

Sequence and structural analysis of 4SNc-Tudor domain protein from *Takifugu rubripes*

Jianzhou Zheng¹, Jian Lu¹, Haijun Liu¹, Jun Li², Keping Chen^{1*}

¹Institute of life science, Jiangsu University, Zhenjiang, China; ²Physical and Chemical Analysis Center, Jiangsu University, Zhenjiang, China; Keping Chen - E-mail: kpchen@ujs.edu.cn; Phone: 86 511 88791923; * Corresponding authors

Received April 01, 2009; Revised May 26, 2009; Accepted June 13, 2009; Published September 30, 2009

Abstract:

The fugu SN4TDR protein belongs to an evolutionarily conserved family, consisting of four repeat staphylococcal nuclease-like domains (SN1-SN4) at the N-terminus followed by Tudor and SN-like domains (TSN). Sequence analysis showed that the C-terminal TSN domain is composed of a complete SN-like domain interdigitated with a Tudor domain. In despite of low level of sequence identities, five SN-like domains have a few conserved amino acids that may play essential roles in the function of the protein. Computer modeling and secondary structural prediction of the SN-like domains revealed the presence of similar structural features of $\beta 1$ - $\beta 2$ - $\beta 3$ - $\alpha 1$ - $\beta 4$ - $\beta 5$ - $\alpha 2$ - $\alpha 3$, which provides a structural basis for oligonucleotides binding. The loop region L₃₀ for binding sites between $\beta 3$ and $\alpha 1$ of SN-like domains are different from human p100, implying the divergence in the structures of binding sites. These results indicate that fugu SN4TDR may bind methylated ligands and/or oligonucleotides through its distant domains.

Keywords: Takifugu rubripes, SN4TDR; SN-like domain; Tudor domain

Background:

4SNc-Tudor domain proteins (SN4TDR) have been identified as highly conserved proteins among species, but not in bacteria [1-2]. There are usually four repeat domains with homology to staphylococcal aureus nuclease (SN) at the N-terminus followed by Tudor domain and SN-like domain at the C-terminus, both of which are defined as TSN domain. The SN4TDR in eukaryotes is a key regulator of gene expression and plays an important role in both transcription and pre-mRNA splicing. The two distinct domains of SN4TDR SN-like domain and Tudor domain act as interaction mediators for nucleic acids and proteins, respectively [3]. The protein was first identified as a coactivator of the Epstein-Barr virus nuclear antigen 2 (EBNA-2) [4] and later found to be able to interact with other transcription factors such as c-Myb [5], Stat5 [6] and Stat6 [7]. Notably, SN4TDR leads to cell proliferation via activating some transcription factors and is linked to autosomal-dominant polycystic kidney disease (ADPKD). The SN4TDR was also found to be one of the members of the RNA-induced gene silencing (RISC) that could bind double-stranded RNA and ultimately degrade them [8], especially for hyper-edited double-stranded RNA containing multiple I-U and U-I pairs [3]. Recently, the SN4TDR was shown to be able to promote both the in-vitro spliceosome complex formation and the first step of pre-mRNA splicing by interacting with U5 snRNP (small nuclear ribonucleoproteins) [9].

The SN4TDR proteins from eukaryotic organisms were classified into 5 categories based on 5th SN-like domain [10]. In contrast, the overall length and sequences of the proteins from fish were highly conserved, suggesting an essential role in this species. Furthermore, the conservation was also observed in SN4TDR proteins between human and fish. Although the human SN4TDR has been studied intensively, the structure and function of fugu (*Takifugu rubripes*) SN4TDR remains unknown. To investigate the function of fugu SN4TDR, here we performed a systemic bioinformatics analysis with a novel HCA (hydrophobic-cluster analysis) method. The analysis reveals that the protein has such a modular architecture including four repeat SN-like domains at the N-terminus, followed by Tudor domain and complete SN-like domain. In addition, some conserved amino acids were detected within fugu SN-like domains, suggesting their important roles in the function of fugu SN4TDR.

Methodology:

Searching database for right sequences

The sequence data of fugu SN4TDR (BAD32626), human SN4TDR (NP_055205), staphylococcal aureus nuclease (2SNS_A) and SMN (the survival of motor neurons) (1MHN_A) for sequence analysis

was obtained directly from the database at National Centre for Biotechnology Information (NCBI).

