ELSEVIER

The whey acidic protein family: A new signature motif and three-dimensional structure by comparative modeling

Shoba Ranganathan,* Kaylene J. Simpson,†‡[1] Denis C. Shaw,§ and Kevin R. Nicholas†

**Australian Genomic Information Centre, University of Sydney, Sydney, New South Wales, Australia*
†Victorian Institute of Animal Science, Attwood, Victoria, Australia
‡School of Agricultural Sciences, La Trobe University, Bundoora, Victoria, Australia
§Protein Biochemistry Group, John Curtin School of Medical Research, Australian National University, Canberra, Australian Capital Territory, Australia

Whey acidic proteins (WAP) from the mouse, rat, rabbit, camel, and pig comprise two "four-disulfide core" domains. From a detailed analysis of all sequences containing this domain, we propose a new PROSITE motif ([KRHGVLN]-X-{PF}-X-[CF]-[PQSVLI]-X(9,19)-C-{P}-X-[DN]-X-{N}-[CE]-X(5)-C-C) to accurately identify new four-disulfide core proteins. A consensus model for the WAP proteins is proposed, based on the human mucous proteinase inhibitor crystal structure. This article presents a detailed atomic model for the two-domain porcine WAP sequence by comparative modeling. Surface electrostatic potential calculations indicate that the second domain of the pig WAP model is similar to the functional human mucous proteinase inhibitor domains, whereas the first domain may be nonfunctional. © 2000 by Elsevier Science Inc.

Keywords: WAP, molecular model, sequence motifs, sequence–structure comparison, molecular electrostatic potential, four-disulfide core

INTRODUCTION

The whey acidic protein (WAP) is the major whey protein in the milk of the mouse (mWAP),[1] rat (ratWAP),[2] rabbit (rabWAP),[3] and camel (cWAP),[4] and was recently identified as a significant component of porcine milk (pWAP).[5] The WAP proteins contain two four-disulfide core (4-DSC) domains, each comprising approximately 50 amino acids and that include eight cysteine residues in a conserved arrangement.[1] The two WAP domains show limited sequence identity both within and between species.[5] The 4-DSC domains are not exclusive to the WAP proteins, with numerous other proteins encoding one or two of these domains. These proteins are typically small, secretory proteins which exhibit a variety of growth- and differentiation-regulatory functions and have been shown to affect extracellular matrix remodeling and carcinoma.[6] A large biological diversity exists between the proteins that contain one or two 4-DSC domains, with many being identified as proteinase inhibitors. These proteins are grouped into families based on their functionality and tissue-specific origins and include the antileukoproteinase (ALKI) family,[7] epididymal[8] and ovulatory[9] specific proteins and elastase inhibitor (elafin) proteins.[10] Therefore, based on the limited sequence similarity with known proteinase inhibitors it has been postulated that WAP may be a proteinase inhibitor.[11] The antibiotic properties of equine neutrophil antibiotic peptide (eNAP-2 fragment[12]) and the growth-inhibitory nature of rat prostate stromal protein (ps20[13]) suggest that the 4-DSC domain is a preferential conformation for the stable folding and action of a class of protein inhibitors of varied function. Currently, no biological activity has been ascribed to the WAP proteins.

In this study, extensive sequence analysis on the 4-DSC family of proteins has been performed with a view to identifying the specific sequence motifs that uniquely constitute the 4-DSC domains and the effect of any missing conserved cysteine residues on the tertiary fold adopted by each domain. On the basis of this analysis, we report a new 4-DSC signature for PROSITE[14] analysis that detects all known domains and eliminates false positives detected by the current PROSITE motif.

The 4-DSC domains are characterized by the conservation of

Color Plates for this article are on pages 134–136.

Corresponding author: Shoba Ranganathan, Australian Genomic Information Centre, C80 ATP, University of Sydney, Sydney NSW 2006, Australia. Tel.: +612 9351 1870; fax: +612 9351 1878.

E-mail address: shoba@angis.org.au (S. Ranganathan)

[1]Present address: Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria 3052, Australia.

Journal of Molecular Graphics and Modelling 17, 106–113, 1999

655 Avenue of the Americas, New York, NY 10010

1093-3263/99/$–see front matter
PII S1093-3263(99)00023-6

sequence motifs and the interaction of specific charged residues involved in stabilizing the elafin fold, based on the experimental structural information available for the two-domain human mucous proteinase inhibitor hSLPI (SWISS-PROT entry ALK1_HUMAN; also known as MPI[15]) and for the one-domain R-elafin (hElaf; SWISS-PROT: ELAF_HUMAN).[16,17] Despite the limited sequence identity, the conserved factors are strongly indicative that the extended planar spiral of the elafin structure, pinned together by four disulfide bridges, is the preferred conformation of the 4-DSC domain.[15–17]

In this study a consensus three-dimensional structural model for the two-domain WAP proteins has been developed, with the modeling of a detailed atomic structure for the mature porcine WAP protein (pWAP) based on the crystal structure of the two-domain hSLPI.[15] The viability of the WAP sequences to adopt this structural model has been examined, even in the absence of one or two of the conserved cysteine residues in some WAP proteins. Domain II of the WAP proteins appears more conserved than domain I,[5] with the surface electrostatic potential of the pWAP domain II being similar to that of the corresponding hSLPI domain, suggesting an as yet undetermined inhibitory activity. Domain I of pWAP is more substituted and less rigid and may carry posttranslational modifications rendering it nonfunctional.

MATERIALS AND METHODS

Sequence Retrieval and Analysis

The 4-DSC domain sequences were retrieved from protein sequence databases at the Australian National Genomic Information Service (using WebANGIS[18]). PROSITE[14] analysis and pattern scans were carried out at the ExPASY web site (http://www.expasy.ch) against the SWISS-PROT (Release 37 and updates up to 01-April-1999 of the database) and TrEMBL databases. Putative serine- and threonine-linked O-glycosylation sites were located using the NetOGlyc[19] server.

