

Chapter 1

MOLECULAR EVOLUTION OF SMAD PROTEINS

STUART J. NEWFELD¹ AND ROBERT G. WISOTZKEY²

¹*School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501, USA,* ²*Ingenuity Systems, Redwood City, CA, USA*

Abstract: To date, Smad family members have been found only in eumetazoan animals. To understand the evolutionary relationship between family members we conducted a phylogenetic analysis. To simplify the analysis but retain its explanatory power, we focused on Smad proteins from organisms in three distinct phyla: human, fly, and nematode. Overall, we found that human and fly proteins always cluster together in four subfamilies while three subfamilies contain only nematode proteins. Sequence alignments of distinct regions of were also analyzed. Data from the alignments confirmed that the MH1 (DNA-binding) and MH2 (protein-protein interaction) domains are highly conserved family-wide. The linker region between these domains is also highly conserved but only within subfamilies. Conservation in the C-terminal receptor phosphorylation region provides new insight into a unique subfamily containing three interacting nematode proteins that signal for DAF-7. From a larger perspective, our analysis strongly supports the traditional view that flies are more closely related to humans than to nematodes.

Key words: multigene family; SMAD proteins; phylogeny; amino acid alignments; evolutionary conservation; developmental-evolution; signal transduction.

1. INTRODUCTION

The evolutionary relationships between members of a multigene family are ascertained through a phylogenetic analysis involving three steps. First, one must calculate the amount of amino acid similarity between each family member by aligning the protein sequences (Thompson et al., 1997). Second, one applies an amino acid similarity matrix and the extent of similarity between each protein and all of the others are prioritized with the most similar proteins clustered together. These clusters are depicted as the familiar

phylogenetic tree (Kumar et al., 2001). Third, the relationships between pairs of proteins are tested for robustness using statistical methods such as bootstrap analysis (Felsenstein, 1985).

Here we describe a new phylogenetic analysis of the Smad protein family. In order to simplify the analysis but retain its explanatory power, we focus on Smad sequences from organisms in three distinct phyla: human (deuterostome), fly (protostome) and nematode (pseudocoelomate; see Raff 1996, for example, for a taxonomic description of these phyla). We include other species as necessary to add confidence to individual results. Our studies of the MH1 and MH2 domains support the long-standing view that they are highly conserved. Our analysis of the linker region between the MH1 and MH2 domains, previously dismissed as a highly divergent and potentially non-functional part of the protein, reveals surprising levels of sequence conservation within Smad subfamilies. This suggests the hypothesis that distinct functions associated with each subfamily involve linker sequences. An analysis of the receptor phosphorylation domain provides new insights into a unique subfamily containing only nematode proteins that signal for the TGF- β /Activin subfamily member DAF-7.

Recently it has become possible to test phylogenetically derived hypotheses using an approach known as functional genomics. In this technique, interspecies experiments are conducted that evaluate the ability of a family member from one species to mimic the activity of another family member either by rescuing mutant phenotypes (*e.g.* Padgett et al., 1993) or in parallel over-expression experiments (*e.g.* Marquez et al., 2001). We have conducted a number of such tests and review those results here.

2. SMAD FAMILY MEMBERS

To date, Smad family members have been found only in animals. Within the animal kingdom they have been identified in eumetazoans (multicellular organisms with many types of cells) but not yet in metazoans such as sponges (multicellular organisms with very few cell types). However, several transmembrane receptors with similarity to both type I and type II TGF- β receptors have been identified in a freshwater sponge (Suga et al., 1999). A phylogenetic analysis showed that the sponge receptors are very similar to the unusual *C. elegans* receptors DAF-1 and SMA-6 that also fall between receptor types (Herpin et al 2004). The similarity between sponge and nematode receptors suggests that Smad-like proteins will eventually be found in sponges. Thus, ancestral TGF- β family members and their signaling pathways predate the metazoan/eumetazoan divergence roughly 1.5 billion years ago (Hedges and Kumar, 2003).

The simplest eumetazoans with definitive Smad family members are cnidarians (animals with two germ layers - diploblasts). A sequence similar to Smad1/Mad in the BMP signaling subfamily has been identified in coral (Samuel et al., 2001) and in hydra (Hobmayer et al., 2001). The simplest eumetazoan with Smad proteins similar to both Smad1/Mad and Smad2/3 is the blood fluke *Schistosoma mansoni* - an acoelomate with three germ layers but no digestive cavity (Beall et al., 2000). From this it is reasonable to conclude that BMP signaling Smads, and by extension their cognate ligands and receptors, represent the oldest of the TGF- β pathways found in higher animals such as flies and mammals.

Nevertheless, one word of caution: gene discovery in simple organisms is not always simple. Insuring that DNA samples are free from contamination from higher organisms is difficult. For example, parasites like *Schistosoma* may be contaminated with human white blood cells or cnidarians may contain shrimp larvae from their last meal. Reproducibility is essential to insuring confidence in these studies.

In order to achieve easily interpretable results but to maintain maximum confidence in our phylogenetic analysis, we focused on three species with fully sequenced genomes. These species belong to three distinct phyla allowing us maximum discriminatory power in the analysis. Humans (deuterostome) and the fruit fly *D. melanogaster* (protostome) belong to sister taxa at the top of the animal kingdom. They are coelomates - animals with three germ layers and a digestive tract with two openings. Our third species is the nematode *C. elegans* (a pseudocoelomate - animals with three germ layers and a digestive cavity with only one opening). *C. elegans* is the simplest organism with a full set of Smad proteins (R-Smad, Co-Smad and I-Smad subfamilies; Newfeld et al., 1999). Molecular evolution studies indicate that the split between deuterostomes and protostomes occurred 990 million years ago and the split between coelomates and pseudocoelomates occurred 1.2 billion years ago (Hedges and Kumar, 2003). Any amino acids conserved over this enormous span of time are clearly subject to strong positive selection that is most likely due to an essential role in either protein structure or function.

Table 1 describes the 19 Smad sequences we examined. These sequences were utilized for the phylogeny (Fig. 1) and for the MH1, MH2 and receptor phosphorylation domain alignments (Figs. 2, 3 and 5). There are eight Smad proteins in humans (hSmads). hSmad1, hSmad5 and hSmad8 (also known as Smad9 in the Entrez Gene database) transduce DPP/BMP subfamily signals. hSmad2 and hSmad3 transduce TGF- β /Activin subfamily signals. hSmad4

Table 1. Representative Smad Family Members^a

Entrez Gene Symbol	GeneID	Reference DNA	Sequence Protein	Synonyms ^b
Human				
SMAD1	4086	NM_001003688	NP_001003688.1	JV4-1, MADH1
SMAD2	4087	NM_001003652	NP_001003652.1	JV18-1, MADH2
SMAD3	4088	NM_005902	NP_005893.1	MADH3, JV15-2
SMAD4	4089	NM_005359	NP_005350.1	DPC4, MADH4
SMAD5	4090	NM_001001419	NP_001001419.1	JV5-1, MADH5
SMAD6	4091	NM_005585	NP_005576.2	MADH6, MADH7
SMAD7	4092	NM_005904	NP_005895.1	MADH7, MADH8
SMAD9	4093	NM_005905	NP_005896.1	SMAD8 , MADH9
Drosophila melanogaster				
DAD	42059	NM_057912	NP_477260	CG5201, EP3196
MAD	33529	NM_057669	NP_477017	CG12399, K00237
MEDEA ^c	43725	NM_079871 NM_170559	NP_524610.1 NP_733438	CG1775 l(3)SG70
SMOX	31738	NM_078524	NP_511079.1	SMAD2 , CG2262
Caenorhabditis elegans				
1J160	187612	NM_059920	NP_492321.1	DAF-8 , R05D11.1
1L81	172930	NM_060345	NP_492746.1	F37D6.6
DAF-14	177908B	NM_069479	NP_501880.1	F01G10.8
DAF-3	180431	NM_001029434	NP_508161.2	F25E2.5
SMA-2	176229	NM_066530	NP_498931	ZK370.2, Cem-1
SMA-3	175955	NM_066092	NP_498493	R13F6.9, Cem-2
SMA-4	175815	NM_065855	NP_498256	R12B2.1, Cem-3
Danio rerio				
SMAD1	30628	NM_131356	NP_571431	madh1, fb39h09
SMAD2	30639	NM_131366	BC044338	madh2, fj43c06
SMAD3a	50892	NM_131571	NP_571646	madh3, smad3
SMAD3b	326283	NM_175083	NP_778258	madh3b, gc92234
SMAD4 ^d	30640	XM_694498	XP_699590	madh4
SMAD5	30641	NM_131368	NP_571443	madh5, fb67b04
SMAD7	326282	NM_175082	NP_778257	madh7
Xenopus tropicalis				
SMAD1	493211	NM_001007480	NP_001007481	MGC89254
Xenopus laevis				
SMAD3	378633	AJ311059	CAC38118	madh3-A
SMAD7	394331	AF026125	AAC17489	madh7, smad10
Anopheles gambiae				
ENSANGG000 00003874	127- 5428	XM_314661	XP_314661	AgMAD
SMAD2	-	AY578801	AAT07306	
ENSANGG000 00016187 ^c	127- 3058	XM_312001 XM_311999	XP_312001 XP_311999	AgMEDEA
Apis mellifera				
LOC409301	409301	XM_392819	XP_392819	AmMAD
LOC412601	412601	XM_396056	XP_396056	AmSMAD2
LOC409321	409321	XM_392838	XP_392838	AmMEDEA
LOC413371	413371	XM_396816	XP_396816	AmDAD

a. Entrez Gene = www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene.

