yeast, causes growth arrest with a terminal phenotype similar to that caused by nitrogen starvation. *Genetics* 155, 611–622
17 Matsumoto, S. *et al.* (2002) Role of the Tsc1–Tsc2 complex in signaling and transport across the cell membrane in the fission yeast *Schizosaccharomyces pombe*. *Genetics* 161, 1053–1063
18 Garami, A. *et al.* (2003) Insulin activation of Rheb, a mediator of mTOR/S6K/4E-BP signaling, is inhibited by TSC1 and 2. *Mol. Cell* 11, 1457–1466
19 Castro, A.F. *et al.* (2003) Rheb binds TSC2 and promotes S6 kinase activation in a rapamycin- and farnesylation-dependent manner. *J. Biol. Chem.* 278, 32493–32496
20 Tee, A.R. *et al.* (2003) Tuberous sclerosis complex gene products, tuberin and hamartin, control mTOR signaling by acting as a GTPase-activating protein complex toward Rheb. *Curr. Biol.* 13, 1259–1268
21 Inoki, K. *et al.* (2003) Rheb GTPase is a direct target of TSC2 GAP activity and regulates mTOR signaling. *Genes Dev.* 17, 1829–1834
22 Zhang, Y. *et al.* (2003) Rheb is a direct target of the tuberous sclerosis tumour suppressor proteins. *Nat. Cell Biol.* 5, 578–581
23 Im, E. *et al.* (2002) Rheb is in a high activation state and inhibits B-Raf kinase in mammalian cells. *Oncogene* 21, 6356–6365
24 Onda, H. *et al.* (2002) Tsc2 null murine neuroepithelial cells are a model for human tuber giant cells, and show activation of an mTOR pathway. *Mol. Cell. Neurosci.* 21, 561–574
25 Kenerson, H.L. *et al.* (2002) Activated mammalian target of rapamycin pathway in the pathogenesis of tuberous sclerosis complex renal tumors. *Cancer Res.* 62, 5645–5650

0968-0004/$ - see front matter
doi:10.1016/j.tibs.2003.09.003

Protein Sequence Motif

The CW domain, a structural module shared amongst vertebrates, vertebrate-infecting parasites and higher plants

Jason Perry[1,2] and Yunde Zhao[2]

[1]Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA
[2]Section of Cell and Developmental Biology, Division of Biological Sciences, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

A previously undetected domain, named CW for its conserved cysteine and tryptophan residues, appears to be a four-cysteine zinc-finger motif found exclusively in vertebrates, vertebrate-infecting parasites and higher plants. Of the twelve distinct nuclear protein families that comprise the CW domain-containing superfamily, only the microrchida (MORC) family has begun to be characterized. However, several families contain other domains suggesting a relationship between the CW domain and either chromatin methylation status or early embryonic development.

The term zinc finger was originally coined to describe a series of zinc-binding modules present in transcription factor IIIA, but its definition has since expanded to include a wide variety of compact domains whose structures are stabilized by the presence of one or more protein-chelated zinc ions [1–3]. Grishin and colleagues have recently undertaken a structural classification of zinc-binding domains, and have found that there are at least eight canonical protein-fold groups that fall under the umbrella of 'zinc finger' with the C_2H_2, the treble-cleft and the zinc-ribbon groups being the most highly represented in the current databases [4]. From a biological perspective, zinc fingers are perhaps even more diverse. Although they are most commonly known for interacting with nucleic-acid templates, such as DNA and RNA, zinc fingers also promote protein–protein interactions and interactions between proteins and small molecules. In these contexts, zinc-finger proteins participate in a host of fundamental biological processes ranging from DNA-based activities, such as transcription, replication and recombination to ubiquitination and the assembly of large protein complexes [3]. Here, we define the CW domain, which is composed of at least four cysteine and two tryptophan residues and is predicted to be another building block in the repertoire of zinc-binding structural modules.