Sequence Alignment

Alignment of all 4-DSC domains was primarily carried out to identify the sequence signature. The WAP sequences (after deletion of the signal peptide sequence) were split into domains and aligned with all available mature 4-DSC protein domain sequences, initially using CLUSTAL W.[20] However, because of the low sequence similarity, the alignment was poor, especially in the N-terminal half of each domain. Further alignments were carried out using MALIGN,[21] with the default scoring matrix derived from multiple-structure alignments. Since there is limited sequence identity among the 4-DSC domains, we consider the alignment of six of the eight conserved cysteine residues (excepting the second and the seventh cysteines, involved in the second disulfide bridge) essential to keep the elafin fold intact and critical to achieving a meaningful alignment. CLUSTAL W[20] results were satisfactory for the alignment of complete 4-DSC sequences. The alignments in each case were visually edited for the alignment of conserved and chemically similar residues.

Three-Dimensional Model Building

The porcine WAP structural model is based on the alignment of mature pWAP and hSLPI sequences. The program Modeller[22] was used to generate the pWAP structure, with constraints to enable the formation of four disulfide bridges within each domain. The initial model was iteratively refined by in-built molecular dynamics with simulated annealing protocols, to improve the structural quality as computed by PROCHECK[23] and the coordinates are available from the Protein Databank (code 1CJH). Electrostatic potentials on the molecular surface of pWAP and hSLPI were computed by the finite-difference Poisson–Boltzmann method, as implemented in GRASP[24] using the simple charge model.

RESULTS AND DISCUSSION

WAP Signature

An alignment of eighty-four 4-DSC domain sequences, derived from WAP and numerous proteins with either confirmed or putative proteinase inhibitory activity, shows the characteristic conservation of the cysteine residues that constitute each domain (Figure 1/Color Plate 1).[25]

Within each domain, the eight conserved cysteine residues (numbered sequentially from C_1 to C_8), are arranged in the following pattern:

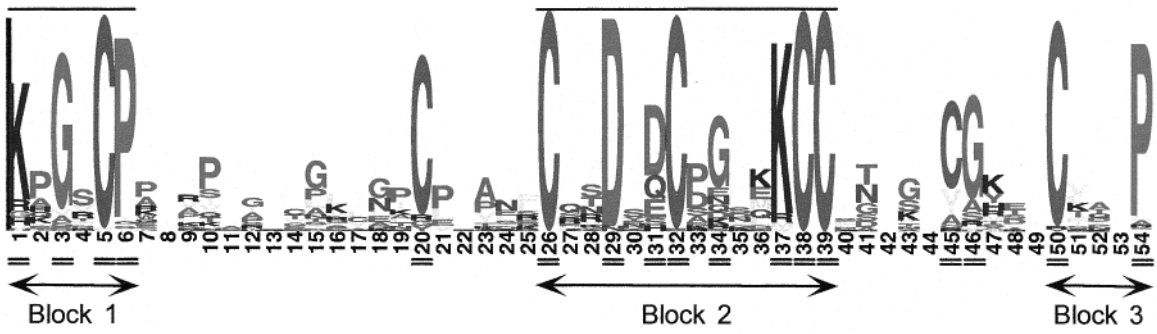
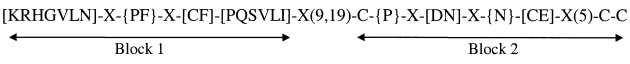


Figure 1/Color Plate 1. Schematic view of the 4-DSC domain alignment. Logo[25] representation of the 84 multiply aligned (using MALIGN[21]) 4-DSC domain sequences, with sequence numbering according to the alignment position and coloring by residue type: H, K, R in blue; D, E in red; N, Q in green; F, I, L, M, V in yellow; A, G, P, S, T, Y, W in magenta; and C in cyan. Horizontal bars indicate the alignment positions constituting the new 4-DSC domain family signature motif. A double line underscores positions showing at least 50% sequence conservation, with conserved sequence blocks labeled.

C_1-(X*n*)-C_2-(X*n*)-C_3-(X5)-C_4-(X5)-C_5-C_6-(X3, X5)-C_7-(X3, X4)-C_8

where X is any residue, X*n* is a stretch of *n* residues and (X*m*, X*n*) represents a variable length of *m*–*n* residues. There are other conserved residues that appear to contribute to a unique 4-DSC domain. The alignment is highly conserved between C_3 and C_8, while the position of C_2 appears variable. This variability in the location of C_2 is understood from the available three-dimensional structures of elafin[16,17] and hSLPI,[15] showing a large highly substituted (and functional) external loop between C_1 and C_3, with the rest of the protein packed into a tight β-turn, with little room for variations in the positions of C_4–C_8. To track the consequences of missing conserved cysteine residues, the disulfide bridges of the elafin fold[15–17] are labeled SS1 (between C_1 and C_6), SS2 (bridging C_2 and C_7), SS3 (linking C_3 and C_5), and SS4 (connecting C_4 and C_8). C_2 can thus occupy any position between C_1 and C_3, provided it can adopt a conformation suited to form SS2 with C_7. The exact location of C_2 in each domain would determine the shape and nature of the protruding loop, which in turn would affect the specificity and function of the domain.

Besides the conserved cysteine residues, several other residues are at least conserved in 50% of the sequences (doubly underlined in Figure 1/Color Plate 1). There are three blocks of conserved residues. The first two blocks, separated by a variable region containing C_2, define the new 4-DSC signature motif:



Block 3 [C-X(2,3)-P] starts at C_8 and extends to the very end of the 4-DSC domain, with a consensus sequence **C**XX**P** (the only exception being domain I of the human epididymal protein terminating in CXXXP). A variable region of between four and nine residues separates block 3 from the end of block 2.

