FOR THE RECORD

The discoidin domain family revisited: New members from prokaryotes and a homology-based fold prediction

STEFAN BAUMGARTNER,¹ KAY HOFMANN,² RUTH CHIQUET-EHRISMANN,³ AND PHILIPP BUCHER²

¹Lund University, Department of Cell & Molecular Biology, Box 94, S-22100 Lund, Sweden

²Swiss Institute for Experimental Cancer Research, Ch. des Boveresses 155, CH-1066 Epalinges, Switzerland ³Friedrich Miescher-Institut, Postfach 2543, CH-4002 Basel, Switzerland

(RECEIVED December 23, 1997; ACCEPTED April 15, 1998)

Abstract: Members of the discoidin (DS) domain family, which includes the C1 and C2 repeats of blood coagulation factors V and VIII, occur in a great variety of eukaryotic proteins, most of which have been implicated in cell-adhesion or developmental processes. So far, no three-dimensional structure of a known example of this extracellular module has been determined, limiting the usefulness of identifying a new sequence as member of this family. Here, we present results of a recent search of the protein sequence database for new DS domains using generalized profiles, a sensitive multiple alignment-based search technique. Several previously unrecognized DS domains could be identified by this method, including the first examples from prokaryotic species. More importantly, we present statistical, structural, and functional evidence that the D1 domain of galactose oxidase whose three-dimensional structure has been determined at 1.7 Å resolution, is a distant member of this family. Taken together, these findings significantly expand the concept of the DS domain, by extending its taxonomic range and by implying a fold prediction for all its members. The proposed alignment with the galactose oxidase sequence makes it possible to construct homology-based three-dimensional models for the most interesting examples, as illustrated by an accompanying paper on the C1 and C2 domains of factor V.

Keywords: DS domain; fold prediction; galactose oxidase; generalized profiles; homology search

The discoidins from the slime mould *Dictyostelium discoideum* were first described as lectins with high affinity for galactose (Poole et al., 1981). When the sequences of the blood coagulation factors V (Jenny et al., 1987) and VIII (Wood et al., 1984) were determined, two C-terminal repeats in these proteins were found to be similar to the N-terminal region of discoidin. This surprising

finding defined a new extracellular module known as DS or F5/8 type C domain. Additional members of this family were later found in milk fat globule (Stubbs et al., 1990), in Xenopus laevis neuronal cell surface antigen A5, recently renamed neuropilin (Takagi et al., 1991; Kawakami et al., 1995), in two subfamilies of mammalian receptor tyrosine kinases (Johnson et al., 1993; Karn et al., 1993), in a pathogen defense protein named hemocytin from Bombyx mori (Kotani et al., 1995), in a mammalian carboxypeptidase termed AEBP (Ohno et al., 1996), in human and Drosophila neurexin IV (Baumgartner et al., 1996), and most recently in XLRS1, a candidate gene for X-linked juvenile retinoschisis (Sauer et al., 1997). Several of these proteins contain tandemly repeated pairs of DS domains (see Fig. 1). One of them, milk fat globule, has subsequently been isolated in several other research contexts, for instance as a zona pellucida-binding protein (Ensslin et al., 1998), or as a ganglioside O-acetyltransferase (Ogura et al., 1996).

Searching the current protein sequence database, we readily identified single DS domains in six additional proteins: SCO-spondin (Gobron et al., 1996), a newly characterized member of the thrombospondin family, CUB1 (Shibata et al., unpubl.), an anonymous human protein, three hypothetical proteins from *Caenorhabditis elegans* and *Caenorhabditis briggsae* encoding receptor protein tyrosine kinases and F47C21.1, a large modular protein also from *C. elegans*. Moreover, a tandem pair of DS domains was found in the Del-1 protein (developmental endothelial locus-1; Hidai et al., 1998), an embryonic endothelial cell protein that binds to $\alpha\nu\beta3$ integrins.