Identification of the CW domain

We first noticed an unusual pattern of cysteine and tryptophan residues upon manual inspection of a predicted protein sequence (CAB83140) that emerged from one of our genetic screens (for auxin mutants) after no conserved domains were found searching the Pfam and SMART databases [5–7]. We next performed iterative PSI–BLAST searches of the non-redundant protein database using the described segment of sequence (of 61 amino acids) as a query with an inclusion cut-off score of 0.005 [8]. From this analysis, 102 sequences were retrieved, 36 of which were above the cut-off score (Figure 1). Of these 36 sequences, there were no false positive identifications. Six additional candidate sequences emerged following manual inspection of the sequences that had been retrieved in the search but had fallen below our prescribed threshold (Figure 1). The search converged after the fourth iteration, and the same group of proteins was retrieved regardless of which newly retrieved sequence was used as the query in subsequent searches. Domain boundaries were delimited by compiling all of the

Corresponding author: Yunde Zhao (yzhao@biomail.ucsd.edu).

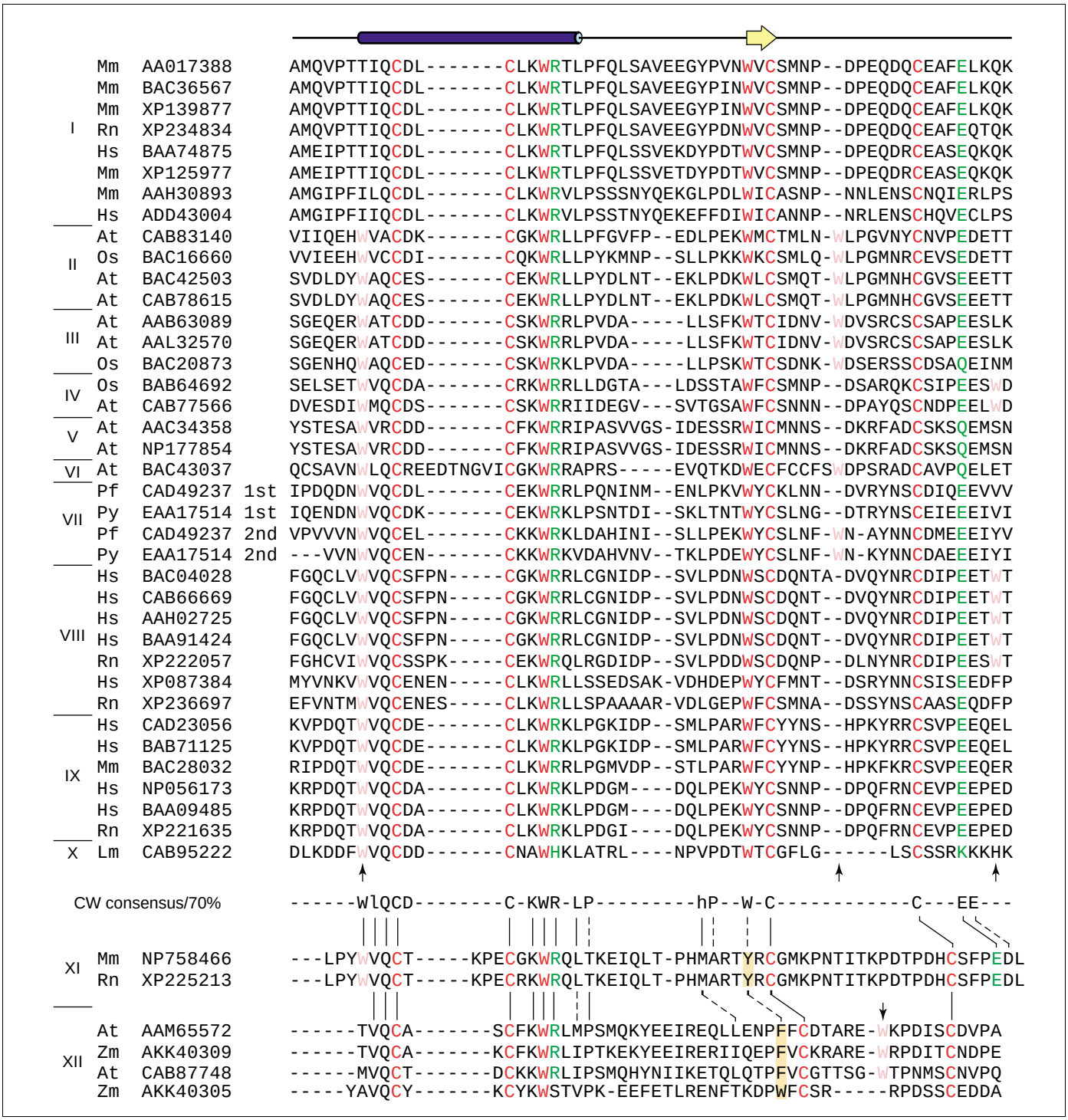

Figure 1. Multiple sequence alignment of the described CW domains. The alignments were constructed with ClustalW (http://us.expasy.org/tools/) and further refined with guidance from the PSI–BLAST pairwise alignments [8]. Those sequences above the CW consensus line (70% of sequences, capital letters are specific amino acid residues, lowercase