Journal of Molecular Graphics and Modelling 27 (2009) 803–812

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Modeling and analysis of MH1 domain of Smads and their interaction with promoter DNA sequence motif

Pooja Makkar [a], Raghu Prasad R. Metpally [a], Sreedhara Sangadala [b], Boojala Vijay B. Reddy [a,*]

[a] *Graduate Center Biochemistry Department and Laboratory of Bioinformatics & in silico Drug Design, Department of Computer Science, Queens College of City University of New York, 65-30 Kissena Blvd, Flushing, NY 11367, United States*
[b] *Atlanta VA Medical Center and the Department of Orthopaedic Surgery, Emory University School of Medicine, Atlanta, Georgia 30329, United States*

ARTICLE INFO

Article history:
Received 3 May 2008
Received in revised form 9 December 2008
Accepted 12 December 2008
Available online 24 December 2008

Keywords:
TGF-β superfamily
Smads
Homology modeling
DNA–Smad interaction

ABSTRACT

The Smads are a group of related intracellular proteins critical for transmitting the signals to the nucleus from the transforming growth factor-β (TGF-β) superfamily of proteins at the cell surface. The prototypic members of the Smad family, *Mad* and *Sma*, were first described in *Drosophila* and *Caenorhabditis elegans*, respectively. Related proteins in *Xenopus*, *Humans*, *Mice* and *Rats* were subsequently identified, and are now known as Smads. Smad protein family members act downstream in the TGF-β signaling pathway mediating various biological processes, including cell growth, differentiation, matrix production, apoptosis and development. Smads range from about 400–500 amino acids in length and are grouped into the receptor-regulated Smads (R-Smads), the common Smads (Co-Smads) and the inhibitory Smads (I-Smads). There are eight Smads in mammals, Smad1/5/8 (bone morphogenetic protein regulated) and Smad2/3 (TGF-β/activin regulated) are termed R-Smads, Smad4 is denoted as Co-Smad and Smad6/7 are inhibitory Smads. A typical Smad consists of a conserved N-terminal Mad Homology 1 (MH1) domain and a C-terminal Mad Homology 2 (MH2) domain connected by a proline rich linker. The MH1 domain plays key role in DNA recognition and also facilitates the binding of Smad4 to the phosphorylated C-terminus of R-Smads to form activated complex. The MH2 domain exhibits transcriptional activation properties. In order to understand the structural basis of interaction of various Smads with their target proteins and the promoter DNA, we modeled MH1 domain of the remaining mammalian Smads based on known crystal structures of Smad3-MH1 domain bound to GTCT Smad box DNA sequence (1OZJ). We generated a B-DNA structure using average base-pair parameters of *Twist*, *Tilt*, *Roll* and base *Slide* angles. We then modeled interaction pose of the MH1 domain of Smad1/5/8 to their corresponding DNA sequence motif GCCG. These models provide the structural basis towards understanding functional similarities and differences among various Smads.

1. Introduction

Transforming growth factor-β (TGF-β) superfamily of proteins consist of more than 40 members of growth and differentiation factors. These include TGF-β, activin, inhibin, nodal, bone morphogenetic proteins (BMPs), mullerian duct inhibiting substance (MIS) and many others. These factors are highly conserved among various species and have a wide role in development, cell differentiation, cell cycle progression, adhesion, neuronal growth, bone morphogenesis, reproductive function, vasculogenesis and angiogenesis [1]. The TGF-β superfamily members transduce the signals from membrane to nucleus by binding to type I (TβR-I) and type II (TβR-II) transmembrane heteromeric serine/threonine kinase receptors. In humans, five type II and seven type I receptors have been identified. All these receptors share a common architecture consisting of a short extracellular cysteine rich N-terminal ligand binding domain, a transmembrane region and an intracellular serine/threonine kinase domain. TβR-I contains a glycine and serine rich domain (GS domain). When specific glycine and serine residues are phosphorylated by TβR-II receptor kinase, TβR-I becomes activated which further activates another group of proteins called Smads which forms a heteromeric complex and transfers to nucleus where they bind to promoter DNA and regulate the transcription of various genes [2].

1.1. The Smads

Genes of Smad proteins were discovered in *Drosophila* and *Caenorhabditis elegans* through genetic screening. The name Smad is a fusion of two gene names, *Drosophila* mothers against dpp (Mad)

* Corresponding author at: Laboratory of Bioinformatics & *in silico* Drug Design, Department of Computer Science, Queens College of CUNY, 65-30 Kissena Blvd, Flushing, NY 11367, United States. Tel.: +1 718 997 3487; fax: +1 718 997 3513.
E-mail address: breddy@qc.cuny.edu (B.V.B. Reddy).

1093-3263/$ – see front matter © 2009 Published by Elsevier Inc.
doi:10.1016/j.jmgm.2008.12.003

and *C. elegans* Sma. Mad and Sma proteins (Smads), are of 42–60 kDa, were discovered as molecules that act as essential factors in downstream of Ser/Thr kinase receptors in TGF-β pathway [3]. In humans, eight Smad proteins have been identified and classified into three groups on the basis of their structure and function as: (i) Receptor-regulated Smads (R-Smads) directly interact with TGF-β receptor kinases. These include Smad1, 2, 3, 5 and 8. Smad1, 5 and 8 share close homology and mediate BMP signaling whereas Smad2 and 3 mediate TGF-β and activin signaling [4]. (ii) Common Smads (Co-Smads) associate with R-Smads forming heteromeric complexes and carry the signal further to nucleus. This group includes only one protein called Smad4 which is similar in structure to R-Smads but is not phosphorylated. Smad4 takes part in TGF-β, activin and BMP signaling pathways along with corresponding R-Smads. (iii) Inhibitory Smads (I-Smads), inhibit the TGF-β signaling mediated by R-Smads and Co-Smads. These include Smad6 and 7. Smad6 inhibits BMP signaling [5] whereas the Smad7 inhibits both the TGF-β and BMP signaling [6].

1.2. Structure of the Smad proteins

Smads consist of two highly conserved terminal domains, Mad Homology 1 (MH1) and Mad Homology 2 (MH2), connected by less conserved linker region (Fig. 1). The MH1 (Mad Homology 1) domain consists of about 130 amino acids and is highly conserved in R-Smads and Co-Smads but not in I-Smads. It binds to DNA and hence is attributed to have had a role in transcriptional activation. There is a highly conserved 11-amino acid residue region in the MH1 domain which forms a β-hairpin that makes contact with major groove of DNA [7]. The MH2 contains about 200 amino acid residues and is responsible for protein–protein interactions [8]. The MH2 domain mediates interactions between R-Smads and type I receptor [9], between R-Smads and Co-Smads [10] and between R-Smads and DNA binding factors (activators and repressors) [11]. The MH2 domain shows three major structural features – a central β-sandwich, an N-terminus loop helix and a C-terminus helix bundle region [12]. The linker region is less conserved and responsible for Smad homo-oligomer formation [11]. Recently it is also shown to be involved in interaction with various DNA binding activator and repressor factors [13].

1.3. The Smad signaling

Smads are present in cytoplasm in basal state. The dimeric ligands of TGF-β superfamily binds to receptor complex. The autophosphorylated type II receptor phosphorylate the type I receptor (Fig. 2). By activation of type II receptors various adaptor proteins are recruited such as Dab-2 and SNX6 that help in recognition of R-Smads. Some GTPases like Rab5 help in moving TGFβR-I to early endosomal compartments, where they come in contact with Smad Anchor for Receptor Activation (SARA) that assist in presentation of R-Smads (Smad2 or 3) to type I receptor kinase. The serine residues in the SXS motif at C-terminal of Smads is phosphorylated leading to a change in conformation of R-Smad and dissociation of type I receptor and SARA. Phosphorylated R-Smads associates with other R-Smads resulting in homo-oligomerization of R-Smads or hetero-oligomerization with Co-Smad. Some proteins like TRAP-1 assist in R-Smad and Co-Smad association. Co-Smad associates with R-Smad in different ratios in different Smads. It forms heterodimer with one Smad2 monomer and forms heterotrimer with two Smad3 monomers [14]. The complex then translocates into nucleus and results in nuclear accumulation of Smads. Several transcription factors in the nucleus join with Smad complex in order to make it successful candidate for DNA binding, resulting in activation of various target genes. In addition, Smads also undergo various post translational modifications like phosphorylation, ubiquitination, acetylation or sumoylation which further change the interaction of Smads and results in regulation of various activities [11]. Inhibitory Smads inhibit the TGF-β signaling pathway by binding to various receptors like STRAP-I, interfering in phosphorylation of R-Smads and competing with R-Smads in binding Smad4. It has been observed that Smad6 inhibits BMP signaling and also TGF-β signaling to some extent and Smad7 inhibits TGF-β signaling. The