Using the more sensitive generalized profile-based search method, we found additional members in microbial species, most notably the crystallized D1 domain of galactose oxidase (Ito et al., 1994) from the fungus *Dactylium dendroides*. This domain was previously found to be similar to noncatalytic extensions of two bacterial sialidases (Bork & Doolittle, 1994). It is also relatively closely related to three internal repeats in ORF 4.7 of AUD1, an amplifiable DNA element from *Streptomyces lividans* (Piendl et al., 1994). In addition, we found homologous sequences in Mu toxin of *Clostridium perfringens* (Canard et al., 1994) and migA of *D. discoideum* (Escalante et al., 1997), a protein involved in chemotaxis to

Reprint requests to: Stefan Baumgartner, Lund University, Department of Cell & Molecular Biology, Box 94, S-22100 Lund, Sweden; e-mail: Stefan.Baumgartner@medkem.lu.se.



Fig. 1. Schematic diagram of the occurrence of DS domains in proteins drawn in scale. Listed are proteins from top to the bottom as they appear in the text. Dark green boxes denote signal peptides, different colors in the boxes denote different modules as they have been discovered or and annotated in the corresponding publication. White areas are portions with no obvious homology. The brown vertical bar represents a cell membrane. TK indicates tyrosine kinase.

cAMP and slug migration. There appears to be a close homologue of migA in *Arabidopsis thaliana*. A complete list of currently known DS domains is given in Table 1, and an alignment of representative members is shown in Figure 2.

Evidence supporting the homology between the galactose oxidase D1 domain and the DS domain: Sequence similarities between previously recognized members of the DS domain family, or between members of the GOase D1 domain subfamily, are readily detected by standard sequence comparison techniques and, thus, need not be further justified. Significant cross-matches between the two groups were detected with the more sensitive profile-based technique (Bucher et al., 1996) and corroborated with a recently introduced robust significance test (Hofmann & Bucher, 1995). With a profile made from the larger eukaryotic subfamily, we obtained a significant match ($P < 10^{-2}$) to one of the GOaserelated repeats in the AUD1 protein (Piendl et al., 1994). With a profile made from all members of the second subfamily, we obtained significant matches to the human milk fat globule (P < 10^{-5}) and to the C2 repeat of bovine factor V ($P < 10^{-2}$). Both profiles also identified a significant match ($P < 10^{-4}$) in Clostridium perfringens Mu toxin, shown to possess hyaluronidase activity (Canard et al., 1994). This match corresponds to the second of three previously reported internal repeats located in the noncatalytic C-terminal region of this protein. Finally, a profile made from the two main subfamilies and the central DS-like repeat of Mu toxin produced a highly significant match ($P < 10^{-5}$) to migA from *D. discoideum*.

The proposed expansion of the DS domain family is further supported by additional structural and functional arguments. For instance, each subtype occurs in at least one protein as tandem repeats of almost identical length of about 150 amino acids. More importantly, the residue conservation pattern observed within the major eukaryotic subfamily is readily explained by structural constraints expected for protein sequences folding into a GOase D1 domainlike structure. This fold has been described as a beta-sandwich where a five-stranded antiparallel beta-sheet (b1-b2-b7-b4-b5) faces another three-stranded antiparallel beta-sheet (b8-b3-b6). A secondary structure prediction (Rost, 1996) made from a multiple alignment of the eukaryotic subfamily only (excluding GOase and migA) is in good agreement with the beta-strand assignments in the GOase D1 structure (Fig. 1). Moreover, the most conserved parts of this multiple sequence alignment correspond to the four strands b2, b3, b4, and b7, located in the center of the two sheets, a conservation pattern reminiscent of other beta-sandwich domains, e.g., fibronectin type III. Finally, virtually all hydrophobic core residue positions in GOase D1 are clearly maintained in the other subgroups. Taken to-

