1
	Fugu rubribes	Eukarya	SINFRUP00000059740
	Xenopus laevis	Eukarya	AAH47161
	Arabidopsis thaliana	Eukarya	AAL86330.1
	Dictyostelium discoideum	Eukarya	CAB40787.1
	Drosophila melanogaster	Eukarya	AAF52942.1
	Anophiles gambiae	Eukarya	EAA14977.1
	Oryza sativa	Eukarya	BAB78619.1
	Deinococcus radiodurans	Bacteria	NP_296223.1
	Shewanella oneidensis	Bacteria	NP_718337.1
	Trichodesmium erythraeum	Bacteria	ZP_00072911.1
	Nostoc sp.	Bacteria	NP_486573.1
	Nostoc punctiforme	Bacteria	ZP_00110050.1
	Flavobacterium meningosepticum	Bacteria	P27028
	Aeromonas punctata	Bacteria	AAD34991.1
	Aeromonas hydrophila	Bacteria	Q06903
	Novosphingobium capsulatum	Bacteria	BAA34052.1
	Novosphingobium aromaticivorans	Bacteria	ZP_00093416.1
	Myxococcus xanthus	Bacteria	AF127082–3
	Thermobifida fusca	Bacteria	ZP_00058751.1
	Pyrococcus abyssi	Archaea	NP_126828.1
	Pyrococcus furiosus	Archaea	NP_578544.1
	Pyrococcus horikoshii	Archaea	NP_143154.1
	Sulfolobus tokodaii	Archaea	NP_375840
DPP IV	Human	Eukarya	CDHU26
	Bovine	Eukarya	P81425
	Cat	Eukarya	Q9N2I7
	Rat	Eukarya	A39914
	Mouse	Eukarya	NP_034204.1
	Xenopus laevis	Eukarya	CAA70136.1
	Fugu rubribes	Eukarya	SINFRUP00000066299
	Anopheles gambiae	Eukarya	EAA05700.1
	Drosophila melanogaster	Eukarya	NP_608961.1
	Aspergillus niger	Eukarya	CAC1019.1
	Scizosaccharomyces pombe	Eukarya	NP_593970.1
	Aspergillus fumigatus	Eukarya	AAC34310.1
	Porphyromonas gingivalis	Bacteria	BAA28265.1
	Flavobacterium meningosepticum	Bacteria	S66261
ACPH	Human	Eukarya	P13798
	Rat	Eukarya	NP_036632.1
	Pig	Eukarya	JU0132
	Caenorhabditis elegans	Eukarya	NP_500647.1
	Fugu rubribes	Eukarya	SINFRUP00000057906
	Basillus subtilis	Bacteria	NP_391103.1
	Oceanobasillus iheyensis	Bacteria	NP_692002.1
	Pyrococcus abyssi	Archaea	NP_127272.1
	Pyrococcus horikoshii	Archaea	NP_142793.1
	Deinococcus radiodurans	Bacteria	NP_293889.1
OB	*Trypanosoma brucei*	Eukarya	AAC80459.1
	Leishmania major	Eukarya	AAD24761.1
	Escherichia coli	Bacteria	E64946
	Shigella flexneri	Bacteria	NP_707707.1
	Salmonella typhimurium	Bacteria	NP_460836.1
	Yersinia pestis	Bacteria	NP_669832.1

Table 1. Continued

Enzyme	Species	Domain of life	Accession number
	Shewanella oneidensis	Bacteria	NP_715786.1
	Xanthomonas axonopodis	Bacteria	NP_640984
	Nostoc sp.	Bacteria	NP_487951.1
	Treponema denticola	Bacteria	AAK39550.1
	Sinorhizobium meliloti	Bacteria	NP_385091.1
	Acrobacterium tumefaciens	Bacteria	NP_353917.1
	Brucella melitensis	Bacteria	NP_540282.1
	Brucella suis	Bacteria	NP_697584.1
	Mycobacterium leprae	Bacteria	NP_302455.1
	Corynebacterium glutamicum	Bacteria	NP_601794.1
	Rickettsia conorii	Bacteria	NP_360014.1
	Rickettsia prowazekii	Bacteria	NP_220665.1
	Bifidobacterium longum	Bacteria	NP_696390.1
	Moraxella lacunata	Bacteria	Q59536
DPP II	Rat	Eukarya	JC7668
	Human	Eukarya	Q9UHL4
	Mouse	Eukarya	Q9ET22
	Arabidopsis thaliana	Eukarya	NP_201377.2
	Anopheles gambiae	Eukarya	EAA04920.1
	Drosophila melanogaster	Eukarya	AAF53897.1
	Caenorhabditis elegans	Eukarya	NP_498718.1

Table 2. Amino acid identity/similarity percentages between POP family enzymes. The identity/similarity percentages of the peptidase domains are shown in brackets.

	POP Human	ACPH Human	DPP IV Human	OB *E. coli*
POP Human	–	9/24 (10/28)	15/30 (17/30)	22/41 (27/46)
ACPH Human		–	10/22 (13/30)	10/24 (14/27)
DPP IV Human			–	11/23 (12/25)
OB *E. coli*				–

and ACPH (95/96%), whereas the similarity between human and rat DPP IV is much lower for both whole sequences and catalytic domains (87/87%). The differences in conservation percentages are even more striking between human/*Fugu rubribes* enzymes and the same kind of conservation order can also be found between human/*Flavobacterium meningosepticum* (55/62% of POP compared to 38/48% of DPP IV) and human/*Pyrococcus abyssi* (42/50% of POP compared to 27/36% of ACPH). OB was excluded from this comparison because it is not found in animals. However, the similarity percentage between OB from *Shewanella oneidensis* and *Nostoc* sp. can be compared to that of POP. Again, POP has the higher conservation percentage: 66/73% compared to 59/69% of OB. This analysis indicates that POP is the most conserved peptidase of these four POP family enzymes, with the highest similarities found between each pair of sequences studied. The differences in conservation degrees between the enzymes are similar when the identity percentages are considered.

