doi:10.1006/jmbi.2001.4486 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 307, 31–37

JMB

COMMUNICATION

MH1 Domain of Smad is a Degraded Homing Endonuclease

Nick V. Grishin

Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323, Harry Hines Blvd, Dallas, TX 75390-9050, USA

Smad proteins are eukarytic transcription regulators in the TGF-β signaling cascade. Using a combination of sequence and structure-based analyses, we argue that MH1 domain of Smad is homologous to the diverse His-Me finger endonuclease family enzymes. The similarity is particularly extensive with the I-PpoI endonuclease. In addition to the global fold similarities, both proteins possess a conserved motif of three cysteine residues and one histidine residue which form a zinc-binding site in I-PpoI. Sequence and structure conservation in the motif region strongly suggest that MH1 domain may also incorporate a metal ion in its structural core. MH1 of Smad3 and I-PpoI exhibit similar nucleic acid binding mode and interact with DNA major groove through an antiparallel β-sheet. MH1 is an example of transcription regulator derived from the ancient enzymatic domain that lost its catalytic activity but retained DNA-binding sites.

Keywords: protein structure classification; Cys-His box; HNH motif; His-Me finger endonucleases; dwarfins

The last five years has yielded significant advances in unraveling of the TGF-β signaling cascade.[1,2] In particular, an elegant system that involves Smad proteins (i.e. dwarfins) has been elucidated.[2–5] TGF-β receptors are membrane-bound protein kinases that are activated upon the TGF-β binding and phosphorylate Smad proteins. Smads translocate into the nucleus and act as transcription factors.

Proteins of the Smad family are about 400-500 amino acid residues long and contain two conserved domains connected by a variable length linker. The N-terminal MH1 domain of most Smads binds DNA.[6] The C-terminal MH2 domain is involved in protein-protein interactions.[7] Three subfamilies of Smads have been characterized.[2] (I) Receptor-activated Smad subfamily (R-Smads) incorporates five paralogs in vertebrates, namely Smads 1-3, 5 and 8. Different paralogs are activated (i.e. phosphorylated) by different receptors and participate in different signaling pathways. Phorphorylation of these Smads correlates with their translocation in the nucleus.[8,9] (II) Common Smads (co-Smads) include one known gene in mammals - Smad4[10] and in *Drosophila* - Medea.[11] These proteins lack phosphorylation sites. They bind to phosphorylated R-Smads through the MH2 domains prior to nuclear translocation.[12] (III) Inhibitory Smads (anti-Smads) are the only subfamily that lost their ability to interact with nucleic acids due to insertions/substitutions in the DNA-binding site of the MH1 domain. Anti-Smads function as negative regulators. Smad6 is a decoy of co-Smads that interacts with activated R-Smads[13] and Smad7 blocks activated receptors.[14,15]

The first Smad family member has been characterized from *Drosophila melanogaster* and called ''mothers against *dpp*'' (Mad) to reflect the maternal-effect enhancement of decapentaplegic (*dpp*) mutations in Mad mutants.[16] *Dpp* encodes a TGF-β superfamily growth factor which is implicated in many developmental events.[17] The subsequent analysis of C*aenorhabditis elegans* TGF-β pathways revealed three genes of Mad homologs: *sma-2*, *sma-3*, and *sma-4* (''*sma*'' from ''small'').[18] Among several developmental problems, mutant alleles of the *sma* genes produced small size worms. These Sma proteins were called ''dwar-

Abbreviations used: PDB, protein data bank; RMSD, root-mean-square deviation; Mad, mothers against decapentaplegic; dpp, decapentaplegic; Smad, ''sma''+''Mad''; MH1, Mad homology 1; MH2, Mad homology 2; FHA, forkhead-associated domain.