b. Synonyms in bold are utilized in this chapter.

c. D. melanogaster and A. gambiae MEDEA have two splice variants.

d. D. rerio SMAD4 reference sequences are unofficial. LOC560317 also has regions of identity with human SMAD4.

participates with the other Smads to transduce signals of both subfamilies. hSmad6 and hSmad7 antagonize signals of both subfamilies (reviewed in Massagué et al., 2000).

There are four Smad proteins in *Drosophila melanogaster* (DmSmads). Mothers Against Dpp (Mad) transduces Dpp signals. DmSmad2 (also known as SMOX in the Entrez Gene database) transduces DmActivin signals. MEDEA (MED) participates with the other Smads to transduce signals of both subfamilies. Daughters Against Dpp (DAD) antagonizes DPP and possibly DmActivin signals (reviewed in Raftery and Sutherland, 1999).

There are seven Smads in the *Caenorhabditis elegans* (CeSmads). SMA-2, SMA-3 and SMA-4 transduce DBL-1 (a BMP subfamily member) signals. DAF-14 and DAF-8 (also known as Ce1J160 in the Entrez Gene database) transduce DAF-7 (a TGF- β /Activin subfamily member) signals. DAF-3 antagonizes DAF-7 signals. Ce1L81 is a predicted open reading frame that has not yet been assigned to a gene (reviewed in Inoue and Thomas, 2000).

As part of this study we conducted the first detailed analysis of sequences in the linker region of Smad family members. Perhaps as an artifact of our inability to align this region of CEM-1, CEM-2 and CEM-3 (*C. elegans* Mad-like genes) with Mad, this domain appeared to us as a highly divergent stretch that was easily dismissed as non-functional (Sekelsky et al., 1995). We now know that CEM-1 (SMA-2), CEM-2 (SMA-3), CEM-3 (SMA-4) and Mad all belong to distinct subfamilies of the Smad family (discussed below). We were able to align the linker region from sequences belonging to the same subfamily and found surprisingly high levels of conservation (Fig. 4). However, no subfamily contains more than four sequences. Therefore, to add confidence to our linker region alignments we added sequences from three vertebrate species (two frogs and zebrafish) and two insect species (mosquito and honey bee) as described in Table 1.

3. SMAD FAMILY TREE

Fig. 1 shows a phylogenetic tree consistent with previous reports (e.g. Newfeld et al., 1999) that there are four distinct subfamilies of Smads. Clusters of sequences corresponding to two subfamilies of R-Smads, a subfamily of Co-Smads and a subfamily of I-Smads are observed. However, seven subfamilies are present overall. Human and fly genes always cluster together in the four known subfamilies while three subfamilies contain only nematode sequences. Further, human and fly proteins belonging to the same subfamily have been shown to function similarly in transgenic experiments (Marquez et al., 2001).

The Smad1/Mad subfamily contains signal transducing R-Smads dedicated to DPP/BMP subfamily ligands. The Smad2/3 subfamily contains signal transducing R-Smads dedicated to TGF- β /Activin subfamily ligands. The hSmad4/MED subfamily contains signal transducing Co-Smads that form complexes with R-Smads of both subfamilies. One nematode protein (SMA-4) also belongs to this subfamily. The hSmad6/7/DAD subfamily contains I-Smads that antagonize signal transduction by R-Smads of both subfamilies. One nematode sequence (Ce1L81) belongs to the I-Smad subfamily.

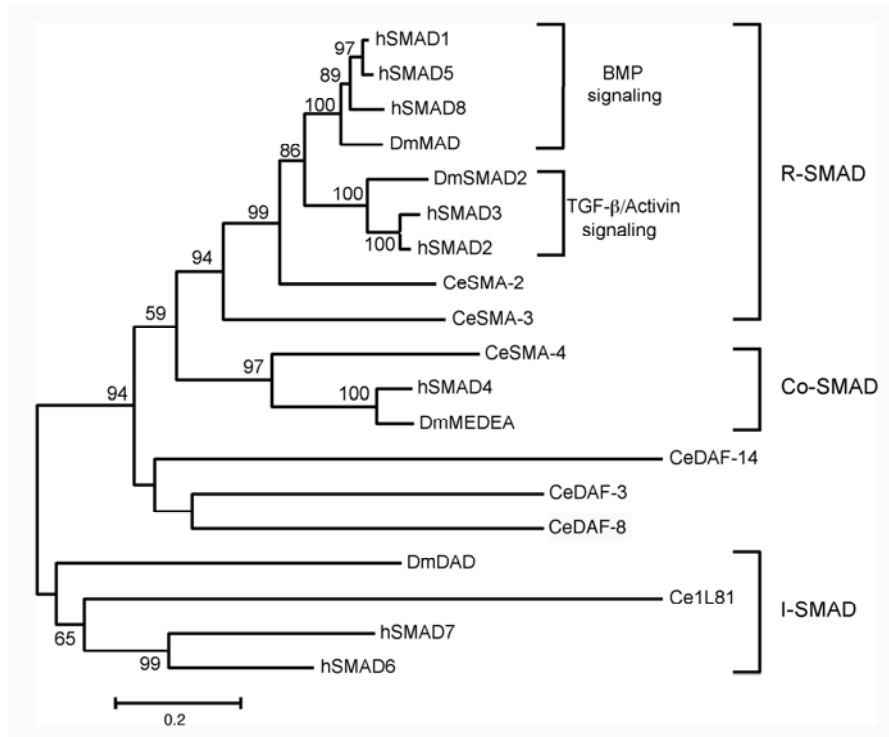


Figure 1. Phylogenetic analysis of the Smad family. Note that human and fly genes cluster together into four major subfamilies: 1) Receptor-associated Smads involved in signaling by DPP/BMP proteins, 2) Receptor-associated Smads involved in signaling by TGF- β /Activin proteins, 3) Co-Smads involved in signaling by DPP/BMP and TGF- β /Activin proteins, 4) Inhibitory Smads. Three subfamilies contain only nematode sequences. Smad sequences were aligned using Clustal-X (Thompson et al., 1997). The neighbor-joining method was utilized (Saitou and Nei, 1987) in the program Mega2 (Kumar et al., 2001) to generate an unrooted phylogeny from the alignment. The length of the alignment (including all unique insertions) is 994 amino acid residues. Branch lengths are drawn to scale. The scale bar shows the number of amino acid substitutions per site between two sequences. Bootstrap values (the percent of trees containing the indicated branch during 1000 trials) above 50 are shown.

Of the three subfamilies that include only nematode proteins two contain a single sequence and the third contains three proteins. Even though they signal for a DPP/BMP subfamily member and they are clearly R-Smads, SMA-2 and SMA-3 are different enough from other R-Smads (and each other) that they each constitute a distinct subfamily. Interestingly, the three-member nematode subfamily contains proteins that cooperate in the same pathway but have distinct functions. Each of these proteins influences dauer formation, an alternative third-stage larva specialized for survival and dispersal activated by environmental stress (Cassada and Russell, 1975). In addition, they all function downstream of the TGF- β /Activin subfamily member DAF-7. The constitutively active DAF-3 antagonizes TGF- β signal transduction by binding to DNA and repressing gene expression (Thatcher et al., 1999), a mechanism not used by other I-Smads. Alternatively, the TGF- β -inducible proteins DAF-8 and DAF-14 stimulate the expression of DAF-7 target genes by inhibiting DAF-3 function (Inoue and Thomas, 2000). This is the only subfamily containing proteins that function as both agonists and antagonists in the same pathway.