HCA plot

Due to the low level of sequence identities, HCA program at the Mobyly server (<http://mobyly.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py?form=HCA>) based on the methods described by Callebaut et al [11] was used to detect similar plots between fugu SN4TDR, SNase and SMN, which could determine the presence of similar three-dimensional folds. The five SN-like domains of fugu SN4TDR detected by HCA analysis were used for multiple sequence alignment by the CLUSTAL W program (<http://www.ebi.ac.uk/clustalw/>) based on the methods described by Thompson et al [12]. They were subsequently used to generate input files to BOXSHADE for direct-viewing multiple alignment output.

Modeling of fugu SN-like domains

The five SN-like domains from fugu SN4TDR were modeled by an automatic web server in Geno 3D home page (<http://geno3d-pbil.ibcp.fr/>) according to the methods described by Combet et al [13]. The crystal structure of SNase (PDB 1SNc) was used as a template for modeling four N-terminal SN-like domains (SN1-SN4), and the crystal structure of p100 co-activator Tudor domain (PDB 2HQE) was used for modeling C-terminal SN-like domain (SN5). PredictProtein program at the ExPASy Molecular Biology Server (<http://cubic.bioc.columbia.edu/>) was performed according to the methods described by Rost et al [14].

Results:

The detection of five SN-like domains and Tudor domain based on HCA plot

By searching the conserved domain against SMART database, we found that fugu SN4TDR sequence contains two distant domains, four repeat SN-like domains (SN1 to SN4) and TSN domain. The later is comprised of the fifth SN domain (SN5) and Tudor domain. The comparison between five SN-like domains shows that they have low sequence identities in the amino acid level. In order to understand structure of fugu SN4TDR, the four SN-like domains and TSN domain were analyzed by hydrophobic cluster analysis (HCA), which aligns protein sequences relying on a two-dimensional (2D) representation of the sequences rather than the sequence similarities. The HCA plot from Mobyly was employed for the comparison of the fugu SN4TDR with staphylococcal aureus nuclease and SMN (the survival of motor neurons) protein sequences for detecting similar motifs.