The first block has the consensus sequence **KXGXCP**, containing C_1. A similar motif of seven conserved amino acids (**KAGRCPW**) was identified from the alignment of pWAP with other WAP sequences and previously defined as a WAP motif by Simpson et al.[5] The function of this **KAGRCPW** motif is unknown, but owing to the high degree of sequence conservation it was thought to be of potential biological significance.[5]

The second block of 14 residues extends from C_3 to C_6 (consensus sequence: **C**XX**D**X**DC**X**G**XX**KCC**), and is separated from block 1 by a variable length spacer region comprising 9 to 19 residues, which contains C_2. The first of the conserved aspartate (**D**) residues (located exactly midway between C_3 and C_4) and the **KCC** at the end of block 2 appear critical in the recognition of the 4-DSC domain.

Blocks 1 and 2, together with the spacer region, are essential in defining the 4-DSC domain signature and in eliminating the false positives (FPs) obtained from the current PROSITE signature (PS00317: C-X-{C}-[DN]-X(2)-C-X(5)-C-C). While PS00317 identifies 36 sequences containing a 4-DSC domain (including the four FPs) from the SWISS-PROT database, the new signature retrieves only the 32 true positives. The first FP corresponds to residues 236–249: CKENTYCMENGSCC in the putative molluscan insulin-related peptide receptor precursor from the great pond snail and lacks the characteristic eight conserved cysteine residues of the 4-DSC family. This sequence in contained in the furin-like cysteine-rich region of MIPR_LYMST in the Pfam database.[26] The second FP from the chicken vasotocin-neurophysin VT precursor (NEUV_CHICK, 92–105: CGSDGRCAANGVCC) is an integral part of the VT neurophysin domain (32–161) from the SWISSPROT feature table. The FPs from the rice oryzain B-chain precursor ORYB_ORYSA (385–398: CDDNFSCPAGSTCC) and the hypothetical *Caenorhabditis elegans* protein YMV2_CAEEL (422–435: CPPDFTCSLSGKCC) lack the block 1 motif as well as the characteristic eight conserved cysteine residues of the 4-DSC proteins.

Other FPs that are not listed by PROSITE are from the TrEMBL database, where PS00317 putatively identifies two 4-DSC domains in lustrin A, a multidomain protein from the shell mantle of Californian red abalone[27] (735–748: CPFNTVCYKGAVCC and 1340–1353: CPSNTYCKSPGICC). Both these sequences correspond to cysteine-rich domains[27] and not the C-terminal 4-DSC domain (defined by Shen et al.[27] as 1384–1414), which is correctly identified as 1382–1414: KPGSCPAVRPDWAGICVVRCFCDNDCRGNLKCC (with two additional residues at the start of the domain), by our signature motif. Thus, compared with the current PROSITE motif, the new signature proposed in this study is able to correctly locate all true positives, while selectively discarding the FPs.

On the basis of the alignment of 84 domain sequences, we propose that the start of a 4-DSC domain be defined from block 1 (the WAP motif, **KXGXCP**), or at least four residues N-terminal to C_1 of the conserved set of eight cysteines and extend up to the end of block 3 (**C**XX**P**), i.e., at least three residues C-terminal to C_8.

One and Two Four-Disulfide Core Domain Proteins

Sequences containing one or two 4-DSC domains were compared with the two-domain WAP proteins using CLUSTAL W.[20] A concise alignment, with all WAP sequences and representative proteins from different families (human epididymal secretory protein, HE4; pig sodium/potassium ATPase inhibitor, pSPAI; rat WDNM1 protein, rWDNM1; human Kallmann syndrome protein, hKall; red sea turtle chelonianin, IBPT; guinea pig caltrin-like protein, calu; purple sea urchin cortical granule protein, urchin) is shown in Figure 2. All sequences containing a single 4-DSC domain aligned with the start of domain II of the proteins containing two 4-DSC domains.

The **KXGXCP** motif is not conserved in domain I of the WAP proteins, along with the conserved lysine in the **KCC** motif at the end of block 2 of the WAP signature. mWAP domain I lacks both residues C_1 and C_8, leaving only two intact disulfide bridges (SS2 and SS3). Also, two cysteine residues are missing from rabWAP domain I, which are either C_2 or C_3, and C_7. From the structure of the elafin fold,[15–17] we suggest that the second disulfide bridge, SS2, has been lost by evolutionary changes in rabWAP, so that both C_2 and C_7 are missing, rather than C_3 and C_7, consistent with mWAP. Consequently, with no unpaired cysteine residues available, the linking of the two rabWAP domains by an interdomain disulfide bridge, as suggested by Baranyi et al.,[28] does not seem possible.