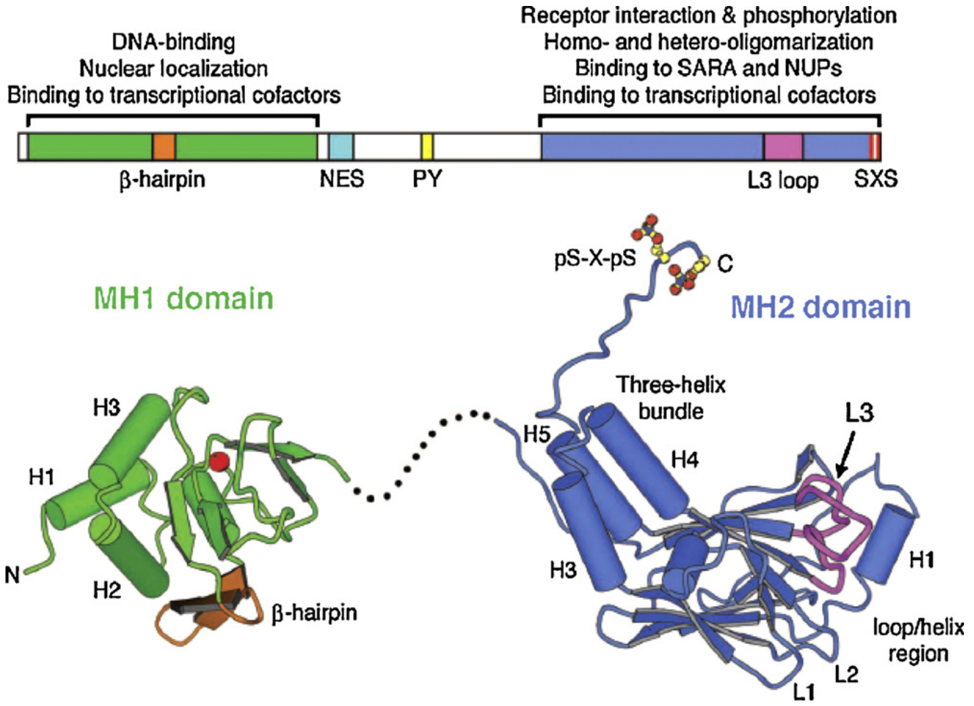

Fig. 1. Typical structure of Smads showing various domain regions and known functions. The known crystal structures of Smad3-MH1 and MH2 domains used as the basis to draw this structure.

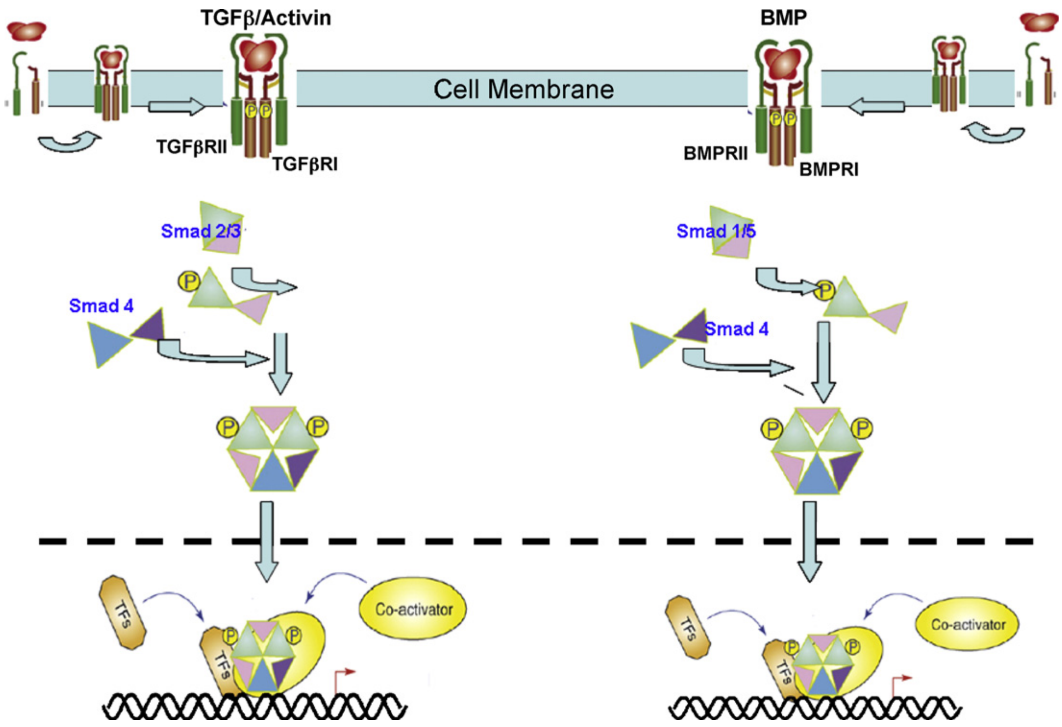

Fig. 2. An overview of TGF-β signaling pathway mediated by Smads. The dimeric TGF-β superfamily of ligands binds to type II receptors which autophosphorylates and transphosphorylates the type I receptor. Various adaptor proteins including R-Smads are recruited and the SXS motif at the C-terminal of R-Smads which promotes homo- and hetero-oligomerization with other R-Smads and Co-Smads. The complex then translocates into the nucleus and joins with several transcription factors to activate the responsive genes.

TGF-β pathway is tightly controlled by the action of various regulatory proteins which act at different points in the pathway [15].

1.4. Smad–DNA interaction

The MH1 domain of Smad is primarily responsible for DNA binding in cooperation with other transcription factors. After the Smad complex is translocated to nucleus, it recognizes and interacts with specific DNA sequences called Smad binding elements (SBEs) [16–18]. The optimal Smad3 binding element is initially observed as four base sequences 5′-GTCT-3′ (or 5′-AGAC-3′) on complementary strand for Smad3 and 4. Optimal binding is observed with the sequence 5′-CAGAC-3′ with extra C at 5′ end although 5′-AGAC-3′ is found to be sufficient for binding. The target genes of TGF-β, activin and BMP often contain such SBEs in their responsive region within promoter sequence. The crystal structure of Smad3 bound to DNA is solved at 2.4 Å resolution. It revealed that DNA binding is mediated by a protruding 11 residue β-hairpin loop in MH1 domain that makes contact with GTCT motif in the major groove of DNA [19]. This β-hairpin loop is highly conserved in all R-Smads and Co-Smad except for the two residues at the turn of β-hairpin in Smad4. TGF-β signaling is not done solely by the four base-pair SBEs alone and the affinity of a Smad-MH1 domain for SBE is in 10^{-7} M range [19] which is very weak for an effective binding. This indicates that additional DNA contacts must be involved in successful binding. More DNA binding is observed as the number of repeats of SBE is increased [20]. Many TGF-β responsive genes like plasminogen activator inhibitor I (PA-I), Jun B, type VII collagen and germline immunoglobulin 1α region contain SBE like sequences [19] and it often appears in multiple copies. The SBE for BMP regulated Smads (Smad1/5/8) was found to be GCCGnCGC that is not being contacted by TGF-β/activin regulated Smads (Smad2 and 3). Binding affinity to this motif depends on the number of repeats of this sequence. Higher the number of repeats, higher is the binding affinity [21]. Smad proteins interact with each other and also with other DNA binding nuclear cofactors forming complexes that assist in achieving specificity and high affinity towards target DNA [13].

In the current manuscript we describe the modeling and analysis of Smad1/5/8-MH1 domains and their interaction with promoter DNA motif based on the known PDB structure of Smad3–DNA complex. We discuss various modeling and model evaluation methods we have employed to select the most energetically favored MH1 domain models of Smad1/5/8 proteins. We further discuss implications of the interaction of each MH1 domain to the Smad binding DNA sequence on the basis of our MH1–DNA interaction modeling.

2. Materials and methods

Amino acid sequences of all Smad-MH1 domains were retrieved from NCBI protein database [22]. The programs FASTA [23] and BLASTP [24] were used for detecting similarities among sequences and to search for a suitable template from the protein structure database (PDB) for homology modeling. Multiple sequence alignment of Smad family of proteins was carried out using CLUSTALX [25].

2.1. Modeling of Smad-MH1 domains

We used the available Smad3-MH1 domain structure to model the structures of remaining Smad-MH1 domains of various Smads using the MODELLER [26] by employing Discovery Studio 1.7 of Accelrys Inc. Between two crystal structures available in PDB, we selected 1OZJ:A as template as it was solved at 2.40 Å as compared to 1mhd:A which was solved at 2.80 Å. MODELLER builds protein tertiary structures by satisfaction of spatial restraints collected from the template structure used for modeling. Target sequences were also submitted to SWISS Model [27] and CPH [28] which are automated model building servers. Ten models were built for each of the template using MODELLER and one model each received from SWISS Model and CPH model building servers. Each of these models was energy minimized using the Optimize protocol of Builder module of Insight-II (Accelrys Inc.) which performs energy minimization by steepest descents, conjugate gradients and quasi-Newton, Newton and truncated-Newton methods sequentially.