Table 1. DS-domain con	ttaining proteins					
Name	# of domains	Accession	Species (gene/variant)	Synonyms	Function	Comments
			Major eu	caryotic subfamily		
Discoidin I	-	P02886 J01283 P02887 P02887	D. discoideum (Dsca) D. discoideum (Dscb) D. discoideum (Dscc) D. discoideum (Dscc)		Lectin; high affinity to galactose; promotes cell aggregation	Lacks signal peptide
Discoidin II	1	P42530	D. discoideum (Dsce)			
Factor V	2	P12259 P28107	H. sapiens B. taurus		Blood coagulation; phospholipid binding	
Factor VIII	7	P00451 AF016234 P12263 Q06194	H. sapiens C. familiaris S. scropha M. musculus		Blood coagulation; phospholipid binding	
Milk fat globule	0	Q08431 Q95114 P79385 P70490 P21956	H. sapiena B. taurus S. scrophu R. norvegicus M. musculs	BA46; PAS-6/7; P47; MFG-E8; AGS	Phospholipid-binding; zona pellu- cida binding; O-acetyl-GD3 synthase	Overexpressed in breast carcinomas
Del-1	2	AF031524	M. musculus	Developmental endo- thelial locus-1	Binds alpha-v beta-3 integrin receptor	
Neuropilin Neuropilin-2	0	AF018956 AF016296 P97333 P79795 P28824 2 AF016297 AF022854	H. sapiens R. norvegicus M. musculus G. gallus X. laevis R. norvegicus M. musculus	A5 antigen	Calcium-dependent cell adhesions; cell-recognition in the nervous system	Neuropilin-2 has at least seven splicing variants
Hemocytin	2	P98092	B. mori	Humoral lectin self-defense		
Discoidin receptor protein tyrosine kinase	_	Q08345 Q63474 Q03146	H. sapiens R. norvegicus M. musculus	DDR; CAK; TRK E; RTK 6	Cell-cell interaction and recognition	Overexpressed in breast carcinomas

Tvro10 recentor	_	016837	H consine	TKT	Marrotenship trenden binana	
tyrosine kinase		Q62371	M. musculus			
Other receptor protein	-	U56248 U39742	C. briggsae (G01D9.2) C. elegans (C25F6.4)			
Tyrosine kinases		U41532	C. elegans (F11D5.3)			
Neurexin IV	-	U87223 P97846 X86685	H. sapiens R. norvegicus D. melanogaster	CASPR (contactin- ass. protein)	Cell adhesion/cell junction	
XLRS1	1	AF014459	H. supiens		Candidate disease gene for X-linked juvenile retinoschisis	
SCO-spondin	I	P98167	B. taurus		Modulation of neural aggregation	
AEBPI	1	D86479 Q61281	H. sapiens M. musculus		Extracellular carboxypeptidase	Murine AEBP1 reported to be transcriptional repressor (?)
CUBI	l	D29810	H. sapiens			
Hypothetical ORF	1	U61946	C. elegans (F47C12.1)			
			Microbi	al subfamily		
Galact. oxidase	1	Q01745	D. dendroides		Galactose oxidase	
Sialidase	1 -	P29767 Q02834	C. septicum M. viridifaciens		Sialidase; neuraminidase	May be pathogenic factor
AUDI ORF 4.7	с,	U22894	S. lividans			Reported similarity to chitinase confined to FN3 domains
			0	utliers		
Mu toxin	ŝ	P26831	C. perfringens	nagH	Hualurono-glucosaminidase	Virulence factor for gas gangrene
MigA	-	U86962 U93215	D. discoideum A. thaliana		Chemotaxis to cAMP; slug migration	U93215 appears to be a migA ortholog