letters are: l, aliphatic hydrophobic; h, any hydrophobic) (protein families I–X) were retrieved from the non-redundant protein sequence database with scores above (less than) the imposed 0.005 threshold. Sequences below the CW consensus line (protein families XI and XII) were extracted as additional candidates following manual inspection of all sequences retrieved in our PSI–BLAST searches. Sequences are listed by GenBank accession number and are from the following organisms: At, *Arabidopsis thaliana*; Hs, *Homo sapiens*; Lm, *Leishmania major*; Mm, *Mus musculus*; Os, *Oryza sativa*; Pf, *Plasmodium falciparum*; Py, *Plasmodium yoelii yoelii*; Rn, *Rattus norvegicus*; Zm, *Zea mays*. Secondary-structure analyses were performed using jpred at the following website: http://www.compbio.dundee.ac.uk/~www-jpred/. This multiple sequence alignment (alignment number ALIGN_000617) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/).

PSI–BLAST pairwise alignments and examining sequence homology and secondary-structure characteristics. In some cases, previously known domains were juxtaposed and connected in linear sequence by low complexity loops of various lengths (Figure 2).

The CW domain is predicted to be a previously undetected four-cysteine zinc-finger motif. The four putative ligands are invariant amongst the retrieved sequences. Starting at the N terminus of CW, the first two cysteine residues are in the classical Cys-Xaa-Xaa-Cys or Cys-Xaa-Xaa-Xaa-Xaa-Cys arrangement in all but one sequence. The arrangement of the other two cysteines is