2.2. Evaluation and selection of best models

Evaluation methods check for whether a model satisfies standard steric and geometric criteria. Each of the tools used in construction of model, template selection, alignment, model building, and refinement was subjected to its own internal measures of quality, but, ultimately the most meaningful criterion for the quality of model was its conformational energy. In order to deal with such a large number of possible conformations, a hierarchical approach to model evaluation was employed in which a set of simplified and easy to evaluate scoring functions were used to rank all the original models so that a subset could be chosen for more detailed and computationally costly evaluation. All the 12 models were evaluated for stereochemical quality using the PROCHECK [29] program. The aim of PROCHECK was to assess how normal, or how unusual, the geometry of the residues in a given protein structure was, as compared to the stereochemical parameters derived from well refined, high resolution structures. Ramachandran plot, a plot of the ϕ–ψ torsion angles for all residues in protein structure, is an important parameter in PROCHECK for assessment of structures. A simple measure of quality that could be used from plot was the percentage of residues in core region and allowed regions to be very high (>90% residues) [29]. Another important factor in structural assessment is Goodness factor or G-factor which shows the quality of dihedral, covalent and overall bond angles. These scores should be above -0.5 for a reliable model.

We also used other model evaluation tools such as Verify3D [30], ERRAT [31] and PROVE [32] from SAVES metaserver [33] to further evaluate our models with different quality factors. ERRAT counts the number of nonbonded interactions between atoms (CC, CN, CO, NN, NO, and OO) within a cutoff distance of 3.5 Å. It gives an overall quality factor for each protein structure which is expressed as the percentage of protein for which the calculated error value falls below 95% rejection limit. Normally accepted model structures produce these values above 50.

The stereochemical criterion implemented in PROVE is regularity or irregularity of atom volume [32]. It provides an average volume *Z*-score of all the atoms. *Z*-score is calculated as the difference between the volume of the atom and the mean atomic volume for the corresponding atom type, divided by the standard deviation of the appropriate distribution. To evaluate the structure as a whole, root mean square deviation of *Z*-score (*Z*-score RMS) is calculated. The value of *Z*-score RMS should be one for a good structure and increases as the structural quality decreases.

Verify3D uses energetic and empirical methods to produce averaged data points for each residue to evaluate the quality of protein structures. This score measures the compatibility of model with its sequence using a scoring function. If more than 80% of the residues exhibit a score of >0.2 then the protein structure is considered of high quality. Negative or less than 0.2 scores were indicative of potential problems. We also used PROSA II [34] for evaluation which gives the interaction energies of each residue based on statistical analysis of mean force potentials from known protein structures, capturing the average properties of native globular proteins in terms of atom pair, and protein–solvent interactions. We evaluated each model using above methods and selected the final model that fits best by criteria of selection in each method.

2.3. Modeling the interaction of Smad-MH1 and its DNA recognition sequence

A double stranded DNA sequence 5′-TCTGCCGCCGCTT-3′ (GCCG motif containing) was taken from a responsive mouse Collagen X gene promoter sequence for modeling A double stranded B-DNA structure was generated [35] by employing average base-pair parameters of *Twist*, *Tilt*, *Roll* and base *Slide* values using Biopolymer module of Insight-II (Accelrys Inc.). The Cα trace of Smad-MH1 domains of Smad1/5/8 and the phosphate backbone of DNA model containing GCCG motif were superposed using Superpose mode of Insight-II onto the Smad3–DNA complex (1OZJ) to obtain initial 3D-coordinates for Smad-MH1/DNA complexes. The two Smad-MH1 domains were associated with the double stranded DNA (TCTGCCGCCGCTT) model, similar to the complex of Smad3-MH1/DNA (TCAGTCTAGACATAC) in the 1OZJ crystal structure.

The complex was soaked in a 15 Å radius spear of explicit water layer at the center of DNA molecule. The entire complex of Smad-MH1/DNA/water was equilibrated with 100 steps of dynamics at 300 K temperature and the resultant complex was energy minimized using the default optimized protocol of Builder module of Insight-II which uses steepest descent, conjugate gradient and Newton rapson approaches sequentially for a total of 1000 steps. This entire process was performed for all complexes of Smad1/5/8-MH1 domains with DNA/water assembly and resultant structures were analyzed to asses their mode of interaction. We have also modeled Smad1/5/8-MH1 interaction with the GTCT motif of DNA (5′-TCAGTCTAGACAT-3′) to compare binding energies with this non-specific motif using the identical procedure.

We used DS modeler with CHARMm force fields to compute binding energy between Smad-MH1 and DNA sequence. We have taken two MH1 domains as receptor and DNA as ligand with implicit distance model using implicit distance dependent dielectrics so that reasonable relative energies could be obtained with dielectric constant 1 and implicit solvent dielectric constant 80. Nonbonded list radius was kept at 14 Å, minimum hydrogen bond radius was kept 1 Å by default. Non-polar surface constant was 0.92 and non-polar surface coefficient was 0.00542 at default values.

3. Results

3.1. Sequence alignments

The sequence alignment of all known human Smad-MH1 sequences is given in Fig. 3 which shows that the MH1 domain of human Smads was an evolutionary conserved domain. The conservation was high both among R-Smads (Smad1, 2, 3, 5 and 8) and I-Smads (Smad6 and 7). Smad1, 5 and 8 showed the highest sequence similarity with the reference template ranging from 81 to 83%. Smad2 showed 72.9% sequence similarity with the reference template 1OZJ:A. Smad6 and 7 showed least sequence similarity with the reference template ranging from 35 to 40% with only 32 and 26 conserved residues, respectively. Table 1 shows the sequence similarity of all known human Smads with the template sequence (Smad3-MH1). The secondary structure prediction showed high conservation among these sequences. The more conserved residues were found in MH1 domains of R-Smads and Co-Smads which included a highly conserved 13 residue stretch LDGRLQVSHRKGLP. Most of these amino acids were however not conserved in Smad6 and 7. The zinc-binding residues, Cys64, Cys109, Cys121 and His126 were invariant among all Smad family members.