The phylogenetic tree of the POP family

The multiple alignment peptidase domains of 72 POP family sequences and seven DPP II sequences were used to construct phylogenetic trees with distance-based (neighbor-joining) and character-based (maximum likelihood) methods. In many cases, these two methods have been shown to be almost equally efficient in obtaining the correct topology [21,22]. The DPP II family was used as an outgroup for phylogenetic constructions. The two tree-building methods gave essentially the same tree topologies and the neighbour-joining tree with bootstrap values and the maximum likelihood tree with support values are shown in Figs 3 and 4. The phylogenetic trees clearly show that each of the four POP family enzymes (POP, DPP IV, OB and ACPH) form a single cluster containing all of the species included in this analysis. Both trees show that OB is the closest relative to POP, not ACPH as was recently stated [23] and that DPP IV is the closest relative to ACPH. In the cases of POP and ACPH the enzyme clusters have members from each of the three domains of the organisms. In this analysis, DPP IV and OB sequences were not found from archaeal species. These four enzyme clusters are supported by high bootstrap values in the neighbor-joining tree and support values in the maximum likelihood tree. The clusters are further divided in subclusters, for example, the POP cluster forms subclusters of archaea (*Pyrococcus horikoshii*, *P. abyssi*, *Pyrococcus furiosus* and *Sulfolobus tokodaii*) and

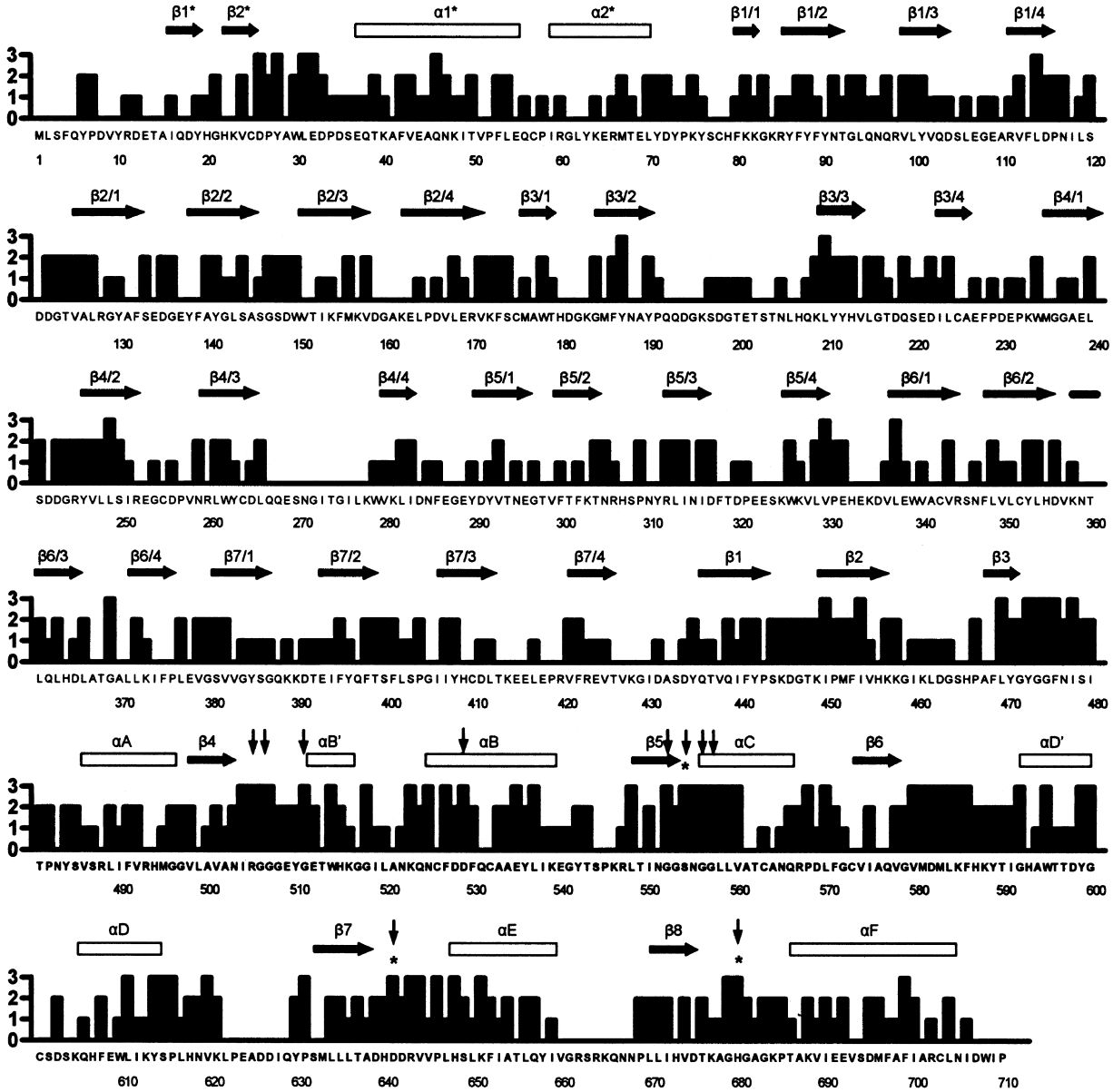


Fig. 1. Conservation profile of 28 POP sequences from different species. The conservation percentage of each amino acid along the pig POP sequence is indicated as 0, ≤49%; 1, between 50 and 74%; 2, between 75 and 99% and 3, 100%. The secondary structure elements of pig POP are indicated by arrows for β-sheets and by boxes for α-helices. The invariant amino acids in each of the 72 analysed POP family sequences are shown by downward arrows and the amino acids of the catalytic triad (Ser554, Asp641 and His680) are indicated by asterisks.

