**Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins**

Isabelle Callebaut¹,*, Jean-Paul Moron¹ and Philippe Monget²

¹Département de Biologie Structurale, IMPMC UMR 7590, Universités Pierre et Marie Curie-Paris6 et Denis Diderot-Paris7, CNRS, Campus Boucicaut, 140 rue de Lourmel, 75015 Paris and ²INRA Physiologie de la Reproduction et des Comportements, UMR6175 INRA – CNRS - Université de Tours - Haras Nationaux, 37380 Nouzilly, France

Received on February 12, 2007; revised on May 9, 2007; accepted on May 10, 2007

Advance Access publication May 17, 2007

Associate Editor: Alex Bateman

**ABSTRACT**

Zona Pellucida (ZP) domains have been found in a wide variety of extracellular proteins, in which they play essential role for polymerization. They are shared by the ZP proteins, which constitute the extracellular coat of animal eggs. Except from ZP3, constituting the primary sperm receptor, the ZP proteins possess, in addition to their C-terminal ZP domains, N-terminal extensions, which are thought to play an important role in the species-specific gamete recognition. Here, we show that these extensions are made of single or multiple copies of a small globular domain, which can be significantly related to the N-terminal region of ZP domains (ZP-N domains). This finding brings new insights into the molecular evolution of ZP proteins, which may have evolved around a common ZP-N architecture, and more generally into the noticeable sequence diversity of ZP-N domains, which can be found as isolated subunits or tightly associated with ZP-C domains to form complete, canonical ZP domains.

**Contact:** isabelle.callebaut@impmc.jussieu.fr

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The extracellular coat surrounding animal eggs, which is called, in mammals, the Zona Pellucida (ZP), is composed of a few proteins (*the ZP proteins*) sharing a common domain (*the ZP domain*) (Bork and Sander, 1992). This domain has also been found in multicellular eukaryotic proteins from various tissues and with a wide variety of functions, from purely structural components to mechanotransducers (Bork and Sander, 1992; Jovine *et al.*., 2005). Proteins of the ZP family are characterized by various overall architectures, containing other types of domains, such as trefoil [ZP protein 1 (ZP1)], CUB (CPR-ductin gene products), EGF/EGF-like (Tamm-Horsfall protein), but the C-terminal location of the ZP domain appears to be a common feature of most of them (Jovine *et al.*, 2005). ZP domains are essential for polymerization of extracellular proteins (Jovine *et al.*, 2002), while other specific functions can generally be ascribed to sequences N-terminal to the ZP domain.

The ZP domain consists in approximately 260 amino acids, with 8 conserved cysteines participating in intramolecular disulfide bridges. Although its 3D structure has not yet been solved, the ZP domain is known to form a bipartite structure, the two halves being linked by a protease sensitive region. As characterized in mouse ZP proteins, the N-terminal subdomain (ZP-N) contains four conserved cysteines, which form disulfide bonds with an invariant 1-4/2-3 connectivity, while differences in connectivity are observed in the C-terminal subdomain (ZP-C) between ZP1/2-like and ZP3-like proteins, the latter ones possessing two additional cysteines (Boja *et al.*, 2003). It has recently been shown that proteins, such as the mammalian placenta-specific protein PLAC1, oocyte-secreted proteins Oosp1-3 and Drosophila Papillote share similarity with the ZP-N sub-domain (also named therefore the PLAC1 homology region), but not with the ZP-C sub-domain (reviewed in Jovine *et al.*, 2006). This suggests that the ZP-N sub-domain could form an independent structural domain, fulfilling a specific function. Jovine *et al.* (2006) recently supported this hypothesis by showing that the ZP-N sub-domain of mouse ZP3 folds and assembles into filaments. A recent analysis of endoglin, using single-particle electron microscopy, suggested that its ZP-N sub-domain would be involved in an extended region of interaction between the two subunits of homodimers (Llorca *et al.*, 2007).

In this study, we analysed further the sequences of the ZP proteins, which play a key role during oogenesis, fertilization and early embryogenesis (Wassarman *et al.*, 2001). With the exception of ZP3, ZP proteins contain additional sequences upstream the ZP domains, which are less conserved among orthologous proteins from different species, suggesting that these regions could play a role in the species-specific gamete recognition (Tian *et al.*, 1999). Using sequence profile searches (Altschul *et al.*, 1997) combined with the sensitive hydrophobic cluster analysis (HCA) approach (Callebaut *et al.*, 1997; Eudes *et al.*, 2007; Gaboriaud *et al.*, 1987), we found that these N-terminal extensions are exclusively made of divergent copies of ZP-N domains, which are repeated in the high molecular weight ZP proteins, such as ZP2 and ZPAX.