Among the two-domain proteins, domain II is more conserved than domain I, with several completely conserved res-

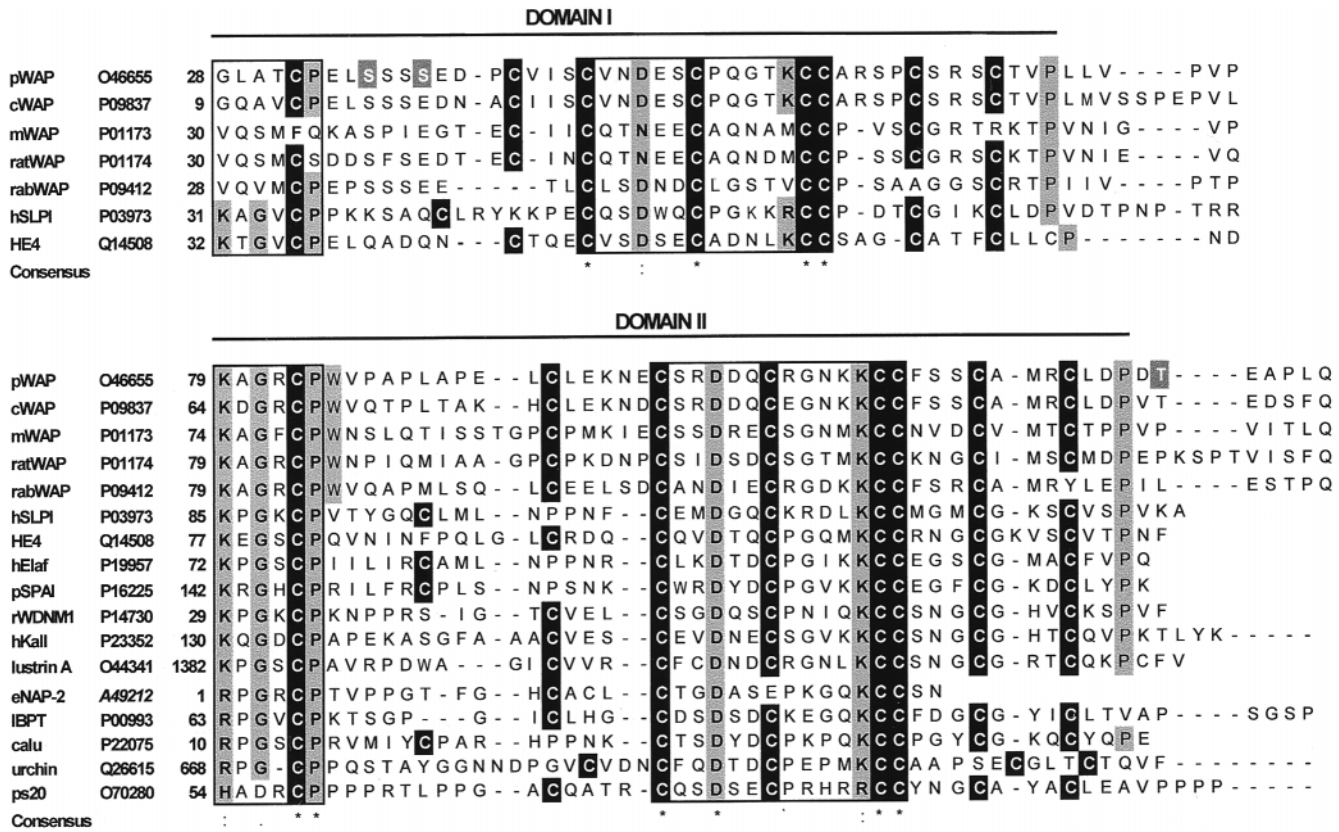


Figure 2. Multiple sequence alignment of one and two 4-DSC domain sequences. Shown are WAP and representative proteins from the CLUSTAL W[20] alignment of 31 sequences, with conserved (*) and conservatively substituted (:) residues in gray background and 4-DSC domain conserved cysteine residues shown in white type with a black background. Alignment positions constituting blocks 1 and 2 of the new 4-DSC family signature motif are shown within boxes. Databases accession numbers correspond to SWISS-PROT or TrEMBL (for the incomplete eNAP-2 sequence alone, PIR database code is given in italics). Residue numbers correspond to the beginning of each domain. Putative O-glycosylation[19] sites for pWAP are shown in white type against a gray background.

idues (Figure 1/Color Plate 1 and Figure 2). The block 1 consensus motif is conserved in all domain II sequences, with the conserved lysine conservatively substituted by arginine in four sequences and by histidine (and by aspartate for the conserved glycine) in ps20 (Figure 2). The **KCC** motif at the end of block 2 is conserved in all domain II sequences, excepting ps20 (with **RCC** instead). The eight conserved cysteines are present in all but two (rabWAP and the incomplete equine sequence eNAP-2) domain II sequences, satisfying the requirements for adopting the elafin fold. Domain II of rabWAP lacks C_8, leaving the C_4 residue unpaired and possibly available for an "interchain" rather than an interdomain[28] disulfide bridge. It is interesting to note that the protein domain database ProDom[29] identifies only domain II of both the WAPs and the epididymal family as "WAP domains" while excluding the two-domain trout antileukoproteinase precursor sequence (accession number: Q91450).

pWAP Model

The structural information available for the 4-DSC domain family[15–17] was analyzed as the first step to model building. The NMR solution structure of R-elafin[17] was not selected since X-ray data were available for both hSLPI[15] and human elafin (hElaf),[16] in accordance with template selection rules for comparative protein modeling.[30] hElaf has only one 4-DSC domain, which superimposes closely with domain II of hSLPI (overall RMS deviation of only 0.83 Å compared with 2.69 Å for domain I) with no insertions or deletions. The inclusion of the hElaf structure in the model building, therefore, does not provide any additional structural information[30] for domain II and has thus not been included in model building.

On the basis of the alignment of the mature porcine WAP sequence (less signal peptide; residues L20–Q132) with that of hSLPI (mature ALKI_HUMAN residues S26–A132) shown in Figure 2, a structural model for pWAP was built by comparative modeling, from the coordinates (residues in the structure being numbered S1–A107) of Grutter et al.[15] Of the pWAP N-terminal residues 20–27 (LAPALNLP), residues A23–P27 were matched with the hSLPI template residues 26–31 (SGKSF), without any gaps. Owing to the limited sequence identity between pWAP and hSLPI (31.6%), the model building was carried out in stages: each domain was modeled separately, and then the interconnecting loop between the two domains (which has three residues less than the corresponding

loop in hSLPI) was added. The model was subjected to refinement as described in Materials and Methods. The overall quality of the refined pWAP model was assessed as satisfactory by PROCHECK,[23] with 97.8% of the residues in allowed backbone conformations.