The tree generates two overall impressions. First, for R-Smads and Co-Smads confidence in the clusters is very high - particularly between human and fly sequences (bootstrap values over 70% are considered statistically significant; Hillis and Bull, 1993). This impression is supported by a study utilizing transgenes expressing human Smad genes in flies. That study showed that human and fly Smads that cluster together in the tree generate the same phenotype (Marquez et al., 2001). Taken together the amino acid similarity and functional conservation studies indicate that human and fly proteins in the same subfamily are encoded by homologous genes. Further, they indicate that one or more gene duplications have occurred in the vertebrate lineage after the split with arthropods leading to multiple human Smad proteins in each R-Smad subfamily.

The second impression is that Smad signaling clearly works the same in flies and humans but is different in many ways in nematodes. For example, there is a nematode specific subfamily composed of agonists and antagonists for the same ligand where the antagonist binds DNA (DAF-3) and the signal transducers (DAF-8 and DAF-14) do not (Thatcher et al., 1999, Inoue and Thomas 2000). This mechanism is the opposite of that utilized by human and fly Smads (signal transducers bind DNA and inhibitors do not).

Overall these two impressions (homology of fly and human Smads and distinctions between Smad signaling mechanisms utilized in humans and flies versus nematodes) strongly argue against the existence of an "Ecdysozoan" phylum containing nematodes and flies (*e.g.* Aguinaldo et al., 1997). All functional genomics and phylogenetics studies of the Smad

family support the traditional view (*e.g.* Hedges and Kumar, 2003) that flies are more closely related to humans than they are to nematodes.

4. SMAD FAMILY DOMAINS

Previous studies have shown that Smad family members that transduce signals (R-Smads and Co-Smads) contain well conserved MH1 domains near their N-terminus and MH2 domains near their C-terminus. Inhibitory Smads have highly divergent MH1 domains but have conserved MH2 domains (*e.g.* Newfeld et al., 1999). This data fits well with experiments showing that the MH1 domain is required for DNA-binding and transcriptional activity while the MH2 domain is involved in a variety of protein-protein interactions including forming multi-Smad complexes (*e.g.* Lagna et al., 1996). Between the MH1 and MH2 domains is a proline-rich linker region not previously characterized in detail. At the C-terminus of R-Smads there is a receptor phosphorylation domain containing serine residues (SSXS) targeted for phosphorylation by TGF- β type I receptor kinases. This domain has typically been included in MH2 domain analyses but in our view it deserves scrutiny as an independent domain. In the analysis that follows, for easy reference, an amino acid residue number described in the text refers to that residue's location in an alignment rather than its location in a given Smad protein.

4.1 MH1 domain

Fig. 2 shows an alignment of the Smad family MH1 domain. Although sequence variation is evident, subsets of the N-terminally located MH1 domain are recognizable in every sequence except for DAF-14. For DAF-14, its N-terminal region is so short that although the MH1 alignment begins at amino acid residue 7, the last 59 amino acid residues of the alignment actually belong to its MH2 domain. We conclude that DAF-14 simply has no MH1 domain. DAD has an extensive amino terminus that only very weakly resembles an MH1 domain. There is just one readily recognizable region in DAF-8: a seven amino acid residue stretch containing the most highly conserved amino acids in the alignment (between amino acid residues 120 and 130). As reported previously (Newfeld et al., 1999), the human I-Smads (hSmad6 and hSmad7) align reasonably well.

The MH1 domain is divided into subregions by unique amino acid insertions in a number of Smads. If there is a biological function for these insertions it is unknown. Two Smad2 proteins, one with and one without the insert, are present in mice. However, mice engineered to express only the

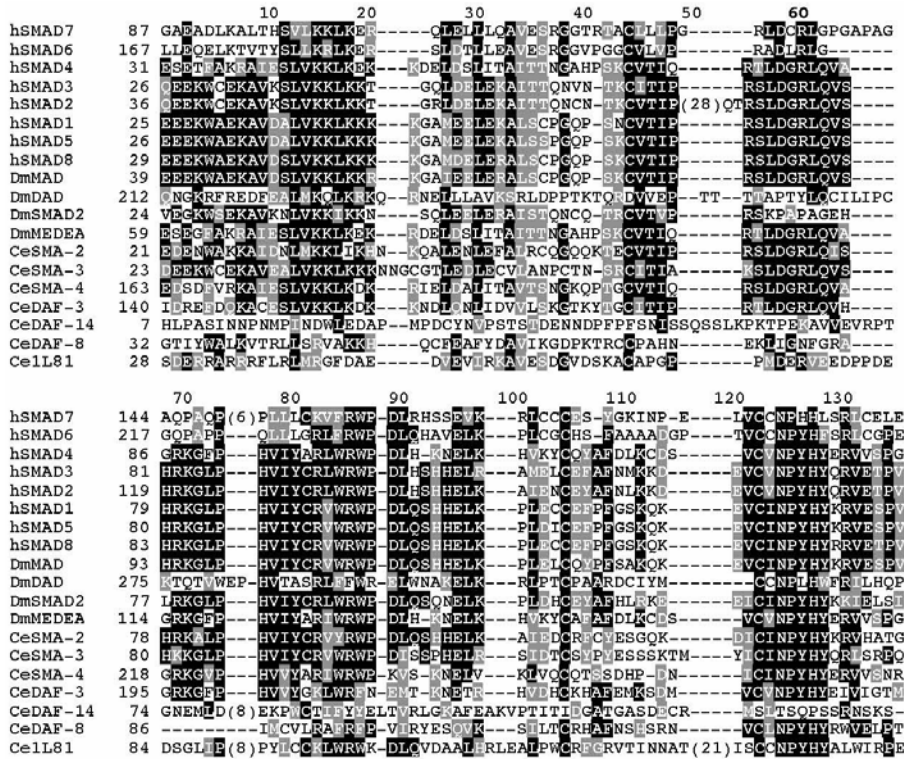


Figure 2. Smad family MH1 domain. This domain is located near the N-terminus of Smad proteins. This domain is highly conserved in R-Smads and Co-Smads. The domain was defined by Pfam (www.sanger.ac.uk/Software/Pfam/index.shtml) based on the crystal structure. Here we show the evolutionarily conserved portion beginning at Glu39 in Mad and ending at Val144 in Mad. The length of the alignment is 194 amino acid residues. Regions were removed when insertions were present in three or fewer sequences and the number of residues shown instead. Residues were shaded if 40% of them were identical (black) or similar (grey) by Boxshade3.21 (www.ch.embnet.org/software/BOX_form.html). Numbers above the alignment begin with the first amino acid and run consecutively. Residue number 60 in bold indicates the location of the DNA-binding region.

short form of Smad2 (without the insert) appear completely normal suggesting that the insert is non-functional (Dunn et al., 2005).

Given its documented role in DNA binding and transcriptional activation (Liu et al., 1997) it is somewhat surprising that there are no absolutely invariant amino acid residues in the MH1 domain. Here we examine the extent of conservation for a number of amino acid residues with known functions. A crystal structure of hSmad3 bound to DNA showed that an 11 amino acid residue region forms a β -hairpin that fits into the major groove of DNA (Shi et al., 1998). The DNA-contacting hairpin is contained

within a conserved 20 amino acid residue region beginning with Arg54 and ending with Pro74. The three residues that contact DNA are Arg61, Gln63 and Lys71. Arg59 is present in all R-Smads and Co-Smads (except DmSmad2, an R-Smad that inexplicably contains a unique stretch of nine amino acid residues in this region) and in DAF-3, hSmad6 and hSmad7. Gln63 and Lys71 are present in all R-Smads and Co-Smads (except DmSmad2) and in DAF-3. All three DNA-contacting residues are absent from DAD, DAF-8 and DAF-14.

A more detailed crystal structure of hSmad3 bound to DNA (Chai et al., 2003) found a bound zinc atom. The zinc-contacting residues are Cys44 (present in all but DAD and DAF-14), Cys105, Cys122 and His127 (these three are present in all but DAF-14). Surprisingly, the four zinc-binding amino acids are more highly conserved than the DNA-contacting residues suggesting that zinc-binding is essential to all Smad functions.