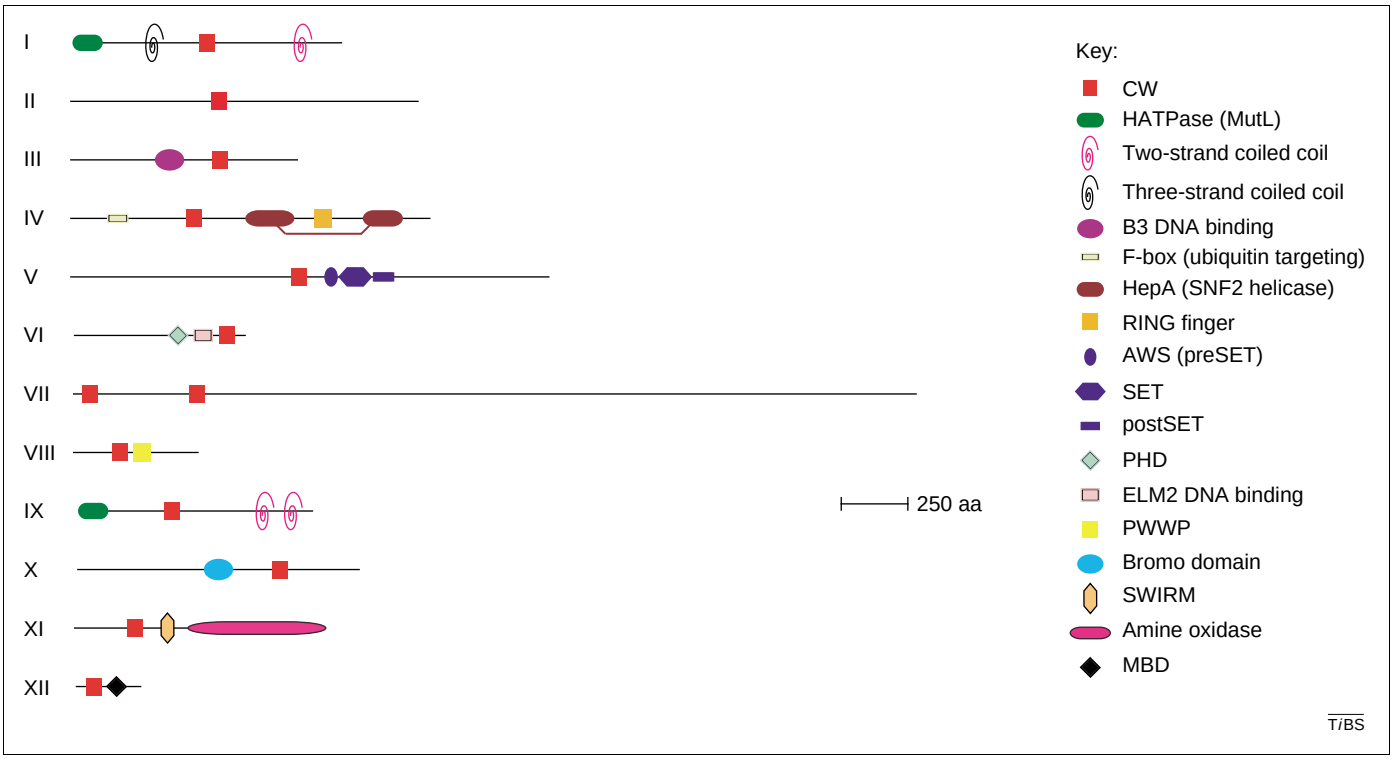

Figure 2. Domain architecture of CW-domain containing proteins. Twelve different CW-domain-containing protein families have been identified. Families I, VIII, IX and XI are from vertebrates including the human; families II, III, IV, V, VI and XII are from higher plants such as *Arabidopsis*, rice and maize; families VII and X are from the *Plasmodium* and *Leishmania* parasites, respectively. All sequences examined also contained putative nuclear localization signals. The diagrams are drawn approximately to scale.

somewhat unusual in that they are separated by a minimum of six and as many as 15 additional residues. All CW domains contain two to four position-conserved tryptophan residues. In most sequences (ten out of 12 protein families, see below), there is a tryptophan-containing motif immediately before the first cysteine (Trp-Val-Gln-Cys). In all sequences, the third residue after the second cysteine is a tryptophan that, in all but two sequences, is flanked by basic residues (Lys-Trp-Arg). There is also always a tryptophan or other aromatic residue positioned two residues before the third cysteine (Trp-Xaa-Cys), a feature reminiscent of rubredoxin (pfam00301.8) and the Sec23/Sec24 zinc finger (pfam04810.2), which are both four-cysteine metal-binding sites. Finally, five of the 12 protein families have a conserved tryptophan between the third and fourth cysteines, and two families share a conserved tryptophan seven positions after the fourth cysteine. The remaining highly conserved elements include an aliphatic residue followed by a proline after the Lys-Trp-Arg motif and a hydrogen-bonding residue in the fourth position after the final cysteine (Figure 1).

The CW domain appears to be more closely related to the four-cysteine site of the binuclear PHD (plant homeodomain, pfam00628) than it is to any other known zinc finger [9]. Most of the below-threshold false positives to emerge from our PSI–BLAST searches were PHDs, and two such sequences had scores of <0.1 (CAB39909 and CAA93898, both from *Schizosaccharomyces pombe*). However, CW domains and PHDs have at least two absolutely distinguishing differences [10]. First, PHDs form binuclear clusters by virtue of their interleaved sets of metal ligands [11]. The first PHD zinc site (3C1H type), including the ligands that would be ordinarily interdigitated within the second (four-cysteine) site, is completely absent in all CW domains. Second, the C-terminal pair of ligands in the four-cysteine site of a PHD always has the Cys-Xaa-Xaa-Cys (occasionally Cys-Xaa-Xaa-His) arrangement, in contrast to the larger separation that is characteristic of CW domains. Nonetheless, one intriguing similarity between the four-cysteine site of PHDs and CW domains is the shared conservation of the rubredoxin-like Trp-Xaa-Cys motif at the third cysteine residue.

