Protein Sequence Motif

BSD: a novel domain in transcription factors and synapse-associated proteins

Tobias Doerks, Saskia Huber, Erich Buchner and Peer Bork

This article describes a novel domain, BSD, that is present in basal transcription factors, synapse-associated proteins and several hypothetical proteins. It occurs in a variety of species ranging from primal protozoan to human. The BSD domain is characterized by three predicted α helices, which probably form a three-helical bundle, as well as by conserved tryptophan and phenylalanine residues, located at the C terminus of the domain.

Initiation of transcription by RNA-polymerase B requires several accessory proteins. Mammalian BTF2 and its yeast homologue TFB1 are essential transcription factors in this initiation event [1,2]. The modular architecture of these proteins is not well understood: no functional details nor domains are characterized. We have analysed different regions of BTF2 to identify possible functional domains. PSI–BLAST [3] searches against the NRDB (non-redundant database) with a central region (conserved residues 180–232, Fig. 1) of BTF2 reveal weak similarity (E value = 0.14) to hypothetical DOS2-like proteins (Box 1) and to a family of proteins with homology (E value = 2.4) to a synapse-associated protein of *Drosophila melanogaster*.

Synapse-associated proteins are expressed specifically in neurons and act as important molecular elements of the nervous system [4]. The similarities of BTF2 to these proteins are just below the default thresholds used by BLAST and are thus indicative of homology, although the significance remains to be proven. Reciprocal PSI–BLAST and PHI–BLAST searches reveal lower E values (e.g. E value = 0.11 using DOS2 as query), although these are still not significant. However, statistically significant similarity is shown by using reciprocal Hidden Markov Model (HMMer) searches [5] as an independent method, starting with the synapse-associated protein family and its related hypothetical plant homologues (including a Ubox-domain-containing protein [6]). The HMMer search reveals significant similarity between members of this family and DOS2-like proteins (E value = 7.2×10^{-7}). In further HMMer iterations (including identified homologues), the suspected homologies of synapse-associated proteins to BTF2-like transcription factors (E value = 2.3×10^{-3}), a BTB/POZ-domain-containing protein [7] and other hypothetical proteins in protozoans (Fig. 2), were confirmed.