eukaryotes. It is interesting to note that according to the POP cluster of the phylogenetic trees, *Drosophila melanogaster* and *Anopheles gambiae* differ more from mammals than do the plants *Oryza sativa* and *Arabidopsis thaliana*. The most probable reason for this apparent discrepancy is that these two insects diverged considerably faster than vertebrates. At the gene sequence level, these two species that diverged 250 million years ago, differ more than even humans and pufferfish *F. rubribes* – species that diverged 450 million years ago [24]. This discovery is valid also with the POP enzyme having sequence identity of 58% between *A. gambiae* and *D. melanogaster* and 74% between human and *F. rubribes*. A similar order of sequence identities can also be seen with DPP IV. The phylogenetic trees were also created using the complete sequences of the enzymes (data not shown). These analyses resulted in the same tree topologies as seen in Figs 3 and 4, except that the branch lengths are slightly longer due to the lower conservation of the β-propeller domains. This shows that the β-propeller domain has been part of this enzyme family for billions of years.

The phylogenetic trees show that the four POP family enzymes were clearly set up before the archaea, prokaryota and eucaryota diverged along their own evolutionary lines between 2000 and 4000 million years ago. This suggests that all POP family proteins are of ancient origin and they were

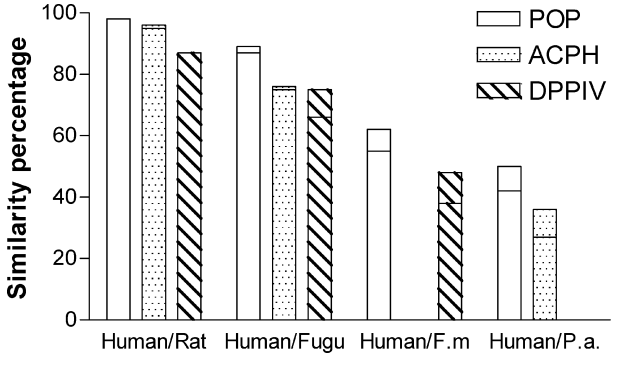


Fig. 2. Amino acid similarity percentages between human–rat, human–*F. rubribes*, human–*F. meningosepticum* and human–*P. abyssi* sequences of POP, ACPH and DPP IV. The whole bar and the lower part of the bar represent the similarity percentages of the catalytic domains and the complete sequences, respectively.

present in the last universal common ancestor (LUCA) of all life forms. Thus, the present enzyme forms are vertically inherited from this ancestor.

The high conservation of POP family enzyme sequences from different species and their presence in the LUCA strongly suggest that these enzymes have important roles in physiological processes. However, the exact roles of these enzymes are more or less unclear at the moment. Evidently there was a need for peptidases that cleave only small peptides specifically after proline, lysine or arginine even during the early days of life.

Conservation profile of POP sequences from different species

The conservation profile of 28 aligned POP sequences is presented in Fig. 1. It is clear that the catalytic domain (residues 1–72 and 428–710) is a much more conserved region than the β-propeller domain (residues 73–427). In the β-propeller domain, only seven amino acids (2.0%) have 100% similarity compared to 65 amino acids (17.8%) in the catalytic domain. Six of the conserved amino acids in the β-propeller are situated in β-sheets and one (Gly369) is located between the β-sheet structures, so that the β-sheets seem to be more conserved than the areas between them. The low homology in the β-propeller domain is not unexpected, as it has been proposed that the β-propeller of *P. furiosus* POP does not perform the same function as the mammalian enzyme, i.e. the exclusion of large peptides from the active site [25]. Clearly the role of the β-propeller has diversified during evolution.

Table 3 lists the conservation percentages of the pig POP active site amino acids that are involved in the substrate binding [4]. The specificity pocket S1 has 100% similary and almost 100% identity among the 28 studied POP sequences. Only Val580 and Tyr599 have some variations among different species. In addition to the amino acids of the catalytic triad and the residues that make hydrogen bonds with substrate, Trp595 is also invariant. This residue is claimed to enhance substrate recognition specificity by ring stacking between the indole ring of Trp595 and the proline ring of the substrate, so that all of the studied POP enzymes can be claimed to be specific for proline [4]. It is surprising that residues Phe476, Val644, Val580 and Tyr599 also have 100% similarities and 89.3–100% identities, as their role in substrate binding is just to provide a hydrophobic environment and appropriate lining for the proline residue [4]. Due to this conservation, it can be predicted that the changes of these residues would dramatically decrease the specificity for, or binding of, the proline residue.