Domain architecture and potential function of CW-containing proteins

Using Pfam hidden Markov model searches (http://pfam.wustl.edu) and multicoil (http://multicoil.lcs.mit.edu/cgi-bin/multicoil) analyses, we have identified 12 different eukaryotic nuclear protein families that comprise the CW domain-containing superfamily [5,12,13] (Figures 1, 2). Families I, VIII, IX and XI are vertebrate proteins found in the human, mouse and rat; families II, III, IV, V, VI and XII are proteins found in higher plants such as *Arabidopsis*, *Oryza sativa* (rice) and *Zea mays* (maize); families VII and X are uncharacterized proteins from the vertebrate-infecting parasites *Plasmodium* and *Leishmania* (Figure 2). Subsequent tBLASTn searches of EST collections (most accessed directly from http://www.ncbi.nlm.nih.gov) found various respective family members in other vertebrates (e.g. frog, fish, chicken and sea squirt), higher plants (e.g. potato, wheat, soybean and lettuce) and *Apicomplexa* obligate intracellular parasites (e.g. *Toxoplasma*, *Cryptosporidium* and *Theileria*). For a current listing, complete

with PSI–BLAST pairwise alignments and statistical valuations, see http://www.biology.ucsd.edu/labs/zhao/. Notably, and by contrast, CW domain proteins were not found in either the proteomes nor expressed sequence tag collections of several commonly studied model organisms including *Drosophila*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *S. pombe* or *Neurospora*. Likewise, no sequences emerged from any of the known prokaryotic genomes.

Only a member of family I – the microrchida or MORC family – has been the subject of even cursory biochemical characterization. The gene encoding MORC is expressed in male germ cells where it is localized to the nucleus and participates in spermatogenesis [14]. *Morc* $^{-/-}$ mouse germ cells arrest during meiotic prophase and undergo apoptosis rather than progressing to the first meiotic division [15]. The MORC polypeptide is composed of a MutL-type ATPase complete with a predicted Bergerat ATP-binding fold [16,17], followed by a predicted three-stranded coiled-coil motif, a CW domain and, finally, a predicted two-stranded coiled coil. Members of family IX are similar to members of the MORC family (family I), but they all lack the predicted three-stranded coiled coil and, instead, have an additional two-stranded coiled-coil motif near their C termini.

Several CW domain-containing families have other domains that are thought to be able to associate with methylated DNA or chromatin. For example, families IV, V, VIII and XII are thought to associate with SNF2 (sucrose non-fermenting) helicases, SET [for Su(var), E(2) and trithorax] domains, PWWP (Pro-Trp-Trp-Pro) domains, and MBDs (methyl-binding domains), respectively. Recently, several mutants have been isolated from both plants and mice illustrating the importance of SNF2 helicases and other chromatin-remodeling-complex components in maintaining cytosine methylation [18,19]. The SNF2 domain found in CW family IV is further specialized by the presence of a RING finger between its N-terminal and C-terminal lobes. Members of this family are also characterized by an ubiquitin-targeting F-box motif near their N termini. Complete SET domains, including the preSET and postSET mini-domains, such as those found in CW family V, are histone methyltransferases that regulate gene silencing, transcription and other processes [20]. In addition, SET-domain-containing proteins have recently been implicated in the maintenance of DNA (cytosine) methylation [21]. Like the CW domain, the PWWP domain (which is known to bind DNA [22]) is exclusive to eukaryotes. Both these domains are found both alone and together in a group of small proteins that comprise CW family VIII. Recent structural comparisons of the PWWP of Dnmt3b (a methyltransferase) with Tudor and Chromo domains show that these three domains can be superimposed in a region suspected of associating with methylated chromatin, suggesting that the PWWP domain might perform an analogous function [23]. Finally, CW family XII consists of small plant proteins with a CW domain immediately followed by the MBD methyl-CpG-binding domain.