*To whom correspondence should be addressed.*

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
2 RESULTS

2.1 A common, repeated domain in the N-termini of ZP1, ZP2, ZP4 and ZPAX proteins

We first focused our investigations on sequences preceding the two known domains of human ZP1 [trefoil domain (aa 234–274); ZP domain: (aa 279–549) — CDD database (Marchler-Bauer et al., 2006)]. A globular domain, including five cysteine and rich in C12-strands, can clearly be predicted using HCA, lying between the signal peptide (aa 1–25) and a hinge region (aa 166–233) (see Supplementary Fig. 1). We thus searched the non-redundant database (nr; 4,198,544 sequences) at NCBI using PSI-BLAST (Altschul et al., 1997) (default parameters) and the corresponding sequence (aa 16–165) as query. The second iteration revealed that this domain is also present in ZP4 proteins (25% identity over 110 amino acids with human ZP4), conserving four out of the five cysteine residues. PSI-BLAST converged by iteration 4, leaving with E-values just above the threshold, alignments with other members of the ZP family of proteins, among which with a N-terminal region of the *Xenopus laevis* egg envelope ZPAX (Lindsay et al., 2001), a high molecular weight homolog of the ZPA (ZP2) protein (115 aa, 18% identity over 115 aa; E-value = 1.1). Note that a ZPAX gene apparently also exists in chicken, but not in mouse and human (Smith et al., 2005).

We integrated this alignment, which show similar hydrophobic clusters and conserve the four invariant cysteines (Fig. 1 and Supplementary Fig. 1), within the pool of significant hits for PSSM construction, and recovered, as iterations progressed, similarities with the N-terminal region of ZP2 proteins with significant E-values (e.g. with human ZP2, iteration 8, E = 3 × 10⁻⁶, 15% identity over 133 amino acids). At convergence by iteration 14, the N-terminal regions of ZP1, ZP2 and ZP4 were thus found to share significant similarities. The hypothesis of a common N-terminal domain was confirmed by reciprocal searches using the corresponding region of some of these proteins, as they consistently recovered ZP1, ZP2 and ZP4 sequences with significant E-values, with no false positives.

For example, the iterative sequence profile search conducted with the *Xenopus tropicalis* ZPAX N-terminal sequence (Q8AWX1; aa 1–140) led to detect with significant E-values ZP2, ZP4 and ZP1 proteins by iteration 5. Comparison of the HCA plots of the conserved N-terminal domains of ZP proteins, which allows to appreciate the conservation of regular secondary structures centered on hydrophobic clusters, and the deduced sequence alignments are shown in the Supplementary Figure 1 and Figure 1, respectively.

When the N-terminal domain of ZP1 was aligned with the N-terminal domain of ZP2 proteins, a second alignment was often observed with a ZP2 region located just downstream,
suggesting that in ZP2, the concerned domain would be duplicated. This second alignment was often quoted with significant E-values by PSI-BLAST convergence [e.g. starting with the X.tropicalis ZPAX N-terminal domain, significant E-values were observed with the X.tropicalis ZPA N-terminal domain, as well as with a second globular domain of the same ZPA protein (aa 20–102; \( E = 2 \times 10^{-6} \) and aa 122–249; \( E = 2 \times 10^{-15} \)]. These last alignments are always characterized by the conservation of hydrophobic positions (Fig. 1). However, only two of the four cysteines are conserved. The simultaneous absence of the second and third cysteines, and presence of the first and fourth ones suggests that these residues are involved in disulfide bonding with a 1–4/2–3 connectivity, similarly to what is observed in ZP-N sub-domains (see above).

Further investigations led to identify additional copies of the conserved N-terminal domain in ZP2 and in the high molecular weight ZPA proteins (total of three and five copies, respectively). Indeed, two shifted significant alignments were observed between the N-terminal extension of X.tropicalis ZPAX (aa 1–610) and human ZP2. The first one (\( E = 6 \times 10^{-24} \), iteration 8) included the two copies of the conserved N-terminal domain, as well as an additional downstream region, which is aligned in the second alignment (\( E = 4 \times 10^{-77} \)) with the second copy of the conserved N-terminal domain, thereby evidencing that the N-terminus of ZPAX, as well as that of ZP2, are made of three consecutive copies of a same domain. Two additional copies of this domain can be found in ZPA. The third ZP2 copy is indeed aligned with a fourth region in ZPAX, whereas a fifth copy can be highlighted in ZPAX, as the corresponding region was significantly aligned with established copies, through other PSI-BLAST searches (e.g. alignment with the second copy of X.laevis MGC86301). Alignments of these additional domains, performed using HCA on the basis of PSI-BLAST results, are also presented in Figure 1. Although the pattern of hydrophobic residues is well conserved among the whole family, the third and fourth copies of ZPAX do not maintain any of the four cysteines, whereas the fifth ZPAX copy contains two of the four cysteines, as the second ZP2 and ZPAX copies.

