Protein Structural Data Mining and Evolutionary Bioinformatic Analysis on Domains of TATA-box Binding Protein-like Fold

Jian-Xiang Liu, Qi Xie, Jun Lin*

Department of Biotechnology, Guilin Medical University, Guilin, 541004, China
* Corresponding author (Tel/fax numbers: 86-773-5890464; E-mail: JLinGLMU@gmail.com)

Abstract: In recent years, data mining has been used widely in the areas of science, such as bioinformatics, genetics and medicine. Studies suggested visual data mining is faster and much more intuitive than is traditional data mining. In the process of X-ray crystallography research, large protein structural data sets have been generated, collected, and stored. A typical TATA-box binding protein-like fold contains a secondary structure arrangement of beta-alpha-beta(4)-alpha (alpha: alpha helix; beta: beta strand). At least three families, including TATA-box binding protein, DNA repair glycosylase and 8-oxoguanine glycosylase, contain this fold. The C-terminal domain of TATA-box binding protein is a duplication region that consists of two clear structural repeats (beta-alpha-beta(4)-alpha-beta-alpha-beta(4)-alpha). In this work, we analyzed the various TATA-box binding protein-like fold using structure and sequence comparisons. Database searching reveals a new homologue, YwmB protein, containing circularly permutated TATA-box binding protein-like fold comparing to the original one. Finally, an extraordinary evolution history of the TATA-box binding protein-like fold was illustrated. Our systematic structural analysis of the crystal structures of TBP-like proteins from the RCSB protein databank, examined the structures of single TBP-like protein domain from various protein families, double TBP-like protein domain containing protein, and double TBP-like protein domain with circular permutation. The TBP-like proteins, including single TBP-like protein domain, double TBP-like protein domain and double circularly permuted TBP-like protein domain, have a common ancestral fold. The C-terminal domain of TATA-box binding protein consists of two structural repeats (beta-alpha-beta(4)-alpha-beta-alpha-beta(4)-alpha) may result from duplication-fusion process .
[Liu JX, Xie Q, Lin J. **Protein Structural Data Mining and Evolutionary Bioinformatic Analysis on Domains of TATA-box Binding Protein-like Fold.** *Life Sci J* 2014;11(2):298-302]. (ISSN:1097-8135). http://www.lifesciencesite.com. 41

Keywords: TATA-box binding protein-like fold; domain duplication-fusion; Circular Permutation (CP); origin of new gene; Evolutionary Bioinformatics

1. Introduction

The analysis of the Knowledge Discovery in Databases process is the computational process of discovering patterns in large data sets. In recent years, data mining has been used widely in the areas of science, such as bioinformatics, genetics and medicine. One example of application of data mining methods are biomedical data facilitated by domain ontologies (Zhu and Davidson, 2007). However, in the process of X-ray crystallography research, large protein structural data sets have been generated, collected, and stored. Studies suggest visual data mining is faster and much more intuitive than is traditional data mining. The overall goal of the data mining process is to extract information from a data set. The TATA-binding protein-like fold (TBP-like fold) could be a transcription factor whose DNA binding fold is composed of a curved anti-parallel beta-sheet (Nikolov, et al., 1992). However, the TATA-binding protein contains a duplication of this fold. Domain duplication is common during protein evolution (Lin, et al., 2009), and the molecule may undergo a gene fusion after a gene duplication event. This fold at least contains three families, and The N terminal domain of DNA glycosylase has only a single copy of the fold, whereas TATA-binding protein (TBP) contains a duplication of this fold. TATA-binding protein-like clan in Pfam (http://pfam.sanger.ac.uk/clan/TBP-like) is composed of a curved antiparallel beta-sheet and this fold is also found in the N terminal region of DNA repair glycosylases, that is AlkA N-terminal domain and 8-oxoguanine DNA glycosylase.