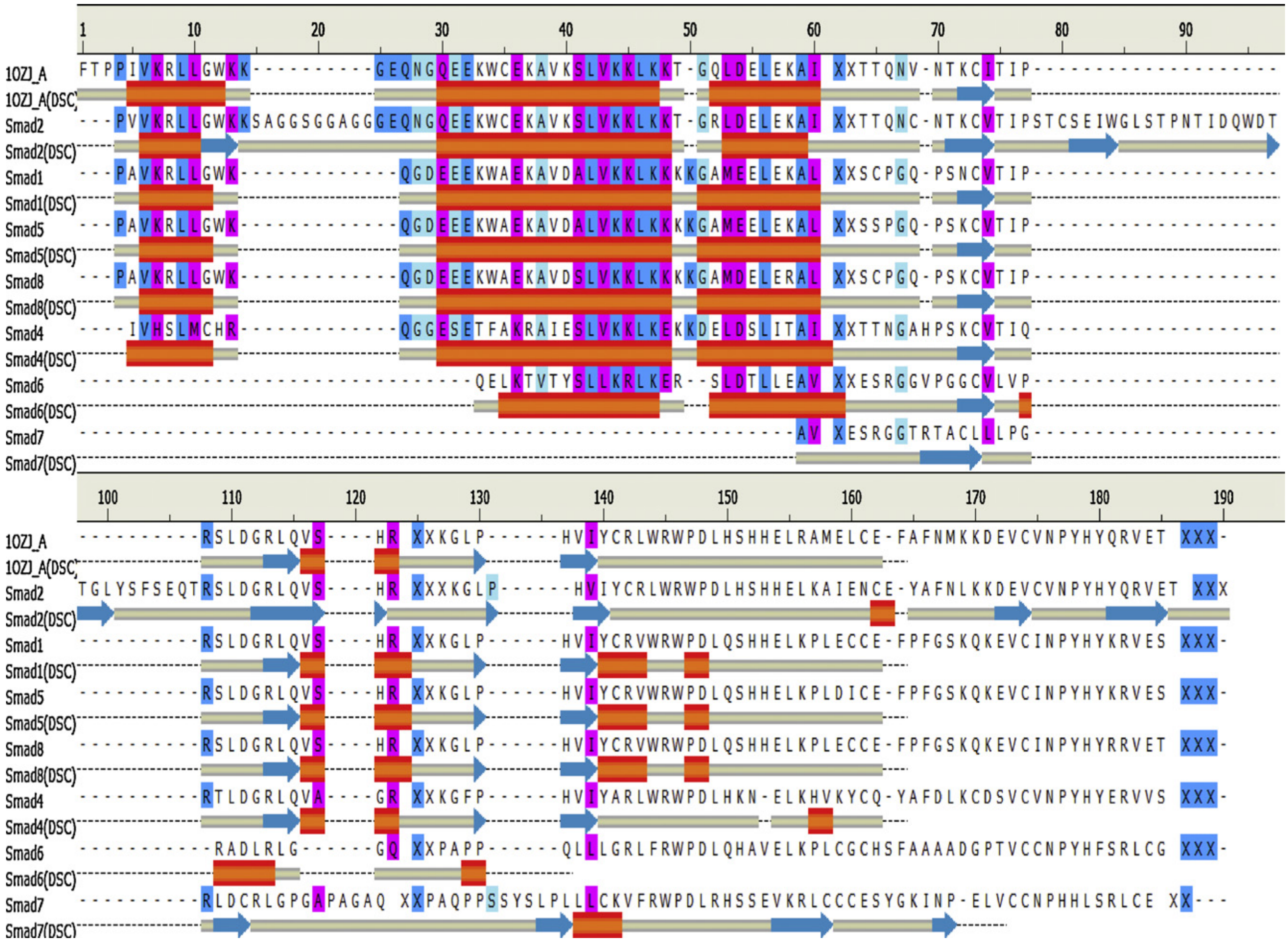
1
10
20
30
40
50
60
70
80
90
1OZJ_A FTPPIVKRLLGWKK-----------GEQNGQEEKWCEKAVKSLVKKLKKT-GQLDELEKAI XXTTQNV-NTKCITIP--------------------
1OZJ_A(DSC)
Smad2 ---PVVKRLLGWKKSAGGSGGAGGGEQNGQEEKWCEKAVKSLVKKLKKT-GRLDELEKAI XXTTQNC-NTKCVTIPSTCSEIWGLSTPNTIDQWDT
Smad2(DSC)
Smad1 ---PAVKRLLGWK-------------QGDEEEKWAEKAVDALVKKLKKKKGAMEELEKAL XXSCPGQ-PSNCVTIP--------------------
Smad1(DSC)
Smad5 ---PAVKRLLGWK-------------QGDEEEKWAEKAVDALVKKLKKKKGAMEELEKAL XXSSPGQ-PSKCVTIP--------------------
Smad5(DSC)
Smad8 ---PAVKRLLGWK-------------QGDEEEKWAEKAVDSLVKKLKKKKGAMDELERAL XXSCPGQ-PSKCVTIP--------------------
Smad8(DSC)
Smad4 ----IVHSLMCHR-------------QGGESETFAKRAIESLVKKLKEKKDELDSLITAI XXTTNGAHPSKCVTIQ--------------------
Smad4(DSC)
Smad6 -------------------------------QELKTVTYSLLKRLKER--SLDTLLEAV XXESRGGVPGGCVLVP--------------------
Smad6(DSC)
Smad7 --AV XESRGGTRTACLLLPG--------------------
Smad7(DSC)
100
110
120
130
140
150
160
170
180
190
1OZJ_A ----------RSLDGRLQVS----HR XXKGLP------HVIYCRLWRWPDLHSHHELRAMELCE-FAFNMKKDEVCVNPYHYQRVET XXX-
1OZJ_A(DSC)
Smad2 TGLYSFSEQTRSLDGRLQVS----HR XXXKGLP------HVIYCRLWRWPDLHSHHELKAIENCE-YAFNLKKDEVCVNPYHYQRVET XXX
Smad2(DSC)
Smad1 ----------RSLDGRLQVS----HR XXKGLP------HVIYCRVWRWPDLQSHHELKPLECCE-FPFGSKQKEVCINPYHYKRVES XXX-
Smad1(DSC)
Smad5 ----------RSLDGRLQVS----HR XXKGLP------HVIYCRVWRWPDLQSHHELKPLDICE-FPFGSKQKEVCINPYHYKRVES XXX-
Smad5(DSC)
Smad8 ----------RSLDGRLQVS----HR XXKGLP------HVIYCRVWRWPDLQSHHELKPLECCE-FPFGSKQKEVCINPYHYRRVET XXX-
Smad8(DSC)
Smad4 ----------RTLDGRLQVA----GR XXKGFP------HVIYARLWRWPDLHKN-ELKHVKYCQ-YAFDLKCDSVCVNPYHYERVVS XXX-
Smad4(DSC)
Smad6 -----------RADLRLG------GQ XXPAPP------QLLLGRLFRWPDLQHAVELKPLCGCHSFAAAADGPTVCCNPYHFSRLCG XXX-
Smad6(DSC)
Smad7 ----------RLDCRLGPGAPAGAQ XXPAQPPSSYSLPLLLCKVFRWPDLRHSSEVKRLCCCESYGKINP-ELVCCNPHHLSRLCE XX---
Smad7(DSC)

Fig. 3. Multiple sequence alignment of MH1 domain sequences with predicted secondary structure elements of all known human Smads. Conserved residues are shown by background color. Blue arrows show the position of β-sheet and red strand shows alpha helix. The sequences of known structure (1OZJ:A) is given in the first row of alignment.

3.2. Homology models and their evaluation

As described in Section 2 we evaluated each of the 12 models that we generated and selected the best model for each Smad-MH1 domain. The PROCHECK, ERRAT and PROVE scores of the finally selected models are given in Table 2. The Ramachandran ϕ and ψ dihedral angles for all the models were above 84% in the chore region of the plot and none of them were in the disallowed region except for a residue of Smad2-MH1 domain. The G-factor of all selected models was grater than -0.11 showing the acceptable quality of various dihedral, covalent and overall bond angles of the models. ERRAT gives an overall quality factor for nonbonded atomic interactions and a higher score of >50 indicates better quality. From the table we could select models that show ERRAT scores greater than 65. PROVE calculates the regularity of the residue volumes and evaluates the structure as a whole. The *Z*-score RMS for all these models was observed to be below 1.76 (Table 2).

Table 1
Percentage sequence identity and similarity of Smad-MH1 domain with respect to template (1OZJ:A). Sequence identity refers to the percentage of matches of the same amino acid residues between two aligned sequences and the similarity refers to the percentage of aligned residues that have similar physiochemical characteristics and can be more readily substituted for each other.

	Identity (%)	Similarity (%)	Bit score	*E*-value
Smad1-MH1	66.9	81.9	161	$4e^{-41}$
Smad2-MH1	68.7	72.9	201	$7e^{-53}$
Smad4-MH1	53.1	75	116	$1e^{-27}$
Smad5-MH1	66.9	83.5	165	$3e^{-42}$
Smad6-MH1	25.2	40.9	53.1	$2e^{-08}$
Smad7-MH1	20.6	34.1	45.4	$4e^{-06}$
Smad8-MH1	69.3	81.9	165	$4e^{-42}$

Fig. 4 shows the verify3D curves for the selected models of Smad1/5/8- and Smad2/4-MH1 domains. We selected the models that show an average score close to template and those with a maximum score above 0.2. The interaction energy values per residue as computed by PROSA II are shown in Fig. 5 for the selected models. The energy profiles were calculated along with the template and they were consistent with a reliable conformation based on their similarity with that of the template.

In summary, the geometric quality of the backbone conformation, the residue interaction, the residue contact and energy profile of the structures are all well within the established limits for reliable structures. All the evaluations suggest that the models picked for further use are of high quality that will allow us to examine their interaction with DNA binding elements.

The homology models of all Smad-MH1 domains that we have used for further analysis are shown in Fig. 6. The overall structure of all Smad-MH1 domains is similar to the template used. The structure of MH1 domain has a novel globular fold, containing four alpha-helices and six short β-strands. The β-strands form two small β-sheets and a β-hairpin that bind with specific DNA sequences and are responsible for DNA recognition. This β-hairpin contains 13 residues namely Leu60, Asp61, Gly62, Arg63, Leu64,

Table 2
Stereochemical quality parameters of the best models of Smad-MH1 proteins selected using PROCHECK, ERRAT and PROVE.

