Minireview

The PWWP domain: a potential protein–protein interaction domain in nuclear proteins influencing differentiation?

Ingrid Stec[a,*], Sylvia B. Nagl[b], Gert-Jan B. van Ommen[a], Johan T. den Dunnen[a]

[a] *MGC-Department of Human and Clinical Genetics, Leiden University Medical Center, Wassenaarseweg 72, 2333 AL Leiden, The Netherlands*
[b] *Bloomsbury Centre for Structural Biology, Department of Biochemistry and Molecular Biology, University College London, London, UK*

Received 28 January 2000; received in revised form 27 March 2000

Edited by Matti Saraste

Abstract Upon characterization of *WHSC1*, a gene mapping to the Wolf–Hirschhorn syndrome critical region and at its C-terminus similar to the *Drosophila* ASH1/trithorax group proteins, we identified a novel protein domain designated PWWP domain. To gain insight into its structure, evolutionary conservation and its potential functional role, we performed database searches to identify other PWWP domain-containing proteins. We retrieved 39 proteins, and a multiple alignment shows that the domain spans some 70 amino acids. It is present in proteins of nuclear origin and plays a role in cell growth and differentiation. Due to its position, the composition of amino acids close to the PWWP motif and the pattern of other domains present, we hypothesize that the domain is involved in protein–protein interactions.

Key words: Protein domain; Multiple alignment; Protein–protein interaction

1. Introduction

Due to the efforts to determine the genomic sequence of many organisms, in particular that of the human genome, many novel genes and their encoded protein products are discovered daily. Here, we report the characterization of a novel protein domain found among these sequences and hypothesize on its potential function. The domain was recently identified upon characterizing a gene located in the Wolf–Hirschhorn syndrome (WHS) critical region on 4p16.3 and, when translocated, involved in lymphoid multiple myeloma (MM) disease, also known as plasmacytoma. Initially, we noticed the domain as an independently conserved segment of the protein present in several proteins with different functions. Based on the central core 'proline–tryptophan–tryptophan–proline', we designated it PWWP domain (Pfam PF00855) [1]. To gain insight into its size, conservation and potential functional role, we performed extensive database searches to identify all proteins containing a PWWP domain (see Section 2).

Thirty-nine proteins are identified to contain a PWWP domain as shown in Fig. 1 (multiple alignment). Fig. 2 presents a general overview of the overall domain structure of the full length proteins. Conservation of the PWWP domain is concentrated on one major and two minor blocks with length differences occurring in between (Fig. 1). Borders of the domain were derived from both decreasing amino acid similarity and the presence of other, directly flanking domains (Fig. 2), both N- and C-terminal. At the genomic level, the major and first minor block are often split by an intron making intron/exon sliding (i.e. insertion/deletion at the protein level) a plausible mechanism behind the observed length variation in this region. Interestingly, PWWP domains also exist where an intron resides inside the major conserved motif (P–WWP), e.g. the second PWWP domain of the *WHSC1* gene.

2. Methods: identification of the domain, database analysis

Database searches were performed using BLASTP, TBLASTN and a position-specific iterated BLAST (PSI-BLAST) ([2], http://www.ncbi.nlm.nih.gov/BLAST/). BLASTP and TBLASTN were run against the non-redundant, EST and htgs sections of GenBank at the National Center for Biotechnology Information (NIH, Bethesda, MD, USA). To pick out more distant similarities, we used PSI-BLAST to compare the sequence of the *WHSC1*-encoded PWWP domain with the non-redundant database. Starting with the first WHSC1 PWWP domain sequence (Fig. 1), the PSI-BLAST search converged at the fifth iteration, with all domains identified giving an *E*-value of $<10^{-6}$ (*E*-value of the iterations 1–4 was 10^{-3}). Finally, we used ClustalX ([3], http://www-igbmc.u-strasbg.fr/BioInfo/) to generate a multiple sequence alignment of all PWWP domain-containing protein segments identified. The final alignment (Fig. 2) contains some minor manual corrections, shifting 'lonely characters'. At this stage, we also removed some false positives clearly missing critical conserved residues (e.g. *Saccharomyces cerevisiae* YMT4 (Q04213)). For domain/motif structure analysis of the proteins, we used the Pfam (http://www.sanger.ac.uk/Pfam/), SMART (http://smart.embl-heidelberg.de/) and Blocks (http://blocks.fhcrc.org/) servers.