The specificity pocket S3 is substantially more variable than the S1 pocket. In pig POP, the S3 pocket ensures that there is a fairly apolar environment. However, this is not common for all POP sequences, because in many species the POP enzyme contains polar and even charged residues (i.e. Asn, Gly, Ser, Asp) at this site. Hence, it seems that only the substrate binding S1 site has remained virtually unchanged throughout the evolution, allowing enhanced flexibility to substrate S2 and S3 residues. There have been attempts to develop species specific POP inhibitors, for example against *Trypanosoma cruzi* [26]. According to our analysis of subsite evolution, the specificity might be achieved by varying the structures of P2 and P3, but not the P1 subsite of the inhibitor.

The most interesting amino acid at the S3 subsite is Cys255, because it is responsible for pig POP inhibition by bulky thiol reagents. *F. meningosepticum*, which has a Thr instead of Cys255, is not inhibited by thiol reagents. In addition to accounting for the inhibition by thiol reagents, Cys255 also improves the catalytic efficacy at pH values above neutrality by increasing the substrate affinity [27]. Therefore it is interesting to note that, of the 28 studied POP sequences, only eukaryotes have cysteine at this site. Most bacterial POP sequences have threonine in place of Cys255 but *Myxococcus xanthus* has tryptophan instead of Cys255. All of the studied archaeal POP enzymes have tryptophan at the same location. This variability of amino acids between the three domains of life is important, because it clearly modifies enzyme properties, i.e. substrate affinity and perhaps also the regulation by oxidation state.

Transmembrane regions and signal peptides in POP sequences

Twenty eight POP sequences were analyzed with TMHMM program to detect transmembraneous regions in the enzyme, because POP has also been characterized in a membrane bound form from bovine brain [6]. Unfortunately, the sequence of this apparently membrane bound POP has not been published. Therefore, it is impossible to conclude whether the enzyme is another form of cytosolic POP or some other enzyme possessing similar properties to POP. The program used in this analysis was recently evaluated to have the best overall performance of the currently available and most widely used transmembrane prediction tools [28]. According to our analysis conducted using the TMHMM program, none of the sequences were predicted to contain transmembrane regions. However, *Novosphingobium capsulatum* POP had a weak possibility (0.45) of a transmebrane region. To decide whether this protein is membrane bound or not, we analyzed this sequence with another transmembrane prediction program, SOSUI [29]. This program also predicted the sequence to be of