Discoidin domain: New members and fold prediction

PH	D predicti	on		EE	EEE		EEEEE	E EEEEEEE
Dd	DisIa	2	.STQGLVQLLANAQCHLR	TSTNYNGV	HTQFNSALNYKNNG. TNT	TIDGSEAWCSSIVD	TNQYIVA	GCEVPRTFMCVALQGRGDADQ.
Dd	DisII	2	.SVPAGSVSCLANALLNLR	SSTDYNAD	HGVKNSILNFSNSKDASE	RFDGSE <mark>SW</mark> SSSVLD	KNQFIVA	GSDSVKHFVAISTQGRGDHDQ.
Hs	fVIII 1	2040	CQTPLGMASGHI.RDFQIT	ASGQYGQ	WAPKLARLHYS	GSINAWSTKE	PFSWIKV	DLLAPMIIHGIKTQGARQKFS.
Hs	fVIII 2	2193	CSMPLGMESKAI.SDAQIT	ASSYFTNMFAT	WSPSKARLHLQ	GRSNAWRPQVNN	PKEWLQV	DFQKTMKVTGVTTQGVKSLLT.
Hs	fV 1	1907	CRMPMGLSTGII.SDSQIK	ASEFLGY	WEPRLARLNNG	GSYNAWSVEKLAAEFA	SKPWIQV	DMQKEVIITGIQTQGAKHYLK.
Hs	fV 2	2066	CSTPLGMENGKI.ENKQII	ASSFKKSW	WGDYWEPFRARLNAQ	GRVNAWQAKANN	NKQWLEI	DLLKIKKITAIITQGCKSLSS.
Mm	MFG 1	148	CSTQLGMEGGAI.ADSQIS	ASYVYMGFMGLQR	WGPELARLYRT	GIVNAWHASNYD	SKPWIQV	NLLRKMRVSGVMTQGASRAGR.
Mm	MFG 2	308	CLEPLGLKNNTI.PDSQMS	ASSSYKTWNLR.AFG	WYPHLGRLDNQ	GKINAWTAQSNS	AKEWLQV	DLGTQRQVTGIITQGARDFGH.
Mm	Dell 1	158	CSGPLGIEGGII.SNQQII	ASSTHRALFGLRK	WYPYYARLNKK	GLINAWTAAEND	RWPWIQI	NLORKMRVTGVITOGAKRIGS.
HS	Dell Z	319	COVERD ICERCEV MODOL	ASSVERTLINMD.MET	WERKARLDRQ	CECCAWLSK FOD	QSQWLQV	DILVPTRVIGILITQGARDIGH.
ns Dm	haut 1	940	CODNYTDIAMCDEDI DUTAFO	LONPEQIVGWI	FARMARINSQ	NSCACSMNDK VNN	DKOVIOV	FLOODEDTYCVILOGSDIFD.
Bm	heyt 2	1116	CTEP LGLIGEL PLENTC	VSSNSEE	KDYLSTN	GNRGWKP LYN	TPGWVMF	DETGERNITGILTKGGNDG
Mm	noje 2	275	CMEALGMESGET.HSDOIT	ASSOYGT	WSVERSRLNY	PENGWTPG	YKEWIOV	DLGILREVTAVGTOGAISKETK
Mm	npl 2	431	CSGMLGMVSGLI.SDSOIT	ASNOADR N	WMPENIRLVT	SRTGWALPPSP HPY	TNEWLOV	DLGDEKIVRGVIIOGGKHREN.
Hs	DDR	31	CRYALGMQDRTI.PDSDIS	ASSSWSDS	TAARHSRLESS	DGDGAWCPAGSV FPK	EEEYLQV	DLQRLHLVALVGTQGRHAGGLG
Hs	TKT	30	CRYPLGMSGGQI.PDEDIT	ASSQWSE	TAAKYGRLDSE	EGDGAWCPEIPV EPDD	LKEFLQI	DLHTLHFITLVGTQGRHAGGHG
Dm	nrx IV	46	CNQPL.MERAVLI	ATSSLTE	RGPDKARLN	GNAAWTPVENT	YNHFLTL	DLGDPRMVRKIATMGRMHTD
Hs	AEBP1	72	CPPIGMESHRI.EDNQIF	ASSMLRHG	LGAQRGRLNMQTGA. TEL	DDYYDG <mark>AW</mark> CAEDDA	RTQWIEV	DTRRTTRFTGVITQGRDSSIH.