Several other CW families contain other domains that are known to associate with chromatin in various contexts including B3 DNA-binding domains (family III), ELM2 (Egl-27 and MTA1 homology 2)DNA-binding domains (VI), Bromo domains (X) and SWIRM (for Sw13p, Rsc8p and Moira) domains (XI). Previously, B3 domains have been found in several transcription factors that are active during plant seed development [24,25]. The ELM2 domain was first recognized in the *C. elegans* protein Egl-27, and like the B3 domain, it is present in several proteins that play a part in embryonic development [26,27]. In addition to the ELM2 and CW domains, family VI proteins also contain the somewhat related PHD motif. This is the only protein family found to contain both a CW and a PHD domain. Bromo domains, such as that in family X, were first found in the *Drosophila* protein Brahma and have since been characterized in a wide range of histone acetyltransferases where they serve as acetyl-lysine-binding modules [28,29]. The SWIRM domain, recently identified by Aravind and Iyer [30], is predicted to be an α-helical domain that promotes protein–protein interactions in the context of chromatin protein complexes. The authors also found that the SWIRM domain could be conjugated to an amine oxidase and suggested that such enzymes could catalyze either lysine or polyamine oxidation to affect chromatin structure. We also find the SWIRM-amine oxidase architecture in CW family XI. Finally, two CW families (II and VII) have no other defined domains. Family VII proteins are large proteins from *Plasmodium* and are the only group in which the CW domain appears to have been duplicated in a single polypeptide. RADAR (rapid automatic detection and alignment of repeats) analyses suggest that there are many repetitive elements present in these proteins including 10–15 unusual asparagine-rich repeats in the middle of the polypeptides (some with weak homology to RNA-binding pumilio repeats) and rather large tyrosine-rich repeats (85 out of 575 residues are tyrosines) at their C-termini [31].

Concluding remarks

The CW domain is predicted to be a highly specialized mononuclear four-cysteine zinc-finger domain that plays a part in DNA binding and/or promoting protein–protein interactions in complicated eukaryotic processes including, but perhaps not limited to, chromatin methylation status and early embryonic development. The limited and unusual distribution of the CW domain suggests that even its evolution has been somewhat peculiar and that it might have been propagated through seemingly disparate lineages by way of ancient lateral gene-transfer events.

Acknowledgements

We thank Job Dekker and Jack Dixon as well as anonymous reviewers for their insightful comments. J.P. is supported by NIH grant RO1-GM25326 to Nancy Kleckner of Harvard University. Y.Z. is supported by NIH grant 1R01GM68631-01.

References

1 Miller, J. *et al.* (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO J.* 4, 1609–1614

2 Frankel, A.D. *et al.* (1987) Metal-dependent folding of a single zinc finger from transcription factor IIIA. *Proc. Natl. Acad. Sci. U. S. A.* 84, 4841–4845