We named the newly discovered region the BSD domain (after the BTF2-like transcription factors, synapse-associated and DOS2-like proteins in which it is found). A multiple sequence alignment was

```
TFB1_a       sc  165  LDDSLSKEKLLTNLKLQQ---SLLKGNKVLMKVFQE---TVINAGLPPSEFWSTRIPLLRAFA   P32776
TFB1_b       sc  243  SENKVNVNLSREKIL------NIFENYPIVKKAYTD----NVPKNFKEPEFWARFFSSKLFRK   P32776
BTF2_a       hs   99  LLPKFKRKANKELEEKN----RMLQEDPVLFQLYKD---LVVSQVISAEEFWANRLNVNATDS   P32780
BTF2_b       hs  180  GCNGLRYNLTSDIIE------SIFRTYPAVKMKYAE----NVPHNMTEKEFWTRFFQSHYFHR   P32780
TFB1dm_a     dm  109  LLPNFKRKVDKDLEDKN----RILVENPNLLQLYKD---LVITKVLTSDEFWATHAKDHALKK   Q9V713
TFB1dm_b     dm  182  GCNGLKYNLTSDVIH------CIFKTYPAVKRKHFE----NVPAKMSEAEFWTKFFQSHYFHR   Q9V713
R02D3.3_a    ce  116  NELAKSVESQSKQVELQAKQKILQEDRNLEKLYQNL----VATKLITPDDFWSDYYQKEGVSE   O44499
R02D3.3_b    ce  231  CKEILKFTIQCEYLTR-----KISRSENYIQKKNLE----LVPHEMSEENFWKKFFQSHYFHR   O44499
F2A19.20_a   at   82  LTPAEQLSMAEFELRF-----KLLRENSELQKLHKQ---FVESKVLTEDEFWSTRKKLLGKDS   Q9M322
F2A19.20_b   at  161  RTNRVTFNLTSEIIF------QIFAEKPAVRQAFIN----YVPKKMTEKDFWTKYFRAEYLYS   Q9M322
SPAC16E8_a   sp   60  RVNSTNLEKDIDLQE------SLLTNNPDLLQTFKE---AVMKGHLSNEQFWSTRLHLLRAHA   O13745
SPAC16E8_b   sp  134  VDNQMKVSLTGQQIH------DMFEQHPLLRKVYDK----HVP-PLAEGEFWSRFFLSKLCKK   O13745
B8B20.390_a  nc  147  WFEDDMLKADVELQQ------SLMKKDKALAHIYND[6]DSLSDASFNSQFWATRISLLRAYA   Q9P5N7
B8B20.390_b  nc  227  ENGELKLNINHEQVQ------LIFQQHPLVKRIYNE----NVP-KLTESEFWSRFFLSRLSKK   Q9P5N7
Hypo47.2     hs  146  WLSQFCLEEKKGEIS------ELLVGSPSIRALYTK----MVPAAVSHSEFWHRYFYKVHQLE   Q9NW68
Y97E10AR.6   ce  294  WISRFNLDEYDGEIN------ILLANNPSLRQMFAN----LVPGSVNHETFWKRYFYAIEVAE   CE27417
F25G13.200   at  207  WSLGLKLEEKRNEIV------ELINGNKGVKEIYEE----IVPVEVDAETFWRRYYYKVYKLE   Q9SV58
F15K9.5      at  179  WESAFSLDGKAEEME------KLLEENGDMKGVYKR----VVPSMVDHETFWFRYFYRVNKLK   Q9ZVT6
HypoBAC      os  409  WRDAFRIDERKEEIE------GVLKESPGLESFVER----LVPSVVDYDMFWCRYFFAVDKLR   Q9LIX9
B23L21.150   nc  463  WVNEFDVDKKTEAIA------ADLDKYPELRATMEK----LVPDQVPYADFWKRYYFLRHGIE   Q9P5L4
SPAC22A12    sp  167  WEKEISIDGKTEEIS------LLLEEYPDLRKQMES----LVPSEVSYDDFWKRFFWHKEVVQ   O13905
DOS2         sc  176  QLDPFDVDEKTEEIC------SILQGDKDISKLMND----IVPHKISYKDFWHIYFLQRNKIL   P54858
HypHS        hs  182  VQFNFDFDQMYPVAL------VMLQEDELLSKMRFA----LVPKLVKEEVFWRNYFYRVSLIK   AAH01468
SAP47        dm  272  VDFEFSYDTAYPTAI------AIMAEDKALETMRFE----LVPKIITEENFWRNYFYRVSLII   Q24503
C16C2.4      ce  174  ANSEYTYEQQQAMAT------LLLKHDPNLANVRFQ----LVPKQVKENQFWQNYFYRIGLIR   O17591
K7P8         at   86  NVKKDLSDWQEKHAV------LVLSKSKELSQLRFK----LCPRVLKEHQFWRIYFQLVRKIV   Q9LRX9
T16K5.150    at  195  FDDFEMTDAQYEHAL------AVENLASSLAALRIE----LCPAYMSEYCFWRIYFVLVHPIF   Q9M2X8
F20B24.15    at  227  IKNLEMSDAQRGHAL------AIERLAPRLAALRIE----LCPCHMSVGYFWKVYFVLLLSRL   Q9SGX8
HypOS        os  161  DENSIISDIQRDHME------AIEKLVPDLASLRAR----LCPSYMDIDVFWKIYFTLLESNL   Q9LWJ8
AT2G10950    at  137  DTEFELSEAQRAHAS------AIEDLVPGLVAVKNQ-----VSSYMDDEHFWLIYFILLMPRL   Q9SKH9
T6L1.21      at  178  NVRKDLSEWQERHAT------LVLGSVKQISKLRYE----LCPRVMKERRFWRIYFTLVSTHV   Q9CAA2
F6H11.10     at  769  FSDFELADAQYEHAL------AVERLAPSLASLRIE----LCPEYMTENCFWRIYFVLVHPKL   O49529
F20N2.15     at  424  STSSEQLSIKELELRF-----KLLREN--RYRLHKQ---FVESKVLTEDEFWATRKKLLGKDS   Q9LFZ6
LMAJFV1      lm  340  WALHSLFDFDRDVQE------GLLASA-EVRAHRYR----LVPARLKEVTFWANYFWKVHCVG   O60968
PFC1055W     pf  302  QKLSKSVEINNELRK------LILCENKELKKLYDY---YIENNILSDSKFWFFLFNNKYSHL   O97305
Consensus   (80%)     ......hp.p...h........lhp....l..hh.p....hss..hp.ppFW.haa..h..h.
Sec.str.pred.         ........hhHHHHHHHHHHHHHHHh.hHHHHHHHH..........hHHHHHHHHHHh...
```