E-mail address of the corresponding author: grishin@chop.swmed.edu

0022-2836/01/010031–7 $35.00/0

fins'' to avoid confusion with unrelated products encoded by other *sma* genes. Although dwarfin is a more suggestive name for these transcription regulators and it was the name first used,[8,18] ''Smad'' (Smad = sma + Mad) has been widely used instead.[2,19]

Crystal structures for both MH1 and MH2 domains of Smad are available.[20,21] The structure of the MH2 domain suggests its common ancestry (i.e. homology) with FHA domain.[22–24] FHAs comprise a large family of nuclear signaling protein-protein interaction domains present in eukaryots and prokaryots.[25,26] The origin of MH1 domain, however, remains enigmatic. Sequence analysis methods detect MH1 (dwarfin A in SMART[27,28]) only in the proteins from the Smad (=dwarfin) family.[29,30] The spatial structure of MH1 domain (PDB code 1mhd) has been considered unique[21] and classified under a fold of its own in SCOP database (1.53 release).[22,24] Structure similarity search programs such as VAST[31,32] and CE[33] do not find MH1 domain similar to any other protein structures in PDB.[34,35] All Smad proteins are eukaryotic,[27,28,30,36] and prokaryotic homologs of MH1 are not known. To fully understand function of the MH1 domain it is important to trace its evolutionary history.

Through a combination of sequence and structure-based analyses we argue that the MH1 domain is homologous to the diverse His-Me finger homing endonuclease family enzymes[22,24] that are present in all kingdoms of life. The structural link with homing endonuclease was hinted by the DALI structure similarity search program[37,38] that employs comparison of distance matrices. DALI finds several protein structures to be similar to MH1 domain. However, only one structure match, namely with intron-encoded endonuclease I-PpoI (PDB entry 1a73), covers significant fraction of both molecules (63 % of 1mhd length and 48 % of 1a73 length). DALI superimposes C^{α} atoms of 78 residues in 1mhd and 1a73 with RMSD of 3.3 Å and *Z*-score of 2.7. The resulting structure-based sequence alignment exhibits 16 % of identity. Likewise, the DALI search started with I-PpoI (1a73) finds only a single structural neighbor: MH1 domain (1mhd) of Smad. Since the Smad MH1 and I-PpoI structures where published within the same year,[21,39] the similarity between them simply could not be noted in the original publications that describe the structure solution and thus it remained unnoticed. To evaluate the DALI results, comprehensive sequence-structure-functional comparison between the MH1 domain and I-PpoI endonuclease has been undertaken.

I-PpoI belongs to the Cys-His box subfamily of homing nucleases.[40] Its structure has been determined in complex with DNA (PDB entry 1a73).[39] The enzyme is composed of three subdomains, two of which have structural equivalents in MH1 Smad (Figure 1(a), blue/yellow and purple/green). The functional segment of the first subdomain is a three-stranded β-sheet cde that binds in the major groove of DNA. The turn between β-strands c and d incorporates active site Arg61.[41,42] The core of the first subdomain is completed by two α-helices A and B. The second subdomain is folded into a left-handed ββα unit with a long Ω-loop between the β-strands f and h. Part of the Ω-loop is in extended conformation and is structured as a β-strand g. This subdomain is present in several related families of intein endonucleases[41-47] and contains catalytic His98.[39] The edge of the β-strand f and an α-helix C form contacts with DNA in the minor groove. The endonuclease active site is situated in the cleft between the two subdomains and substrate DNA chain fits into the cleft. The core of the I-PpoI molecule is unusual, since it is structured around a Zn^{2+}.[39] Three of the zinc ligands are contributed by the Ω-loop and the fourth is donated by a long twisted β-hairpin ab inserted in the N-subdomain. This family of nucleases was termed Cys-His-box due to the presence of the conserved zinc ligands.[40]

From the global structural comparison (Figure 1(a) and (b)), it is clear that MH1 and I-PpoI possess the same fold, since they have the same secondary structural elements in the same spatial arrangement and with the same topological connections. Indeed, out of 11 abovementioned regular secondary structural elements of I-PpoI, nine have structural equivalencies in MH1 domain (Figure 1(a) and (b)). The first subdomain contains a reduced version of the three-stranded β-sheet cde and two α-helices A and B. The β-sheet (β-hairpin de in particular) fits into the DNA major groove.[21] The second subdomain consists of the β-Ω-loop-β unit, which ends the region of MH1 with determined X-ray structure.[21] The C-terminal α-helix C is not covered by this region. The second subdomain of MH1 is not close to DNA since the DNA segment present in the crystal is too short. Endonuclease active site residues His, Arg, and Asn are not present in MH1. Arg61 and Asn123 of I-PpoI are mapped to the regions lacking in MH1 structure (Figure 1(c)) and His98 aligns with Ala107. Despite the absence of these active site residues, DNA-binding modes of I-PpoI and MH1 are similar (Figure 1(a) and (b)). The β-hairpin de binds to the DNA major groove in both structures.