3. PWWP domain encoding genes

WHSC1 is one of a series of genes mapping to the critical gene region of WHS. At its C-terminus, the *WHSC1*-encoded protein is highly similar to the *Drosophila melanogaster* trithorax protein ASH1. Trithorax group proteins are found to be part of energy-dependent chromatin remodeling machines, multi-protein complexes which are thought to be involved in

*Corresponding author. Fax: (31)-20-8753784.
E-mail: i.stec@lumc.nl

PII: S0014-5793(00)01449-6

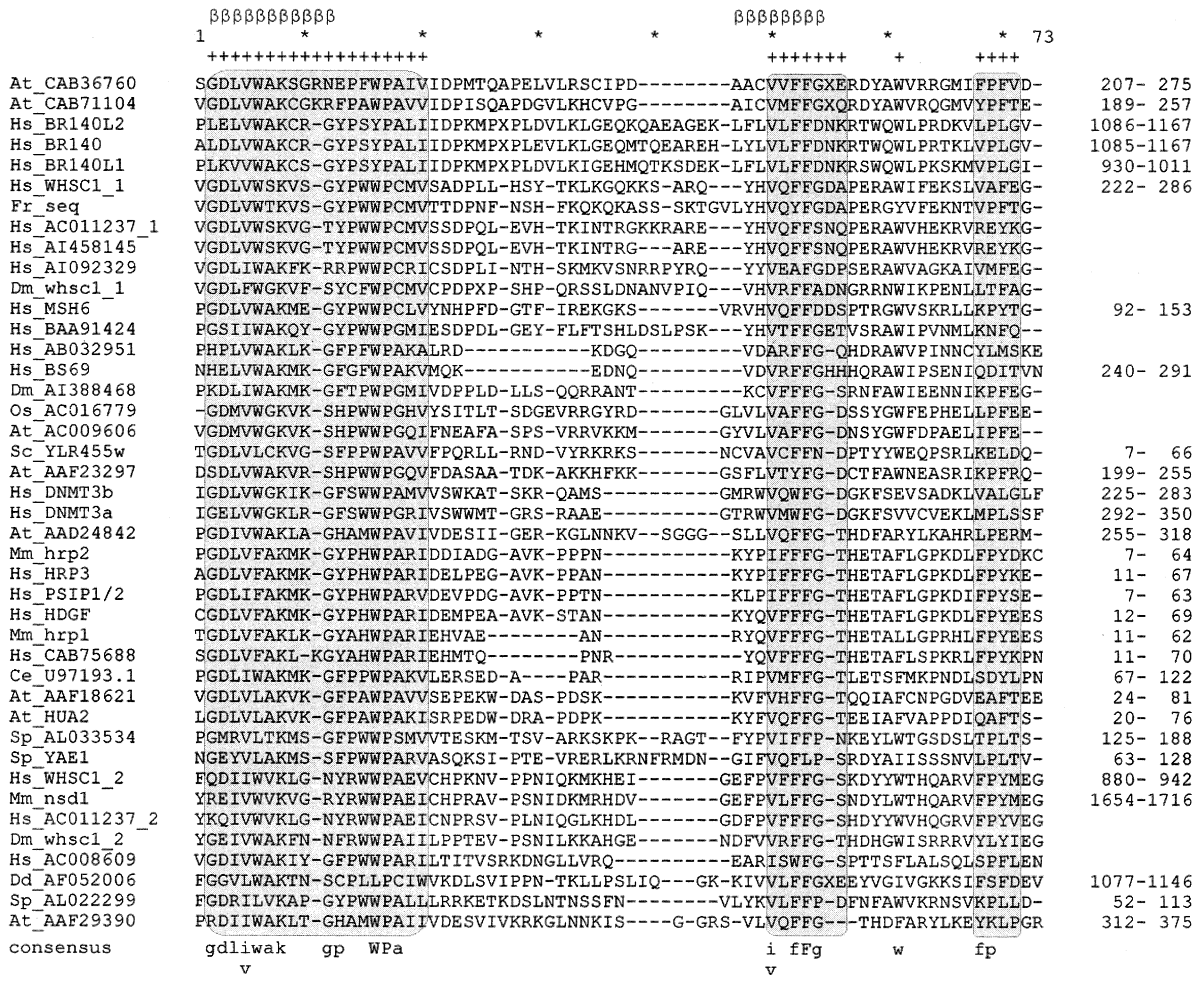

Fig. 1. Multiple alignment of PWWP domain-containing proteins. The homologs for bovine HDGF and HRP3, murine MSH6, HDGF, DNMT3a, DNMT3b, EST AW413965, rat Rn EST111365 and zebra fish DNMT3 are not included in the alignment. The human PSIP1 and PSIP2 sequences are differential splice products of one gene and are therefore listed only once. The bottom line lists the consensus amino acid sequence; lower case 50–89%, capital letters >90% conserved. Organism abbreviations are: At, *A. thaliana*; Ce, *C. elegans*; Dd, *D. discoideum*; Dm, *D. melanogaster*; Dr, *Danio rerio*; Fr, *F. rubripes*; Hs, *H. sapiens*; Mm, *M. musculus*; Os, *O. sativa*; Rn, *R. norvegicus*; Sc, *S. cerevisiae*; Sp, *S. pombe*; Tb, *Trypanosoma brucei*. Proteins are listed with their names or database identification numbers at the left (complete list, see Section 2). At the right, the amino acid positions in the respective proteins are provided (for those whose sequence are complete). Abbreviations: *DNMT*, DNA-cytosine-5-methyltransferase; *HDGF*, heptoma-derived growth factor; HRP, HDGF-related protein; *MSH6*, MutS homologue 6; Nsd1, NR-binding SET domain-containing protein; *PSIP2*, PC4 and SFRS1 interacting protein 2; *WHSC1*, WHS candidate 1. BS69, BR140, HUA2 and CGI-142 are not specified in the respective publications.