The TATA-box binding protein is required for the initiation of transcription by RNA polymerases I, II and III, from promoters with or without a TATA box. The TBP can initiate transcription from different RNA polymerases. There are several related TBPs (Brindefalk, et al., 2013), including TBP-like (TBPL) proteins. The C-terminal core of TBP is highly conserved and contains two 77-amino acid repeats that produce a saddle-shaped structure that straddles the DNA; the sequence binds to the TATA box and interacts with transcription factors and regulatory proteins. The 3-methyladenine-DNA glycosylase II (AlkA) is a base excision repair glycosylase from *Escherichia coli* and this domain is also found at the N-terminus of

bacterial AlkA (Bruner, Norman and Verdine, 2000). Moreover, N-terminal domain of 8-oxoguanine DNA glycosylase (Fromme, Banerjee, and Verdine, 2004) is found in archaea, bacteria and eukaryotic species, and is specifically responsible for the process which leads to the removal of 8-oxoguanine residues. The region featured in this family is the N-terminal domain, which is organized into a single copy of a TBP-like fold.

In this study, the evolution of TBP-like fold was investigated. Our work focused on the protein domain architecture evolution after the duplication-fusion events, which may help to further understand a gene or protein domain's evolution after gene duplication events and the origin of new gene containing TBP-like fold.

2. Material and Methods

The structural data are from the Protein Data Bank (PDB), and the PDB codes of three-dimensional structures are PDB:1CDW (TATA-box binding protein, C-terminal domain from Homo sapiens), PDB:1MPG (N-terminal domain of DNA repair glycosylase), PDB:1EBM(8-oxoguanine glycosylase from *Homo sapiens*), PDB:1F7W(Cell-division protein ZipA, C-terminal domain from *Escherichia coli*) and PDB:2FPN(a putative exported protein YwmB from *Bacillus subtilis,* an unpublished work from PDB database). All structural classification of structures that obtained from the PDB database was done by SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop/). To obtain an alignment, we linked that some segment to its C-terminal and some large insertion regions in sequence have been deleted. The results of structure superposition and structure-based alignments obtained from different programs are checked by graphics.

The sequences of from PDB are taken as the seeds with which to search the Uniprot database by PSI-Blast. E-values $\leqslant$0.005 are used in the database searches. The sequences hit by seeds are analyzed and aligned by the ClustalW program (http://www.ebi.ac.uk/Tools/msa/clustalw2/). All fragments are aligned by ClustalW, and they are linked in different orders. Circular permutation definition followed SCOP database annotation, and these fragments with the different orders are aligned with the different sequences. Structure based searching was performed by using the FATCAT program (Veeramalai, Ye, and Godzik, 2008). The structures were visualized by Raswin program (http://rasmol.org/) and structure superimpose was performed by VMD software (http:// www. ks.uiuc.edu/ Research/ vmd/). The structural alignment analysis was carried out using the multiple structural alignment program VMD, with a parameters npass=2, scanscore=0 and scanslide=2; and the software Multiseq in VMD based on structural similarity measure QH which takes into account the effects of gap on the aligned protein was employed.

3. Results and Discussions

3.1 Identification of distantly homologue

TATA-binding protein is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II. The transcription factor TFIID repeat signature holds consensus pattern: Y-x-[PK]- x(2)-[IF] -x(2)-[LIVM] (2)-x-[KRH]-x(3)-P-[RKQ]-x(3)-L-[LIVM]-F-x-[STN]-G-[KR]-[LIVMA]-x(3)-G-[TAGL]-[KR]-x(7) -[AGCS]-x(7)-[LIVMF] (http://prosite.expasy.org/ cgi-bin/prosite/ nicedoc.pl?PS00351). At least, three families hold this conserved core of TATA-box binding protein-like fold, including C-terminal domain TATA-box binding protein and DNA repair glycosylase.

As noted above, N-termnal domain of AlkA is structurally homologous to one conserved tandem repeat of the TATA-binding protein. In addition, a protein YwmB from Bacillus subtilis also contains this fold (Figure 1).