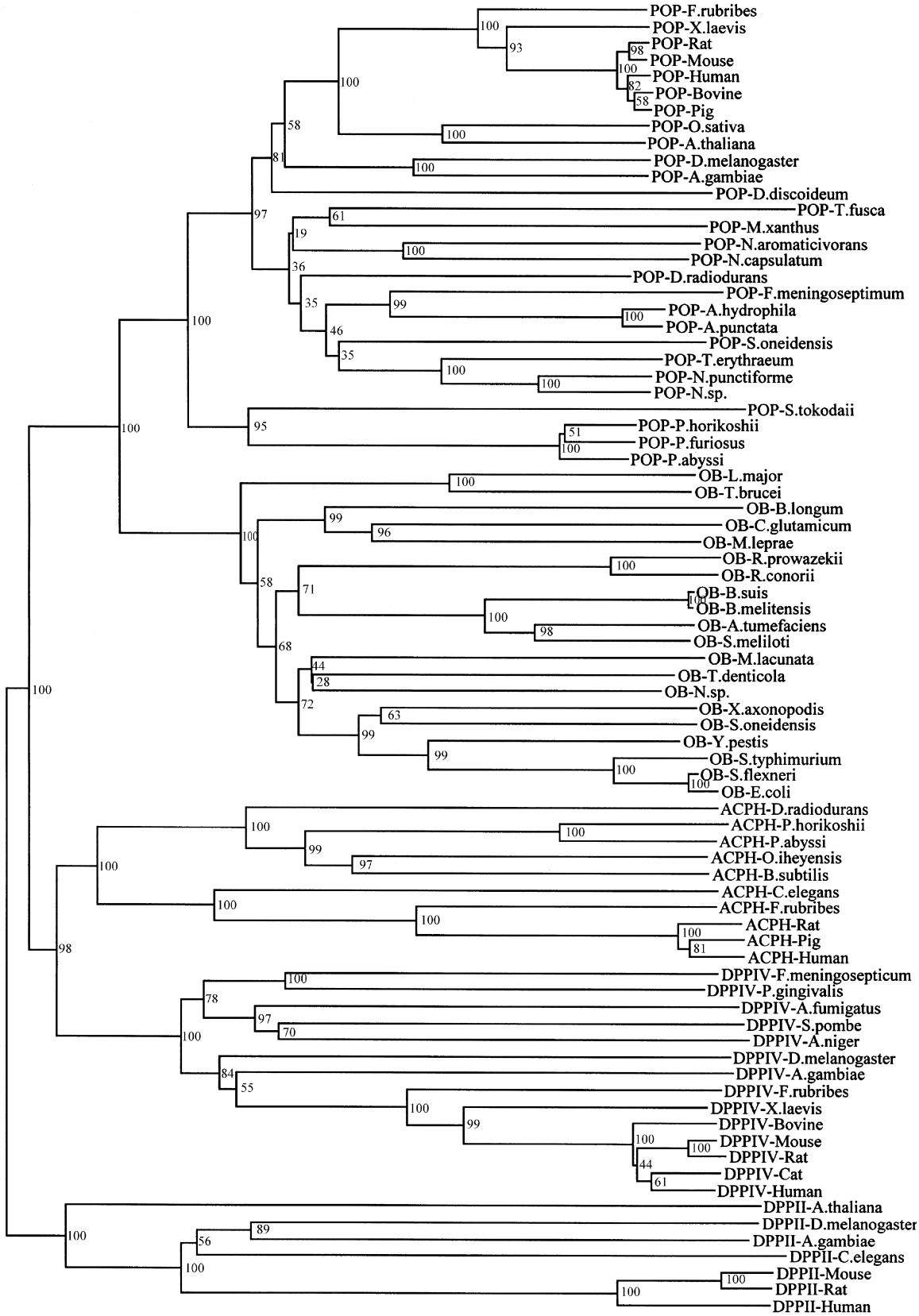


Fig. 3. The neighbor-joining tree of POP family enzymes. Protein sequences were aligned with T-COFFEE and CLUSTALX programs and the tree with bootstrap values was then constructed with CLUSTALX program. DPP II sequences were used as outgroups and numbers represent the percentages of 1000 bootsraps. The tree was then visualized with NJPLOT program.

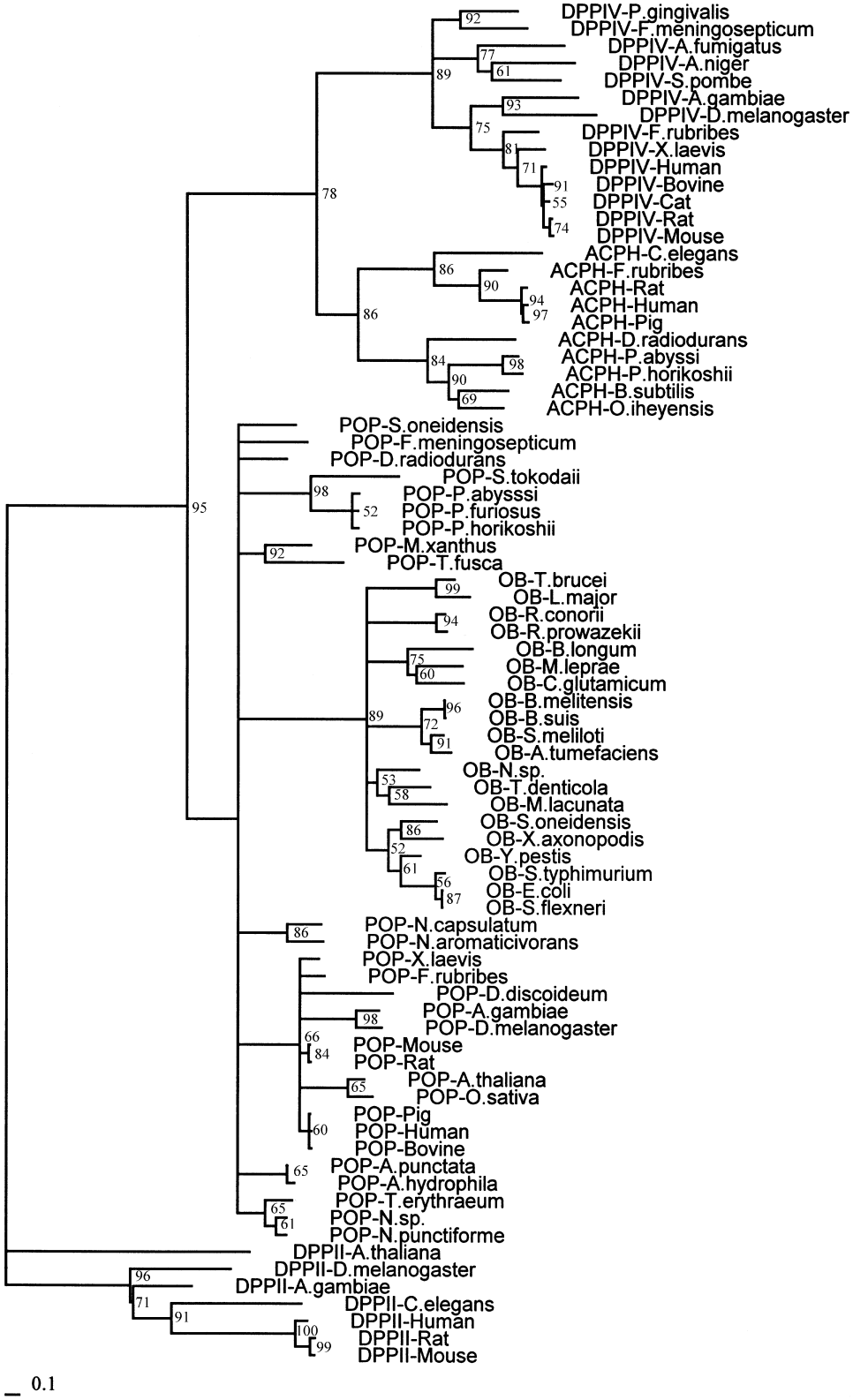


Fig. 4. The maximum likelihood tree of POP family enzymes. Protein sequences were aligned with T-COFFEE and CLUSTALX programs and the maximum likelihood tree with support values was calculated using TREE-PUZZLE version 5.0. DPP II sequences were used as outgroups and the tree was visualized with TREEVIEW program.