counteracting repressive chromatin structures allowing accessibility of transcription factors to target sequences.

Patients affected by WHS suffer from developmental defects affecting the face, brain and several inner organs. The *WHSC1* gene encodes at least three different protein isoforms, 62 kDa, 64 kDa and 136 kDa, which all encode a PWWP domain within the N-terminus (central motif with PWWP) and a HMG-box downstream. The 136 kDa WHSC1 product possesses a second PWWP domain (central motif with residues RWWP, Fig. 1), several PHD-type zinc fingers and at the C-terminal extension a SET domain. The SET domain has been shown to be a protein–protein interaction site controlling dual-specific phosphatases and anti-phosphatases, thereby coordinating development [4,5]. The encoded domain pattern of *WHSC1* (Fig. 2), containing a nuclear localization signal, a DNA-binding HMG-box and PHD-type zinc fingers implicated to be involved in chromatin regulation [6], makes it a likely transcription (co)factor regulating developmental processes [1]. Very interestingly, a closely related *WHSC1* gene is present in human and other organisms, e.g. in mouse and the fruitfly, the human one maps on chromosome 8p11.2 (GenBank human PAC sequences AC024242, AC011237 and AF214633). We named it *WHSC1L1* for *WHSC1L1 like 1* (see UniGene transcript Hs.27721 and Fig. 2). Indeed, it shows homology over nearly the full length of WHSC1 and is the only gene found so far which, like *WHSC1*, encodes two PWWP domains.

WHSC1 genes from other organisms have not been described so far. The C-terminal region of the mouse gene can be derived from dbEST but does not cover the PWWP domain. The currently available genomic sequence of *D. melanogaster* contains a WHSC1 ortholog (GenBank AC009249, AC009388) spanning both conserved PWWP domains (Fig. 1). Nsd1 (NR-binding SET-domain-containing protein [7]) is