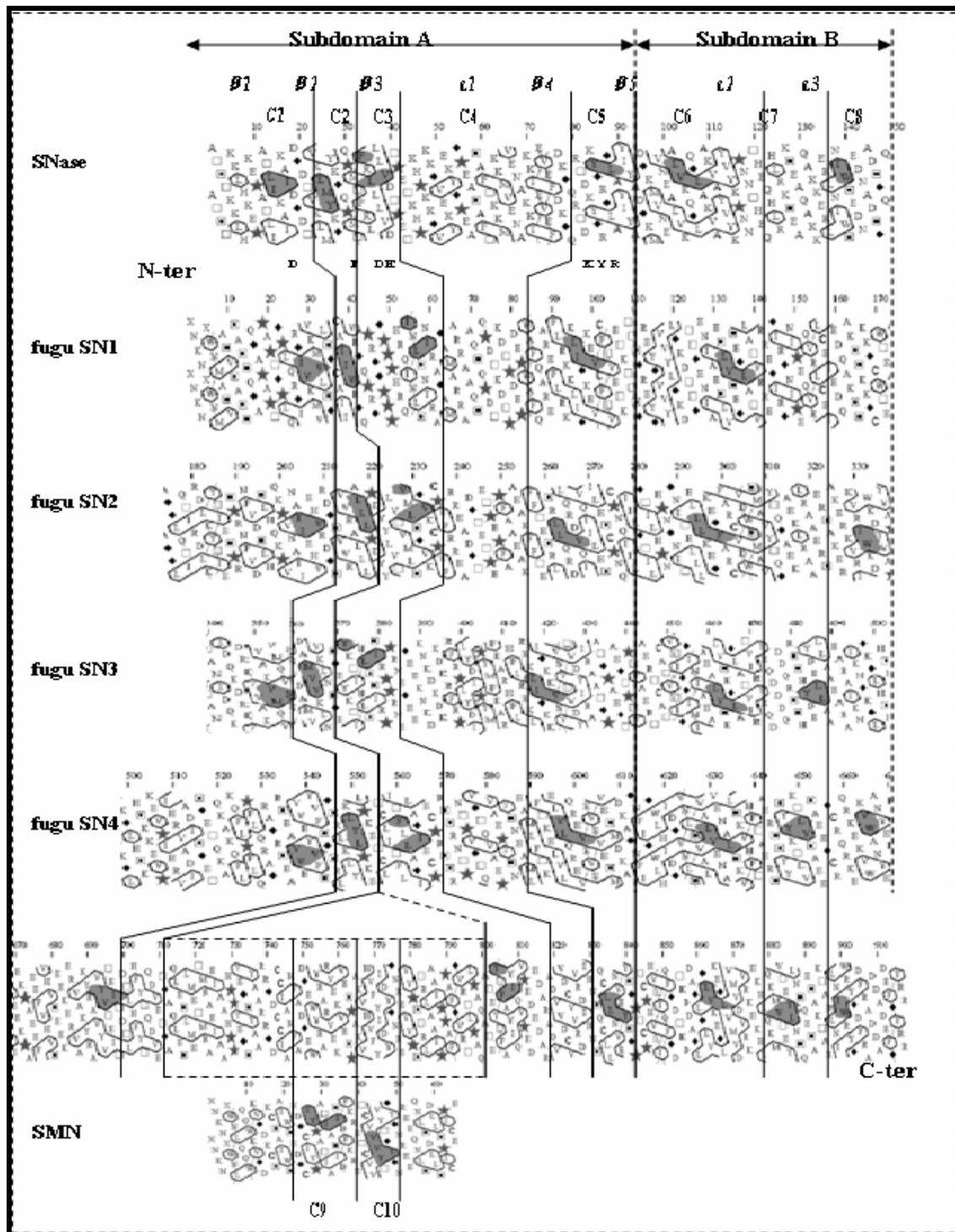


Figure 1: HCA plots of SN-like domains, SNase and SMN showing their structural homology. In the HCA plots, the residues are shown on a duplicated α -helical net and the clusters of hydrophobic residues are automatically drawn. One-letter corresponds to one amino acid, except for proline, glycine, serine and threonine which are represented by the symbols \star , \blacklozenge , \square , \square respectively. The catalytic sites (D21, F34, D40, E43, K84, Y85 and R87) of SNase described in Christopher [17] and its secondary structure predicted by PredictProtein (β 1- β 2- β 3- α 1- β 4- β 5- α 2- α 3) are shown in HCA plot of SNase. SNase and SN-like domains are divided into subdomain A and subdomain B. Hydrophobic clusters with similar characteristics between SNase and SN-like domains are designated as motifs C1 to C8, while similar clusters between SMN and Tudor domain are denoted as motifs C9 to C10. Vertical lines indicate the putative correspondences between sequences; clusters with identical characteristics are darkly shaded. SNase = staphylococcal nuclease; SMN = the survival of motor neurons.