Hs	CUB1	224	CYGTLGMESGVI.ADPQIT	ASSVLEWTDHRARDS	WKPKKARLKKTW	TALGAFATD	EYQWLQI	DLNKEKKITGIITTGHHGEHN.
Bt	SCO SP	334	CYSP LGLARLP EGSLE	ASSQQLEHP	AWAAILRPAP	GAPGWSPVEHADTQGHT	PPPYLQL	DLLQPRNLTGIIVQGAGSSDWI
Ce	F47C12 1	940	CVLSEDVGIVSGFV.PDGAFA	DNSDSTNLG	YEPHHARMG	SSGWCGAKE	DFIFLSV	DLQRIYTLTTLRIAGVAGSGH.
Ср	mu	799	.DSNSGVEFNPSLIRSE	SWQVYEG	NEANLLDGDD	NTGVWYKTLNGDTSI	AGEFIGL	DLGKEIKLDGIRFVIGKNGGGS
Ds	migA	659	.NEKWSNPHSTSKIKI	SSSIDKG	NLYDIVELT	PNAFWTKD	PASWVMI	DLGPNRTVVPMYYTIRHGLSY
S 1	AUD1 1	33	.ATAAGGPNIAVGDATA	ASSHGEY	GAANITDGN	QGTYWQSGGGS	LPQWVQT	DLGATERIDEVVLRLPAGWE.
S 1	AUD1 2	174	.AGESSVDLAAGRRL	ASSTGAY	TPGNGNDGN	RATYWESANN	LPQWLQA	DLGSARRVNRVVLRLPDGWP.
Μv	Neu	501	.VGLLDQARMSI.ADVDSE	ETAREDG	RASNVIDGN	PSTFWHTEWSRA.DAPG	YPHRISI	DLGGTHTISGLQYTRRQNSAN .
Cs	Neu	43	.SQPIAEKLVPR.SEIQAS	ATSALTGE	GPEKAIDGN	TSTLWHTPWAGV.DIQI	NPQSLTI	KLGKTRNISSICVTPRQEGTN
Dd	e GOase	43	.SAPIGSAISRNNWAV	CDSAQSGN	ECNKAIDGN	KDTFWHTFYGAN.GDPF	(PPHTYTI	DMKTTQNVNGLSMLPRQDGNQ
GO	ase 3D st	ruct.	bbł	dd			bbbbb	dddddddd dd
			EEEEE EEEE	E	EEEEEE	EEEEEEEE	E	PHD prediction
	WVTSYKI	RYSLD	VSWFEYRNGAAVTGVTD	R.NTVVNHFFDTPIR.	ARSIAIHP	LTWNGHISLRCEFYT.	151 E	d DisIa
	WVTSYKL	RYTLD	VNWVEYNNGEIINANKDI	R.NSIVTINFNPPIK.	ARSIAIHP	QTYNNHISLRWELYA.	153 D	Dd DisII
	SLYISQFII	MYSLD	KKWQTYRGNSTGTLMVFFGNVDS	S.SGIKHNIFNPPII.	ARYIRLHP	THYSIRSTLRMELMGC	2188 H	Is fVIII 1
	SMYVKEFLI	SSSQD	HQWTLFFQNGKVKVFQGNQDS	S.FTPVVNSLDPPLL.	TRYLRIHP	QSWVHQIALRMEVLGC	2345 H	Is fVIII 2
	SCYTTEFYV	AYSSN	QINWQIFKGNSTRNVMYFNGNSDA	A.STIKENQFDPPIV.	ARYIRISP	TRAYNRPTLRLELQGC	2061 H	Is fV 1
	EMYVKSYTI	HYSEQ	VEWKPYRLKSSMVDKIFEGNTN	C.KGHVKNFFNPPII.	SRFIRVIP	KTWNQSITLRLELFGC	2221 H	Is fV 2
	AEYLKTFKV	AYSLD	RKFEFIQD.ESGGDKEFLGNLD	NSLKVNMFNPTLE.	AQYIRLYP	VSCHRGCTLRFELLGC	303 N	Am MFG 1
•	IQYVESYKV	AHSDD	VQWTVYEEQGSSKVFQGNLD	N.NSHKKNIFEKPFM.	ARYVRVLP	VSWHNRITLRLELLGC	463 N	Im MFG 2
	PEYIKSYKI.	AYSND	KTWAMYKVKGTNEEMV F RGNVDI	N.NTPYANSFTPPIK.	AQYVRLYP	QICRRHCTLRMELLGC	314 N	Am Dell 1
