Structure and Function of Whey Acidic Protein (WAP) Four-Disulfide Core (WFDC) Proteins

Towards defining the complement of mammalian WFDC-domain-containing proteins

Colin D. Bingle
Academic Unit of Respiratory Medicine, Department of Infection and Immunity, University of Sheffield, Royal Hallamshire Hospital, Glossop Road, Sheffield S10 2JF, U.K.

Abstract

WFDC (whey/four-disulfide core)-domain-containing proteins are defined by the possession of one or more 40–50 amino acid domains that include eight conserved cysteine residues linked by four characteristic intramolecular disulfide bonds. Some also contain other structural domains, whereas in many the WFDC-domain is the only domain present. The WFDC-domain is not limited to mammals but is widespread across all lineages. There is increasing evidence to suggest that mammalian WFDC-domain-containing proteins are undergoing rapid molecular evolution and as might be expected they exhibit low levels of sequence similarity coupled with multiple examples of species-specific gene acquisition and gene loss. The characteristic structural domain (that is generally encoded by a single exon) makes these proteins relatively easy to identify in databases. This review will outline the repertoire of such domains within the mouse, but similar principles can be applied to the identification of all proteins within individual species.

The term ‘FDC (four-disulfide core)’ was originally used to describe a structural domain shared between a group of small disulfide-rich proteins isolated from a diverse range of sources, including wheat germ agglutinin, snake venom postsynaptic neurotoxins and ragweed pollen allergen [1]. This domain contained approximately 40–50 amino acids including eight conserved cysteine residues that generate four intramolecular disulfide bonds, hence the term FDC. At the same time, the major whey protein of mouse milk, WAP (whey acidic protein), was identified [2,3] and subsequently this has become the prototypic member of a subfamily of the FDC-domain-containing proteins; members of this subfamily are now termed WFDC (whey/four-disulfide core) proteins [4]. The WFDC domain is not unique to WAP proteins, but is found in numerous other proteins and perhaps confusingly not all of these proteins are present in whey/milk [4].

The WFDC signature was initially defined based on systematic analysis of over 80 FDC domains from a wide range of species [4] and clearly identified the characteristic bonding arrangements between the cysteine residues. The number of proteins containing a WFDC domain has increased dramatically, particularly in invertebrates where multiple new proteins have been described over the last 10 years [5]. The fact that the vast majority of WFDC domains are encoded by a single exon has allowed for the relatively easy recognition of new domains in a wide range of species based on predictions from genomic DNA. We recently defined an updated WFDC domain signature based on the analysis of the 26 human domains that we were able to demonstrate were shared between 18 different human proteins [6]. This signature and the disulfide-bonding arrangement are shown in Figure 1(A).

It is well recognized that the majority of human WFDC-domain-containing proteins are encoded by a locus present on human chromosome 20 (referred to here as the WFDC locus)
Figure 1 | Spacing differences in the conserved cysteine residues of mouse WFDC domains

(A) Schematic representation of the mammalian WFDC motif illustrating the position of the conserved disulfide bonds.

(B) All mouse WFDC domains were aligned using the multiple alignment tool, ClustalW. Individual WFDC domains from multi-domain-containing proteins are designated −1, −2, −3, etc. according to the position within each protein. Manual adjustments were made so that the spacing of the second and third cysteine residues is variable relative to the position of the remaining six residues and the alignment order is organized such that the spacing between these same two residues decreases. The positions of the eight cysteine residues are shown by * below the Figure. Identical amino acids are shown as white on black, and conserved residues are shown as black on grey.

[7]. They are small secreted proteins in which the WFDC domain is either the only known structural domain or else it is associated with a Kunitz domain [6,7]. Proteins encoded by genes present in this locus include the well-studied proteins elastin, SLPI (secretory leucocyte proteinase inhibitor) and WFDC2 [8–10]. Outside the chromosome 20 cluster are four additional genes that encode proteins with a WFDC domain, WFIKKN1 [11], WFIKKN2 [12], KAL1 [13] and WFDC1 [14], which are significantly larger proteins and the first three also contain a number of other structural domains.

Comparative analysis has shown that there is a significant level of diversity within the WFDC locus that has given rise to a unique complement of WFDC proteins in different mammalian species [7,15,16]. Indeed, a number of studies have shown that this locus is one of the most rapidly evolving mammalian gene loci [16,17] and that genes within this locus also have increased complexity due to a significant level of alternative splicing [10,18]. Such rapid molecular evolution is a common feature of proteins, such as WFDC proteins, that play roles in reproduction, immunity and host defence.
[17]. Although clear one to one orthologues exist for many of the WFDC genes, there are multiple examples of species-specific gene loss and gene duplication within the family. Perhaps the most striking examples are for the SLPI gene, where there are four Slpi-related genes in rats compared with one in mice [15], and also in the PI3 (elafin/Trappin-2) gene, which is absent from humans [15] and has given rise to an expanded repertoire of paralogous genes in many mammalian species [19]. A number of systematic comparative studies have attempted to identify these genes in multiple species through analysis of genomic and EST (expressed sequence tag) databases [19,20] and, although this has mostly involved genes within the WFDC locus, it is also clear that there is significant species-specific diversity outside this region.