The eight disulfide bridges in pWAP are shown in Figure 3/Color Plate 2,[31] along with the limited secondary structural elements present in this model. On the basis of the pWAP structure, homology models for the camel and rat sequences (Figure 2: cWAP and ratWAP) have been generated, containing all four disulfide bridges in each domain. Domain II of mWAP and domain I of rabWAP (lacking SS2) have been modeled from the corresponding pWAP domains. Models for domain I of mWAP (lacking both SS1 and SS4) and domain II of rabWAP (missing SS4) reveal the onset of disorder in the elafin fold, when the critical disulfide bridge set (comprising SS1, SS3, and SS4) is incomplete.

Mapping the conserved sequence blocks defined in Figure 1/Color Plate 1 onto the structural model for pWAP (shown in Figure 3/Color Plate 2) identifies the outer loop, and the inner β-sheet and β-turn as the least conserved part of the structure. These conserved sequence blocks also serve to delineate the extent of the 4-DSC domain. Blocks 1, 2, and 3 are spatially localized, as they contain the conserved cysteine residues essential for stabilizing the elafin fold: these cysteines form the critical set of disulfide bridges (SS1, SS3, and SS4). The current PROSITE motif overlaps with block 2 (shown in magenta in Figure 3/Color Plate 2), while the new signature motif from this study extends from block 1 (shown in white in Figure 3/Color Plate 2) upto the end of block 2. It is interesting to note that the side chain of the tryptophan residue (W85 in Figure 3/Color Plate 2), conserved in all WAP domain II sequences, is solvent accessible and is probably involved in protein–protein interactions. The role of the conserved charged residues in stabilizing the model structure is discussed below in the light of hydrogen-bonding interactions.

A detailed analysis of possible hydrogen bonds[32] formed by pWAP is listed in Table 1. Both domains of pWAP retain the inner tight two-residue β-turn of hSLPI,[15] with five main-chain hydrogen bonds in each domain stabilizing the two-strand antiparallel β-sheet (residues K58–R62 and S66–T70 in domain I and K112–S116 and R119–L123 in domain II, shown as underlined in Table 1). The observed secondary structure in each domain comprises a β-hairpin (domain I: K58–T70 and domain II: K112–L123) and a five-residue 3,10-helical segment (domain I: N49–C53 and domain II: R103–C107, shown in boldface in Table 1) in the first half of block 2 and terminating with C_4. Apart from these, there are remarkably few backbone hydrogen bonds: only four in domain I and five in domain II, as was noted in the case of hSLPI.[15] Besides the four disulfide bridges formed in each domain, the pWAP fold derives additional stability from the main chain–side chain interactions, totaling 9 in domain I and 10 in domain II. Grutter et al.[15] noted that both domains in hSLPI contained a charged interaction between the conserved (first) aspartate and lysine residues in block 2, probably involved in stabilizing the β-hairpin. This interaction is observed only in domain II of pWAP (between D104 and K112) and not domain I, indicating that domain II is more compactly folded than domain I. The second charged interaction noted in each domain of hSLPI,[15] between the side chain of the positively charged residue in the block 1 (**K** of KXGXCP) and three or four backbone backbone oxygen atoms of block 2, is observed only in domain II of pWAP (interactions 31 and 32 of Table 1). Domain I has a glycine residue present at the start of block 1, instead of the consensus lysine residue, and thus has no side chain to interact with block 2 backbone oxygen atoms. While there is no functional significance for this second charged interaction, its pres-