3 Matthews, J.M. and Sunde, M. (2002) Zinc fingers – folds for many occasions. *IUBMB Life* 54, 351–355
4 Krishna, S.S. *et al.* (2003) Structural classification of zinc fingers. *Nucleic Acids Res.* 31, 532–550
5 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266
6 Ponting, C.P. *et al.* (1999) SMART: identification and annotation of domains from signaling and extracellular protein sequences. *Nucleic Acids Res.* 27, 229–232
7 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234
8 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
9 Schindler, U. *et al.* (1993) HAT3.1, a novel *Arabidopsis* homeodomain protein containing a conserved cysteine-rich region. *Plant J.* 4, 137–150
10 Marchler-Bauer, A. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 31, 383–387
11 Pascual, J. *et al.* (2000) Structure of the PHD Zinc finger from human Williams–Beuren syndrome transcription factor. *J. Mol. Biol.* 304, 723–729
12 Sonnhammer, E.L. *et al.* (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26, 320–322
13 Wolf, E. *et al.* (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* 6, 1179–1189
14 Inoue, N. *et al.* (1999) New gene family defined by MORC, a nuclear protein required for mouse spermatogenesis. *Hum. Mol. Genet.* 8, 1201–1207
15 Watson, M.L. *et al.* (1998) Identification of *morc* (microrchida), a mutation that results in arrest of spermatogenesis at an early meiotic stage in the mouse. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14361–14366
16 Mushegian, A.R. *et al.* (1997) Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5831–5836
17 Dutta, R. and Inouye, M. (2000) GHKL, an emergent ATPase/kinase superfamily. *Trends Biochem. Sci.* 25, 24–28
18 Dennis, K. *et al.* (2001) Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes Dev.* 15, 2940–2944
19 Jeddeloh, J.A. *et al.* (1999) Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat. Genet.* 22, 94–97
20 Marmorstein, R. (2003) Structure of SET domain proteins: a new twist on histone methylation. *Trends Biochem. Sci.* 28, 59–62
21 Malagnac, F. *et al.* (2002) An *Arabidopsis* SET domain protein required for maintenance but not establishment of DNA methylation. *EMBO J.* 21, 6842–6852
22 Qiu, C. *et al.* (2002) The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. *Nat. Struct. Biol.* 9, 217–224
23 Maurer-Stroh, S. *et al.* (2003) The tudor domain 'Royal Family': tudor, plant agenet, chromo, PWWP and MBT domains. *Trends Biochem. Sci.* 28, 69–74
24 Luerben, H. *et al.* (1998) FUSCA3 encodes a protein with a conserved VP1/ABI3-like B3 domain which is of functional importance for the regulation of seed maturation in *Arabidopsis thaliana*. *Plant J.* 15, 755–764
25 Stone, S.L. *et al.* (2001) LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. *Proc. Natl. Acad. Sci. U. S. A.* 98, 11806–11811
26 Solari, F. *et al.* (1999) The *Caenorhabditis elegans* genes *egl-27* and *egr-1* are similar to *MTA*1, a member of a chromatin regulatory complex, and are redundantly required for embryonic patterning. *Development* 126, 2483–2494
27 Ding, Z. *et al.* (2003) Human MI-ER1 alpha and beta function as transcriptional repressors by recruitment of histone deacetylase 1 to their conserved ELM2 domain. *Mol. Cell. Biol.* 23, 250–258
28 Tamkun, J.W. *et al.* (1992) brahma: a regulator of *Drosophila* homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. *Cell* 68, 561–572
29 Zeng, L. and Zhou, M-M. (2002) Bromodomain: an acetyl-lysine binding domain. *FEBS Lett.* 513, 124–128
30 Aravind, L. and Iyer, L.M. (2002) The SWIRM domain: a conserved module found in chromosomal proteins points to novel chromatin-modifying activities. *Genome Biol.* 3 RESEARCH0039
31 Herger, A. and Holm, L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* 41, 224–237

0968-0004/$ - see front matter
doi:10.1016/j.tibs.2003.09.007

Ode to a calcium pump

Take a rabbit, mince the meat,
Marinate with salt and sweet.
Pour off supernatant liquor,
Spin it faster, faster, quicker.
Lucent bubbles form a pellet
Of the amber muscle membranes,
Packed with pumps to shift the Calcium,
Billions per square millimetre.

Freeze and thaw the little bubbles,
Treat with soap and magic toxins
Till the bubbles fuse to tubules
And the pumps line up in rows
Winding round them in a helix.
Scan them with electron beams,
Moving in a powerful field,
Focussing the scattered charges
On a photographic plate.
Turn the black spots into numbers,
Feed into a large computer -
Forming unexpected shapes,
Not unlike a lumpy cow head
Or a sheep with scraggy neck.

Whence has come this bovine object
Permeating all our meat?
Nothing but a stringy protein,
Coded by some DNA,
Folded to a complex structure,
Pumping calcium all the day.
When the nerve excites the muscle,
Calcium floods in at a jump,
Fibre grabs a hold on fibre,
Tetanus could seize the rump, Butt!
Calcium sparks its own removal
Priming its specific pump.
Ponin recombines with Actin,
Muscles can relax again.
Contract-relax! Contract-relax!
Contract-relax! Much gain -
No pain!

Michael Green