The repertoire of mouse WFDC domains

Simple iterative database searches, coupled with searches of protein domain websites, are highly informative for the identification of WFDC-domain-containing proteins. Because of the way in which these databases are compiled, it is important to reduce redundancy by manual curation of retrieved sequences. In the mouse genome, for example, the InterPro database [21] suggests the existence of 47 proteins containing a WFDC domain. These sequences, however, are redundant with regard to the number of independent genes as they contain multiple copies of individual genes as well as splice variants. The figure is reduced to 19 in the output generated by the SMART tool [22] when searched in the ‘genomic’ mode. Importantly, neither of these estimations is correct and illustrates how care must be taken in their use as domain-counting tools. Manual curation of mouse sequences retrieved using iterative BLAST searches, along with analysis of domain architecture tools, yields a figure of 31 WFDC domains distributed between 24 individual proteins (compared with 19 found in humans) [6]. A multiple sequence alignment of these domains is shown in Figure 1(B). Fifteen of these proteins have a WFDC-based nomenclature and 16 are present in a locus on chromosome 2 that is syntenic with the chromosome 20 locus in humans. Ten proteins have one to one orthologues in humans, whereas species-specific duplications have occurred in three genes, Wfd6 and Wfdc15 in mice and WFDC10 in humans. Outside the WFDC locus on chromosome 11, mice also contain a gene for Wap itself as well as three further closely related genes expi, gm11428 (or activated macrophage/microglia WAP domain protein precursor) and wdnm1-like [23,24]. Wfikkn1, Wfikkn2, Wfdc1 and UmodL1 (uromodulin-like 1) encode the four other murine proteins that contain WFDC domains.

Interestingly, in three proteins, domains appear to be incomplete with regard to the conservation of the number of cysteine residues. For example, mouse UmodL1 [25] is lacking cysteine 2, the first cysteine is lacking from wdnm1-like [24] and the first WFDC domain of WAP is lacking cysteine residues 1 and 8 [3]. Conservation of other residues within these domains do, however, support the view that they are bona fide WFDC domains (or perhaps remnants thereof) and it must be assumed that they have either evolved distinct functions or else the structural constraints that are presumed to be important for the overall structure of an individual domain are dispensable on occasions. In the previous analysis of human WFDC domains, it was shown that human WFDC6 was the only human protein that lacked a cysteine residue (number 7) [6]; this is lacking from most primate orthologues, but is present in the rodent orthologues (Figure 1B).

Mouse UmodL1 is a large multi-domain protein containing a single WFDC domain that was originally identified during a screen for KAL-1 orthologues in mice [25]. KAL-1 (or anosmin-1) is itself a 680-amino-acid glycosylated, ECM (extracellular matrix)-associated protein that also contains a WFDC domain [13,26]. Mutations in the multi-domain protein were shown to be the cause of X-linked Kallmann syndrome, a disease characterized by olfactory bulb dysgenesis and disrupted neuronal development [27]. Although KAL-1 is present in species as far back as invertebrates [28], a rodent orthologue has not been described. UmodL1 shares some similarities with KAL-1 and is expressed in the olfactory epithelium, the vomeronasal organ and some olfactory neurons [25], locations where KAL-1 function might be required. The existence of a human orthologue of UmodL1 [29] suggests that it may not be the functional equivalent of KAL-1. The WFDC domain of UmodL1 is lacking the cysteine residue in all orthologous genes (certainly as far back as fish) suggesting that this is not a recent evolutionary event and confirming that the loss of this residue is not detrimental to the (as yet unknown) function of the protein. Structural characterization of this domain may reveal interesting structural differences between this and other ‘classical’ WFDC domains.

Wdnm1-like, a gene recently identified in adipocytes, is located immediately adjacent to expi and gm11428 on chromosome 11 and appears to have evolved from a recent duplication event [24]. As well as lacking the first cysteine residue an additional eight to ten amino acids are lacking from what must be the extreme N-terminal end of the domain. Additionally, the protein ends prematurely relative to other WFDC domains. Although endogenous protein has not been detected, it is secreted from transfected cell lines, suggesting that it has sufficient structural stability even in the absence of a complete WFDC domain.

The final murine ‘outlier’ WFDC domain is the first domain from WAP. This is lacking two cysteine residues, the first and the eighth [3]. The loss of these will result in the loss of two of the four disulfide bonds found in other WFDC domains and will probably cause some structural disorder in the protein fold. It is worth noting that this loss is also seen in the sequence from the western wild mouse [30], whereas the corresponding domain in more divergent rodent species, the deer mouse and the rat, each contain eight cysteine residues as do domains from pig and camel WAP (Figure 2). Rabbit WAP, however, also lacks two cysteine residues [31] and although it is residues 3 and 7 that are lacking, the result remains the loss of two of the four disulfide bonds from the domain. The rabbit domain is also lacking...
Figure 2 | Multiple sequence alignment of the first WFDC-domain of mammalian WAP sequences reveals the loss of conservation of cysteine residues in a number of species

Sequences were aligned using ClustalW. Identical amino acids are shown as white on black, and conserved residues are shown as black on grey. Spaces introduced during the alignment are shown by —. The position of the cysteine residues that make up the FDC of the WAP domain are illustrated by °. Letters to the left of the alignments correspond to individual species. Mm, Mus musculus, mouse; Ms, Mus spretus, western wild mouse; Pm, Peromyscus maniculatus bairdii, deer mouse; Rr, Rattus norvegicus, Norway rat; Oc, Dryctolagus cuniculus, rabbit; Ss, Sus scrofa, pig; Cd, Camelus dromedarius, Arabian camel.

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence Alignment</th>
</tr>
</thead>
</table>
| Mm      | VQxMFlKAPExXhteXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXeXa

References


©The Authors Journal compilation ©2011 Biochemical Society.