Most importantly, however, the three Zn^{2+}-binding residues in the Ω-loop regions are conserved between MH1 and I-PpoI sequences (Figure 1(a) to (c)). Moreover, the cysteine residue from the inserted β-hairpin ab in I-PpoI structure superimposes with MH1 cysteine placed in a structurally equivalent insertion A′b. Such conservation of the three cysteine residues and a histidine residue is surprising, since metal ion is not modeled in the MH1 structure[21] and zinc or other structure-stabilizing metal ions were not mentioned in experimental studies of Smad proteins. To address the question about the potential metal binding site in MH1 domain of Smads, local structural similarity around the zinc-binding site of I-PpoI was examined. Local superposition of the zinc-binding site

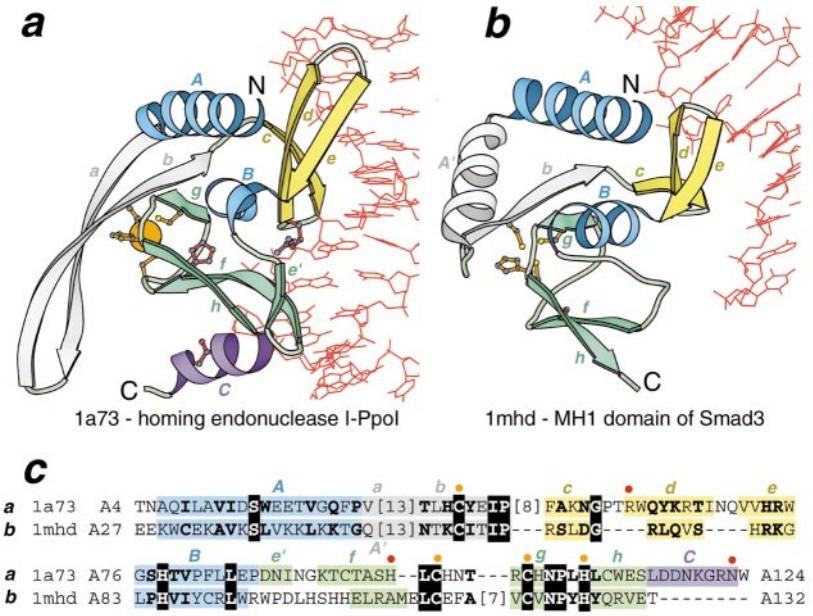

Figure 1. Global structural similarity between homing endonuclease and MH1 domain of Smad. The ribbon diagrams of (a) homing endonuclease I-PpoI from *Physarum polycephalum* (PDB entry 1a73, residues A7-A125) and (b) human Smad3 MH1 domain (PDB entry 1mhd, residues A29-A132) in complex with DNA were drawn by Bobscript,[57] a modified version of MOLSCRIPT.[58] The structures were superimposed and then separated for clarity. N and C termini are labeled. The spatially equivalent structural elements are colored correspondingly in the two structures. α-Helices/β-strands in the N and C-terminal subdomains are colored in blue/yellow, and in purple/green, respectively. Insertion in the N-terminal domain is shown in gray. The DNA chains are red. The side-chains of active site residues (red) and zinc ligands (orange) in homing endonuclease and corresponding residues in MH1 Smad are shown in ball-and-stick presentation. Zinc ion is shown as orange ball. (c) Structure-based sequence alignment of endonuclease I-PpoI (1a73) and Smad3 MH1 (1mhd) generated by DALI[37,38] and modified manually. The panel label, PDB entry name, starting and ending residue numbers are given for each protein. Invariant residues are boxed with black and conserved substitutions are shown in bold letters. The numbers of residues omitted from the alignment are shown in brackets. Color shading and labels of secondary structure elements correspond to those shown in (a) and (b). The active site residues and zinc ligands in endonuclease are marked above the alignment with red and orange dots, respectively, and their side-chains are displayed on the (a) and (b).