	VQEVGSYKL	AYSND	SEHWMVHQDEKQRKDKVFQGNFDI	.DTHRKNVID PPI Y.	ARFIRILP	WSWYGRITLRSELLGC	476 F	is Dell 2
	.EWMTKYSV	QYRTD	SRLNWIYYKDQTGNNRVFYGNSD	K. TSTVQNLLRPPII.	SKEIRLIP	LGWHVRIAIRMELLEC	219 1	15 XLRSI
	.QIVISIEL	TYTCD	ENTESTVDG. PDGRPKIERGPID	TUSUTNETUDDID	ARVVKIRP	OWWWITELBIEDIGC	1254 5	am heyt 1
		DICON	POWISIK FONKATIFOGNEN	D TOWLCVPSKPLT	TO FUE TKD	VSWETCISMPFEUVCC	1234 E	fm np] 1
r	KVEMPKEKT	AVENN	SOMETIMODSKOKAKS FEGNNN		BETRIVP	FRATHSCLCLEMELLCC	583 N	fm npl 2
1	KEESBSYRL	RYSRD	BRWMGWK DRWGOEVISGNED	P. EGVVI.KDLGPPMV	ARLVRFYP	BADRYMSVCLRVELYGC	185 F	IS DDB
	TEFAPMYKI	NYSRD	TRWISWR. NRHGKOVLDGNSN	P.YDIFLKDLEPPIV.	ARFVRFIP	VTDHSMNVCMRVELYGC	185 H	IS TKT
	EFVTEXIV	OYSDD	GEFWRSYVN. PTSEPODVOGNSD	G.NSIHYNVFEVPII.		TRWHDRISMRVELYGC	184 I	Om nrx IV
	DDFVTTFFV	GFSND	SQTWVMYT NGYEEMTFHGNVD	K.DTPVLSELPEPVV	ARFIRIYP	LTWNGSLCMRLEVLGC	227 H	Is AEBP1
	.YYVSAYRI	LYSDD	GOKWTVYREPGREQDKIFQGNKR	IITRMVRNNFLPPII.	ARFIRVNP	TOWOOKIAMKMELLGC	380 H	Hs CUB1
ç	VSSDGLHWH	SYRDI	HGTQ PAPQLFPKNWN	GPSTVWMFARMVQ	ARHVRVWPSDGHHQA	APSSDANLDGPLRVELLGC	502 H	Bt SCO SP
1	LKGHVTKFQL	FYKTE	NQHTHKPYPVEFESPAG	NHNAMHHFDLKFPLR.	ARYILFGV	TEYEGNPCMKFDLMGC	1085 0	Ce F47C12 1
	SDKWNKFKL	EYSLD	NESWTTIKEYDKTGAPAG	KDVIEESFETPIS	AKYIRLTNME	NINKWLTFSEFAIVS.	945 0	Cp mu
	SDSLRTWDF	OGSTN	GEQWTVLKRHTNDPSLNYKY	ATHSWPVTGCETA	FRYFRILQTGK	.NSNNRNFLVIGGLEIYG.	803 I	Ds migA
	SRDOTISV	OGSAD	GTGEST LKTSATYAFDP.	GSGNTVTIAFPAT	.OTREVEVDITA	NTGWOAAOLSELEVHA	173	SI AUDI 1
	ARSOTIKT	OASDN	GSDFTDLTAAKAYTFDA.	AGGOSATVTFDAA	TARYVRVLVTA	. NTGOPAAOISELEVYG	313	SI AUDI 2
	. EOVADYET	YTSLN	GTTWDGPVASGRETTS	. LAPORAVE PARDA	RYIRLVALS	. EQTGHKYAAVAELEVEG	656	Mv Neu
	GMITDYKI	YSGDD	VIAEGKWKSD	SSDKYVVFDNPIS	TDNIRIEAISTVG	DENNKHASIAEVEVYE.	185 0	Cs Neu
	NGWIGRHEV	YLSSD	GTNWGSPVASGSWFAD	STTKYSNFETRPA	RYVRLVAIT	.EANGQPWTSIAEINVFQ.	188 1	Dde GOase
	bbbb	bbb	dddd	bbbbbbbbb	bbbbbb	bbbbbb	(GOase 3D struct.