Fig. 1. Multiple sequence alignment of BSD domains of BTF2-like transcription factors (TFB1, BTF2, TFB1dm, R02D3.3, F2A19.20, SPAC16E8, B8B20.390), DOS2-like proteins (Hypo47.2, Y97E10AR.6, F25G13.200, F15K9.5, HypoBAC, B23L21.150, SPAC22A12, DOS2), proteins related to a synapse-associated protein (HypHS, SAP47, C16C2.4, K7P8, T16K5.150, F20B24.15, HypOS, AT2G10950, T6L1.21 and, with an N-terminal Ubox, F6H11.10), BTB domain-containing protein (F20N2.15), and other hypothetical proteins (LMAJFV1, PFC1055W). First column: protein names (multiple domains in the same protein are labelled a and b); second column: species names; third column: start of the domain in the respective sequences; rightmost column: database accession numbers. Partially conserved (>50%) negatively charged residues are shown in red; conserved hydrophobic residues in blue; conserved aromatic residues in bold blue, and other conserved residues in bold black. The consensus sequence (conserved in 80% of the sequences) is shown below: h, p, s, l and a indicate hydrophobic, polar, small, aliphatic and aromatic residues, respectively. The predicted secondary structure taken from the consensus of the alignment (H, helix predicted with expected average accuracy >82%; h, helix predicted with expected average accuracy <82%) [9]. Several expressed sequence tags exist in various eukaryotes, supporting the idea that the BSD domain is found in a wide species range (data not shown). The domain boundaries shown have been predicted after analysis of all family members. For all subfamilies, PHD only predicts a coherent secondary structure in the region displayed here. Similarities beyond the region (if existing at all) are confined to individual subfamilies; other subfamilies contain deletions or segments of low complexity in regions preceding or succeeding the predicted boundaries [e.g. N terminus of T16K5.150 (185–199) or C terminus of DOS2 (240–265)]. Abbreviations: at, *Arabidopsis thaliana*; BSD, found in BTF2-like transcription factors, synapse-associated and DOS2-like proteins; BTB, Broad-complex, Tamtrack and Bric a Brac; ce, *Caenorhabditis elegans*; dm, *Drosophila melanogaster*; hs, *Homo sapiens*; lm, *Leishmania major*; nc, *Neurospora crassa*; os, *Oryza sativa*; pf, *Paramecium falciparum*; sc, *Saccharomyces cerevisiae;* sp, *Saccharomyces pombe*; Ubox, a modified Ring finger domain associated with ubiquitination.