		Selected MH1 domains of various Smads						
		Smad1	Smad2	Smad4	Smad5	Smad6	Smad7	Smad8
PROCHECK: Ramachandran plot	Core	87.50	85.90	84.30	87.50	90.20	84.50	88.50
	Allow	9.60	10.60	13.90	9.60	7.30	12.70	8.70
	Gen allow	2.90	2.80	1.90	2.90	2.40	2.80	2.90
	Disallow	0.00	0.70	0.00	0.00	0.00	0.00	0.00
G-factors	Overall	−0.01	−0.05	−0.11	−0.07	−0.03	−0.16	−0.03
ERRAT		82.301	65.333	73.214	68.142	73.404	73.171	75.221
PROVE	*Z*-score mean	0.281	0.143	0.205	0.311	−0.062	0.097	0.202
	Z-score SD	1.650	1.676	1.63	1.686	1.617	1.764	1.536
	Z-score RMS	1.6	1.68	1.641	1.712	1.615	1.762	1.548
RMSD of the final models with template		0.57	0.2	0.76	0.58	0.37	2.32	0.6

Gln65, Val66, Ser67, His68, Arg69, Lys70, Gly71 and Leu72 in R-Smads (Smad1, 5 and 8). The residues in β-hairpin are represented in green while the backbone is represented with green ribbon. The residues are highly conserved in this region where all have similar residues among all R-Smad. Smad4-MH1 has unique residues (Ala67 and Gly68) at the turn of β-hairpin. A 30 residue insert was observed in Smad2 before the β-hairpin suggesting the weak binding of Smad2 to DNA, as previously reported by other studies [36]. The helix 2 containing many basic residues is conserved among Smads. R-Smads share about 99% of the same sequence in this region. The nuclear localization signal (KKLKK), which constitutes N-terminal alpha helix 2, is conserved in all Smads as was shown earlier [19]. Structure of Smad3-MH1 domain contains four key residues which binds a zinc atom; Cys44, Cys105, Cys122 and His127 [19]. These residues are highly conserved in all Smads and are shown in ball and stick (green) in Fig. 6. The differences between each model with the template are represented by dark pink in Fig. 6(F).

3.3. Smad–DNA interactions

As described in Section 2 we modeled Smad1/5/8–DNA (5′-TCTGCCGCCGCTT-3′) interaction models using the Smad3–DNA (GTCT) structure as template. The resulting models were analyzed for their interactions with DNA and are shown in Fig. 7 with sense (S) and anti-sense (A) strands. The interaction pattern shows that in Smad1-MH1, Lys32 binds to S-A12, Arg63 with A-G5 and S-G4, Lys70 with S-G9 and Gln65 reacts with A-G8. Smad5/DNA model shows interactions of following residues Lys21-(A-C7), Lys29-(A-G8), Lys29-(S-C8), Arg63-(S-G4) and Arg63-(S-G10). Smad8 shows maximum interactions with Arg63-(A-G5), Lys21-(A-C7), Lys29-(A-G8), Gln65-(A-C9), Lys32-(A-T13), Gln65-(S-C8), Lys70-(S-G9),

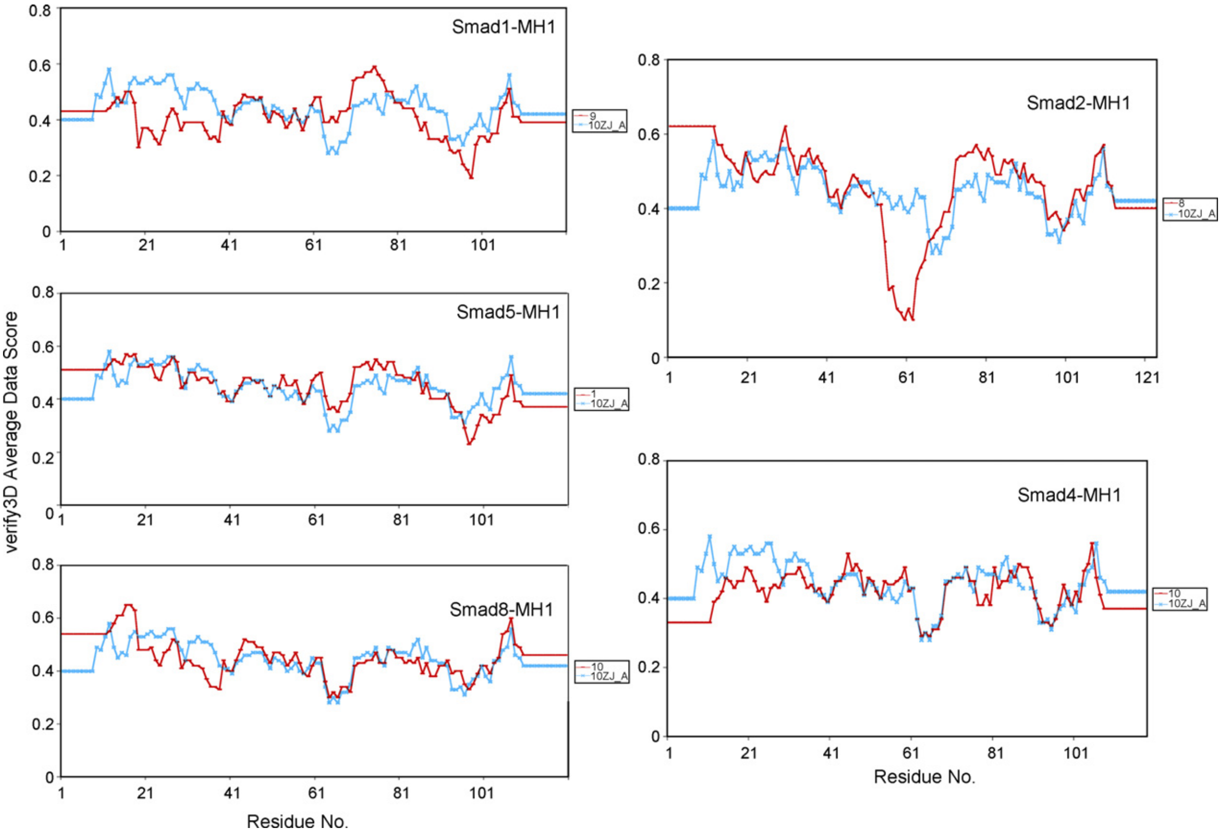

Fig. 4. Verify3D curves of best selected models. This score measures the compatibility of model with its sequence using a scoring function. If more than 80% of the residues have a score of >02 then the protein structure is considered to be of high quality.

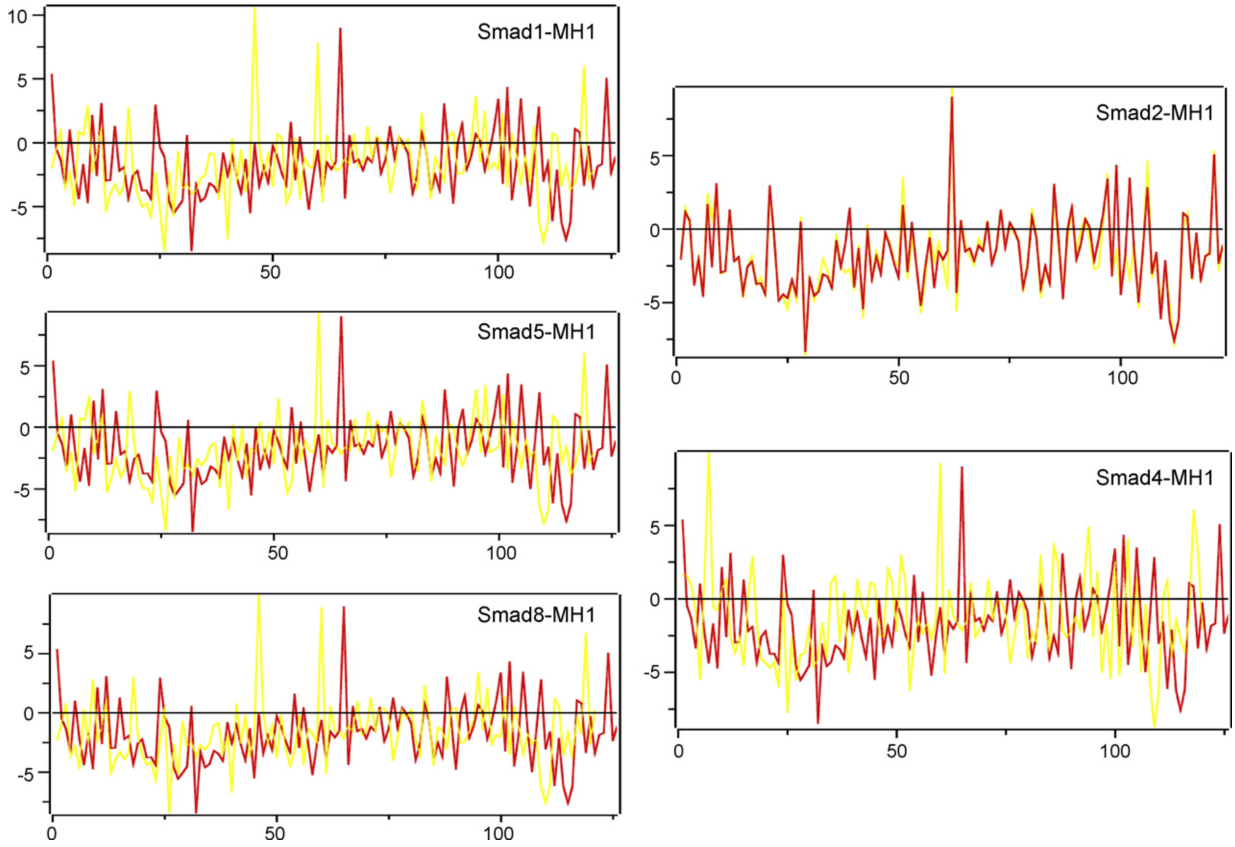

Fig. 5. PROSA II energy graphs of selected models (yellow) and template (red). PROSA captures the average properties of native globular proteins in terms of atom pair, and protein–solvent interactions which shows consistent with a reliable conformation based on its similarity with that of the template.