Fig. 2. Alignment of representative DS domains. Sequences are listed from the top to the bottom as they appear in the text. Conserved residues are colored red; those that appear in more than 50% of the cases are shown in bold red. Each domain sequence is identified by a SWISS-PROT or EMBL accession number, and by the starting and ending positions within the protein sequence. Several putative frame-shifts in the human CUB1 sequence were corrected using information from the EST sequence M91216. On top of the alignment the secondary structure prediction for eukaryotic DS domains obtained from the PHD server (Rost, 1996) is shown. E stands for "extended structure." The eight b-strands in the crystal structure of galactose oxidase are indicated as "b" below the alignment.

gether, these arguments strongly suggest that all members of the enlarged DS domain family have the same overall fold.

Evolutionary and functional implications: The relative degrees of sequence conservation among different members of the discoidin domain family suggest that this module has been transferred once or several times between eukaryotes and prokaryotes. Bork and Doolittle (1994) have already proposed horizontal transmission as the most likely explanation for the high similarity between the GOase D1 domain and its bacterial homologues. The identification

of these domains as distant members of the DS domain family provides a stronger quantitative argument supporting this hypothesis: The DS domains of the bacterial sialidases are sequence-wise clearly more similar to the DS domain of GOase than to the DS domains of Mu toxin; however, this bacterial enzyme is functionally and evolutionary more closely related to bacterial sialidases than to fungal GOase. At least two other DS domain-containing proteins appear to have exchanged other parts of their sequences with distantly related organisms, rendering horizontal gene transfer of DS domains even more plausible. The C-terminal sequences of the discoidins share significant sequence similarity only with one other sequence in the current sequence database, a hypothetical protein from *Rohdopseudomonas blastica* (SWISS-PROT accession P05450). The AUD1 protein from *Streptomyces lividans* contains two fibronectin type 3 domains located between DS domains, which presumably are of eukaryotic origin.

There is also a common functional theme to proteins harboring the DS domain: binding to cell surface-attached carbohydrate residues. The discoidins and hemocytin biochemically behave as lectins. The other functionally characterized proteins from higher eukaryotes, i.e., the blood coagulation factors, neuropilin, receptor tyrosine kinases, and neurexin IV, all appear to be implicated in cell surface-mediated regulatory events. Recent data have suggested that neuropilins bind semaphorins via the DS domain (He & Tessier-Lavigne, 1997), thus the DS domain appears to be involved in protein-protein interaction (possibly dependent on post-translationally attached carbohydrate residues). Another interesting case of a DS domain protein mediating cellular interactions is the apparent involvement of P47 (identical to milk fat globule) in fertilization. This protein was detected on the acrosomal cap of testicular sperm and on spermatozoa bound to zona pellucida (Ensslin et al., 1998) suggesting an active role in binding of the sperm to the zona pellucida. Finally, the D1 domain of GOase, which was shown to have weak galactosebinding activity, was proposed to function as an anchor fixing the enzyme to carbohydrates of the cell walls of a tree, the natural habitat of the fungus from which the protein was purified.

DS domains occur in a number of medically important proteins including blood coagulation factors V and VIII, and the recently isolated X-linked juvenile retinoschisis gene XLRS1 (Sauer et al., 1997). The possibility of homology-based three-dimensional structure modeling of their DS domains based on the known crystal structure of galactose oxidase opens new perspectives for studying their function, as well as for designing therapies against diseases caused by mutation of the corresponding genes. Examples are homology-modeled structures of the C1 and C2 domains of factor V presenting new insights on blood coagulation (Villoutreix et al., in prep.). The XLRS1 protein, which is almost exclusively composed of a DS domain, would be another obvious target for such an approach. XLRS1 is a genetic disease causing retinal degradation in males (Sauer et al., 1997). Not surprisingly, all sequenced mutant alleles from patients show changes in phylogenetically conserved amino acids of the DS domain. The previously discussed zona pellucida binding protein represents another example where structural inferences based on the fold prediction reported in this paper may lead to applications. Finally, the bacterial DS domains are also relevant from a medical perspective, as they all occur in proteins that were shown or hypothesized to be virulence factors of human pathogens.