0968-0004/02/$ – see front matter PII: S0968-0004(01)02042-4

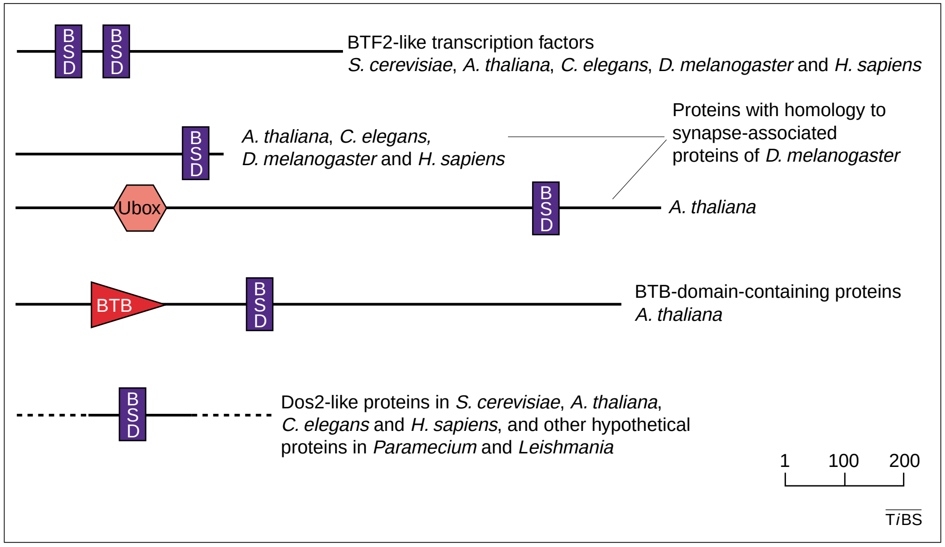
BSD
BSD
BTF2-like transcription factors
S. cerevisiae, *A. thaliana*, *C. elegans*, *D. melanogaster* and *H. sapiens*
BSD
A. thaliana, *C. elegans*, *D. melanogaster* and *H. sapiens*
Proteins with homology to synapse-associated proteins of *D. melanogaster*
Ubox
BSD
A. thaliana
BTB
BSD
BTB-domain-containing proteins
A. thaliana
BSD
Dos2-like proteins in *S. cerevisiae*, *A. thaliana*, *C. elegans* and *H. sapiens*, and other hypothetical proteins in *Paramecium* and *Leishmania*
1
100
200
TiBS

Fig. 2. Domain architecture of proteins containing the BSD domain. Only proteins with distinct modular organizations are shown. The domain names are according to those in the Simple Modular Architecture Research Tool [13] (http://smart.embl-heidelberg.de). Note that the name synapse-associated protein might be misleading as synaptic localization has so far only been demonstrated for *Drosophila melanogaster* [4], and close homologues are found in species without synapses. Abbreviations: BSD, found in BTF2-like transcription factors, synapse-associated and DOS2-like proteins; BTB, Broad-complex, Tamtrack and Bric a Brac; Ubox, a modified Ring finger domain associated with ubiquitination.

generated [8] for all candidates to identify the potential domain boundaries (Fig. 1). From this, the BSD domain appears to be ~60 amino acids in length. Secondary structure prediction with PHD [9] indicates the presence of three α helices, which probably form a three-helical bundle in small domains. The third predicted helix contains neighbouring phenylalanine and tryptophan residues – less common amino acids that are invariant in all the BSD domains identified and that are the most striking sequence features of the domain (Fig. 1). The BSD domain is found in a variety of species from primal protozoan to human, indicating a conserved, probably important, function.

Although the BSD domain occurs in very different protein families (e.g. synapse-associated proteins, hypothetical proteins and transcription factors), the presence of the novel domain in transcription factors suggests a role in chromatin-associated processes. The domain architectures of additional BSD domain-containing proteins are consistent with this assumption, but also make other functions feasible. For example, the BSD domain can co-occur with Ubox domains (Fig. 2), which are known to be involved in ubiquitination [6]. Although several proteins involved in the ubiquitination process are known to be associated with chromatin [10,11], this is not a prerequisite. BSD domains can also precede a BTB domain (Fig. 2), a protein–protein interaction domain that frequently occurs in transcription factors, in which it is succeeded by ZnF_C2H2 DNA-binding domains [7,12]. These findings suggest that the BSD domain could have a role in DNA binding, although it should be noted that neither synapse-associated proteins nor DOS2-like proteins are known to be associated with chromatin.