of I-PpoI with the structurally equivalent regions in MH1 revealed a very close match in main chain conformations: RMSD of 1.05 Å for 152 backbone atoms of 19 residues (Figure 2(a) to (c)). The region covered by the superposition includes not only the two β-strands of the hairpin fh (Figure 1(a) and (b)), but also the Ω-loop with the short inserted β-strand g that forms hydrogen bonds with the β-strand b. This particular structural unit does not occur frequently in proteins. Moreover, the conformations of conserved cysteine and histidine side-chains are very similar between the two structures (Figure 1(a)-(c)), and the four residues with metal-chelating properties in MH1 domain are arranged to form a potential metal-binding site. Additionally, the distance geometry of C109 and C121 in 1mhd deviates from ideal, for example in C109, C^{β}-S^{γ} distance is 1.91 instead of being close to 1.82. Furthermore, the PDB header record (1mhd) shows the presence of two disulfide bonds: C64-C109 and C109-C121. It appears that C109 is simultaneously involved in two different disulfide bonds, which is not possible. It should be noted, however, that Shi *et al.*[21] did not assign any disulfide bonds to the MH1 domain and the assignments in the PDB header are done automatically on the basis of short distances between the corresponding sulfur atoms. In summary, structural comparison of MH1 and I-PpoI and deviations from ideal geometry around the potential metal-binding site in MH1 domain, allow us to speculate that a metal ion may be present in the MH1 molecules, but additional experimental data are needed to clarify the question.

Residue conformations in MH1 and I-PpoI near the zinc-binding site are very similar (Figure 2(a)-(c)). To approach the question if this structural similarity is reflected in the amino acid sequences, multiple alignment of these regions was constructed (Figure 2(d)). The sequences of Smad and endonuclease homologs were retrieved in iterative PSI-BLAST and PHI-BLAST searches.[29,43–45] The searches were performed against the non-redundant protein sequence database (nr, Oct 17, 2000, 574,979 sequences;