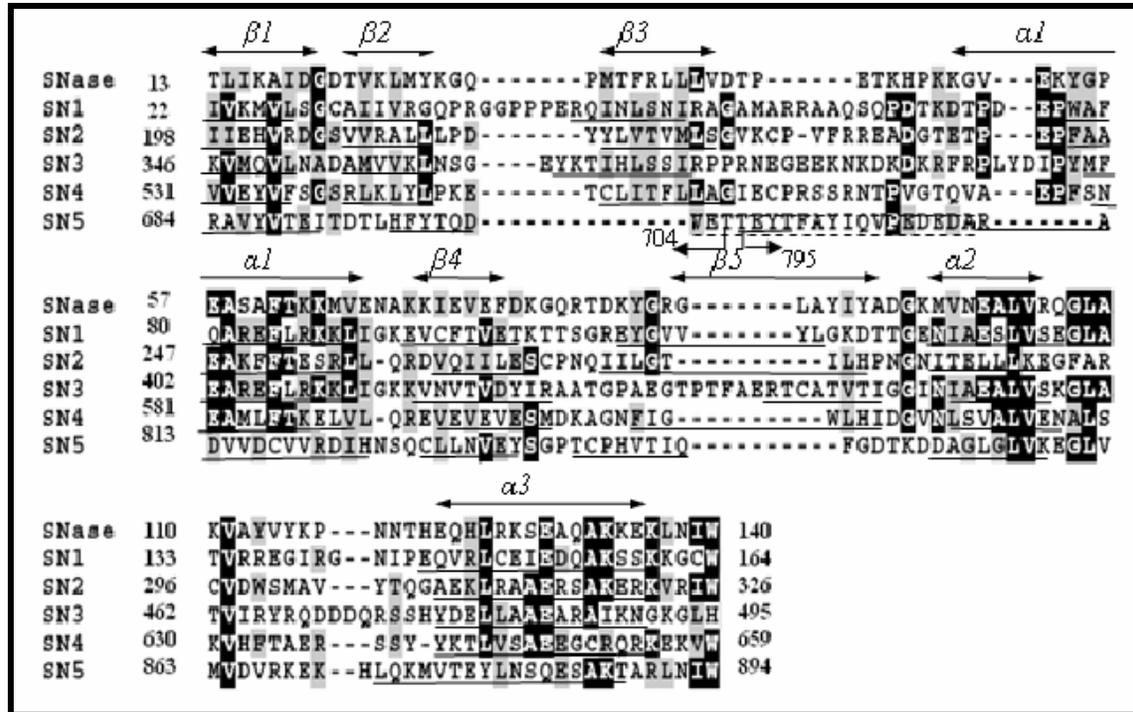


Figure 2: Multiple sequence alignment of SNase and five SN-like domains based on HCA analysis. The protein sequence alignment is performed by CLUSTAL W program using default parameters, and then handled using BOXSHADE. The secondary structure of SNase protein sequences predicted by PredictProtein are denoted as $\beta 1$, $\beta 2$, $\beta 3$, $\alpha 1$, $\beta 4$, $\beta 5$, $\alpha 2$ and $\alpha 3$. The corresponding secondary structures of SN-like domains are displayed as underline at the bottom of the sequences. The region $\beta 3$ between $\beta 2$ and $\alpha 1$ in SN5 domain is predicted by modeling SN5 and denoted as dashed line. The region $\beta 3$ of SN5 domain is analyzed according to crystal structure of human SN4TDR protein.