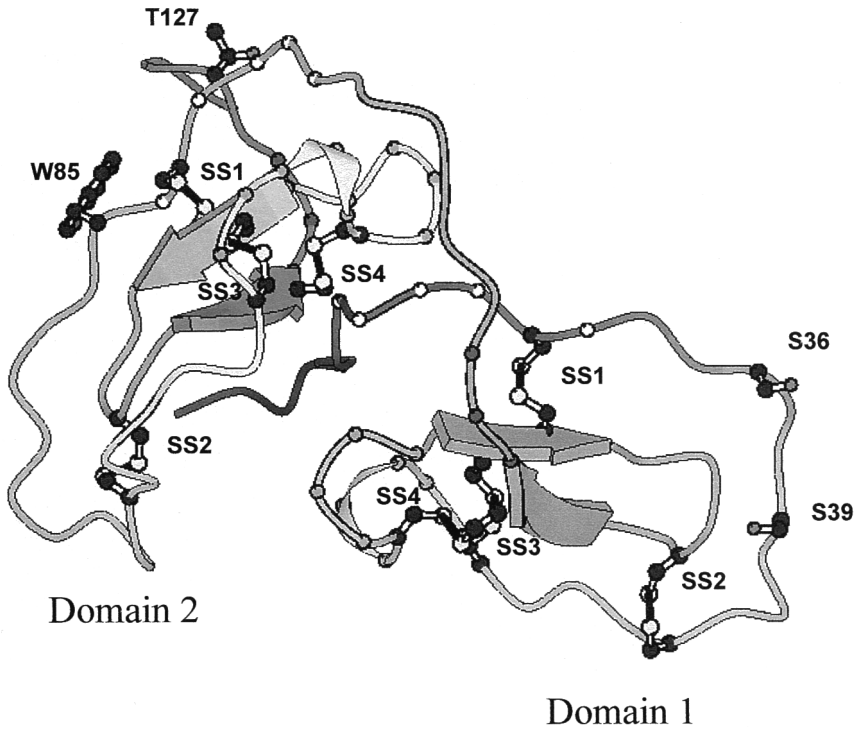


Figure 3/Color Plate 2. Structural model for pWAP. MOLSCRIPT[31] diagram, colored from blue (N terminus) to red (C terminus), with the two domains (as defined in Figure 2) labeled. Arrows represent β strands and spirals indicate helical segments. The four disulfide bridges (labeled SS1–SS4) in each domain are shown as black bars. Heavy atoms of the conserved cysteine residues, the tryptophan residue conserved in WAP domain II (W85 of pWAP), and putative O-glycosylation[19] sites (S36, S39, and T127) are shown in ball-and-stick representation and colored by atom type (oxygen, red; nitrogen, blue; sulfur, yellow; carbon, black). The C_α atoms of the conserved residue blocks (shown in Figure 1/Color Plate 1) are shown as small spheres: block 1, white; block 2, magenta; block 3, purple.

Table 1. Hydrogen bonding analysis of the pWAP model[a]

No.	Domain	Type	Donor	Acceptor	No.	Domain	Type	Donor	Acceptor
1	I	MM	S46:N	R67:O	24	II	MM	C114:N	R121:O
2	I	MM	**S52:N**	**N49:O**	25	II	MM	S116:N	A119:O
3	I	MM	**C53:N**	**D50:O**	26	II	MM	R121:N	C114:O
4	I	MM	K58:N	T70:O	27	II	MM	L123:N	K112:O
5	I	MM	C60:N	S68:O	28	II	MR	E100:N	E100:OE1
6	I	MM	R62:N	S66:O	29	II	MR	N110:N	N110:ODI
7	I	MM	S66:N	C43:O	30	II	MR	C113:N	D104:OD1
8	I	MM	S68:N	C60:O	31	II	RM	K79:NZ	G99:O
9	I	MM	T70:N	K58:O	32	II	RM	K79:NZ	K111:O
10	I	MR	S36:N	S63:OG	33	II	RM	N99:ND2	N99:O
11	I	MR	E40:N	E40:OE1	34	II	RM	R103:NH1	S102:O
12	I	MR	S52:N	N49:OD1	35	II	RM	D104:OD1	C113:O
13	I	MR	C59:N	D50:OD1	36	II	RR	D104:OD2	K112:NZ
14	I	RM	S39:OG	S39:O	37	II	RR	R121:NE	D104:OD1
15	I	RM	S39:OG	S63:O	38	Inter	MM	***L26:N***	***C122:O***
16	I	RM	S52:OG	S52:O	39	Inter	MM	T31:N	L73:O
17	I	RM	T57:OG1	P54:O	40	Inter	MM	L73:N	T31:O
18	I	RM	T70:OG1	V71:O	41	Inter	MM	R82:N	D126:O
19	I	RR	K58:NZ	C32:SG	42	Inter	MM	D126:N	R82:O
20	II	MM	E100:N	M120:O	43	Inter	MM	***C122:N***	***L26:O***
21	II	MM	**Q106:N**	**R102:O**	44	Inter	RM	T31:OG1	L73:O
22	II	MM	**C107:N**	**D104:O**	45	Inter	RR	***K98:NZ***	***Q55:OE1***
23	II	MM	K112:N	L123:O					

[a] Hydrogen bonds classified by domain (I, II: see Figure 2) or interdomain (Inter) and as MM (main chain-to-main chain), MS (main chain-to-side chain), RM (side chain-to-main chain), or RR (side chain–side chain). Donor and acceptor atoms for each hydrogen bond are specified by residue type and number, followed by the atom type. β-hairpin segments are underlined; 3,10-helical sections are in boldface and interdomain links are in boldface italic type.

ence could be correlated with a functional role for the corresponding domain, since all 4-DSC domain sequences possessing the consensus block 1 motif are potentially capable of this side chain-to-main chain interaction, and are functional (WAP sequences excepted).

Besides the intradomain interactions, four of the main chain (Table 1, interactions 39–42) and the only side chain-to-main chain (Table 1, interaction 44) hydrogen-bonding interactions involve acceptor residues within a few amino acids of the domain ends for pWAP (see Figure 2) and are thus not strictly interactions between domain I and domain II. Interactions 39, 40, and 44 (between T31 and L73) add additional stability to domain I while interactions 41 and 42 (between R82 and D126) stabilize domain II. The interactions between domains I and II (shown in boldface italic in Table 1) are the two main-chain hydrogen bonds between L26 and C122 (corresponding to the F30–C126 interaction in hSLPI) and a side-chain interaction between K98 and Q55. The two pWAP domains thus have little interaction and would be capable of independent function, as has been observed[15] for the two hSLPI domains.

A structural overlay of pWAP and hSLPI is presented in Color Plate 3, with parts a and c showing "top" views and parts b and d showing "side" views focusing on domains I and II, respectively. Between the domains, from C_8 of domain I to C_1 of domain II, pWAP has 10 residues compared with 13 residues in hSLPI. Despite this difference, the two structures superimpose reasonably well, with an RMS deviation of 1.30 Å over 85 of the 125 C_α atoms of pWAP (68%). The two pWAP domains are oriented 153° to each other, with a translation of 16.4 Å (based on May and Johnson's method for protein structure comparison),[33] which is similar to the disposition of the two hSLPI domains (about 150° to each other, with a translation of about 16 Å).[15] While SS1, SS3, and SS4 of both domains of pWAP and hSLPI overlap closely, the second disulfide bridge (SS2) of each pWAP domain is not located at the top of the outer loop as in hSLPI, but to one side of it (shown by arrows in Color Plate 3a and c).