An alignment of the DNA-binding domain of the NFI/CTF family and the MH1 domain of the Smad family identified 22 highly conserved residues, including the four zinc-binding residues (Stefancsik and Sarkar, 2003, Sadreyev and Grishin, 2003). The nuclear factor I (NFI) and CCAAT box-binding transcription factor (CTF) family is composed of vertebrate nuclear proteins that bind a palindromic DNA sequence. The 22 conserved amino acid residues are present in all NFI/CTF family members and all DNA-binding Smads (except DmSmad2). All 22 residues are present in hSmad6 and hSmad7 but not in fly or nematode I-Smads. The majority of conserved amino acids are located in two regions. Four are located between residues 10 and 20 and eleven are located between residues 80 and 100. In contrast, none of the three Smad DNA-binding residues are conserved in the NFI/CTF family. The authors' data is consistent with their hypothesis that the NFI/CTF family diverged from the Smad family after the split between flies and mammals.

A number of conserved residues in the MH1 domain have had their functional importance demonstrated via mutation. For example, we recently conducted a transgenic analysis of MH1 point mutations in Mad (DNA-binding residue), Med (Zinc-binding residue) and hSmad4 (two residues conserved in the NFI/CTF family). We showed that they elicit a variety of mutant phenotypes (Takaesu et al., 2005).

To explicitly examine the relationship between Smad MH1 domains we generated a phylogenetic tree (not shown) from the MH1 domain alignment in Fig. 2. The MH1 domain tree places most *C. elegans* sequences in subfamilies distinct from their placement in the full-length tree shown in Fig. 1. First, SMA-2 and SMA-3 were unique R-Smad subfamilies in the full-length tree but in the MH1 tree they now cluster with the Smad1/Mad subfamily. This fits with the fact that their ligand (DBL-1) belongs to the

DPP/BMP subfamily. Second, the subfamily containing the three DAF-7 signaling pathway components (DAF-3, DAF-8 and DAF-14) in the full-length tree breaks up. In the MH1 domain tree, DAF-3 now clusters with SMA-4 in the Co-Smad subfamily. This fits with the fact that both proteins can bind DNA. In the MH1 domain tree, DAF-8 and DAF-14 each form unique subfamilies due to their highly divergent or absent MH1 domains respectively. Another difference is that DAD moves out of the I-Smad subfamily to become a unique subfamily in the MH1 tree.

4.2 MH2 domain

Fig. 3 shows an alignment of the Smad family MH2 domain. First identified as essential for Smad homo-trimer formation (Shi et al., 1997) this C-terminally located domain is now known as a versatile protein-protein interaction module essential for many Smad activities. Functions associated with the MH2 domain are: 1) formation of homo-trimers of R-Smads and Co-Smads, 2) formation of hetero-trimers containing two R-Smads and one Co-Smad, 3) interaction of R-Smads with the SARA adapter protein and TGF- β type I receptor kinases and 4) interaction of R-Smad/Co-Smad hetero-trimers with transcriptional activators and repressors (see Moustakas and Heldin, 2002, for a review). Phylogenetically and functionally, the MH2 domain is the core of the Smad family and is present in all members.

Note that the C-terminal receptor phosphorylation region was included in many previous analyses of the MH2 domain (including our own; Newfeld et al., 1999). However, we exclude the receptor phosphorylation region from this analysis based on structural data showing that prior to phosphorylation this C-terminal region extrudes from the MH2 domain and is not involved in homo-trimer formation (*e.g.* Wu et al., 2001). This distinction should be kept in mind when comparing data reported here with previous studies.

An examination of the MH2 alignment reveals that 24% of the amino acid residues are extremely well conserved (at least 17 of the 19 sequences have an identical or similar amino acid at a particular position). Eleven of the 47 highly conserved residues are identical in all sequences and 13 are very well conserved (a similar amino acid residue in all sequences). Many of the highly conserved residues are contained in six small regions. The largest of these regions (166-193) corresponds to the L3 loop near the C-terminus of the MH2 domain. Here 10 residues are similar or identical in all Smads and 7 are well conserved. As discussed below the L3 loop is involved in two well-documented protein-protein interactions.

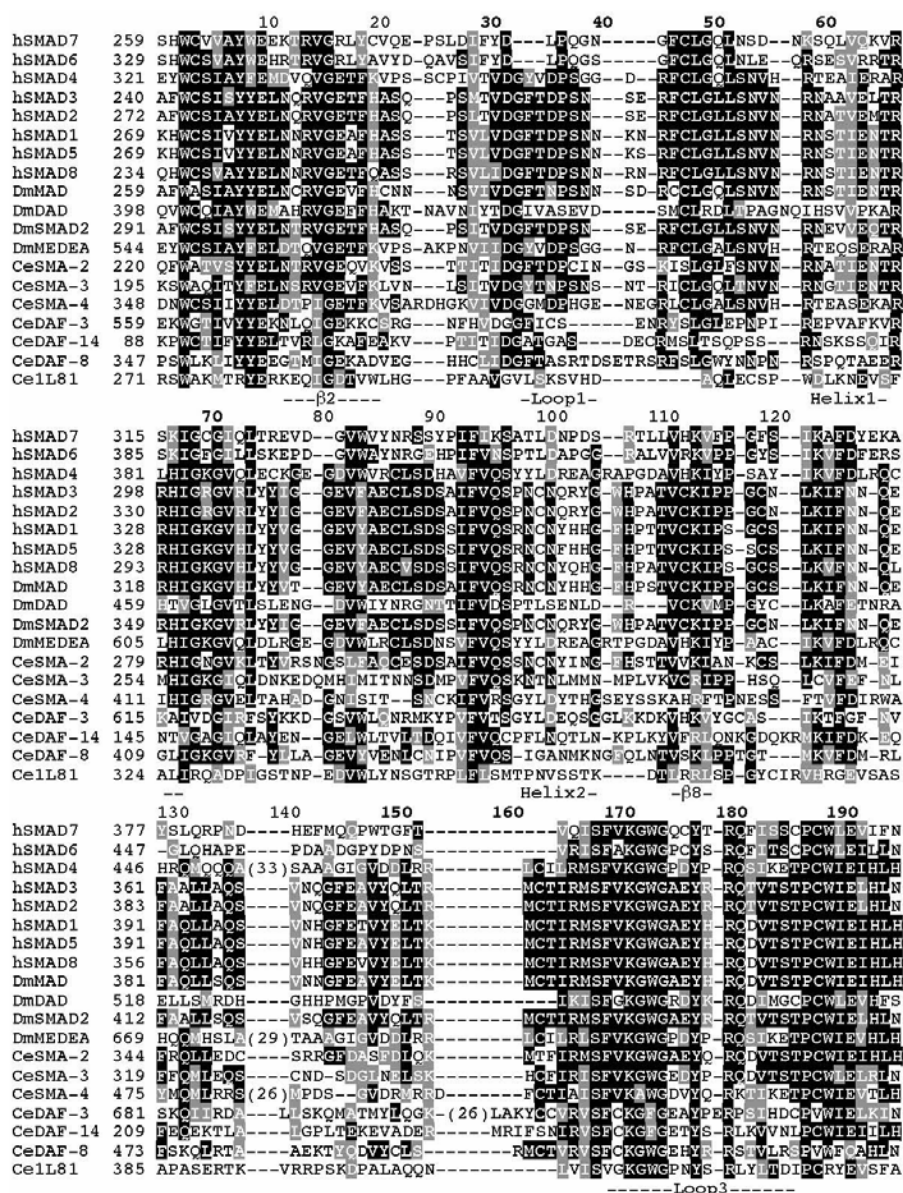


Figure 3. Smad family MH2 domain. This domain is located near the C-terminus of Smad proteins and functions in protein-protein interactions. This domain is highly conserved in all family members with DAD and Ce1L81 the most divergent. Domain extent, the representation of insertions, alignment numbering and shading are as described in Fig. 2. The evolutionally conserved portion of the MH2 domain begins at the invariant Trp261 in Mad and ends at His431 in Mad. The length of the alignment is 222 residues. Bold numbers 30-60 indicate the loop-helix region and 150-190 indicate the helix-bundle region. Several structural features are indicated (note that Helix1 extends two amino acids beyond the alignment break).

The overall structure of an hSmad MH2 domain homo-trimer reveals three subdomains. There is a central β -sandwich, a loop-helix region near the amino-terminus and a helix-bundle region at the C-terminus that extends into the receptor phosphorylation region. In unphosphorylated R-Smad homo-trimers (hSmad3; Chacko et al., 2001) and in Co-Smad homo-trimers (hSmad4; Shi et al., 1997), Loop1 of the loop-helix region of one monomer packs with Helix5 of the helix-bundle region of the adjacent monomer. However, residues identified as essential for homo-trimer formation (*e.g.* Arg46) are not conserved in I-Smads suggesting that heteromeric interactions may involve other features.