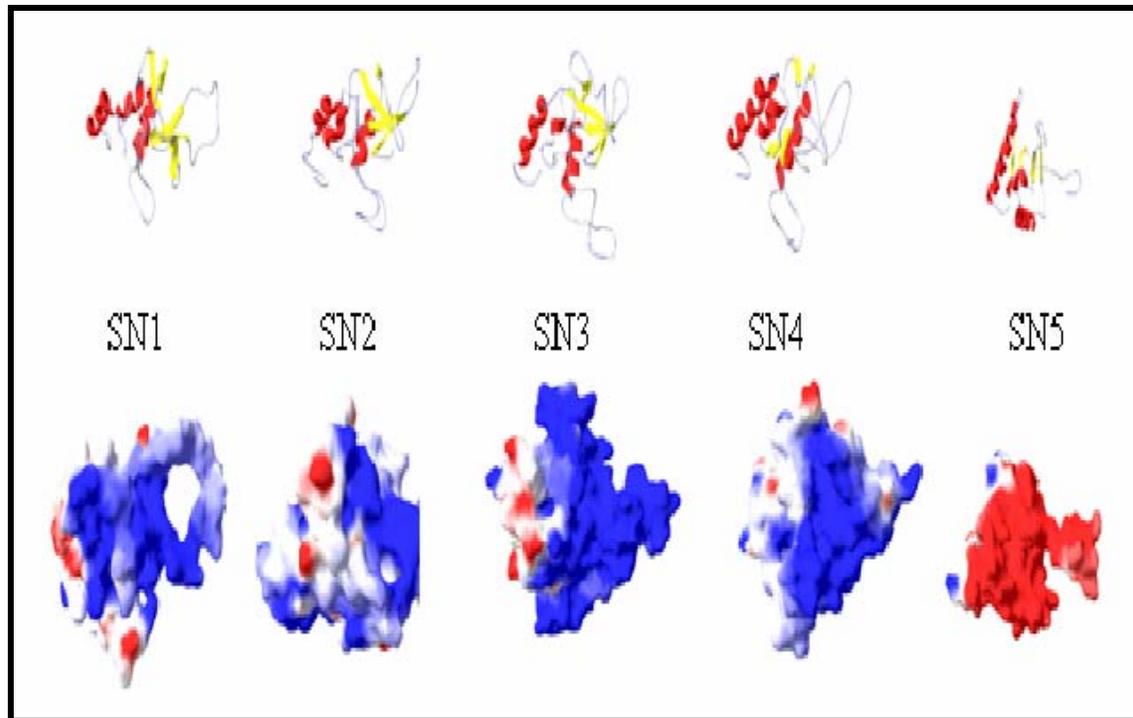


Figure 3: Structure of fugu SN-like domains modeled using Geno 3D. (A) Cartoon illustration of five SN-like domains structures. Red, α -helix; yellow, β -sheet; (B) Surface electrostatic potential calculated by pdbviewer show positively charged surfaces as blue and negatively charged surface as red. The classification of residues is as follows: positively charged, Arg and Lys; negatively charged, Asp and Glu.

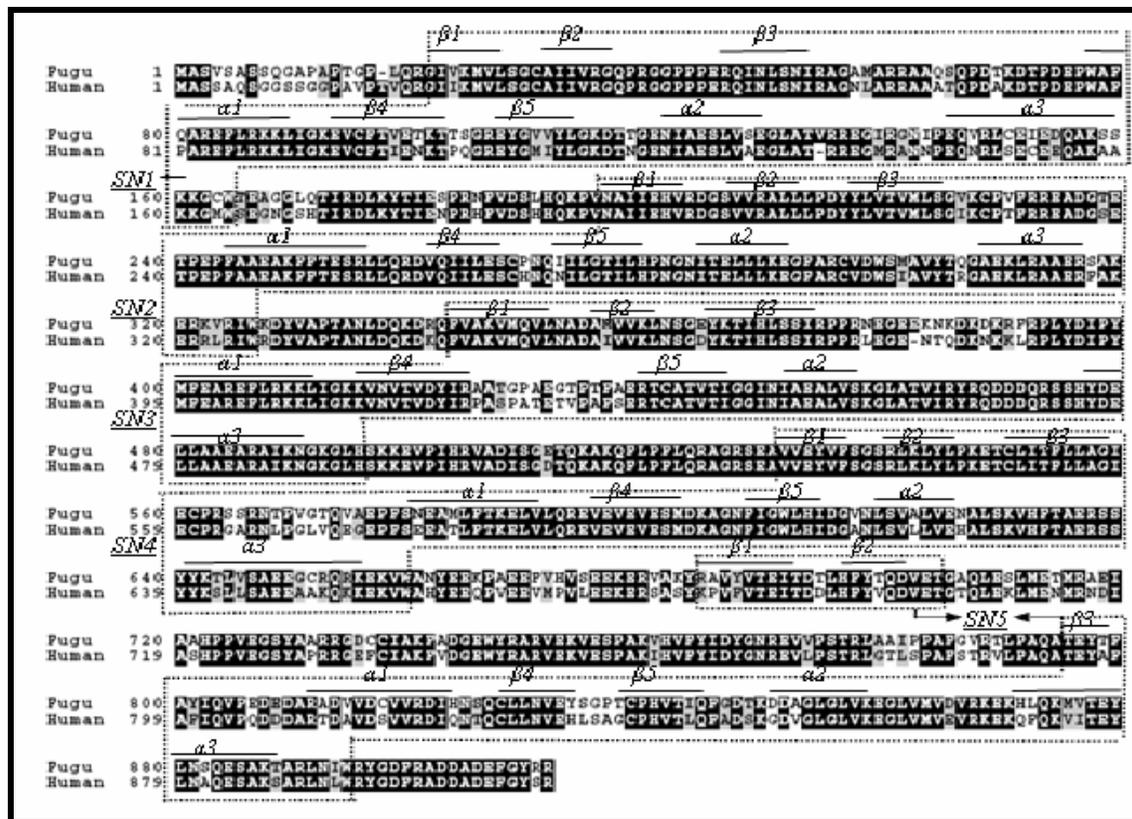


Figure 4: Comparison of fugu and human SN4TDR. The protein sequence comparison is performed by CLUSTAL W program using default parameters, and then handled using BOXSHADE. The SN-like domains are denoted using dashed line. The secondary structures of five SN-like domains are displayed as upper line.