Lys32-(S-A12) and Lys28-(S-A12). We also observed that the hydrogen bonds between water and protein residues play key role in interaction with DNA which became apparent as several hydrogen bonds stabilized the structure. The key residues forming hydrogen bonds with water are Glu14, Lys33, Leu60, Ser67, His68 and Pro57. Smad1 has additional residues interacting with water which are Lys17, Lys21, Lys28, Lys32, Lys33 and Lys70, Glu20, Asp24 and Asp61, Ser51, Gln65, Leu72 and Arg79. Residues in Smad5 that form hydrogen bonds with water are Glu15, Glu20 and Glu108, Lys21, Lys26, Lys28, Lys29, Lys31, Lys32, Lys70 and Lys107, Asp61, Gly62, Gln65, Arg69, Leu72, Tyr77 and His89. Smad8 differs from Smad5, in containing Tyr18, Asp24, Ser59, Arg79 and Lys105 residues that interact with water. Table 3 shows the contact area of amino acids with the DNA as computed by PSA program [37] which was used to calculate the relative solvent-accessible surface area of all residues in a protein. The difference between PSA values of protein–DNA complex and the individual binding partner provides us the key amino acid residues involved in direct interaction.

Table 4 depicts the binding energies of Smad1/5/8-MH1 interactions with GCCG motif of DNA. We find more or less similar stabilizing binding energies with all Smad1/5/8-MH1 binding. However the binding energies with GTCT, a non-specific, motif are drastically high with some of them having positive values.

4. Discussion

Based on biochemical studies to date Smads appear to be important members in TGF-β pathway whose deficiency is reported in serious diseases such as lung and colorectal cancer, osteoporosis and bone deficiency diseases [19]. We wanted to understand the structural basis for functional similarities and differences observed among different Smad-MH1 domains. In the present study we report homology modeling and evaluation of MH1 domain of Smad1, 2, 4, 5 and 8 using the available crystal

Table 3
Amino acids of Smad1, 5 and 8 interacting with DNA. Total surface area in angstroms (SA) and percentage of surface area (%) of contact is given for each interacting amino acid residue.

Complex	Pos.	AA	SA	%
Smad1_DNA: Chain S	32	Lys	8.44	13.7
	63	Arg	21.94	30.2
	65	Gln	19.13	37.1
Smad1_DNA: Chain A	63	Arg	26.92	37.1
	70	Lys	19.85	32.3
Smad5_DNA: Chain S	21	Lys	11.74	19.1
	29	Lys	11.02	18.0
	63	Arg	22.88	31.5
	65	Gln	19.22	37.2
Smad5_DNA: Chain A	29	Lys	8.63	14.0
	63	Arg	28.14	38.7
Smad8_DNA: Chain S	21	Lys	10.63	17.3
	28	Lys	20.78	33.8
	29	Lys	9.74	15.9
	32	Lys	12.68	20.7
	65	Gln	18.50	35.8
Smad8_DNA: Chain A	32	Lys	11.45	18.6
	63	Arg	30.58	42.1
	65	Gln	7.35	14.3
	70	Lys	23.34	37.9

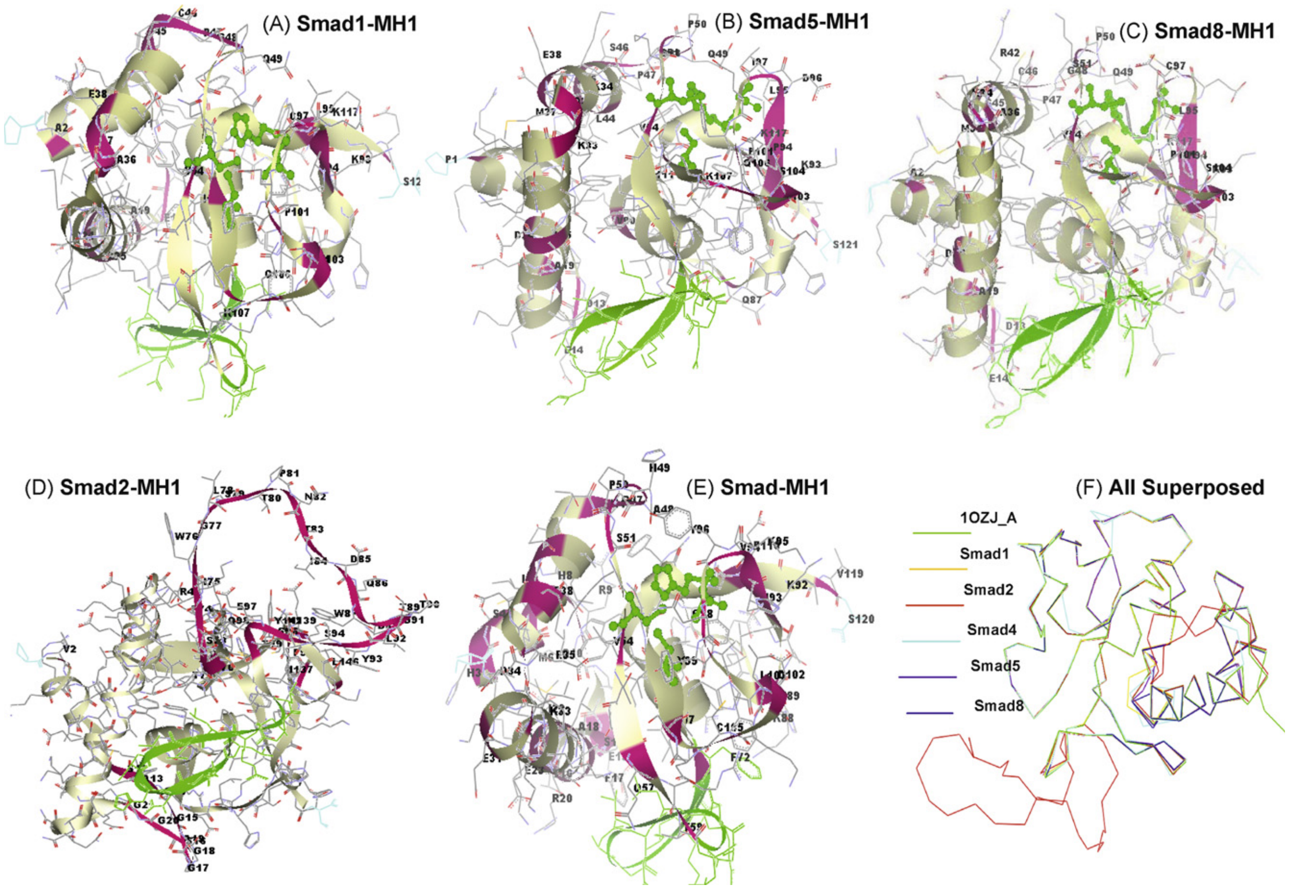

Fig. 6. Cα trace of the models of MH1 domain of various Smads. All Smad proteins show four alpha-helices and six short β-strands. The β-hairpin responsible for DNA binding is shown in green. Note that residues in this strand are conserved in R-Smads but not in Co-Smad and I-Smads. The residues different from template are highlighted and the respective region is depicted by dark pink. The zinc-binding residues are conserved in all Smads and are shown in green ball and stick.

structure of Smad3-MH1. These models can be used to establish phylogenetic relationships as well as identifying characteristic features to design specific inhibitors. Such structural analysis would help to explain the high level of specific DNA binding ability of Smads. Multiple sequence alignments reveal functionally important residues within a protein family. They can be particularly useful for the identification of key residues that determine functional differences between protein subfamilies. The sequence alignment of human Smad-MH1 is shown in Fig. 3 which indicates that the MH1 domain of human Smads is evolutionarily conserved proteins. This conservation is high among R-Smads (Smad1, 2, 3, 5 and 8) and among I-Smads (Smad6 and 7). The secondary structure elements seem to be more conserved than other elements. This is in agreement with the earlier study done on Smad5-MH1 modeling [38]. Smad6 and 7 do not have many conserved residues indicating a similar function related to receptor-regulated and co-mediator Smads.

Protein–DNA interactions play an important role in transcription control. The knowledge and identification of key residues taking part in the interaction and investigation of the mechanism of binding affinity play an important role in underlying macromolecular functions and hence drug discovery [39]. So we further

Table 4
Binding energies (kcal/mol) of Smad1/5/8 interaction with DNA models.