Studies of phosphorylated hSmad2/hSmad4 hetero-trimers (Wu et al., 2001) and phosphorylated hSmad1/hSmad4 hetero-trimers (Qin et al., 2001) identified four amino acids as essential to complex formation based on their role in positioning the phosphorylated C-terminal serine residues within the trimer. These are either conserved (Lys114 in β 8 of the β -sandwich region) or identical in all species (Lys172, Tyr178 and Arg181 in the L3 loop of the helix-bundle region). The extraordinary conservation suggests that complexes containing a phosphorylated R-Smad and any other Smad (Co-Smad or I-Smad) are assembled via the same mechanism. The absolute conservation of these residues in all Smads fits with the hypothesis that competition between Co-Smads and I-Smads to form complexes with R-Smads (functional and non-functional respectively) is an essential aspect of I-Smad inhibition (Hayashi et al., 1997).

The SARA adapter protein facilitates physical interactions between TGF- β /Activin subfamily signaling R-Smads and their type I receptors (Tsukazaki et al., 1998). Residues in hSmad2 and hSmad3 essential for interactions with SARA are located in the central β -sandwich region and flank Lys114 suggesting that phosphorylation by the receptor disrupts the Smad/SARA complex (Wu et al., 2000). The residues of hSmad2/3 that interact with SARA are Ile77, Phe84, Tyr104, Trp107 and Asn121. These amino acids are also present in DmSmad2 but not in any other sequence. Alternatively, the residues in these positions are identical in all DPP/BMP signaling R-Smads. This dichotomy suggests that an adaptor molecule specific to DPP/BMP signaling will be identified.

Two interactions between the MH2 domain of hSmad3 and the TGF- β /Activin type I receptor have been identified. Two amino acids in β 2 near the amino-terminus of the MH2 domain (Asn12 and Gln13) and two just downstream in Helix1 of the Loop-Helix region (Arg66 and His67) together form a basic surface. This surface is attracted to an acidic loop created by phosphorylation of the type I receptor GS domain by the ligand-binding type II receptor (Qin et al., 2002). Of these basic amino acids the pair in Helix1 is better conserved. Arg66 and His67 are present in all

R-Smads, except that SMA-3 has instead Met66 and His67. On the other hand, Asn12 and Gln13 are only present in hSmad2 and hSmad3 while two asparagine residues are present in hSmad1, hSmad5 and hSmad8. The distinction suggests that the basic residues in Helix1 are important for the interaction of all R-Smads with their receptors and the residues upstream mediate pathway specific interactions. There is at least one basic amino acid residue in both the upstream and Helix1 locations in all I-Smads, except Ce1L81 has two basic amino acid residues in Helix1. The presence of basic residues at these locations in I-Smads fits with the hypothesis that competition between R-Smads and I-Smads for type I receptor binding (to form functional and non-functional complexes, respectively) is a second aspect of I-Smad inhibition (Nakao et al., 1997, Hayashi et al., 1997).

A pathway-specific interaction between the MH2 domain of R-Smads and their cognate type I receptors has been identified in a study of hSmad1 and hSmad2 (Chen et al., 1998). Two residues in the L3 loop (Arg179 and Thr183) of hSmad2 interact with the L45 loop of TGF- β /Activin type I receptors but not DPP/BMP receptors. Alternatively, His179 and Asp183 in this region of hSmad1 interact only with DPP/BMP receptors. Conservation of the hSmad1 configuration in all BMP signaling R-Smads and the hSmad2 configuration in all TGF- β /Activin signaling R-Smads (plus DAF-8) supports these results. This pair of pathway-specific residues is sandwiched between the invariant Tyr178 and Arg181 involved in positioning the R-Smad phospho-serine residues in the R-Smad/Co-Smad hetero-trimer. Perhaps in addition to their role in hetero-trimer formation the tyrosine and arginine residues also act as signposts for type I receptors in their quest to identify the correct R-Smad to phosphorylate.

To date only a few of the many interactions between Smads and their transcriptional partners (activators or repressors) have been mapped. Pathway specific interactions between hSmad2/hSmad4 complexes and the transcriptional activator FAST-1 were localized to the hSmad2 MH2 domain (Chen et al., 1998). Specifically, six residues in Helix2 of the β -sandwich region that are not shared with hSmad1 are responsible for insuring that FAST-1 only interacts with hSmad2. At these positions, five are distinct between DPP/BMP signaling R-Smads and TGF- β /Activin signaling R-Smads supporting their results (for hSmad2 the residues are Pro98, Gln102, Arg103, Tyr104 and Trp107). Tyr104 and Trp107 are also essential for pathway specific interactions between TGF- β /Activin signaling R-Smads and SARA.

Pathway specific interactions between hSmad3/hSmad4 complexes and the transcriptional repressor Ski were localized to the hSmad3 MH2 domain (Qin et al., 2002). Specifically, several of the residues involved in pathway specific interactions between TGF- β /Activin signaling R-Smads and SARA

and pathway specific interactions with FAST-1 also bind Ski. These include (for hSmad2) Phe84, Tyr104, Trp107. These residues are not present in any BMP signaling R-Smad but at these positions all DPP/BMP signaling R-Smads have the same amino acid. This dichotomy suggests that these residues in DPP/BMP signaling R-Smads mediate interactions with their transcriptional partners.

Near the C-terminal end of the MH2 domain, two insertions are present in the alignment. One insert is unique to DAF-3, but the other is found in all Co-Smads (hSmad4, MED and SMA-4). The insertion in the Co-Smads contains a run of alanine and glutamine residues of differing length encoded by CAG tri-nucleotide repeats (data not shown). CAG repeats are frequently found in transcription factors and the encoded monotonous stretches of the protein are thought to be evolutionarily variable, unstructured spacer regions (Newfeld et al., 1993, 1994). The role of this region in Co-Smads is unknown.

To explicitly examine the relationship between Smad MH2 domains we generated a phylogenetic tree (not shown) from the MH2 domain alignment in Fig. 3. The clusters of R-Smads and Co-Smads are identical in the MH2 tree and the full-length tree shown in Fig. 1. This further supports the hypothesis that the MH2 domain is the fundamental feature of Smad family proteins. We noted two differences between the trees. First, the subfamily containing only DAF-7 signaling pathway components breaks up with DAF-14 becoming a unique subfamily located between the Co-Smads and the clustered DAF-3 and DAF-8. Second, the I-Smad subfamily also breaks up with DAD and Ce1L81 forming highly divergent unique subfamilies.

Overall, the MH2 domain analysis provides further evidence against the existence of an Ecdysozoan phylum. Residues underlying pathway-specific interactions (with SARA, with receptors and with transcription factors) are always identical for human and fly members within each R-Smad subfamily but are rarely conserved in nematode Smads.

One issue concerning the similarity of the MH2 domain to domains in other proteins should be addressed here. Structural similarities between the MH2 domain and a forkhead-associated domain have been reported (Durocher et al., 2000; Lee et al., 2003). In addition, the structure of an autoinhibitory domain in interferon regulatory factors (IRFs) has similarities to the MH2 domain (Qin et al., 2003; Takahashi et al., 2003). It should be noted that the amino acid residues in these three domains are completely dissimilar but they are all capable of binding phospho-serine or phospho-threonine (two structurally very similar amino acids). Some, but not all, of these investigators clearly point out that structural similarities between the domains derive not from evolutionary conservation of amino acid sequences present in a common ancestor (homology) but from the fact that they

perform the same function (convergence). In other words, the relationship between these domains is the same as the relationship between the fin of a fish and the fin of a dolphin - two structures with completely different origins that have evolved for highly efficient swimming. To date, unlike the case for the MH1 domain, there are no domains in other proteins that are considered homologous to the Smad MH2 domain.

4.3 Linker region

Fig. 4 shows alignments of Smad subfamily linker regions. As described above, this domain has received scant attention. It is not possible to generate a meaningful alignment of the entire Smad family and there are just a few abbreviated alignments of this domain in mammalian R-Smads in the literature. Here we discuss alignments containing only Smad subfamily members that identify considerable amino acid conservation.