The results indicated that the fugu SN4TDR has a modular architecture. At the N-terminus, four complete SN-like domains are located in residues 22-164, 198-326, 346-495 and 531-659, and each SN-like domain includes four similar repetitive hydrophobic domains. At the C-terminus, the SN-like domain are located in residues 684-704 and residues 795-896, both of which interdigitate with the Tudor domain (residues 705-794) (Figure 1).

Similar to SNase, all five SN-like domains of fugu SN4TDR protein contain a similar hydrophobic cluster that includes eight similar hydrophilic motifs (designated as motif C1 to C8, Figure 1). Relative to other motifs, C1, C3, and C6 are well-conserved in all SN-like domains (Figure 1). The motifs C1 and C2 in SN-like domains are linked by a loop (L₁₂), which contributes a conserved glycine necessary for nuclease binding. An exception is observed in SN3 domain, where glycine is replaced by an alanine. For loop L_{3a} (linking β3 and α-helix), the sequence and length are different between five modeled SN-like domains and SNase.

The HCA analysis showed that Tudor domain of TSN domain (residues 705-794) (Figure 1) is similar to that of SMN (MHN), which contains a typical β-barrel domain with four β-sheets. Furthermore, the hydrophobic cluster of Tudor domain (designated as motif C9 and C10) is also similar to that of SMN Tudor domain (Figure 1). In motifs C9 and C10, the residues Phe741, Try747, Try764, and Try767 correspond to the residues Trp102, Try109, Try127 and Try130 in SMN Tudor domain. The four residues of SMN Tudor domain have been proved to form an aromatic cage that is associated with the protein-protein interactions by enclosing a dimethylated arginine ligand to the cage [15]. Previous studies described a similar mechanism for recognition and binding of methylated amino acids residues in the Tudor domains of JMJD1A [16] and 53BP1 [8].

Secondary structure and modeling of fugu SN-like domains

The secondary structure elements of all SN-like domains were predicted by PredictProtein. The result indicates that five SN-like domains closely resemble the overall structure of staphylococcal nuclease, which mainly includes a five-stranded β-barrel capped by an α-helix between β3 and β4, followed by two α-helices (α2 and α3, termed as subdomain B). It is of note that two amino acids (Asp21 and Asp40) in SNase required for catalysis are missing in the structure of five SN-like domains. In spite of the similarity of secondary structure elements, the surface residues of SN-like domains differ from those of SNase. Model analysis of five N-terminal SN-like domains indicates that SN1-SN4 domains have positively charged surfaces that provide a solvent accessible surface, whereas SN5 has negatively charged surface unable to interact with nucleic acid (Figure 3).

Discussion:

4SNc-Tudor domain protein (SN4TDR) was first identified as a transcriptional coactivator of the Epstein-Barr virus nuclear antigen 2, coactivating gene expression by interacting with the EBNA-2 acidic domain [4]. Previous studies indicated that SN4TDR was conserved in model structures among a wide variety of organisms. It contains four repeat SN-like domains at the N-terminus; followed by a Tudor domain and a variety of SN-like domain (TSN). The high degree of structural conservation in many eukaryotic organisms may imply that SN4TDR plays an indispensable role in eukaryotic cells. Different from the model structures, the amino acid sequence of SN4TDR is less conserved (below 30%). Therefore, HCA analysis was used to detect the presence of the conserved structure within N-terminal SN-like domains and TSN domain. For the diversity of the fifth SN-like domain (SN5), SN4TDR is classified into 5 categories based on homology and the status of SN5 domain [10]. In addition, the fugu SN5 domain contains a full-length secondary structure formed by residues 687-704 (β1 and β2) and residues 795-896 (β3, α1, β4, β5, α2 and α3). Multiple sequence

alignment showed there is a low identity between five SN-like domains and SNase. However, when compared with diverse eukaryotes, the domains of fugu SN4TDR are conserved, suggesting that they are functionally important [17]. In addition, the secondary structural analysis by PredictProtein (Figure 2) indicated that SN-like domains contain the typical OB-folds, implying a function of binding oligonucleotides or oligosaccharides [18]. It is notable that subdomain B ($\alpha 2$ and $\alpha 3$) has more similarities than subdomain A ($\beta 1$, $\beta 2$, $\beta 3$, $\alpha 1$, $\beta 4$ and $\beta 5$) (Figure 2). The conserved residues in subdomain B are considered to be necessary for its stability [17], and the subdomain A appears to contribute to oligonucleotides binding due to the presence of typical OB-fold [18]. In subdomain A, the L_{3a} loop regions of fugu SN4TDR between $\beta 3$ and $\alpha 1$ are rich in Arg and Lys, implying a potential to bind phosphates. Interestingly, the overall structures of SN4TDR between fugu and human are conserved in sequence and length, except for L_{3a} where residues are different, suggesting different substrates for binding sites (Figure 4). Furthermore, the charged properties of L_{3a} of SN-like domains except for SN5 domain have no significant difference between fugu and human.