In summary, the delineation of the BSD domain and its boundaries (Fig. 1) should allow directed structural studies to test the involvement of this domain in chromatin-associated or more general processes.

Box 1. Artificial *in silico* support of function prediction*

Currently, one of the proteins in the alignment, DOS1, is annotated in most databases to be involved in single-copy DNA replication and ubiquitination. This assumed function would match our findings that the BSD domain is present in basal transcription factors and could have a role in DNA-binding.

The functional description is based on a mutation in a region around the yeast open reading frame YDR068W (dating back to 1995), and the gene was originally named DOS1 in *Saccharomyces cerevisiae*. Further studies revealed that the mutation is localized in an adjacent gene; therefore, YDR068w was renamed DOS2 in GenBank at a later stage, without functional description. However, in the meantime, the old name and functional implications of the gene were imported into different public databases, where the erroneous entry can often still be found (Table I).

Table I. Variable nomenclature found in databases

Date of information retrieval	Correct database entries		Erroneous database entries		Erroneous database entry
	Gene names: DOS2, YDR068W,YD9609.22, YD8554.01, D4267 Function: hypothetical		Wrong gene names: DOA4,DOS1, UBP4, SSV7, NPI2, YDR069C		Correct gene name, but functional prediction of DOS1
September –October 2001	GenBank:	AAA66522 NP_010353	Sptrembl:	O13905 Q9P5L4	
	EMBL:	CAB16584 (one representative)	SWISSPROT:	P54858	
			EMBL:	CAB91683	
			PIR:	S54052 T49702	
28 November 2001	GenBank:	AAA66522 NP_010353	Sptrembl:	O13905, Q9P5L4	SWISSPROT: P54858
	EMBL:	CAB16584 (one representative)	EMBL:	CAB91683	
			PIR:	S54052 T49702	

*Tim Formosa (University of Utah, Salt Lake City, UT 84112-5330, USA) was a coauthor of Box 1.

References

1 Fischer, L. *et al.* (1992) Cloning of the 62-kilodalton component of basic transcription factor BTF2. *Science* 257, 1392–1395

2 Gileadi, O. *et al.* (1992) Cloning of a subunit of yeast RNA polymerase II transcription factor b and CTD kinase. *Science* 257, 1389–1392

3 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs *Nucleic Acids Res.* 25, 3389–3402

4 Reisch, D. *et al.* (1995) The sap47 gene of *Drosophila melanogaster* codes for a novel conserved neuronal protein associated with synaptic terminals. *Brain Res. Mol. Brain Res.* 32, 45–54

5 Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* 14, 755–763

6 Aravind, L. and Koonin, E.V. (2000) The U box is a modified RING finger – a common domain in ubiquitination. *Curr. Biol.* 10, R132–R134

7 Zollman, S. *et al.* (1994) The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 91, 10717–10721

8 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680

9 Rost, B. *et al.* (1994) PHD an automatic mail server for protein secondary structure prediction *CABIOS* 10, 53–60

10 Hershko, A. and Ciechanover, A. (1998) The ubiquitin system. *Annu. Rev. Biochem.* 67, 425–479

11 Ciechanover, A. (1998) The ubiquitin–proteasome pathway: on protein death and cell life. *EMBO J.* 17, 7151–7160

12 Bardwell, V.J. and Treisman, R. (1994) The POZ domain: a conserved protein–protein interaction motif. *Genes Dev.* 15, 1664–1677

13 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234

Tobias Doerks*
Peer Bork
EMBL, 69012 Heidelberg, Meyerhofstr. 1, and
Max-Delbrueck-Centrum, Berlin, Germany.
*e-mail: doerks@embl.heidelberg.de

Saskia Huber
Erich Buchner
Lehrstuhl für Genetik und Neurobiologie,
Biozentrum der Universität, Am Hubland,
97074 Würzburg, Germany.