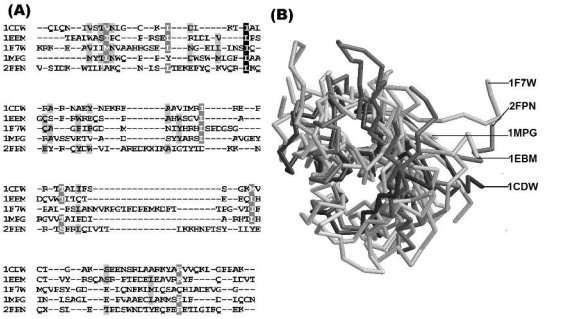

Figure 1. Sequence alignment and structural superimpose of TATA-box binding protein-like fold (A) Sequence alignment of TATA-box binding protein-like fold based on structural information (B) Structural superimpose of TATA-box binding protein-like fold, PDB: 1CDW, TATA-box binding protein, C-terminal domain from Homo sapiens; PDB: 1EBM, 8-oxoguanine glycosylase; PDB: 1F7W, Cell-division protein ZipA, C-terminal domain from Escherichia coli; PDB: 1MPG, 3-Methyladenine DNA glycosylase II (gene *alkA* or *aidA*) from Escherichia coli; PDB: 2FPN, putative exported protein YwmB from *Bacillus subtilis.*

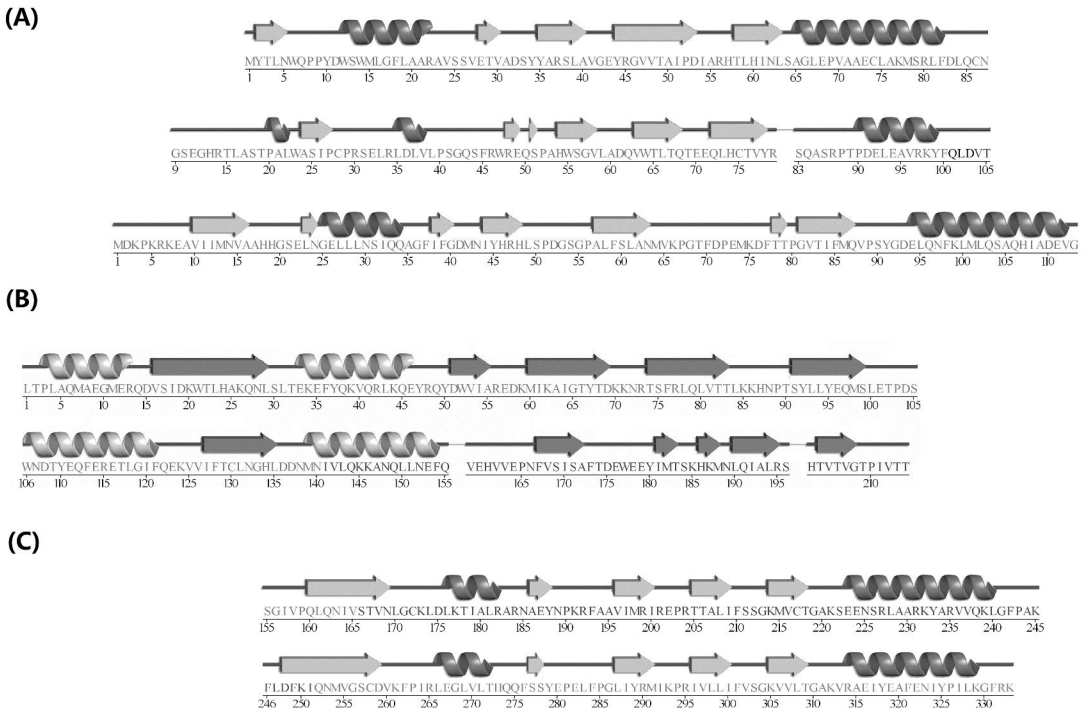

Figure 2 Secondary structures of several TATA-box binding protein-like fold in various proteins
(A) the single TBP-like protein domains from N-terminal domain of 8-oxoguanine glycosylase, N-terminal domain of DNA repair glycosylase and Cell-division protein ZipA respectively (B) the two repeats of putative exported protein YwmB; two obvious circularly permutated structures linked in 2FPN pdb dataset (C) the two repeats of C-terminal domain of TATA-box binding protein