Tudor domains have been found in many eukaryotic organisms and are involved in protein-protein interactions. Recent study on mutagenesis of p100 TSN domain revealed that methylated ligands are trapped inside a cage which is composed of at least three aromatic amino acids residues [19]. The Tudor domain of the survival of motor neurons (SMN) was demonstrated to bind to dimethylated arginines of arginine-glycine (RG) rich sequences at the C-terminus of Sm proteins. Here, we indicated that four amino acids (Phe741, Try747, Try764, and Try767) in Tudor domain are invariant and likely to form conserved aromatic cage to bind methylated ligands. The model analysis of five SN-like domains shows that there are highly conserved secondary structures and differences between N-terminal SN1-SN4 and C-terminal SN5. The similar secondary structures and positively charged surfaces between fugu and human SN-like domains suggest that fugu SN1-SN4 may bind DNA and/or double-stranded RNA in the same way as human SN1-SN4 [20]. In contrast, the SN5 domain is unable to bind phosphoric acids due to the presence of negatively charged surface. In addition, it is demonstrated that TSN domain of p100

could interact with snRNP complexes and promote pre-mRNA splicing [19]. Thereby charged surface of SN5 domain may compromise methylated ligands binding of Tudor domain.

References:

- [1] S Abe *et al.*, *Mar Biotechnol.* **7**:677 (2005) [PMID: 16132464]
- [2] RA Howard-Till, MC Yao, *Eukaryot Cell.* **6**:1795 (2007) [PMID: 17715366]
- [3] AD Scadden *Mol Biol.* **12**:489 (2005) [PMID: 15895094]
- [4] X Tong *et al.*, *Mol Cell Biol.* **15**:4735 (1995) [PMID: 7651391]
- [5] T Valineva *et al.*, *J Biol Chem.* **280**:14989 (2005) [PMID: 15695802]
- [6] K Paukku *et al.*, *Mol Endocrinol.* **17**:1805 (2003) [PMID: 12819296]
- [7] J Yang *et al.*, *EMBO J.* **21**:4950 (2002) [PMID: 12234934]
- [8] Y Huyen *et al.*, *Nature* **432**:406 (2004) [PMID: 15525939]
- [9] J Yang *et al.*, *Nucleic Acids Res.* **35**:4485 (2007) [PMID: 17576664]
- [10] S Abe *et al.*, *J Exp Bot.* **54**:971 (2003) [PMID: 12598568]
- [11] I Callebaut *et al.*, *Cell Mol Life Sci.* **53**:621 (1997) [PMID: 9351466]
- [12] JD Thompson *et al.*, *Nucleic Acids Res.* **22**:4673 (1994) [PMID: 7984417]
- [13] C Combet *et al.*, *Bioinformatics* **18**:213 (2002) [PMID: 11836238]
- [14] B Rost *Nucleic Acids Res.* **32**:321 (2004) [PMID: 15215403]
- [15] PR Nielsen *et al.*, *Nature* **416**:103 (2002) [PMID: 11882902]
- [16] Y Huang *et al.*, *Science* **312**:748 (2006) [PMID: 16601153]
- [17] PP Christopher *Protein Sci.* **6**:459 (1996) [PMID: 9041650]
- [18] AG Murzin *EMBO J.* **12**: 861 (1993) [PMID: 8458342]
- [19] S Neil *et al.*, *Nat Struct Mol Biol.* **14**:779 (2008) [PMID: 17632523]
- [20] CL Li *et al.*, *Nucleic Acids Res.* **36**: 3579 (2008) [PMID: 18453631]

Edited by V. S. Mathura

Citation: Zheng *et al.*, *Bioinformatics* 4(3): 127-131 (2009)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.