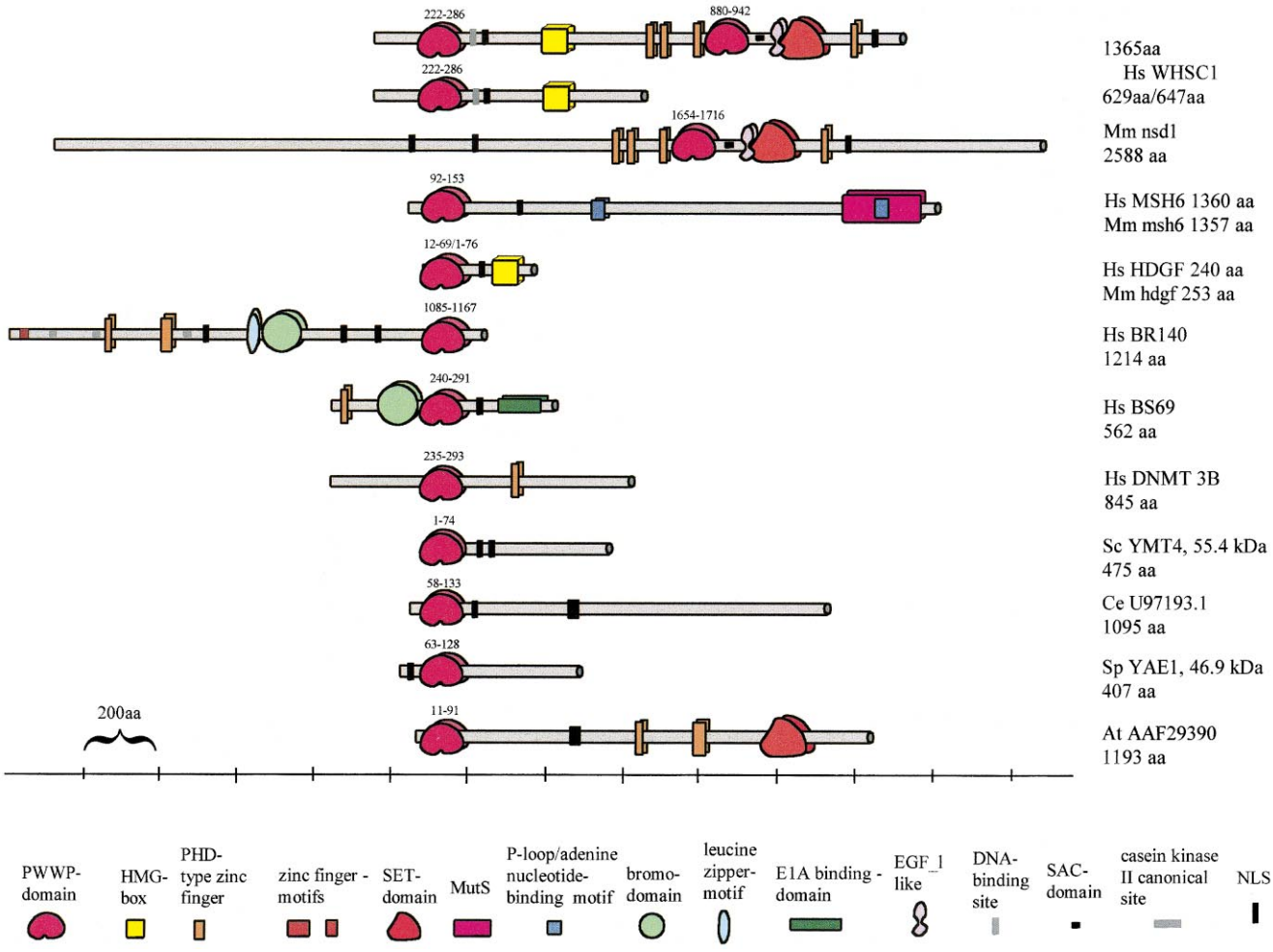

Fig. 2. PWWP domain-containing proteins and their domain and motif structures. The PWWP domain is indicated by a pink circle shape, other domains are shown in colored boxes and other shapes (bottom). Protein length and amino acid positions of PWWP domains are given for each protein domain. Species abbreviations and accession numbers are as in Fig. 1. Other abbreviations: SAC, SET domain associated Cys-rich domain; EGF, epidermal growth factor. Only relevant domains, motifs, signals and binding sites are shown. Pfam accession numbers for the domains are: HMG-box: PF00505; PHD-type zinc finger: PF00628; zinc finger motifs: PF00096 (C2H2), PF01530 (C2HC); SET domain: PF00856; MutS: PF00488; bromodomain: PF00439 (references are given properly at Pfam and SMART).

a mouse protein of which the C-terminal half (amino acids 661–1435) is highly homologous (66% identity, 79% similarity) to the WHSC1 protein. Nsd1 is implicated in transcription control at the chromatin level. No corresponding human homolog has yet been described, although dbEST does contain homologous human transcripts (UniGene Hs.156865).