The significant structural and sequence similarities between PDB: 2FPN (a putative exported protein YwmB) and classic TATA-box binding protein have been found using the jFATCAT-rigid algorithm as well as PSI-blast. Moreover, when the two domains' sequences were aligned based on structures, the result showed that 22% positive residues of the region overlapped between two structures, having a RMSD value equal to 5.43 angstroms while P-value 2.95e-03. The 2FPN is a distantly homologues to the classic TATA-box binding protein. All members of this fold are conserved both in sequence and in structure. The structures of TATA-box binding protein-like fold were superimposed by using the VMD program (Figure1B). There is a structural conservation of about 50 residues in TBP-like proteins from various sources.

3.2 Secondary elements analysis

The TATA-box binding protein-like fold contains a secondary structure arrangement of beta-alpha-beta(4)-alpha (alpha: alpha helix; beta: beta sheet). And N-terminal domain of 8-oxoguanine glycosylase, N-terminal domain of DNA repair glycosylase and Cell-division protein ZipA hold this typical secondary structure arrangement respectively (Figure 2A). The first structure of N-terminal domain of 8-oxoguanine glycosylase contains a single copy of this fold with original structures; the second structure of N-terminal domain of DNA repair glycosylase holds a single copy of this fold and an extra alpha helix at the N-terminus; and the third structure of Cell-division protein ZipA has a single copy of this fold with a beta strand insertion before and after the first helix. However, the YwmB-like protein contains a duplicated tandem repeat of two circularly permuted motifs, alpha-beta-alpha-beta(4)-alpha-beta-alpha-beta(4) (Figure 2B), assembled as in the TATA-box binding domain, and forms a swapped dimer by exchanging the equivalent parts of the C-terminal motifs. Compared to the N-terminal domain, the C-terminal fold decorated with additional beta strand structure at the C-terminus. The C-terminal domain of TATA-box binding protein is a duplication region that consists of two clear structural repeats (beta-alpha-beta(4)-alpha-beta-alpha-beta(4)-alpha) (Figure 2C). TATA-binding protein is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II. The C-terminal core region of TBP binds to the

TATA consensus sequence, recognizing minor groove determinants and inducing a dramatic DNA deformation. And the N-terminal domain shows little or no conservation among different organisms, and is largely unnecessary for transcription in certain yeast strains (Nikolov, et al., 1996).