### 2.2 Relationship to ZP-N subdomains

Just below the threshold value of our iterative searches using the conserved domains present in the N-terminal extensions of ZP proteins as queries, marginal similarities were often observed with ZP-N subdomains [e.g. the ZP-N subdomain of the CUB and ZP-like domain-containing protein 1 (CUZD1_HUMAN, aa 274–359) was aligned with the first copy of the repeated domain of X.tropicalis ZPAX; E-value = 3.7, 18% of identity]. These alignments were supported at the 2D level using HCA (Supplementary Fig. 1). Moreover, secondary structure predictions performed using the multiple alignment of the repeated conserved domain described in this study led to propose an all \( \beta \)-fold, with eight \( \beta \)-strands, consistently with the prediction made for ZP-N sub-domains ([Jovine et al., 2005] and Fig. 1). The PSI-BLAST results described above (second alignment between the N-terminal extension of ZPAX and ZP2, \( E = 4 \times 10^{-77} \)) showed that the fifth repeated domain of ZPAX proteins (aa 507–602, X.tropicalis ZPAX) can be significantly aligned with the ZP-N sub-domains of ZP2 proteins (aa 371–467). The alignment with various ZP-N subdomains, performed using the PSI-BLAST results and HCA, is reported in Figure 1 (also see Supplementary Fig. 1), showing the conservation of the core hydrophobic residues, as well as of the four cysteines present in the first repeated domains of ZP proteins.

### 3 DISCUSSION

We have shown here that the N-terminal extensions of ZP proteins, found in ZP1, ZP2, ZP4 and ZPAX, share a common domain, which is present in single or multiple copies (Fig. 2) and which can be related to the ZP-N family of domains. This structural relationship, at high level of divergence, provides new insights into the molecular evolution of ZP proteins, which may have evolved around a common ZP-N architecture. The N-terminal repeated ZP-N domains may play a role in the specificity of the biological role of ZP1, ZP2, ZP4 and ZPAX in comparison with ZP3, which is considered as the primary ligand of sperm, recognizing specific O-linked oligosaccharides attached to the polypeptide [reviewed in (Wassarman, 2005; Wassarman et al., 2005)]. In contrast, ZP2 is thought to constitute the secondary ligand, which binds to acrosome reacted sperm and triggers events important for the prevention of polyspermy, whereas ZP1 was proposed to contribute to the structural integrity of the ZP matrix. The importance of the N-terminus of human ZP2 and X.laevis gp69/64 (the homolog of mammalian ZP2) in sperm-binding functions has been evidenced by two independent studies, showing the inhibition of spermatozoa binding and penetration into egg envelope by ZP2 antisera against the N-terminal region (Koyama et al., 1991) as well as by proteolytic cleavage of the gp69/64 amino-terminal region (Tian et al., 1999). It may thus be interesting to investigate further the role of the N-terminal repeated ZP-N domains in both specific processes of sperm binding and blockage of polyspermy.

Until now, ZP-N domains were known to be intimately associated with ZP-C sub-domains to form complete ZP domains, with a few exceptions of proteins, such as PLAC1, Oosp1 and Papillote in which they occur as isolated units (Jovine et al., 2006). However, the ZP-N domain sequences of these last proteins, in contrast to those of ZP proteins, can readily be identified through profile searches (Fig. 1). In particular, the four cysteine, as well as a glycine, which constitute the hallmarks of canonical ZP-N domains.
(Jovine et al., 2005), are found in these proteins with isolated ZP-N domains (Fig. 1). Although not necessarily possessing the typical sequence hallmarks described above, the divergent ZP-N domains of ZP proteins described in this study however conserve hydrophobicity in positions that would participate to the ZP-N core structure. It appears, among others, from this ZP-N domain extension that the four cysteine residues invariably present in canonical ZP-N sub-domains are dispensable to the ZP-N fold stability. Consistently, the construction of a tree based on alignment shown in Figure 1 (Guindon and Gascuel, 2003) clearly indicated that isolated ZP-N domains form separated groups, distinct from canonical ZP-N domains. The first one, including PLAC1 and Oosp1, is more proximal to canonical ZP-N domains than the second one, gathering isolated ZP-N domains from ZP proteins (data not shown). The extended ZP-N profile, which can be build based on this analysis, might thus help to predict such isolated sub-domains, but also to specify its limits within complete ZP domains, in which it appear tightly associated with ZP-C sub-domains.

Currently, no experimental 3D structure is available for ZP-N domains. Moreover, the use of several fold-recognition algorithms (Kelley et al., 2000; Shi et al., 2001) did not lead to reliable predictions, except for the fact that they would consist in an all β-fold. This suggests that ZP-N domains might adopt a novel protein fold, as already noted by Jovine et al. (2005). However, among the different, non-significant, alignments provided by fold-recognition methods, in which the immunoglobulin fold has an important place, the CUB domain fold retained more particularly our attention. Indeed, an HCA-based alignment between ZP2 and spermadhesin (pdb 1sfp) (Supplementary Fig. 2) put in correspondence the extended ZP-N profile, which can be build based on this analysis, might thus help to predict such isolated sub-domains, but also to specify its limits within complete ZP domains, in which it appear tightly associated with ZP-C sub-domains.

REFERENCES


Conflict of Interest: none declared.