The largest group of homologous proteins are those related to hepatoma-derived growth factor (HDGF). HDGF is an extra-cellular heparin-binding acidic, nuclear polypeptide with mitogenic activity [8]. Its PWWP domain is encoded within the extreme N-terminus. HDGF is expressed in various normal tissues as well as amplified in certain tumor cell lines. Izumoto et al. characterized two *HDGF*-related cDNAs encoding the proteins hdgfrp1 (HRP1) and hdgfrp2 (HRP2) [9] and mentioned the region around the PWWP domain as the '*hath* region'. Several other related proteins have recently been identified, including PSIP2 (PC4 and SFRS1 interacting protein 2) [10] and CGI-142 [11].

BS69 is another nuclear human protein that interacts with E1A, specifically suppressing E1A-activated transcription [12]. BS69 is highly expressed in colon carcinoma. An N-terminal isoform of BS69 has been described as BRAM1 (BMP receptor associated molecule 1) lacking the PWWP domain and localized outside the nucleus [13] underlining the role of the PWWP domain in nucleic functions.

PWWP domain-containing nuclear proteins of the BR140 family (peregrin and BR140-like proteins 1 and 2) contain the most distantly related PWWP domain (Fig. 1). BR140 is the only protein with a C-terminal PWWP domain. It is ubiquitously expressed, though most abundant in testes and spermatogonia. BR140 shows similarity to various transcriptional co-activators such as TAF 250 (TATA-binding protein associated factors), a subunit of TFIID which is essential for RNA polymerase II-mediated transcription [14]. BR140, like the related proteins AF10 and AF17, is part of a family of regulatory proteins which have been found to be disrupted by chromosomal translocations in myeloid leukemias. A similar mechanism is seen in the WHSC1 protein of which the PWWP domain is disrupted by t(4;14) translocations causing lymphoid MM [1]. These observations strengthen the suggested role of these proteins in transcriptional processes and regulation of development. In this respect, an interesting ques-

tion arises whether the position of the translocation on 4p, i.e. N-terminal, inside or C-terminal of the PWWP domain, has an effect on the clinical characteristics of MM patients.

MSH6 is one of three MutS (bacterial mismatch repair gene) homologs which is essential for mismatch-binding activity [15]. Mutations in *MSH6* predispose to hereditary non-polyposis colorectal cancer (HNPCC) and carcinoma of the endometrium [16,17]. Mice, homozygous for a null mutation in *msh6*, develop gastrointestinal tumors and B- and T-cell lymphomas [18]. MSH6 contains a PWWP domain at its very N-terminus; interestingly, msh6 orthologs found in *S. cerevisiae*, *Caenorhabditis elegans* and *Arabidopsis thaliana* do not contain this domain.

Very recently, mutations were found in another PWWP-containing gene involved in human genetic disease. This disease, the ICF syndrome, a malformation syndrome affecting the face with immune deficiency and chromosomal instability, was reported to be caused by mutations in the *DNMT3B* gene (DNA-cytosine-methyl transferase gene) [19]. Conspicuously, although mutations in HNPCC and ICF syndrome are spread throughout the *MSH6* and *DNMT3B* genes, respectively, no mutations (or polymorphisms) were reported in the PWWP domain. This may indicate that such mutations have other and/or more severe phenotypic consequences.

3.1. Accession numbers of PWWP-containing sequences

Homo sapiens WHSC1: AAD21771, BR140: P55201, BR140L1: Z84485, BR140L2: Z98885, BS69: S56145, DNMT3a: AAD33084, DNMT3b: CAB53070, HDGF: P51858, HRP3: BAA90477, MSH6: U28946, PSIP2: NP_004673, AB032951, AC008609, AC011237, EST AI458145, EST AI092329, BAA91424, CAB75688; *Bos taurus* hdgf: CAB40626 and hrp1: CAB40348; *Mus musculus* dnmt3a: AAC40177, dnmt3b: AF068628, hdgf: P51859, hrp1: JC5661, hrp2: D63707, hrp3: BAA90478, msh6: U42190, nsd1: AF064553, EST AW413965, EST AI391173; *Rattus norvegicus* EST111365: H35312; *Brachydanio rerio* dnmt3: AF135438; *Fugu rubripes* FG:199L16bE9; *A. thaliana* HUA2: AF116556, AAD24842, AAF18621, AAF23297, AAF29390, AC009606, CAB36760, CAB71104; *Oryza sativa* AC016779; *D. melanogaster* EST AI388468, whsc1: AC009249 and AC009388; *C. elegans* T25520; *Dictyostelium discoideum* C-module-binding factor: AF052006, *Schizosaccharomyces pombe* YAE1 - Q09842, AL033534, AL022299 and *S. cerevisiae* YLR455w: Q06188.