Domains I and II of hSLPI show antitryptic and antichymotryptic activity, respectively.[15] Grutter et al.[15] suggest that the scissile bond of domain I is R45–Y46 (corresponding to R20–Y21 of the structure). The residues of pWAP domain I, located similarly at the tip of the outer loop of the elafin fold, are D41 and P42 (shown in Color Plate 3c). The presence of the negatively charged residue in the place of R45 dramatically alters the surface electrostatic potential of the outer loop of domain I of pWAP *vis-à-vis* that of hSLPI (Color Plate 3e), suggesting a possibility of loss of function.

In the antichymotrypsin domain II of hSLPI, the scissile bond has been identified[15] to be L97–M98 (L72–M73 of the structure) sequence. The residues of pWAP domain II that are spatially similarly located on the outer loop of the fold are L90 and A91 (Color Plate 3d). The surface electrostatic potential of the outer loop of the two domains is thus similar and hydrophobic (Color Plate 3f), suggesting a possible functional role for domain II of pWAP. In the absence of any experimental

proteinase inhibitory activity,[5] the exact functional nature of this domain is still under study.

The pWAP protein appears to be glycosylated from molecular weight considerations.[5] There are no N-glycosylation sites,[5] but there are two potential O-glycosylation[19] sites at amino acid positions S39 and T127 (shown in the pWAP sequence in Figure 2 and in the model in Figure 3/Color Plate 2). T127 is located close to the C terminus of pWAP, beyond C_8 of domain II, and its side chain is exposed to the solvent. Thus it is likely that this residue is glycosylated.[19] S39 is positioned on the outer loop of domain I with its side chain facing the inner β hairpin (Figure 3/Color Plate 2), in its energetically most favorable conformation. This side chain is exposed in other less favorable conformations and it is therefore possible that the hydroxyl group of S39 becomes solvent exposed under glycosylating conditions. Another likely glycosylation position, which falls just short of the threshold value,[19] is S36, whose side chain is exposed to solvent. The outer loop of pWAP is more flexible than the loop in hSLPI,[15] since the disulfide bridge SS2 pinning the two loops together is not in a central position, but displaced to one side (Color Plate 3a and c) and hence can adopt slightly different backbone conformations, enabling either S36 or S39 to be glycosylated. Van den Steen et al.[34] have shown that O-linked oligosaccharide chains occur in regions containing clustered serine residues. The occurrence of such a four-residue cluster (S36–S39) in domain I of pWAP and similar serine clusters (Figure 2, between C_1 and C_2) in other WAP domain I sequences suggests a possible O-glycosylation site in the outer loop of this domain. Experimental determination of the glycosylation positions of pWAP is currently under investigation.

CONCLUSIONS

A new four-disulfide core domain signature motif has been proposed to correctly identify new members of this "WAP-type"[13] family. We have also suggested, on the basis of consensus sequence patterns, that this domain be defined from the KXGXCP motif (including the first conserved cysteine residue) to the CXXP motif (involving the last or eighth conserved cysteine residue), to avoid ambiguous domain extents defined in the literature. The two-domain WAP proteins can adopt the structure of hSLPI, with only minor variations in the loop conformations, but with the second disulfide bridge displaced in both domains. In the case of mWAP domain I, two disulfide bridges (the first and the last) cannot form, leading to possible disorder in the fold. RabWAP domain I is possibly missing only the second disulfide bridge: in the absence of C_7, it is likely that C_2 and not C_3 is also missing and the exterior loop is thus mobile. Domain II of rabWAP lacks only the last disulfide bridge, but is possibly disordered because of disruption of the covalent network essential for holding the inner and outer loops of the elafin fold together.[15] The detailed atomic model for pWAP reveals that the absence of the KXGXCP motif in domain I leads to the loss of a specific charged interaction, involved in stabilizing the overall fold and with possible functional implications. The electrostatic nature of the protruding exterior loop in domain I and the location of putative O-glycosylation sites in this segment of the protein chain may account for its lack of putative antitryptic behavior, while that of domain II and the lack of any putative O-glycosylation sites suggest that this domain is possibly functional. On the basis of the lack of sequence similarity in the antiproteinase region of other 4-DSC proteins with known function, WAP appears likely to be targeting proteinases specific to the mammary gland or gut of the young.

ACKNOWLEDGMENTS

We thank Prof. Wolfram Bode (Max-Planck Institut fur Biochimie, Martinsreid, Germany) for kindly providing us the coordinates of hSLPI. This work was supported by the award of a SPIRT grant (C09804978) from the Australian Research Council to AGIC, with industry partners SGI Australia Pty. Ltd. and MSI Australia Pty. Ltd.

REFERENCES

1 Hennighausen, L.H., and Sippel, A.E. Mouse whey acidic protein is a novel member of the family of "four-disulphide core" proteins. *Nucleic Acids Res.* 1982, **10,** 2677–2684

2 Campbell, S.M., Rosen, J.M., Hennighausen, L.G., Strech-Jurk, U., and Sippel, A.E. Comparison of the whey acidic protein genes of the rat and mouse. *Nucleic Acids Res.* 1984, **12,** 8685–8697

3 Devinoy, E., Hubert, C., Jolivet, G., Thepot, D., Clergue, N., Desaleux, M., Dion, M., Servely, J.L., and Houdebine, L.M. Recent data on the structure of rabbit milk protein genes and on the mechanism of the hormonal control of their expression. *Reprod. Nutrition Dev.* 1988, **28,** 1145–1164

4 Beg, O.U., Von Bahr-Lindstrom, H., Zaidi, Z.H., and Jornvall, H. A camel milk whey protein rich in half-cysteine: Primary structure, assessment of variations, internal repeat patterns, and relationships with neurophysin and other active polypeptides. *Eur. J. Biochem.* 1986, **159,** 195–201