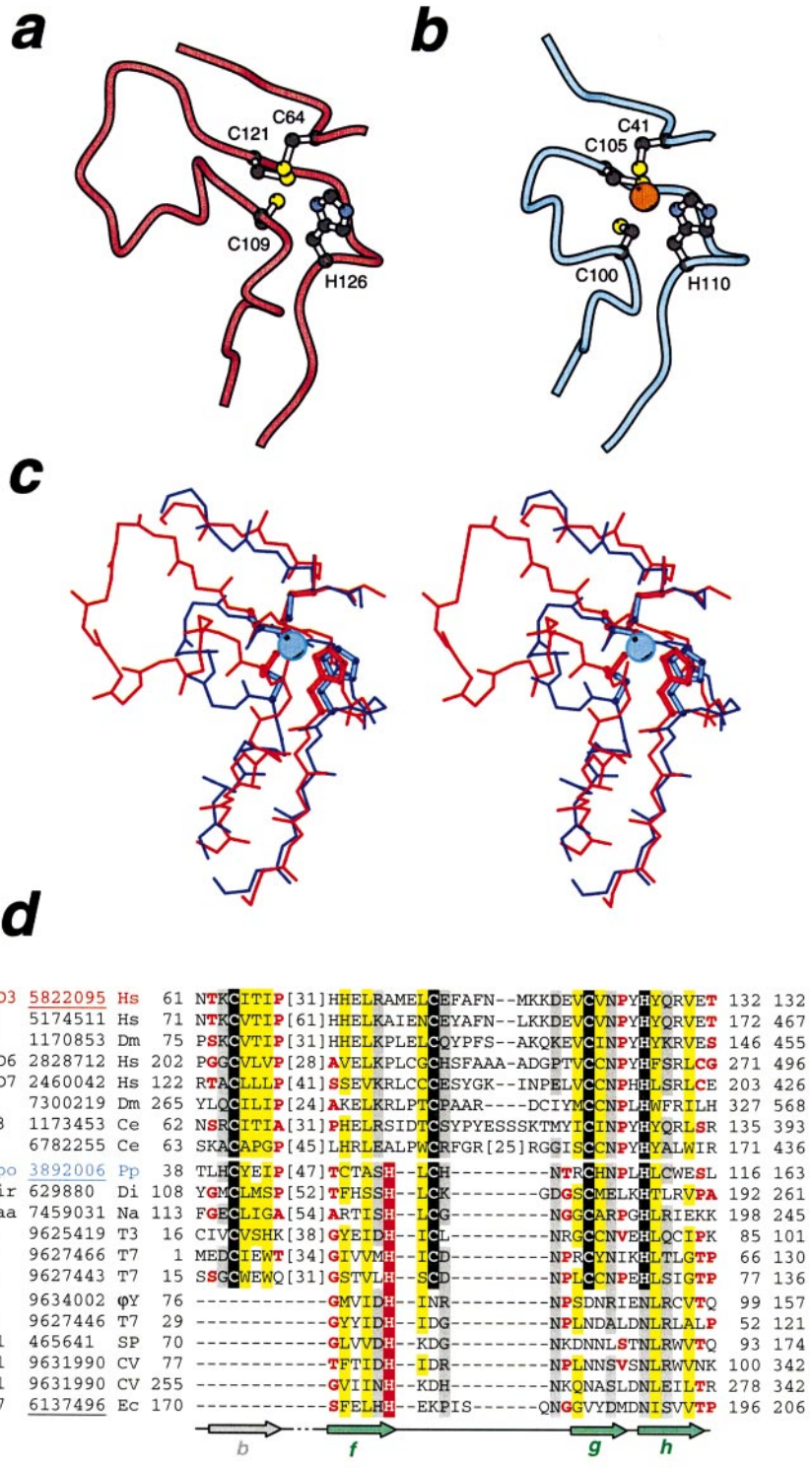

Figure 2. Zinc binding site in MH domain of MAD and homing endonucleases. The ribbon diagrams of a zinc binding site (Cys-His box) in (a) human Smad3 MH1 domain (PDB entry 1mhd, residues A63-A68, A102-A130) and (b) homing endonuclease I-PpoI from *Physarum polycephalum* (PDB entry 1a73, residues A40-A45, A95-A114). Zinc ligands in homing endonuclease and their structural equivalents in MH1 domain are labeled and shown in ball-and-stick presentation. Zn^{2+} is shown as orange ball. (c) The stereo diagram of superimposed Zn-binding sites of MH1 (red) and endonuclease (blue). Protein backbones, side chains of zinc ligands, and Zn^{2+} are shown. Superposition was performed using InsightII package (MSI Inc) according to the DALI alignment.[37,38] A total of 152 backbone atoms of 19 residues from the two molecules superimpose with RMSD of 1.05 Å; the following segments from chains A were superimposed: 1a73, 40-43, 96-99, 104-114 *versus* 63-66, 103-106, 120-130 of 1mhd). (d) Sequence alignment of the Cys-His box region in representative sequences of Smad MH1 (top) and HNH endonuclease families (bottom). Protein name, gene identification (GI) number of the NCBI/Genbank protein sequence database, organism abbreviation, first and the last residue numbers and the total number of residues are shown for each sequence. Names for Smad3 MH1 (PDB 1mhd) and endonuclease I-PpoI (PDB 1a73) are shown in red and blue, respectively. GIs of the sequences with known spatial structure are underlined: 6137496 corresponds to PDB entry 7cei, chain B. Some name abbreviations: HP, hypothetical protein; CoE7, endonuclease domain of colicin E7. The species name abbreviations are: Ce, *Caenorhabditis elegans*; CV, *Paramecium bursaria Chlorella* virus 1; Di, *Didymium iridis*; Dm, *Drosophila melanogaster*; Ec, *Escherichia coli*; Hs, *Homo sapiens*; Ni, *Naegleria andersoni*; φY, bacteriophage phi-YeO3-12; Pp, *Physarum polycephalum*; SP, bacteriophage SPO1; T3, bacteriophage T3; T7, bacteriophage T7. Only sequence segments that are in close proximity for the zinc-binding site are shown. Long insertions are not displayed: the number of omitted residues is specified in brackets. Potential zinc ligands are boxed with black, active-site histidine residue in endonucleases is boxed with red, the uncharged residues (all amino acids except D,E,K,R) in mostly hydrophobic sites are highlighted in yellow, the non-hydrophobic residues (all amino acids except W,F,Y,M,L,I,V) at mostly hydrophilic sites are highlighted in light gray, and the small residues (G,P,A,S,C,T,V) at positions occupied by mostly small residues are shown in red letters. Secondary structure consensus is shown below the alignment. The β-strands are displayed as arrows colored and labeled according to the scheme from the Figure 1.