4. Computational analysis of the WHSC1 protein

Computational analysis using PSIPRED predicts the WHSC1 protein to contain large randomly coiled regions. The PWWP domain itself may be partly disordered. Region 239–261 in human WHSC1 (amino acids 21–43 in Fig. 1), just C-terminal of the PWWP motif, has a low content of the aromatic residues tyrosine, tryptophan and phenylalanine (4%), and few cysteines and histidines (4%). At the same time, the region is rich in the charged residues glutamate, aspartate and lysine, and in serines (30%). This type of amino acid composition is typical of disordered regions that function as protein–protein interaction sites [20–22]. Disordered regions play an important role in protein function via disorder to order transitions upon complex formation. It is thought that the resulting increases in free energy when flexible regions solidify upon binding enable high specificity without exceedingly high affinity.

5. The PWWP domain throughout evolution

Scanning the genomes of those organisms which have been completely sequenced shows that the PWWP domain is not present in prokaryotes. The yeast *S. cerevisiae* harbors one PWWP protein, Ylr455wp, with unknown function and, besides the PWWP domain, devoid of any other recognizable domain. The partial genome sequence of the yeast *S. pombe* shows the presence of three PWWP domain proteins, allmost similar to the *S. cerevisiae* protein Ylr455wp. The nematode *C. elegans*, the first multi-cellular eukaryote fully sequenced, contains only one PWWP protein, U97193. This protein contains two domains, an N-terminal PWWP domain, most similar to those of the HDGF-like proteins, and a C-terminal SIR2 domain. Although several SIR2 domain-containing human proteins have been described recently, none of these contains a PWWP domain. The partial sequence of the plant *A. thaliana* thus far reveals the presence of eight PWWP domain proteins, amongst these HUA2, a putative transcription factor [23] and AAF2930 and AAD24842 containing a SET domain and PHD-type zinc fingers as well (Fig. 2). The available *Drosophila* sequence thus far contains two PWWP proteins, amongst which a clear homolog of *WHSC1* possessing two PWWP domains. Remarkably, in double PWWP domain-containing proteins (*Drosophila*, mouse and WHSC1L1), the second domain seems to be stronger conserved than the first PWWP domain, suggesting that the proteins which interact with these regions are different.

6. Conclusions

Our results illustrate that the PWWP domain defines a new structural unit which is found in a wide variety of proteins playing a role in cell division, growth and differentiation. Several of these proteins are linked to cancer and certain diseases or are growth factors. PWWP domain proteins for which data are available appear to be nuclear, often DNA-binding proteins that function as transcription factors regulating developmental processes. Because of its position at either the N- or C-terminus, the composition of amino acids close to the PWWP motif, and the fact that other regions of the protein are responsible for nuclear localization and DNA-binding, we hypothesize its function to be a site for protein–protein interaction, influencing chromatin remodeling and thereby facilitating fine tuning of transcriptional processes. The hypothesized role of the PWWP domain in protein–protein interaction provides a direct handle for future experimentation. Several techniques are available to directly test this role and, e.g. using yeast two-hybrid analysis, to isolate and characterize the interacting partners. These experiments should unravel the biochemical pathways in which the respective proteins are involved, thereby raising our understanding about their precise function. The interactions are potential candidates to be involved in human genetic disease.

Note added in proof:

Garcia et al. (Nucl. Acids Res. 28:1692) just published the *A. thaliana* homolog of the *ATM* gene mutated in ataxia

telangiectasia with homologs reported in yeast, *Drosophila*, *Xenopus* and mouse. Interestingly, only the *ATM* gene of *Arabidopsis* contains an (N-terminal) PWWP domain. Next to the *MSH6* gene, this is the second example where a protein appears to have been N-terminally extended by a PWWP domain, potentially adding a protein-protein interaction site used for fine regulation of its function.

Acknowledgements: We thank Alex Bateman for helpful discussions and critical reading of the paper.