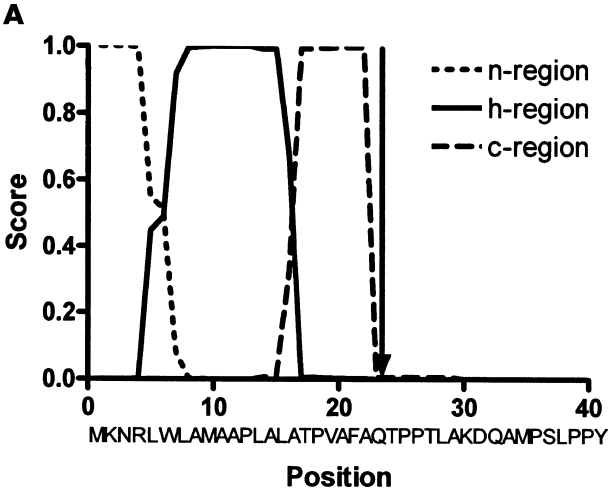
Table 3. Conservation percentages of the pig POP amino acids involved in substrate binding.

Location	Amino acid	Role	Identity/ similarity (%)
S1-Pocket	Ser554	Catalysis	100/100
	Asp641	Catalysis	100/100
	His680	Catalysis	100/100
	Trp595	Ring stacking	100/100
	Asn555	H-bond with S	100/100
	Tyr473	H-bond with S	100/100
	Phe476	Lining	100/100
	Val644	Lining	100/100
	Val580	Lining	92.9/100
	Tyr599	Lining	89.3/100
S2-Pocket	Arg643	H-bond with S	100/100
S3-Pocket	Trp595	H-bond with S	100/100
	Phe173	Lining	75.0/82.1
	Met235	Lining	28.6/32.1
	Cys255	Lining	42.9/42.9
	Ile591	Lining	71.4/78.6
	Ala594	Lining	57.1/57.1

a soluble protein so we believe that this enzyme is not membrane bound.

Proteins can also be membrane bound even if they do not possess a transmembrane sequence, if they contain a lipid anchor. In that case the protein is post-translationally modified with a glycosylphosphatidylinositol (GPI) moiety and anchored on the extracellular side of the plasma membrane [18]. The entry to the GPI-modification route is directed by a C-terminal sequence signal, consisting of about 20 amino acids. These signal sequences were searched with the BIG PI program. None of the eukaryotic and bacterial sequences possessed lipid anchor sequences, but archaeal POP enzymes *P. horikoshi*, *P. abyssi* and *P. furiosus* seemed to contain the signal sequence with false positive probabilities of 0.0147, 0.0172 and 0.0173, respectively. The predicted attachment sites of the GPI moiety were Ala594, Ala596 and Ala595 which all correspond to the Gly683 of pig POP. The search was carried out using the metazoa prediction function of the program and it is unclear whether the result is valid for archaeal sequences. However, GPI-linked proteins closely related to eukaryotes have also been found from archaeal sources [30], suggesting that the prediction may be correct. Naturally, this result will need to be verified experimentally, but to our knowledge, this is the first hint of a possible mechanism by which POP could be attached to the cell membrane.

Sequence analysis with the SIGNALP program resulted in the identification of four bacterial POP sequences that contain signal peptide sequences, i.e. the enzymes are secreted through the cell membrane. The POP forms are secreted from Gram negative bacterias *F. meningosepticum*, *N. capsulatum*, *Novosphingibium aromaticivorans* and *Shewanella oneidensis*. The calculated signal peptide probabilities of these enzymes varied from 0.971 to 1.000. The SIGNALP output of *N. capsulatum* POP is presented in Fig. 5A. The output contains n-, h- and c-region probabilities and the most likely cleavage site,



B

Flavobacterium meningosepticum MKYNKLSVAVAAFAFAAVSAQNSNVLKYPETKKVSHTDTY
Novosphingobium capsulatum MKNRLWLAMAAPLALATPVAFAQTPPTLAKDQAMPSLPPY
Novosphingobium aromaticovorans MATMVWEMDRKRRIAAWALLLLAGSRHAMSEEAALPAASV
Shewanella oneidensis MALRFRLAQQGLLVATLGMALAACQSGNTPEEAAAQQISA

Anopheles gambiae LQPFAVGAARVCAFSLPTSSGTSGKVCMHKLATSLLVRSC
Xenopus laevis SHLYLTAKRVLPVTVSARLLCHAGVFPRPSLCSASLSISG

0 10 20 30 40

Fig. 5. The secreted POP sequences. (A) The SIGNALP output of *Novosphingobium capsulatum* POP. Predicted n-, h- and c-regions are shown and the predicted cleavage site between residues 22 and 23 is shown with a downward arrow. (B) The amino acid sequences of secreted POP forms, the predicted cleavage sites are shown with underlined letters.

which is between residues 22 (alanine) and 23 (glutamine). The cleavage sites of *F. meningosepticum*, *N. aromaticivorans and S. oneidensis* signal peptides were predicted to be between residues 20–21 (alanine-glutamine), 30–31 (serine-glutamic acid) and 33–34 (alanine-alanine), respectively. The signal sequences and their potential cleavage sites are presented in Fig. 5B.

SIGNALP predicted correctly the *F. meninosepticum* POP signal peptide, as this enzyme has been shown experimentally to be periplasmic, the cleavage site of the signal peptide being between residues 20 (alanine) and 21 (glutamine) [31]. This correct prediction increases the reliability of SIGNALP results. The biological relevance of the periplasmic POP activity is not clear. However, secretion of POP in bacterial sources seems to be quite common, as four of the studied 12 bacterial sequences (33%) contained the signal sequence.