	Ligand energy	Protein energy	Complex energy	Binding energy
DNA with GCCG motif				
Smad1MH1(chainS)–DNA	−2478.83	−4637.56	−7830.88	−714.49
Smad1MH1(chainA)–DNA	−2478.83	−4894.63	−8150.73	−777.27
Smad5MH1(chainS)–DNA	−2636.78	−4524.37	−7833.87	−672.73
Smad5MH1(chainA)–DNA	−2636.78	−4521.86	−7853.76	−695.12
Smad8MH1(chainS)–DNA	−2863.43	−4990.80	−8182.52	−828.29
Smad8MH1(chainA)–DNA	−2863.43	−4580.58	−7999.22	−555.21
DNA with GTCT motif				
Smad1MH1(chainS)–DNA	1423.45	−5075.03	−3781.38	−129.81
Smad1MH1(chainA)–DNA	1423.45	−5028.78	−3648.09	−42.75
Smad5MH1(chainS)–DNA	−867.21	−5241.71	−4707.16	1401.76
Smad5MH1(chainA)–DNA	−867.21	−5038.72	−6030.68	−124.74
Smad8MH1(chainS)–DNA	5067.19	−4804.34	98.62	−164.24
Smad8MH1(chainA)–DNA	5067.19	−4782.65	816.82	532.27

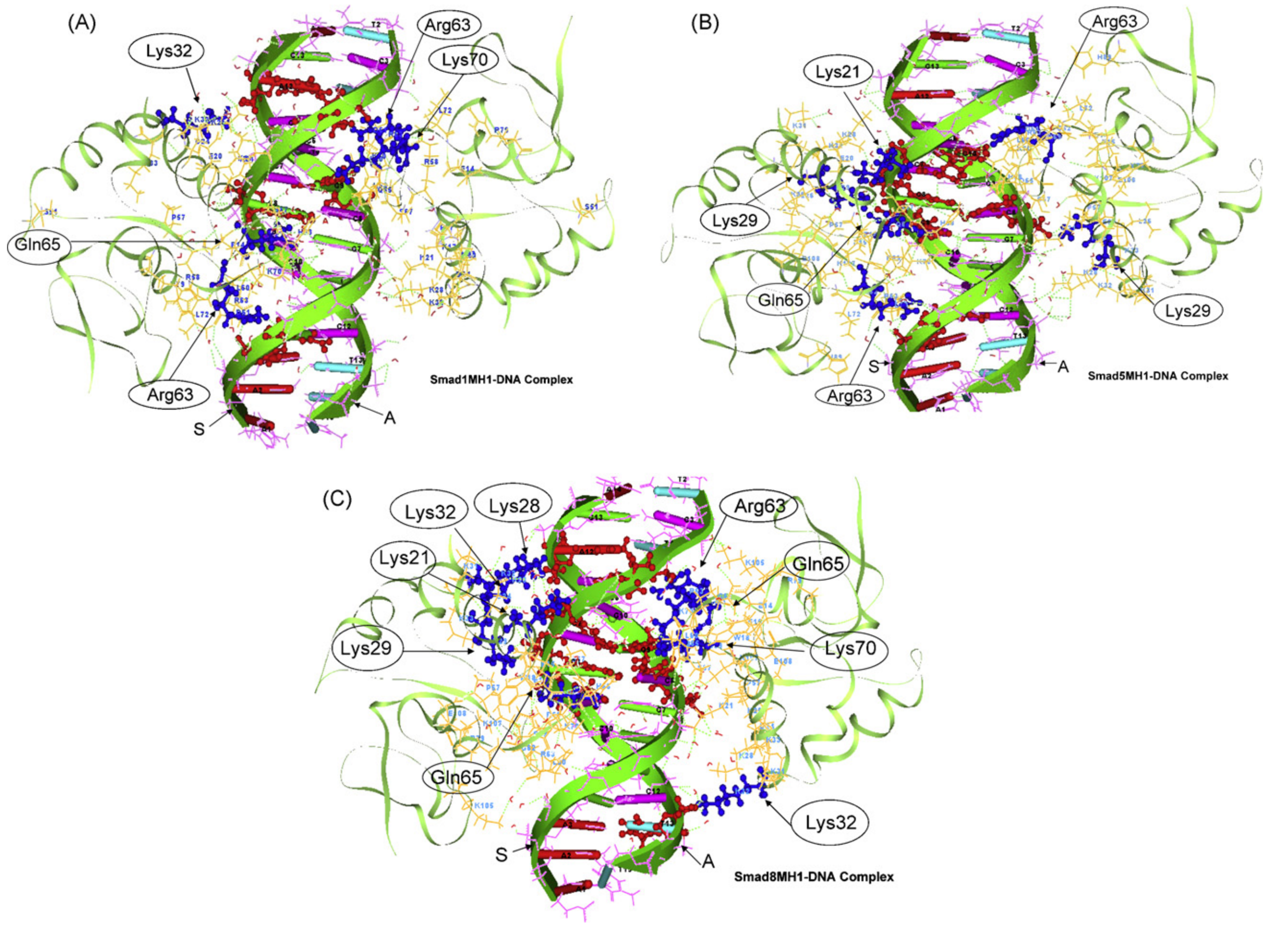

Fig. 7. Smad-MH1/DNA interactions for Smad1, 5 and 8. Residues forming hydrogen bonding with DNA and water are shown. Key residues forming interactions with DNA are shown in dark blue ball and stick. Nucleotides showing interactions with protein are shown in red ball and stick. Hydrogen bonds are shown in green.

studied the interaction between Smad proteins and DNA. It was demonstrated from previous studies that Smad3 and 4 binding sequences yielded a consensus binding site of two inverted repeats of GTCT. Further studies also revealed that Smad–DNA binding require GnCn repeat and it was found that Smad3 and 4 which carry TGF-β and activin signal interact with ‘GTCT’ sequence called Smad box on DNA whereas Smad1, 5 and 8 which mediate BMP signals come in contact with ‘GCCG’ sequence. Since no experimental information was available for BMP signal mediated Smads we modeled the Smad1/5/8–DNA interaction using crystal structure of Smad3-MH1 bound to a specific DNA sequence ‘GTCT’. We use different word that a reliable model of DNA binding domain of Smads and interaction of Smad1/5/8-MH1 domain with DNA would guide further biochemical and genetic efforts in its evaluation as a potential therapeutic target.

The overall structures of all Smad-MH1 domains are similar to Smad3-MH1 and have a novel globular fold, containing four alpha-helices and six short β-strands. The β-strands form two small β-sheets and β-hairpin, which binds with specific DNA sequence and is responsible for DNA recognition. β-Hairpin contains 13 residues namely Leu-Asp-Gly-Arg-Leu-Gln-Val-Ser-His-Arg-Lys-Gly-Leu72 are highly conserved in this region. Slight difference of residues at the turn of β-hairpin in Smad4 and large differences in I-Smads (Smad6 and 7) indicate a small and large functional difference respectively from R-Smads. Non conservation of these residues in I-Smads explains the absence of binding affinity of I-Smads to DNA. Smad7 lacks the β-hairpin. MH1 domain of Smad7 has more differences in sequence from R-Smads when compared to Smad6. Smad6- and 7-MH1 domains also have less similarity in sequence (about 35%). This can account for Smad7 as a more potent inhibitor for TGF-β/activin pathway as compared to Smad6. A 30 residue insert is observed in Smad2 before β-hairpin suggesting the weak binding of Smad2 to DNA as previously reported by other studies.

Helix 2 which contains many basic residues is also conserved among Smads. R-Smads share about 99% of sequence similarity in this region. The nuclear localization signal (KKLKK) which constitutes N-terminal alpha helix 2 is conserved in all Smads as previously shown [19]. It has been demonstrated that helix2 in MH1 domain supports DNA binding through β-hairpin and plays a key role in DNA binding [17]. The structure of Smad3-MH1 domain contains four key residues which binds a zinc atom. These are Cys44, Cys105, Cys122 and His127 [19]. These residues are highly conserved in all Smads. The high conservation in these residues suggests the importance of this metal in Smad proteins.

Smad1/5/8 interacts with ‘GCCG’ motif on SBE rather than ‘GTCT’ to which Smad2/3 interacts. More basic residues in helix2 present next to DNA binding β-hairpin can account for this interaction. The resulting models were analyzed for their interactions with DNA and are shown in Table 3 below. Our models show that Smad1/5/8 interacts mostly at GCCGCCG sequence and the key residues which make direct interaction with DNA are Arg-63, Gln-65 and Lys-70 of Smad1-MH1. It can be seen from Table 3 that above residues show more solvent accessible contact area (SACA) loss as they are involved in interaction between Smad1/5/8 and DNA. These residues are same in Smad1, 5 and 8 and also same DNA interacting residues are found in the Smad3-MH1 crystal structure. However the binding energies of Smad1/5/8-MH1 interaction with the non-specific GTCT motif DNA have comparatively high energy values indicating that

Smad1/5/8 may not recognize the GTCT motif. Therefore, the functional differences in Smads and their preferences for transcription of various genes may depend on yet unknown motif they recognize and also on the other factors such as binding of co-activators and transcription factors to Smad proteins.