Sec61β – a component of the archaeal protein secretory system

Lisa N. Kinch, Milton H. Saier, Jr and Nick V. Grishin

Sec61p/SecYEG complexes mediate protein translocation across membranes and are present in both eukaryotes and bacteria. Whereas homologues of Sec61α/SecY and Sec61γ/SecE exist in archaea, identification of the third component (Sec61β or SecG) has remained elusive. Using PSI–BLAST, the archaeal counterpart of Sec61β has been detected. With the identification of the Sec61β motif, functions for a universal family of archaeal proteins can be predicted and the archaeal translocon system can be definitively detected.

Detection and rationalization of motifs in membrane proteins are more difficult than in soluble proteins because of their biased amino acid composition that is restricted to mostly hydrophobic residues and to a limited number of available spatial structures. The single transmembrane-spanning protein Sec61β (Sbh1p in yeast) interacts with two other integral membrane proteins (Sec61α and Sec61γ) to form the core of the eukaryotic protein translocation machinery (reviewed in Refs [1,2]). The bacterial counterpart of this machine consists of a similar complex (SecYEG), with SecY and SecE representing homologues of Sec61α and Sec61γ, respectively [3]. The third bacterial membrane protein, SecG, differs somewhat from Sec61β in both the number of membrane-spanning regions and residue conservation. This divergence brings into question the evolutionary origins of this third subunit, although SecG and Sec61β both function to stimulate protein translocation activities and are thought to be homologous [4,5]. Although archaeal homologues of SecY/Sec61α and SecE/Sec61γ exist, the identification of an archaeal homologue to either Sec61β or SecG has remained elusive. The Sec61p/SecYEG system is universally present in all eukaryotes and bacteria for which completely sequenced genomes are available (T. Cao and M.H. Saier, Jr, unpublished). Thus, the absence of a Sec61β/SecG homologue in archaea is puzzling. Based on sequence analyses, we have identified the third component of the archaeal translocation machinery. The archaeal counterpart resembles eukaryotic Sec61β, suggesting an overall functional similarity between the translocation apparatus of archaea and the eukaryotes. Although this functional similarity awaits experimental conformation, it mimics similarities displayed in other universal processes such as DNA replication, transcription and translation [6], and provides additional data for the studies of archaeal evolutionary origin [7].

We first detected a possible archaeal counterpart (gi | 15920503) to the human Sec61β sequence (gi | 5803165) using PSI–BLAST [8] (parameters described in Fig. 1). Upon searching protein databases for related archaeal sequences, we found hits in all but two of the completely sequenced archaeal genomes. Searches against the nucleotide databases of these genomes suggested that these sequences (AE000914 and AE006662) were missed in gene prediction efforts. To substantiate the link to eukaryotic Sec61β sequences, we generated a position-specific scoring matrix with a multiple sequence alignment of the archaeal Sec61β. Using this matrix, we initiated PSI–BLAST searches with each sequence from the alignment as a query. Two archaeal sequences used as queries (gi | 15920503 or RAP00437) identified the eukaryotic Sec61β sequence (gi | 15239337) with significant statistics (E-value 0.002). This E-value, representing the estimated number of alignments with scores no less than that of a given alignment that is expected to occur in a database search by chance [8], falls below the threshold observed for distant homologues (E = 0.01) [9].

The short 45-residue motif identifying Sec61β consists of a single, mostly hydrophobic stretch of ~20 amino acids preceded by a region of similar size that starts with several small amino acids and displays a particular residue conservation pattern (Fig. 1). The hydrophobic segment is predicted to form a transmembrane helix, with the C terminus of the helix defined by small and positively charged residues. The sequence of the helix incorporates a small residue at the beginning of the third turn and a relatively conserved histidine in the last turn. We suggest that the most conserved residue in the motif (proline) forms part of the N-terminal cap structure of this helix. The sequence between this helix and the stretch of small residues at the N terminus of the motif is characterized by four predominantly charged positions, having two negative charges surrounded by positive charges on either side. The archaeal sequences additionally contain conserved positively charged residues N-terminal to the transmembrane helix,

0968-0004/02/$ – see front matter PII: S0968-0004(01)02055-2