180,825,488 total letters) maintained at the National Center for Biotechnology Information using different sequences as queries. The parameters used were: an *E*-value threshold of 0.01, BLOSUM62 matrix,[46] no low complexity filtering[47,48] and no composition-based statistics.[49] The representative sequences from the MH1 Smad family include not only DNA-binding Smads, but also anti-Smads, such as Smad 6 and 7 (Figure 2(d)). The potential zinc ligands that comprise the core of the MH1 domain are invariant in all of the family members and demonstrate that the MH1 domain is conserved in anti-Smads as well. Endonuclease sequences constitute two groups. The first group contains Cys-His box motif[40] that should be able to bind zinc. The I-PpoI protein belongs to this group. The members of the second group lack the zinc ligands and are known as the NHN family.[50,51] NHN group is typified by the *Escherichia coli* DNase domain of colicin E7 with known structure[52] (PDB entry 7cai, Figure 2d, CoE7). The catalytic histidine residue is invariant in all active endonucleases (Figure 2(d)). Bacteriophage T sequences Y53, Y77 and Y28 (Figure 2(d)) with Cys-His box motif were attributed to the NHN family previously.[51] Thus the Cys-His box motif sequences are a subfamily within the NHN family.[53–55] Due to the absence of a cysteine residue and significant sequence divergence, detection of the β-strand b in the proteins that lack Cys-His box motif is challenging, and this region is not shown in the alignment (Figure 2(d)). Comparison of the MH1 and endonuclease families reveals conservation of the Cys-His box motif that is not limited to the invariant zinc ligands. The patterns of hydrophobicity and the distribution of small residues are conserved as well (Figure 2(d)). Sequence similarity between the families is the strongest in the gh region (Figure 2(d)). For example, the eight residue string CCNPHHLS in human Smad7 corresponds to the string CCNPEHLS in Y28 bacteriophage T7 endonuclease with a single mismatch. Thus the sequences of MH1 and endonucleases show similarity in the Cys-His box region.

The overall fold resemblance, significant local structural match near the core of both molecules assembled around Zn^{2+}, local sequence similarity in the zinc-binding site region combined with functional similarity in DNA-binding modes strongly suggest that I-PpoI and MH1 shared a common ancestor and are homologous.[23] His-Me endonucleases[22,24] and MH1 domain of Smads (dwarfin A) should be classified within the same superfamily. Since most members of this superfamily are endonucleases that are present in all major phylogenetic lineages and MH1 domains are exclusively eukaryotic, it is likely that MH1 is a modified endonuclease that was recruited as a transcription regulator. Thus MH1 of Smad represents another example of an ancient enzymatic domain that lost its catalytic activity and functions as a transcription factor in eukaryots.[56]

Acknowledgments

The author is grateful to Hong Zhang for critical reading of the manuscript and helpful comments.