5. Conclusion

Genetic and biochemical studies have shown that Smads play an important role as effectors for TGF-β super family in controlling cell fate. MH1 domain of Smad proteins consist of highly conserved 11 residue beta hairpins which bind consensus sequence in DNA and hence is vital for gene activation. Here we modeled structures of DNA binding domains (MH1) of Smad1, 2, 4, 5, 6, 7 and 8 based on crystal structure of Smad3. Due to high target-template similarity, the homology models we generated were of sufficient quality as shown by their assessment results. These reliable models of DNA binding domain of human Smads serve as structural basis for studying biological functions of these proteins. Our further use of these models to study their mode of interaction with DNA provided insight into the functions of the Smads at a molecular level. Smad1/5/8 have same DNA binding residues indicating that differences in their functions may be due to different interacting proteins. Our results also clearly show that Smad1/5/8 do not have similar specificity to bind to GTCT, Smad3 specific, motif. This investigation of crosstalk of Smad proteins with SBEs will allow us to understand the underlying cause of many clinical issues related to Smads.

Acknowledgements

This work was partly supported by a grant from Howard Hughes Medical Institute to Queens College, CUNY and Queens College Research Enhancement Grant. We also acknowledge Dormitory Authority, State of New York.

References

[1] S.J. Lin, T.F. Lerch, R.W. Cook, T.S. Jardetzky, T.K. Woodruff, The structural basis of TGF-beta, bone morphogenetic protein, and activin ligand binding, Reproduction 132 (2006) 179.

[2] J. Massague, TGF-beta signal transduction, Annu. Rev. Biochem. 67 (1998) 753.

[3] P. ten Dijke, K. Miyazono, C.H. Heldin, Signaling inputs converge on nuclear effectors in TGF-beta signalling, Trends Biochem. Sci. 25 (2000) 64.

[4] J.M. Graff, A. Bansal, D.A. Melton, Xenopus Mad proteins transduce distinct subsets of signals for the TGF beta superfamily, Cell 85 (1996) 479.

[5] T. Imamura, M. Takase, A. Nishihara, E. Oeda, J. Hanai, M. Kawabata, K. Miyazono, Smad6 inhibits signalling by the TGF-beta superfamily, Nature 389 (1997) 622.

[6] A. Nakao, M. Afrakhte, A. Moren, T. Nakayama, J.L. Christian, R. Heuchel, S. Itoh, M. Kawabata, N.E. Heldin, C.H. Heldin, P. ten Dijke, Identification of Smad7, a TGFbeta-inducible antagonist of TGF-beta signalling, Nature 389 (1997) 631.

[7] Y. Shi, Y.F. Wang, L. Jayaraman, H. Yang, J. Massague, N.P. Pavletich, Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signalling, Cell 94 (1998) 585.

[8] G. Lagna, A. Hata, A. Hemmati-Brivanlou, J. Massague, Partnership between DPC4 and SMAD proteins in TGF-beta signalling pathways, Nature 383 (1996) 832.

[9] M. Macias-Silva, S. Abdollah, P.A. Hoodless, R. Pirone, L. Attisano, J.L. Wrana, MADR2 is a substrate of the TGFbeta receptor and its phosphorylation is required for nuclear accumulation and signalling, Cell 87 (1996) 1215.

[10] A. Hata, R.S. Lo, D. Wotton, G. Lagna, J. Massague, Mutations increasing autoinhibition inactivate tumour suppressors Smad2 and Smad4, Nature 388 (1997) 82.

[11] F. Liu, C. Pouponnot, J. Massague, Dual role of the Smad4/DPC4 tumor suppressor in TGFbeta-inducible transcriptional complexes, Genes Dev. 11 (1997) 3157.

[12] Y. Shi, A. Hata, R.S. Lo, J. Massague, N.P. Pavletich, A structural basis for mutational inactivation of the tumour suppressor Smad4, Nature 388 (1997) 87.

[13] J. Massague, D. Wotton, Transcriptional control by the TGF-beta/Smad signaling system, EMBO J. 19 (2000) 1745.

[14] A. Moustakas, Smad signalling network, J. Cell Sci. 115 (2002) 3355.

[15] S. Itoh, P. ten Dijke, Negative regulation of TGF-beta receptor/Smad signal transduction, Curr. Opin. Cell Biol. 19 (2007) 176.

[16] S. Dennler, S. Itoh, D. Vivien, P. ten Dijke, S. Huet, J.M. Gauthier, Direct binding of Smad3 and Smad4 to critical TGF beta-inducible elements in the promoter of human plasminogen activator inhibitor-type 1 gene, EMBO J. 17 (1998) 3091.

[17] K. Kusanagi, M. Kawabata, H.K. Mishima, K. Miyazono, Alpha-helix 2 in the amino-terminal mad homology 1 domain is responsible for specific DNA binding of Smad3, J. Biol. Chem. 276 (2001) 28155.

[18] J.M. Yingling, M.B. Datto, C. Wong, J.P. Frederick, N.T. Liberati, X.F. Wang, Tumor suppressor Smad4 is a transforming growth factor beta-inducible DNA binding protein, Mol. Cell. Biol. 17 (1997) 7019.

[19] J. Chai, J.W. Wu, N. Yan, J. Massague, N.P. Pavletich, Y. Shi, Features of a Smad3 MH1–DNA complex. Roles of water and zinc in DNA binding, J. Biol. Chem. 278 (2003) 20327.

[20] K. Johnson, H. Kirkpatrick, A. Comer, F.M. Hoffmann, A. Laughon, Interaction of Smad complexes with tripartite DNA-binding sites, J. Biol. Chem. 274 (1999) 20709.

[21] K. Kusanagi, H. Inoue, Y. Ishidou, H.K. Mishima, M. Kawabata, K. Miyazono, Characterization of a bone morphogenetic protein-responsive Smad-binding element, Mol. Biol. Cell 11 (2000) 555.

[22] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 35 (2007) D5.

[23] W.R. Pearson, Comparison of methods for searching protein sequence databases, Protein Sci. 4 (1995) 1145.

[24] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389.

[25] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876.

[26] A. Sali, T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, J. Mol. Biol. 234 (1993) 779.

[27] M.C. Peitsch, ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling, Biochem. Soc. Trans. 24 (1996) 274.

[28] O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, S. Brunak, Protein distance constraints predicted by neural networks and probability density functions, Protein Eng. 10 (1997) 1241.

[29] A.L. Morris, M.W. MacArthur, E.G. Hutchinson, J.M. Thornton, Stereochemical quality of protein structure coordinates, Proteins 12 (1992) 345.

[30] D. Eisenberg, R. Luthy, J.U. Bowie, VERIFY3D: assessment of protein models with three-dimensional profiles, Methods Enzymol. 277 (1997) 396.

[31] C. Colovos, T.O. Yeates, Verification of protein structures: patterns of nonbonded atomic interactions, Protein Sci. 2 (1993) 1511.

[32] J. Pontius, J. Richelle, S.J. Wodak, Deviations from standard atomic volumes as a quality measure for protein crystal structures, J. Mol. Biol. 264 (1996) 121.

[33] M. Li, B. Wang, Homology modeling and examination of the effect of the D92E mutation on the H5N1 nonstructural protein NS1 effector domain, J. Mol. Model. 13 (2007) 1237.

[34] M.J. Sippl, Recognition of errors in three-dimensional structures of proteins, Proteins 17 (1993) 355.

[35] B.V. Reddy, V. Gopal, D. Chatterji, Recognition of promoter DNA by subdomain 4.2 of *Escherichia coli* sigma 70: a knowledge based model of -35 hexamer interaction with 4.2 helix-turn-helix motif, J. Biomol. Struct. Dyn. 14 (1997) 407.

[36] M. Simonsson, M. Kanduri, E. Gronroos, C.H. Heldin, J. Ericsson, The DNA binding activities of Smad2 and Smad3 are regulated by coactivator-mediated acetylation, J. Biol. Chem. 281 (2006) 39870.

[37] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379.

[38] R. Hariharan, M.R. Pillai, Homology modeling of the DNA-binding domain of human Smad5: a molecular model for inhibitor design, J. Mol. Graph. Model. 24 (2006) 271.

[39] A. Hoglund, O. Kohlbacher, From sequence to structure and back again: approaches for predicting protein–DNA binding, Proteome Sci. 2 (2004) 3.