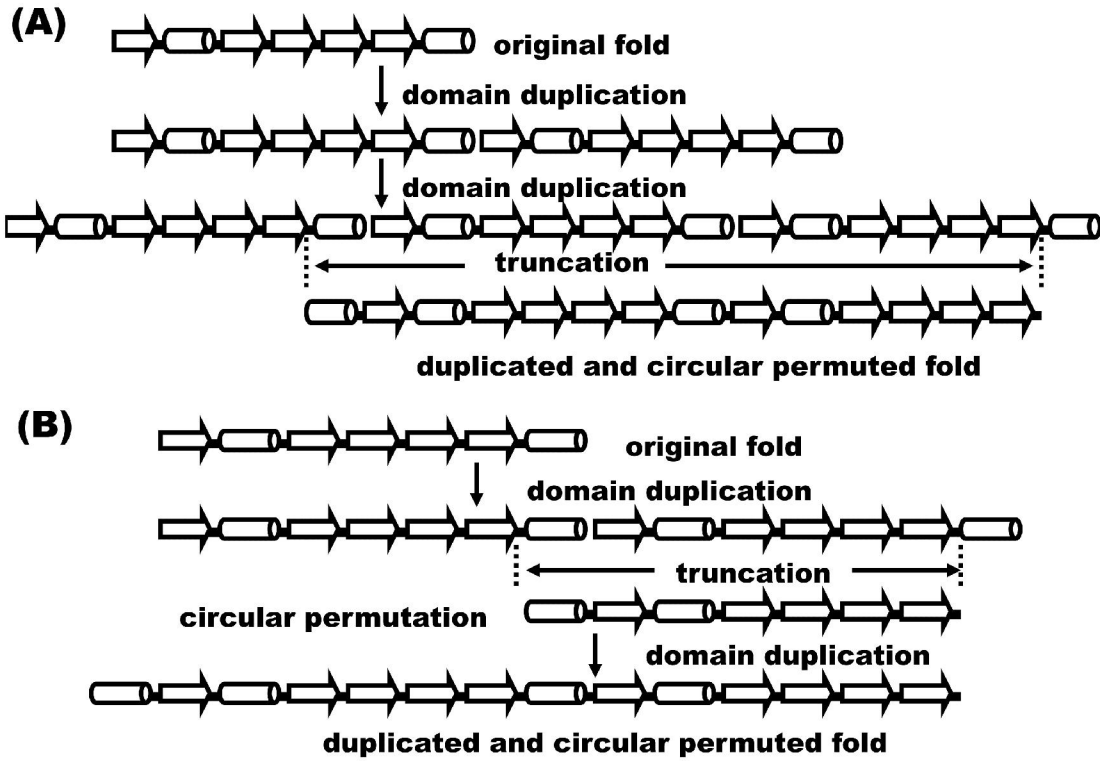

Figure 3 two evolutionary models of the repeats
(A) truncation after two-round domain duplication (duplication-duplication-truncation model)
(B) circular-permuted domain duplication (duplication-truncation-duplication model)

TBP has a conserved C-terminal domain containing two direct repeats (Figure 2C). The secondary structural analysis of crystal structure of original domain of TATA-binding protein domain and circular permuted domain from YwmB protein proved that the two domains have similar structures and possess a core of beta(4) elements. The sequence signature binds specifically to the TATA box promoter element which lies close to the position of transcription initiation. The intra-molecular symmetry generates a saddle-shaped structure that sits astride the DNA (Crowley, et al., 1993). The structure of YwmB protein consists of two structurally similar domains, and the N-terminal domain holds a fold of circular permutation, while the C-terminal domain has a circularly permuted fold with a little different both in secondary and tertiary structure. To our knowledge, all the single domains hold an arrangements of typical TATA-box binding protein-like fold of "beta-alpha-beta(4)-alpha", and single copy of circularly permuted version, alpha-beta-alpha-beta(4), was not found. Therefore, The single TBP-like protein domain with beta-alpha-beta(4)-alpha secondary structure may be the original/ ancestral fold of TBP-like proteins.

3.3 Evolutionary scenarios

Circular permutations are a frequent event in molecular evolution, and they have been observed in many protein families (Lindqvist, and Schneider, 1997). The evolutionary mechanisms of circular permutation in DNA methyltransferase genes (Jeltsch, 1999) and tyrosine phosphatase superfamilies (Huang, 2003) have been illustrated. However, circular permutation can perturb local tertiary structure, resulting in improved protein catalytic activity. On the basis of sequence and structure analyses, we have derived two evolutionary patterns of two circularly permuted fold repeats (Figure 3). The first evolutionary scenario is duplication-duplication-truncation model (Figure 3A), an original protein domain undergoes a two-round duplication-fusion event, and three or four tandem repeats are formed. It's easy to infer that a duplication-fusion event may result in the double

TBP-like protein. All TBP-like protein families may have resulted from a common ancestral gene by a series of duplication, fusion, and circular permutation. And the circular permutations may result from different gene truncated reading frames (Huang, 2003). Processing truncation after two-round domain duplication, the double circularly permuted TBP-like protein comes into being. The second possible evolutionary pathway is duplication of circular-permuted domain (Figure 3B). A single circularly permutated copy of TBP-like protein formed by duplication, fusion and truncation, thus this single copy version doubled by duplication-fusion process. We believe that this complicated evolutionary pathway is less possible, because this evolution process is not maximum parsimony. On the other hand, we found no evidence of single circularly permutated domain version in TBP-like proteins.