References

1. Massague, J. (1998). TGF-beta signal transduction. *Annu. Rev. Biochem.* **67**, 753-791.
2. Massague, J. & Wotton, D. (2000). Transcriptional control by the TGF-beta/Smad signaling system. *EMBO J.* **19**, 1745-1754.
3. Heldin, C. H., Miyazono, K. & ten Dijke, P. (1997). TGF-beta signaling from cell membrane to nucleus through SMAD proteins. *Nature,* **390**, 465-471.
4. Kretzschmar, M. & Massague, J. (1998). SMADs: mediators and regulators of TGF-beta signaling. *Curr. Opin. Genet. Dev.* **8**, 103-111.
5. Attisano, L. & Wrana, J. L. (1998). Mads and Smads in TGF beta signaling. *Curr. Opin. Cell. Biol.* **10**, 188-194.
6. Kim, J., Johnson, K., Chen, H. J., Carroll, S. & Laughon, A. (1997). *Drosophila* Mad binds to DNA and directly mediates activation of vestigial by Decapentaplegic. *Nature,* **388**, 304-308.
7. Wu, R. Y., Zhang, Y., Feng, X. H. & Derynck, R. (1997). Heteromeric and homomeric interactions correlate with signaling activity and functional cooperativity of Smad3 and Smad4/DPC4. *Mol. Cell. Biol.* **17**, 2521-2528.
8. Yingling, J. M., Das, P., Savage, C., Zhang, M., Padgett, R. W. & Wang, X. F. (1996). Mammalian dwarfins are phosphorylated in response to transforming growth factor beta and are implicated in control of cell growth. *Proc. Natl Acad. Sci. USA,* **93**, 8940-8944.
9. Kretzschmar, M., Liu, F., Hata, A., Doody, J. & Massague, J. (1997). The TGF-beta family mediator Smad1 is phosphorylated directly and activated functionally by the BMP receptor kinase. *Genes Dev.* **11**, 984-995.
10. Hahn, S. A., Schutte, M. & Hocque, A. T. *et al.* (1996). DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science,* **271**, 350-353.
11. Wisotzkey, R. G., Mehra, A., Sutherland, D. J., Dobens, L. L., Liu, X., Dohrmann, C., Attisano, L. & Raftery, L. A. (1998). Medea is a *Drosophila* Smad4 homolog that is differentially required to potentiate DPP responses. *Development,* **125**, 1433-1445.
12. Lagna, G., Hata, A., Hemmati-Brivanlou, A. & Massague, J. (1996). Partnership between DPC4 and SMAD proteins in TGF-beta signaling pathways. *Nature,* **383**, 832-836.
13. Hata, A., Lagna, G., Massague, J. & Hemmati-Brivanlou, A. (1998). Smad6 inhibits BMP/Smad1 signaling by specifically competing with the Smad4 tumor suppressor. *Genes Dev.* **12**, 186-197.
14. Hayashi, H., Abdollah, S. & Qiu, Y. *et al.* (1997). The MAD-related protein Smad7 associates with the TGF-beta receptor and functions as an antagonist of TGF-beta signaling. *Cell,* **89**, 1165-1173.
15. Nakao, A., Afrakhte, M. & Moren, A. *et al.* (1997). Identification of Smad7, a TGFbeta-inducible antagonist of TGF-beta signaling. *Nature,* **389**, 631-635.
16. Sekelsky, J. J., Newfeld, S. J., Raftery, L. A., Chartoff, E. H. & Gelbart, W. M. (1995). Genetic