For the R-Smad subfamily that signals for DPP/BMP proteins (Smad1/Mad subfamily – Fig. 4A) this region is well conserved in humans and flies along its entire length. For the R-Smad subfamily that signals for TGF- β /Activin proteins (Smad2/3 subfamily – Fig. 4B) this domain shows the same level of conservation. However, this domain is not alignable between these two subfamilies (nor can SMA-2 or SMA-3 be aligned with either subfamily). For the Co-Smad subfamily (Fig. 4C) two small regions are well conserved in humans and flies. One region adjoins the MH1 domain and the other adjoins the MH2 domain. Other regions (data not shown) in the Co-Smad alignment are well conserved either within vertebrates or within insects. No regions in this subfamily are conserved with SMA-4. For I-Smads one conserved region was identified that adjoins the MH1 domain. This domain is well conserved in vertebrates, moderately conserved between vertebrates and insects and only very weakly conserved in Ce1L81.

There is a biochemical interaction associated with developmental functions of vertebrate Smad family members that has been mapped to this domain. Erk kinases belong to the Mitogen-Activated Kinase (MAP) kinase family of Ser/Thr kinases. Four consensus Erk phosphorylation sites PX(S/T)P were identified in the linker region of hSmad1. Subsequently, two Erk sites were identified in hSmad3 and one of these sites is present in hSmad2. All Erk sites are phosphorylated in mammalian cells (Kretzschmar et al., 1997, 1999). Recent studies of Smad1 proteins with mutations in these sites revealed a developmental function for Erk phosphorylation in neural induction in *Xenopus* (Pera et al., 2003) and germ cell development in mice (Aubin et al., 2004).

Examination of our alignments shows that conservation of the four Erk sites in hSmad1 is highly variable. The first (in hSmad1 beginning with

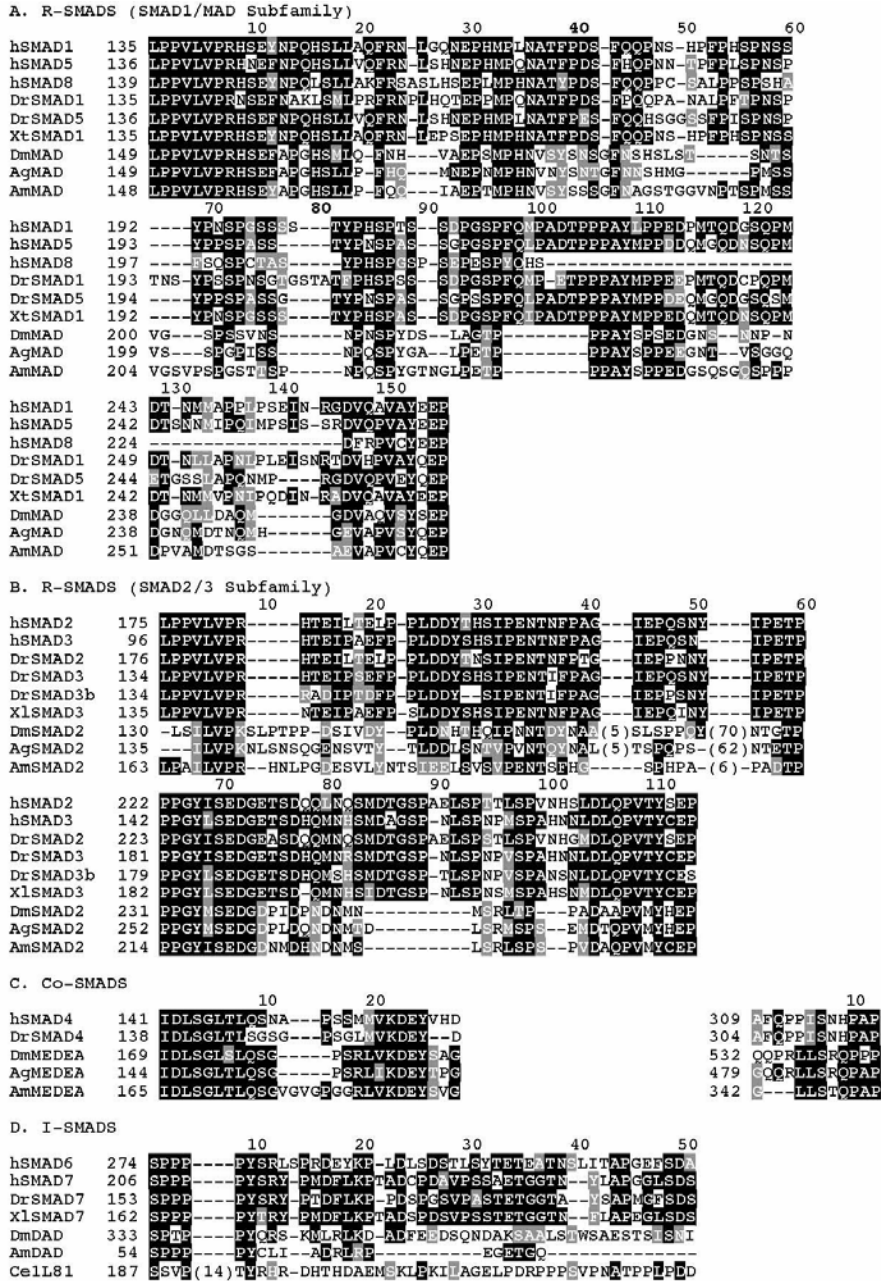


Figure 4. Smad subfamily linker region. This region encompasses all residues between the MH1 and MH2 domains. Conserved stretches were identified in alignments for each of the four major subfamilies. A) R-Smads involved in signaling by DPP/BMP proteins - the Smad1/Mad subfamily. Amino acid residue number 40 in bold indicates the location of the

fully conserved GSK3 β site (TFPDS in hSmad1) and number 80 the fully conserved Erk site (PHSP in hSmad1). B) R-Smads involved in signaling by TGF- β /Activin proteins - the Smad2/3 subfamily. C) Co-Smads involved in signaling by DPP/BMP and TGF- β /Activin proteins. Left side - adjoins the MH1 domain. Right side - adjacent to the MH2 domain. The DrSmad4 sequence was assembled from two partial sequences to generate a contiguous sequence with the greatest agreement to the linker region of hSmad4. D) Inhibitory Smads. One conserved region that begins three amino acid residues downstream of the MH1 domain was identified. Insertions, alignment numbering and amino acid shading are as described in Fig. 2. Additional species are included in each alignment to document the extent of conservation.

Pro54) is present in vertebrates and bees, the second (Pro70) is present only in vertebrates, the third (Pro84) is present in all sequences and the fourth (Pro94) is present in all except flies. Given this pattern, the most parsimonious explanation is that all four Erk sites were present in the common ancestor of human and insect Smad1/Mad and that individual sites were lost at various times in insect lineages after the divergence from vertebrates. There is even less conservation of the Erk sites in hSmad2 and hSmad3. The common site (in hSmad2 beginning at Pro57) is not present in insects and the unique site in hSmad3 (Pro96) is only present in zebrafish (DrSmad3). No Erk sites are present in the linker region of any other Smad family member. Overall, Erk phosphorylation of R-Smad function may be relevant outside vertebrates but this cannot be assumed based on the pattern of conservation.

The presence of one fully conserved Erk site in the Smad1/Mad subfamily led us to examine the linker region for other conserved phosphorylation sites. We discovered numerous consensus sites (S/T)XXX(S/T) for the glycogen synthase kinase3 β (GSK3 β) Ser/Thr kinase (Fiol et al., 1987) in this subfamily and in SMA-3. Further, two of these sites are conserved between vertebrates and insects. The first (beginning at hSmad1 Thr38) is present in all sequences and the second (Ser72) is present in all vertebrate Smad1/Smad5 sequences and in Mad. There are two sites in SMA-3 separated by the same number of amino acid residues but the region surrounding these sites is too degenerate to say with confidence that the Thr38 and Ser72 sites are conserved. No GSK3 β sites are present in the linker region of any other Smad family member.

The presence of a fully conserved GSK3 β site in the Smad1/Mad subfamily is intriguing because GSK3 β (and its *Drosophila* homolog Zeste White3) are antagonists of Wnt family growth factor signaling. In vertebrate and insect systems TGF- β and Wnt pathways interact frequently to influence developmental processes (*e.g.* Takaesu et al., 2005) but the mechanism underlying many of these interactions is unknown. The conservation of a GSK3 β site suggests that phosphorylation of Smad1/Mad subfamily members may be a mechanism utilized for growth factor "crosstalk". This

hypothesis awaits experimental verification and in such experiments it is important to remember that GSK3 β phosphorylation is typically a secondary event; it occurs when different serine residues in the target protein have been phosphorylated by another Ser/Thr kinase. For example, CREB is phosphorylated first at Ser133 by cAMP-dependent kinase and then at Ser129 by GSK3 β (Fiol et al., 1994).