In addition to bacteria, secretion signal sequences were also found from eukaryotes *A. gambiae* and *Xenopus laevis* with probabilities of 0.905 and 0.808, respectively. The cleavage sites were predicted to be between residues 24–25 (glycine-lysine) and 34–35 (alanine-serine). To our knowledge, these are the first eukaryotic POP enzymes that are thought to be secreted out of the cell. It is interesting to note the difference of POP localization between the fruit fly *D. melanogaster* and the malaria transmitting mosquito *A. gambiae*. Despite the different localization and rather low sequence identity (58%), the POP proteins of *A. gambiae* and *D. melanogaster* are likely to have similar catalytic

properties because their amino acids involved in substrate binding (Table 3) are identical. *A. gambiae* has only one *POP* gene but *D. melanogaster* has an extra *POP*-like gene (NP_610129) in addition to the *POP* sequence used in this study (AAF52942). These proteins have sequence identity and similarity percentages of 60% and 73% and their substrate binding residues are identical with one important exception: the C-terminal part starting from Val660 has been deleted from NP_610129 and hence the third member of the catalytic triad (His680) is missing. It is probable that this protein is inactive or has a different function than POP and that the extra *POP*-like gene is a product of gene duplication in *D. melanogaster*.

A. gambiae and *D. melanogaster* belong to the same taxonomic order, but have different lifestyles. Due to blood feeding, *A. gambiae* is exposed to parasites such as *Plasmodium falciparum*, the human malaria parasite. *A. gambiae* efficiently combats the *P. falciparum* infection and therefore an understanding of the immune system of *A. gambiae* could be a very useful way to obtain clues to controlling malaria. This has been done by comparing the differences between immune-related genes of *A. gambiae and D. melanogaster* [32]. Interestingly, POP has been claimed to play a role in immunopathological processes associated with lupus erythematosus and rheumatoid arthritis [33]. Furthermore, several serine proteases have been shown to regulate invertebrate defense responses such as antimicrobial peptide synthesis [34]. Therefore, it is possible that the secreted POP might play a role in the immune responses of *A. gambiae*.

In summary, POP family enzymes were found to be of ancient origin, as they were already present in the last universal common ancestor of life. With respect to the studied enzymes of the POP family, POP seems to be the most conserved enzyme. Ten conserved amino acids were found at the active site of the enzyme of each of the studied POP family enzymes, indicating that those residues are probably critical to the enzyme function. In POP, the S1 specificity pocket was found to be highly conserved, compared to the more variable S3 specificity pocket. This finding may help to develop species-specific POP-inhibitors. Signal sequences were found in one third of bacterial POP sequences and also in two eukaryotic species. Lipid anchor sequences were found from three archaeal sources, indicating that the POP enzyme in these species is membrane bound.

Acknowledgements

This work was supported by National Technology Agency of Finland and Ministry of Education of Finland (to J. I. V.). We wish to thank Prof. Dan Larhammar, University of Uppsala, for his advice and extremely helpful comments on the manuscript and Dr Ewen MacDonald for linguistic advice.

References

1. Kanatani, A., Masuda, T., Shimoda, T., Misoka, F., Lin, X.S., Yoshimoto, T. & Tsuru, D. (1991) Protease II from *Escherichia coli*: sequencing and expression of the enzyme gene and characterization of the expressed enzyme. *J. Biochem. (Tokyo)* **110**, 315–320.
2. Rawlings, N.D., Polgár, L. & Barrett, A.J. (1991) A new family of serine-type peptidases related to prolyl oligopeptidase. *Biochem. J.* **279**, 907–908.
3. Polgár, L. (2002) The prolyl oligopeptidase family. *Cell. Mol. Life Sci.* **59**, 349–362.
4. Fülöp, V., Böcskei, Z. & Polgár, L. (1998) Prolyl oligopeptidase: an unusual beta-propeller domain regulates proteolysis. *Cell* **94**, 161–170.
5. Hiramatsu, H., Kyono, K., Higashiyama, Y., Fukushima, C., Shima, H., Sugiyama, S., Inaka, K., Yamamoto, A. & Shimizu, R. (2003) The structure and function of human dipeptidyl peptidase IV, possessing a unique eight-bladed beta-propeller fold. *Biochem. Biophys. Res. Commun.* **302**, 849–854.
6. O'Leary, R.M., Gallagher, S.P. & O'Connor, B. (1996) Purification and characterization of a novel membrane-bound form of prolyl endopeptidase from bovine brain. *Int. J. Biochem. Cell. Biol.* **28**, 441–449.
7. Yoshimoto, T., Kado, K., Matsubara, F., Koriyama, N., Kaneto, H. & Tsura, D. (1987) Specific inhibitors for prolyl endopeptidase and their anti-amnesic effect. *J. Pharmacobio-dyn.* **10**, 730–735.
8. Atack, J.R., Suman-Chauhan, N., Dawson, G. & Kulagowski, J.J. (1991) *In vitro* and *in vivo* inhibition of prolyl endopeptidase. *Eur. J. Pharmacol.* **205**, 157–163.
9. Marighetto, A., Touzani, K., Etchamendy, N., Torrea, C.C., De Nanteuil, G., Guez, D., Jaffard, R. & Morain, P. (2000) Further evidence for a dissociation between different forms of mnemonic expressions in a mouse model of age-related cognitive decline: effects of tacrine and S 17092, a novel prolyl endopeptidase inhibitor. *Learn. Mem.* **7**, 159–169.
10. Morty, R.E., Troeberg, L., Powers, J.C., Ono, S., Lonsdale-Eccles, J.D. & Coetzer, T.H. (2000) Characterisation of the antitrypanosomal activity of peptidyl alpha-aminoalkyl phosphonate diphenyl esters. *Biochem. Pharmacol.* **60**, 1497–1504.
11. Hughes, T.E., Mone, M.D., Russell, M.E., Weldon, S.C. & Villhauer, E.B. (1999) NVP-DPP728 (1-[[[2-[(5-cyanopyridin-2-yl) amino]ethyl]amino]acetyl]-2-cyano-(S)-pyrrolidine), a slow-binding inhibitor of dipeptidyl peptidase IV. *Biochemistry* **38**, 11597–11603.
12. Conarello, S.L., Li, Z., Ronan, J., Roy, R.S., Zhu, L., Jiang, G., Liu, F., Woods, J., Zycband, E., Moller, D.E., Thornberry, N.A. & Zhang, B.B. (2003) Mice lacking dipeptidyl peptidase IV are protected against obesity and insulin resistance. *Proc. Natl Acad. Sci. USA* **100**, 6825–6830.
13. Polgár, L. (1992) Structural relationship between lipases and peptidases of the prolyl oligopeptidase family. *FEBS Lett.* **311**, 281–284.
14. Notredame, C., Higgins, D.G. & Heringa, J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
15. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. (1997) The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.
16. Schmidt, H.A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.
17. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
18. Eisenhaber, F., Eisenhaber, B., Kubina, W., Maurer-Stroh, S., Neuberger, G., Schneider, G. & Wildpaner, M. (2003) Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Pi, NMT and PTS1. *Nucleic Acids Res.* **31**, 3631–3634.

19. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.
20. Barrett, A.J. & Rawlings, N.D. (1992) Oligopeptidases, and the emergence of the prolyl oligopeptidase family. *Biol. Chem. Hoppe–Seyler* **373**, 353–360.
21. Tateno, Y., Takezaki, N. & Nei, M. (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* **11**, 261–277.
22. Leitner, T., Escanilla, D., Franzen, C., Uhlen, M. & Albert, J. (1996) Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl Acad. Sci. USA* **93**, 10864–10869.
23. Rosenblum, J.S. & Kozarich, J.W. (2003) Prolyl peptidases: a serine protease subfamily with high potential for drug discovery. *Curr. Opin. Chem. Biol.* **7**, 496–504.
24. Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M., Mueller, H.-M., Dimopoulos, G., Law, J.H., Wells, M.A., Birney, E., Charlab, R., Halpern, A.L., Kokoza, E., Kraft, C.L., Lai, Z., Lewis, S., Louis, C., Barillas-Mury, C., Nusskern, D., Rubin, G.M., Salzberg, S.L., Sutton, G.G., Topalis, P., Wides, R., Wincker, P., Yandell, M., Collins, F.H., Ribeiro, J., Gelbart, W.M., Kafatos, F.C. & Bork, P. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159.
25. Harris, M.N., Madura, J.D., Ming, L.J. & Harwood, V.J. (2001) Kinetic and mechanistic studies of prolyl oligopeptidase from the hyperthermophile *Pyrococcus furiosus*. *J. Biol. Chem.* **276**, 19310–19317.
26. Vendeville, S., Goossens, F., Debreu-Fontaine, M.A., Landry, V., Davioud-Charvet, E., Grellier, P., Scharpe, S. & Sergheraert, C. (2002) Comparison of the inhibition of human and *Trypanosoma cruzi* prolyl endopeptidases. *Bioorg. Med. Chem.* **10**, 1719–1729.
27. Szeltner, Z., Renner, V. & Polgár, L. (2000) The noncatalytic beta-propeller domain of prolyl oligopeptidase enhances the catalytic capability of the peptidase domain. *J. Biol. Chem.* **275**, 15000–15005.
28. Möller, S., Croning, M.D. & Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653.
29. Hirokawa, T., Boon-Chieng, S. & Mitaku, S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378–379.
30. Kobayashi, T., Nishizaki, R. & Ikezawa, H. (1997) The presence of GPI-linked protein(s) in an archaeobacterium, *Sulfolobus acidocaldarius*, closely related to eukaryotes. *Biochim. Biophys. Acta* **1334**, 1–4.
31. Chevallier, S., Goeltz, P., Thibault, P., Banville, D. & Gagnon, J. (1992) Characterization of a prolyl endopeptidase from *Flavobacterium meningosepticum*. Complete sequence and localization of the active-site serine. *J. Biol. Chem.* **267**, 8192–8199.
32. Christophides, G.K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., Collins, F.H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N.T., Hoffmann, J.A., Kanzok, S.M., Letunic, I., Levashina, E.A., Loukeris, T.G., Lycett, G., Meister, S., Michel, K., Moita, L.F., Muller, H.-M., Osta, M.A., Paskewitz, S.M., Reichhart, J.-M., Rzhetsky, A., Troxler, L., Vernick, K.D., Vlachou, D., Volz, J., von Mering, C., Xu, J., Zheng, L., Bork, P. & Kafatos, F.C. (2002) Immunity-related genes and gene families in *Anopheles gambiae*. *Science* **298**, 159–165.
33. Cunningham, D.F. & O'Connor, B. (1997) Proline specific peptidases. *Biochim. Biophys. Acta* **1343**, 160–186.
34. Gorman, M.J. & Paskewitz, S.M. (2001) Serine proteases as mediators of mosquito immune responses. *Insect Biochem. Mol. Biol.* **31**, 257–262.

Supplementary material

The following material is available from http://blackwellpublishing.com/products/journals/suppmat/EJB/EJB4199/EJB4199sm.htm

Fig. S1. Multiple sequence alignment of studied POP family enzymes.