References

[1] Stec, I., Wright, T.J., van Ommen, G.-J.B., de Boer, P.A.J., van Haeringen, A., Moorman, A.F.M., Altherr, M.R. and den Dunnen, J.T. (1998) Hum. Mol. Genet. 7, 1071–1082.
[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[3] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) Nucleic Acids Res. 15, 4876–4882.
[4] Cardoso, C., Timsit, S., Villard, L., Khrestchatisky, M., Fontes, M. and Colleaux, L. (1998) Hum. Mol. Genet. 7, 679–684.
[5] Cui, X., De Vivo, I., Slany, R., Miyamoto, A., Firestein, R. and Cleary, M. (1998) Nat. Genet. 18, 331–337.
[6] Aasland, R., Gibson, T.J. and Stewart, A.F. (1995) Trends Biochem. Sci. 20, 56–59.
[7] Huang, N., vom Baur, E., Garnier, J.M., Lerouge, T., Vonesch, J.L., Lutz, Y., Chambon, P. and Losson, R. (1998) EMBO J. 17, 3398–3412.
[8] Nakamura, H., Izumoto, Y., Kambe, H., Kuroda, T., Mori, T., Kawamura, K., Yamamoto, H. and Kishimoto, T. (1994) Biol. Chem. 269, 25143–25149.
[9] Izumoto, Y., Kuroda, T., Harada, H., Kishimoto, T. and Nakamura, H. (1997) Res. Commun. 238, 26–32.
[10] Ge, H., Si, Y. and Roeder, R.G. (1998) EMBO J. 17, 6723–6729.
[11] Lin, C.-W. (1999) Institute of Biomedical Sciences, Academia Sinica, No. 128, Sec. II, Academia Road, Taipei.
[12] Hateboer, G., Gennissen, A., Ramos, Y.F.M., Kerkhoven, R.M., Sonntag-Buck, V., Stunnenberg, H.G. and Bernards, R. (1995) EMBO J. 14, 3159–3169.
[13] Kurozumi, K., Nishita, M., Yamaguchi, K., Fujita, T., Ueno, N. and Shibuya, H. (1998) Cells 3, 257–264.
[14] Thompson, K.A., Wang, B., Agraves, W.S., Giancotti, F.G., Schranck, D.P. and Ruoslahti, E. (1994) Biochem. Biophys. Res. Commun. 198, 1143–1152.
[15] Palombo, F., Gallinari, P., Iaccarino, I., Lettieri, T., Hughes, M., D'Arrigo, A., Truong, O., Hsuan, J.J. and Jiricny, J. (1995) Science 30, 1912–1914.
[16] Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Yasuno, M., Igari, T., Koike, M., Chiba, M. and Mori, T. (1997) Nat. Genet. 17, 271–272.
[17] Wijnen, J., de Leeuw, W., Vasen, H., van der Klift, H., Moller, P., Stormorken, A., Meijers-Heijboer, H., Lindhout, D., Menko, F., Vossen, S., Moslein, G., Tops, C., Brocker-Vriends, A., Wu, Y., Hofstra, R., Sijmons, R., Cornelisse, C., Morreau, H. and Fodde, R. (1999) Nat. Genet. 23, 142–144.
[18] Edelmann, W., Yang, K., Umar, A., Heyer, J., Lau, K., Fan, K., Liedtke, W., Cohen, P.E., Kane, M.F., Lipford, J.R., Yu, N., Crouse, G.F., Pollard, J.W., Kunkel, T., Lipkin, M., Kolodner, R. and Kucherlapati, R. (1997) Cell 14, 467–477.
[19] Hansen, R.S., Wijmenga, C., Luo, P., Stanek, A.M., Canfield, T.K., Weemaes, C.M. and Gartler, S.M. (1999) Proc. Natl. Acad. Sci. USA 7, 14412–14417.
[20] Orengo, C.A., Pearl, F.M., Bray, J.E., Todd, A.E., Martin, A.C., Lo Conte, L. and Thornton, J.M. (1999) Nucleic Acids Res. 1, 275–279.
[21] Romero, P., Obradovic, Z., Kissinger, K., Villafranca, J.E. and Dunker, A.K. (1997) Int. Conf. Neural Netw. 1, 90–95.
[22] Xie, Q., Arnold, G.E., Romero, P., Obradovic, Z., Garner, E. and Dunker, A.K. (1998) Genome Inform. 9, 193–200.
[23] Chen, X. and Meyerowitz, E.M. (1999) Mol. Cell. 3, 349–360.