characterization and cloning of mothers against dpp, a gene required for decapentaplegic function in Drosophila melanogaster. *Genetics,* **139**, 1347-1358.
17. Raftery, L. A. & Sutherland, D. J. (1999). TGF-beta family signal transduction in *Drosophila* development: from Mad to Smads. *Dev. Biol.* **210**, 251-268.
18. Savage, C., Das, P., Finelli, A. L., Townsend, S. R., Sun, C. Y., Baird, S. E. & Padgett, R. W. (1996). *Caenorhabditis elegans* genes sma-2, sma-3, and sma-4 define a conserved family of transforming growth factor beta pathway components. *Proc. Natl Acad. Sci. USA,* **93**, 790-794.
19. Liu, F., Hata, A., Baker, J. C., Doody, J., Carcamo, J., Harland, R. M. & Massague, J. (1996). A human Mad protein acting as a BMP-regulated transcriptional activator. *Nature,* **381**, 620-623.
20. Shi, Y., Hata, A., Lo, R. S., Massague, J. & Pavletich, N. P. (1997). A structural basis for mutational inactivation of the tumour suppressor Smad4. *Nature,* **388**, 87-93.
21. Shi, Y., Wang, Y. F., Jayaraman, L., Yang, H., Massague, J. & Pavletich, N. P. (1998). Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. *Cell,* **94**, 585-594.
22. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
23. Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380-387.
24. Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **28**, 257-259.
25. Hofmann, K. & Bucher, P. (1995). The FHA domain: a putative nuclear signaling domain found in protein kinases and transcription factors. *Trends Biochem. Sci.* **20**, 347-349.
26. Wang, P., Byeon, I. J., Liao, H., Beebe, K. D., Yongkiettrakul, S., Pei, D. & Tsai, M. D. (2000). II. Structure and specificity of the interaction between the FHA2 domain of rad53 and phosphotyrosyl peptidesdagger. *J. Mol. Biol.* **302**, 927-940.
27. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA,* **95**, 5857-5864.
28. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucl. Acids Res.* **28**, 231-234.
29. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
30. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.
31. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377-385.
32. Wang, Y., Addess, K. J., Geer, L., Madej, T., Marchler-Bauer, A., Zimmerman, D. & Bryant, S. H. (2000). MMDB: 3D structure data in Entrez. *Nucl. Acids Res.* **28**, 243-245.
33. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739-747.
34. Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol.* **277**, 556-571.
35. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.
36. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260-262.
37. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.
38. Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.* **25**, 231-234.
39. Flick, K. E., Jurica, M. S., Monnat, R. J., Jr & Stoddard, B. L. (1998). DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. *Nature,* **394**, 96-101.
40. Johansen, S., Embley, T. M. & Willassen, N. P. (1993). A family of nuclear homing endonucleases. *Nucl. Acids Res.* **21**, 4405.
41. Friedhoff, P., Franke, I., Meiss, G., Wende, W., Krause, K. L. & Pingoud, A. (1999). A similar active site for non-specific and specific endonucleases. *Nature Struct. Biol.* **6**, 112-113.
42. Friedhoff, P., Franke, I., Krause, K. L. & Pingoud, A. (1999). Cleavage experiments with deoxythymidine 3′,5′-bis-(p-nitrophenyl phosphate) suggest that the homing endonuclease I-PpoI follows the same mechanism of phosphodiester bond hydrolysis as the non-specific Serratia nuclease. *FEBS Letters,* **443**, 209-214.
43. Altschul, S. F. & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444-447.
44. Aravind, L. & Koonin, E. V. (1999). Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287**, 1023-1040.
45. Zhang, Z., Schaffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucl. Acids Res.* **26**, 3986-3990.
46. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.
47. Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269-285.
48. Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.
49. Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. & Altschul, S. F. (1999). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics,* **15**, 1000-1011.
50. Gorbalenya, A. E. (1994). Self-splicing group I and group II introns encode homologous (putative)

DNA endonucleases of a new family. *Protein Sci.* **3**, 1117-1120.
51. Shub, D. A., Goodrich-Blair, H. & Eddy, S. R. (1994). Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends Biochem. Sci.* **19**, 402-404.
52. Ko, T. P., Liao, C. C., Ku, W. Y., Chak, K. F. & Yuan, H. S. (1999). The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. *Structure Fold. Des.* **7**, 91-102.
53. Dalgaard, J. Z., Klar, A. J., Moser, M. J., Holley, W. R., Chatterjee, A. & Mian, I. S. (1997). Statistical modeling and analysis of the LAGLIDADG family of site- specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucl. Acids Res.* **25**, 4626-4638.
54. Kuhlmann, U. C., Moore, G. R., James, R., Kleanthous, C. & Hemmings, A. M. (1999). Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined? *FEBS Letters,* **463**, 1-2.
55. Aravind, L., Makarova, K. S. & Koonin, E. V. (2000). Survey and summary: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucl. Acids Res.* **28**, 3417-3432.
56. Aravind, L. & Koonin, E. V. (1998). Eukaryotic transcription regulators derive from ancient enzymatic domains. *Curr. Biol.* **8**, R111-R113.
57. Esnouf, R. M. (1997). An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* **15**, 133-138.
58. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Edited by J. Thornton

(Received 25 October 2000; received in revised form 19 January 2001; accepted 22 January 2001)