Overall, conservation of the Linker domain within but not between R-Smad subfamilies suggests that pathway-specific functions likely involve amino acids in this region. In addition, sequence conservation in this domain follows the pattern noted previously: humans and flies are similar or identical with nematodes highly divergent.

4.4 Receptor phosphorylation region

Fig. 5 shows an alignment of the Smad family receptor phosphorylation region. At the N-terminus of the region is a stretch of amino acid residues that is well conserved in human and fly R-Smads that may function with the MH2 domain in protein-protein interactions. At the C-terminus of the region is the SSXS motif in human and fly R-Smads, the most C-terminal two serine residues of which are phosphorylated by the type I receptor to stimulate signal transduction. An examination of the alignment reveals that the second amino acid from the C-terminus in this motif is either a valine or a methionine in all human and fly R-Smads, or a conserved isoleucine in SMA-3. In addition, DAF-8 has the sequence SSRT at its terminus indicating it could possibly be recognized by a type I receptor and phosphorylated like an R-Smad. In I-Smads, this motif is essentially absent, whereas Co-Smads show some conservation of this region. While hSmad4 and Med have no C-terminal serine residues, SMA-4 has two. Interestingly, DAF-14 has two and SMA-3 has a serine and a threonine.

As mentioned above, the presence or absence of an MH1 domain is not an accurate predictor of Smad function for nematode sequences (the DNA-binding antagonist DAF-3 has an MH1 domain and no C-terminal serine while the positively signaling DAF-8 has a nearly unrecognizable MH1 domain and three C-terminal serines). Given the conservation pattern, perhaps the number of C-terminal serine residues is a better predictor of function. From this perspective, the nematode specific subfamily composed of DAF-7 signaling pathway components (DAF-3, DAF-8 and DAF-14) can be assigned the following roles: DAF-3 is an I-Smad (no serine), DAF-14 is a Co-Smad (2 serines, like SMA-4) and DAF-8 is an R-Smad (SSRT). Given these roles, it appears that the members of this unique subfamily, a pathway unto themselves, are co-evolving to maintain their ability to interact. If this

prediction is validated by experiments, this nematode subfamily is a truly unique example of developmental pathway evolution.

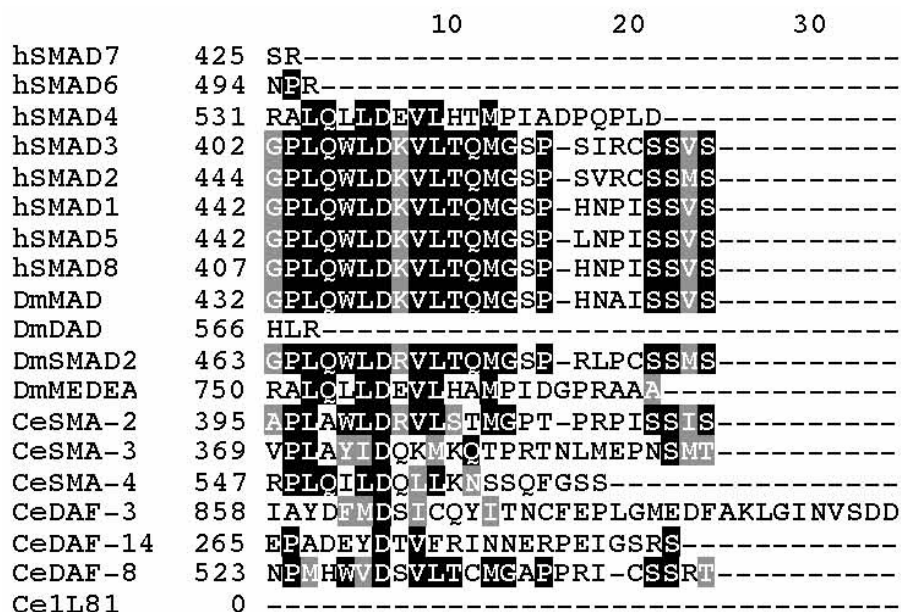


Figure 5. Smad family receptor phosphorylation region. The region begins immediately after the end of the MH2 domain (Gly432 in Mad) and ends at the C-terminal amino acid residue (Ser455 in Mad). This domain is highly conserved in R-Smads and weakly conserved in Co-Smads. Alignment numbering and amino acid shading are as described in Fig. 2.

5. FUTURE PERSPECTIVES

One important area for future research is to investigate the diversification of the Smad family into its four major subfamilies. At present, we know that the ancestral R-Smad split into two R-Smad subfamilies after the divergence of diploblasts (cnidarians) and acoelomates (*Schistosoma*). However, the origin of Co-Smads and I-Smads is still unknown as both subfamilies are already present in nematodes (a pseudocoelomate). Additional acoelomate and pseudocoelomate species need to be surveyed to fill this gap.

A second area for future research is to test our hypotheses about the subfamily specific to the DAF-7 signaling pathway of *C. elegans*. For example, what are the biochemical interactions that underlie the inhibition of DAF-3 activity by DAF-8 and DAF-14? How do these unusual Smads interact with the equally unusual TGF- β receptor DAF-1?

In summary, our phylogenetic analysis of Smad family proteins has provided hypotheses for experimental testing and also provided explanations for experimental results that were previously difficult to interpret. In our view there is no impediment to extending positive feedback between experimental and phylogenetic studies to other signaling pathways. In fact, in addition to continuing our studies of the Smad family we have begun a phylogenetic analysis of families that participate in the Wnt signaling pathway.

ACKNOWLEDGEMENTS

We thank Peter ten Dijke and Sudhir Kumar for valuable discussions. Research in the Newfeld lab is supported by the U.S. National Institutes of Health (NCI and NHGRI).

REFERENCES

- Aguinaldo, A., Turbeville, J., Linford, L., Rivera, M., Garey, J., Raff, R., and Lake, J., 1997, Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387: 489-493.
- Aubin, J., Davy, A., and Soriano, P., 2004, In vivo convergence of BMP and MAPK signaling pathways: impact of differential Smad1 phosphorylation on development and homeostasis. *Genes Dev* 18: 1482-1494.
- Beall, M., McGonigle, S., and Pearce, E., 2000, Functional conservation of *Schistosoma mansoni* Smads in TGF- β signaling. *Mol Biochem Parasitol* 111: 131-142.
- Cassada, R., and Russell, R., 1975, The dauer larva, a post-embryonic developmental variant of the nematode *C. elegans*. *Dev Biol* 46: 326-342.
- Chacko, B., Qin, B., Correia, J., Lam, S., de Caestecker, M., and Lin, K., 2001, The L3 loop and C-terminal phosphorylation define Smad protein trimerization. *Nat Struct Biol* 8: 248-253.
- Chai, J., Wu, J., Yan, N., Massagué, J., Pavletich, N., and Shi, Y., 2003, Features of Smad3 MH1-DNA complex: roles of water and zinc in DNA binding. *J Biol Chem* 278: 20327-20331.
- Chen, Y., Hata, A., Lo, R., Wotton, D., Shi, Y., Pavletich, N., and Massagué, J., 1998, Determinants of specificity in TGF- β signal transduction. *Genes Dev* 12: 2144-2152.
- Dunn, N., Koonce, C., Anderson, D., Islam, A., Bikoff, E., and Robertson, E., 2005, Mice exclusively expressing the short isoform of Smad2 develop normally and are viable and fertile. *Genes Dev* 19: 152-163.
- Durocher, D., Taylor, I., Sarbassova, D., Haire, L., Westcott, S., Jackson, S., Smerdon, S., and Yaffe, M., 2000, The molecular basis of FHA domain: phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms. *Mol Cell* 5: 1169-1182.
- Felsenstein, J., 1985, Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.