4. Conclusions

Our systematic structural analysis of the crystal structures of TBP-like proteins from the RCSB protein databank, examined the structures of single TBP-like protein domain from various protein families, double TBP-like protein domain containing protein, and double TBP-like protein domain with circular permutation. The conclusions reached in this paper are that

(I) The single TBP-like protein domain with beta-alpha-beta(4)-alpha secondary structure may be the original/ancestral fold of TBP-like proteins, because of most of protein families holding this version of fold and no single TBP-like protein domain with circular permutation found.

(II) The TBP-like proteins, including single TBP-like protein domain, double TBP-like protein domain and double circularly permuted TBP-like protein domain, have a common ancestral fold.

(III) The C-terminal domain of TATA-box binding protein consists of two structural repeats (beta-alpha-beta(4)-alpha-beta-alpha-beta (4)-alpha) may result from duplication-fusion process. However, it's more possible that truncation after two-round domain duplication (duplication-duplication-truncation model) resulted in double circularly permuted TBP-like protein domain.

Acknowledgements:

Foundation item: The project was supported by the Fund to Jun Lin from the State Key Laboratory of Genetics Resources and Evolution of China (Grant No.GREKF10-11), National Science Fund of China (Grant No. 31100963) and partly supported by Program for Innovative Research Team of Guilin Medical University (PIRTGMU).

Corresponding Author:
Dr. Jun Lin
Department of Biotechnology
Guilin Medical University
Guilin, Guangxi 541004, China
E-mail: JLinGLMU@gmail.com

References

1. Nikolov DB, Hu SH, Lin J, et al. TATA-box binding protein. Nature 1992; 360: 40-46.
2. Lin J, Hu Y, Tian V, et al. Evolution of double MutT/Nudix domain-containing proteins: similar domain architectures from independent gene duplication-fusion events. J Genet Genomics 2009; 36: 603-610.
3. Brindefalk B, Dessailly BH, Yeats C, et al. Evolutionary history of the TBP-domain superfamily. Nucleic Acids Res 2013; 41: 2832-2845.
4. Bruner SD, Norman DP, Verdine GL. Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. Nature 2000; 403: 859-866.
5. Fromme JC, Banerjee A, Verdine GL. DNA glycosylase recognition and catalysis. Curr Opin Struct Biol 2004; 14: 43-49.
6. Rose PW, Beran B, Bi C, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 2011; 39: 392-401.
7. Veeramalai M, Ye Y, Godzik A. TOPS++ FATCAT: fast flexible structural alignment using constraints derived from TOPS+ Strings Model. BMC Bioinformatics 2008; 9: 358.
8. Nikolov DB, Chen H, Halay ED, et al. Crystal structure of a human TATA box-binding protein/TATA element complex. Proc Natl Acad Sci USA 1996; 93: 4862-4867.
9. Crowley TE, Hoey T, Liu JK, et al. A new factor related to TATA-binding protein has highly restricted expression patterns in Drosophila. Nature 1993; 361: 557-561.
10. Lindqvist Y, Schneider G. Circular permutations of natural protein sequences: structural evidence. Curr Opin Struct Biol 1997; 7: 422-427.
11. Jeltsch A. Circular permutations in the molecular evolution of DNA methyltransferases. J Mol Evol 1999; 49: 161-164.
12. Huang JF. Different protein tyrosine phosphatase superfamilies resulting from different gene reading frames. Mol Biol Evol 2003; 20: 815-820.

1/25/2014