5 Simpson, K.J., Bird, P., Shaw, D., and Nicholas, K. Molecular Characterisation and hormone-dependent expression of the porcine whey acidic protein gene. *J. Mol. Endocrinol.* 1998, **20,** 27–34

6 Dear, T.N., and Kefford, R.F. The WDNM1 gene product is a novel member of the "four-disulfide core" family of proteins. *Biochem. Biophys. Res. Commun.* 1991, **176,** 247–254

7 Heinzel, R., Appelhans, H., Gassen, G., Seemuller, U., Machleidt, W., Fritz, H., and Steffens, G. Molecular cloning and expression of cDNA for human antileukoproteinase from cervix uterus. *Eur. J. Biochem.* 1986, **160,** 61–67

8 Kirchhoff, C., Habben, I., Ivell, R., and Krull, N. A major human epididymis-specific cDNA encodes a protein with sequence homology to extracellular proteinase inhibitors. *Biol. Reprod.* 1991, **45,** 350–357

9 Garczynski, M.A., and Goetz, F.W. Molecular characterisation of a ribonucleic acid transcript that is highly up-regulated at the time of ovulation in the brook trout (*Salvelinus fontinalis*) ovary. *Biol. Reprod.* 1997, **57,** 856–864

10 Wideow, O., Schroder, J.M., Gregory, H., Young, J.A., and Christophers, E. Elafin: An elastase-specific inhibitor of human skin. Purification, characterisation and complete amino acid sequence. *J. Biol. Chem.* 1990, **265,** 14791–14795

11 McKnight, R.A., Burdon, T., Pursel, V.G., Shamany, A.,

Wall, R.J., and Hennighausen, L. The whey acidic protein. In: *Genes, Oncogenes, and Hormones: Advances in Cellular and Molecular Biology of Breast Cancer* (Dickson, R.E., and Lippman, M.E., eds.). Kluwer Academic, Boston, 1991, pp. 399–412
12 Couto, M.A., Harwig, S.S.L., and Lehrer, R.I. Selective inhibition of microbial serine proteases by eNAP-2, an antimicrobial peptide from equine neutrophils. *Infect. Immun.* 1993, **61,** 2991–2994
13 Larsen, M., Ressler, S.J., Lu, B., Gerdes, M.J., McBrides, L., Dang, T.D., and Rowley, D.R. Molecular cloning and expression of ps20 growth inhibitor: A novel WAP-type "four-disulfide core" domain protein expressed in smooth muscle. *J. Biol. Chem.* 1998, **273,** 4574–4584
14 Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res.* 1999, **27,** 215–219
15 Grutter, M.G., Fendrich, G., Huber, R., and Bode, W. The 2.5 Å X-ray crystal structure of the acid-stable proteinase inhibitor from human mucous secretions analysed in its complex with bovine alpha-chymotrypsin. *EMBO J.* 1988, **7,** 345–351
16 Tsunemi, M., Matsuura, Y., Sakakibara, S., and Katsube, Y. Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 Å resolution. *Biochemistry* 1996, **35,** 11570–11576
17 Francart, C., Dauchez, M., Alix, A.J.P., and Lippens, G. Solution structure of R-elafin, a specific inhibitor of elastase. *J. Mol. Biol.* 1997, **268,** 666–677
18 Littlejohn, T.G., Bucholtz, C.A., Campbell, R.M.M., Gaeta, B.A., Huynh, C., and Kim, S.H. Computing for biotechnology—WebANGIS. *Australasian Biotech.* 1996, **6,** 211–217
19 Hansen, J.E., Lund, O., Tolstrup, N., Gooley, A.A., Willaims, K.L., and Brunak, S. NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj. J.* 1998, **15,** 115–130
20 Thompson, J.D., Higgins, D.G., and Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994, **22,** 4673–4680
21 Johnson, M.S., Overington, J.P., and Blundell, T.L. Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.* 1993, **231,** 735–752
22 Sali, A., and Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 1993, **234,** 779–815
23 Laskowski, R.A., McArthur, M.W., Moss, D.S., and Thornton, J.M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 1993, **26,** 283–291
24 Nicholls, A., Sharp, K.A., and Honig, B. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* 1991, **11,** 281–296
25 Schneider, T.D., and Stephens, R.M. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* 1990, **18,** 6097–6100
26 Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* 1999, **27,** 260–262
27 Shen, X., Belcher, A.M., Hansma, P.K., Stucky, G.C., and Morse, D.E. Molecular cloning and characterisation of lustrin A, a matrix protein from shell and pearl nacre of *Haliotis rufescens. J. Biol. Chem.* 1997, **272,** 32472–32481
28 Baranyi, M., Brignon, G., Anglade, P., and Ribedeaudumas, B. New data on the proteins of rabbit (*Oryctolagus cuniculus*) milk. *Comp. Biochem. Physiol. B* 1995, **113,** 407–415
29 Corpet, F., Gouzy, J., and Kahn, D. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.* 1999, **27,** 263–267
30 Sali, A., and Overington, J.P. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* 1994, **3,** 1582–1596
31 Kraulis, P.J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 1991, **24,** 946–950
32 Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graphics* 1990, **8,** 52–56
33 May, A.C.W., and Johnson, M.S. Improved genetic algorithm-based protein structure comparisons: Pairwise and multiple superpositions. *Protein Eng.* 1995, **8,** 873–882
34 Van den Steen, P., Rudd, P.M., Proost, P., Martens, E., Paemen, L., Kuster, B., van Damme, J., Dwek, R.A., and Opdenakker, G. Oligosaccharides of recombinant mouse gelatinase B variants. *Biochim. Biophys. Acta* 1998, **1425,** 587–598