- Fiol, C., Mahrenholz, A., Wang, Y., Roeske, R., and Roach, P., 1987, Formation of protein kinase recognition sites by covalent modification of the substrate: molecular mechanism for the synergistic action of casein kinase II and glycogen synthase kinase 3. *J Biol Chem* 262: 14042-14048.
- Fiol, C., Williams, J., Chou, C., Wang, Q., Roach, P., and Andrisani, O., 1994, A secondary phosphorylation of CREB at Ser129 is required for the cAMP-mediated control of gene expression: a role for glycogen synthase kinase-3 in the control of gene expression. *J Biol. Chem.* 269: 32187-32193.
- Hayashi, H., Abdollah, S., Qiu, Y., Cai, J., Xu, Y., Grinnell, B., Richardson, M., Topper, J., Gimbrone, M., Wrana, J., and Falb, D., 1997, The MAD-related protein Smad7 associates with the TGF- β receptor and functions as an antagonist of TGF- β signaling. *Cell* 89: 1165-1173.
- Hedges, S., and Kumar, S., 2003, Genomic clocks and evolutionary timescales. *Trends Genetics* 19: 200-206.
- Herpin, A., Lelong, C., and Favrel, P., 2004, TGF- β -related proteins: an ancestral and widespread superfamily of cytokines in metazoans. *Dev Comp Immunol* 28: 461-485.
- Hillis, D., and Bull, J., 1993, An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol* 42: 182-192.
- Inoue, T., and Thomas, J., 2000, Targets of TGF- β signaling in *C. elegans* dauer formation. *Dev Biol* 217: 192-204.
- Kretschmar, M., Doody, J., and Massagué, J., 1997, Opposing BMP and EGF signaling pathways converge on the TGF- β family mediator Smad1. *Nature* 389: 618-622.
- Kretschmar, M., Doody, J., Timokhina, I., and Massagué, J., 1999, A mechanism of repression of TGF- β / Smad signaling by oncogenic Ras. *Genes Dev* 13: 804-816.
- Kumar, S., Tamura, K., Jakobsen, I., and Nei, M., 2001, MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17: 1244-1245.
- Lagna, G., Hata, A., Hemmati-Brivanlou, A., and Massagué, J., 1996, Partnership between DPC4 and Smad proteins in TGF- β signaling pathways. *Nature* 383: 832-836.
- Lee, G., Ding, Z., Walker, J., and Van Doren, S., 2003, NMR structure of the forkhead-associated domain from the Arabidopsis receptor kinase-associated protein phosphatase. *Proc Natl Acad Sci U S A* 100: 11261-11266.
- Liu, F., Hata, A., Baker, J., Doody, J., Carcamo, J., Harland, R., and Massagué, J., 1996, A human MAD protein acting as a BMP-regulated transcriptional activator. *Nature* 381: 620-623.
- Liu, F., Poupponnot, C., and Massagué, J., 1997, Dual role of the Smad4/DPC4 tumor suppressor in TGF- β -inducible transcriptional complexes. *Genes Dev* 11: 3157-3167.
- Marquez, R., Singer, M., Takaesu, N., Waldrip, W., Kraytsberg, Y., and Newfeld, S., 2001, Transgenic analysis of the Smad family of TGF- β signal transducers in *Drosophila* suggests new roles and interactions between family members. *Genetics* 157: 1639-1648.
- Massagué, J., Blain, S., and Lo, R., 2000, TGF- β signaling in growth control, cancer and heritable disorders. *Cell* 103: 295-309.
- Moustakas, A., and Heldin, C.-H., 2002, From mono- to oligo-Smads: the heart of the matter in TGF- β signal transduction. *Genes Dev* 16: 1867-1871.
- Nakao, A., Afrakhte, M., Morén, A., Nakayama, T., Christian, J., Heuchel, R., Itoh, S., Kawabata, M., Heldin, N.-E., Heldin, C.-H., and ten Dijke, P., 1997, Identification of Smad7, a TGF- β -inducible antagonist of TGF- β signaling. *Nature* 389: 631-635.
- Newfeld, S., Schmid, A., and Yedvobnick, B., 1993, Homopolymer length variation in the *Drosophila* gene *mastermind*. *J Mol Evol* 37: 483-495.
- Newfeld, S., Tachida, H., and Yedvobnick, B., 1994, Drive-selection equilibrium: homopolymer evolution in the *Drosophila* gene *mastermind*. *J Mol Evol* 38: 637-641.

- Newfeld, S., Wisotzkey, R., and Kumar, S., 1999, Molecular evolution of a developmental pathway: phylogenetic analyses of TGF- β family ligands, receptors and Smad signal transducers. *Genetics* 152: 783-795.
- Padgett, R., Wozney, J., and Gelbart, W., 1993, Human BMP sequences confer normal dorsal ventral patterning in the *Drosophila* embryo. *Proc Natl Acad Sci U S A* 90: 2905-2909.
- Pera, E., Ikeda, A., Eivers, E., and De Robertis, E., 2003, Integration of IGF, FGF, and anti-BMP signals via Smad1 phosphorylation in neural induction. *Genes Dev* 17: 3023-3028.
- Qin, B., Chacko, B., Lam, S., de Caestecker, M., Correia, J., and Lin, K., 2001, Structural basis of Smad1 activation by receptor kinase phosphorylation. *Mol Cell* 8: 1303-1312.
- Qin, B., Lam, S., Correia, J., and Lin, K., 2002 Smad3 allosteric links TGF- β receptor kinase activation to transcriptional control. *Genes Dev* 16: 1950-1963.
- Qin, B., Liu, C., Lam, S., Srinath, H., Delston, R., Correia, J., Derynck, R., and Lin, K., 2003, Crystal structure of IRF-3 reveals mechanism of autoinhibition and virus-induced phosphoactivation. *Nat Struct Biol* 10: 913-921.
- Raff, R., 1996, *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. Univ. Chicago Press, Chicago, IL, USA.
- Raftery, L., and Sutherland, D., 1999, TGF- β family signal transduction in *Drosophila* development: from MAD to Smads. *Dev Biol* 210: 251-68.
- Sadreyev, R., and Grishin, N., 2003, COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* 326: 317-336.
- Saitou, N., and Nei, M., 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406-425.
- Samuel, G., Miller, D., and Saint, R., 2001, Conservation of a DPP/BMP signaling pathway in the nonbilateral cnidarian *Acropora millepora*. *Evol Dev* 3: 241-250.
- Sekelsky, J., Newfeld, S., Raftery, L., Chartoff, E., and Gelbart, W., 1995, Genetic characterization and cloning of *Mothers against dpp*: a gene required for *decapentaplegic* function in *Drosophila*. *Genetics* 139: 1347-1358.
- Shi, Y., Hata, A., Lo, R., Massagué, J., and Pavletich, N., 1997, A structural basis for mutational inactivation of the tumor suppressor Smad4. *Nature* 388: 87-93.
- Shi, Y., Wang, Y., Jayaraman, L., Yang, H., Massagué, J., and Pavletich, N., 1998, Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF- β signaling. *Cell* 94: 585-594.
- Stefancsik, R., and Sarkar, S., 2003, Relationship between the DNA binding domains of Smad and NF1/CTF transcription factors defines a new superfamily. *DNA Seq* 14: 233-239.
- Suga, H., Ono, K., and Miyata, T., 1999, Multiple TGF- β receptor related genes in sponge and ancient gene duplications before the parazoan-eumetazoan split. *FEBS Lett* 453: 346-350.
- Takaesu, N., Herbig, E., Zhitomersky, D., O'Connor, M., and Newfeld, S., 2005, DNA-binding domain mutations in Smad genes yield dominant negative proteins or a neomorphic protein that can activate Wg target genes in *Drosophila*. *Development* 132: 4883-4894.
- Takahashi, K., Suzuki, N., Horiuchi, M., Mori, M., Suhara, W., Okabe, Y., Fukuhara, Y., Terasawa, H., Akira, S., Fujita, T., and Inagaki, F., 2003, X-ray crystal structure of IRF-3 and its functional implications. *Nat Struct Biol* 10: 922-927.
- Thatcher, J., Haun, C., and Okkema, P., 1999, The DAF-3 Smad binds DNA and represses gene expression in the *C. elegans* pharynx. *Development* 126: 97-107.
- Thompson, J., Gibson, T., Plewniak, F., Jeanmougin, F., and Higgins, D., 1997, The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25: 4876-4882.

- Tsukazaki, T, Chiang, T., Davison, A, Attisano, L., and Wrana, J., 1998, SARA, a FYVE domain protein that recruits Smad2 to the TGF- β receptor. *Cell* 95: 779-791.
- Wu, G., Chen, Y., Ozdamar, B., Gyuricza, C., Chong, P., Wrana, J., Massagué, J., and Shi, Y., 2000, Structural basis of Smad2 recognition by SARA. *Science* 287: 92-97.
- Wu, J., Hu, M., Chai, J., Seoane, J., Huse, M., Li, C., Rigotti, D., Kyin, S., Muir, T., Fairman, R., Massagué, J., and Shi, Y., 2001, Crystal structure of a phosphorylated Smad2: recognition of phosphoSerine by the MH2 domain and insights on Smad function in TGF- β signaling. *Mol Cell* 8: 1277-1289.