The profile describing the DS domain has been added to PROSITE (Bairoch et al., 1996) under the accession number PS50022.

Acknowledgments: We would like to thank Mary Stewart for critical reading of the manuscript. S.B. was supported by a grant from the Per-Eric and Ulla Schyberg Foundation. K.H. was supported by grant 31-49669.96 from the Swiss National Research Foundation.

References

- Bairoch A, Hofmann K, Bucher P. 1996. The PROSITE database, its status in 1995. Nucleic Acids Res 23:189–196.
- Baumgartner S, Littleton JT, Broadie K, Bhat MA, Harbecke R, Lengyel JA, Chiquet-Ehrismann R, Prokop A, Bellen HJ. 1996. A Drosophila neurexin

- Bork P, Doolittle RF. 1994. Drosophila kelch motif is derived from a common enzyme fold. J Mol Biol 236:1277-1282.
- Bucher P, Karplus K, Moeri N, Hofmann K. 1996. A flexible motif search technique based on generalized profiles. *Comput Chem* 20:3-24.
- Canard B, Garnier T, Saint-Joanis B, Cole ST. 1994. Molecular genetic analysis of the nagH gene encoding a hyaluronidase of *Clostridium perfringens*. Mol Gen Genet 243:215-224.
- Ensslin M, Vogel T, Calvete JJ, Thole HH, Schmidtke J, Matsuda T, Töpfer-Petersen E. 1998. Molecular cloning and characterization of P47, a novel boar sperm-associated zona pellucida-binding protein homologous to a family of mammalian secretory proteins. *Biol Reprod* 52:1057–1064.
- Escalante R, Wessels D, Soll DR, Loomis WF. 1997. Chemotaxis to cAMP and slug migration in *Dictyostelium* both depend on migA, a BTB protein. *Mol Biol Cell* 9:1763–1775.
- Gobron S, Monnerie H, Meiniel R, Creveaux I, Lehmann W, Lamalle D, Dastugue B, Meiniel A. 1996. SCO-spondin: A new member of the thrombospondin family secreted by the subcommissural organ is a candidate in the modulation of neuronal aggregation. J Cell Sci 109:1053–1061.
- He Z, Tessier-Lavigne M. 1997. Neuropilin is a receptor for the axonal chemorepellent semaphorin III. Cell 90:739–751.
- Hidai C, Zupancic T, Penta K, Mikhail A, Kawana M, Quertermous EE, Aoka Y, Fukagawa M, Matsui Y, Platika D, Auerbach R, Hogan BLM, Snodgrass R, Quertermous T. 1998. Cloning and characterization of developmental endothelial locus-1: An embryonic endothelial cell protein that binds to the αvβ3 integrin receptor. *Genes Dev* 12:21–33.
- Hofmann K, Bucher P. 1995. The FHA domain: A putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem Sci* 20:347–349.
- Ito N, Phillips SE, Yadav KD, Knowles PF. 1994. Crystal structure of a free radical enzyme, galactose oxidase. J Mol Biol 238:794–814.
- Jenny RJ, Pittmann DD, Toole JJ, Kriz RW, Aldape RA, Hewick RM, Kaufman RJ, Mann KG. 1987. Complete cDNA and derived amino acid sequence of human factor V. Proc Natl Acad Sci USA 84:4846–4850.
- Johnson JD, Edman JE, Rutter WJ. 1993. A receptor tyrosine kinase found in breast carcinoma cells has an extracellular discoidin I-like domain. Proc Natl Acad Sci USA 90:5677–5681.
- Karn T, Holtrich U, Brauninger A, Bohme B, Wolf G, Rubsamen-Waigmann H, Strebhardt K. 1993. Structure, expression and chromosomal mapping of TKT from man and mouse: A new subclass of receptor tyrosine kinases with a factor VIII-like domain. *Oncogene* 8:3433–3440.
- Kawakami A, Kitsukawa T, Takagi S, Fujisawa H. 1995. Developmentally regulated expression of a cell surface protein, neuropilin, in the mouse nervous system. J Neurobiol 29:1–17.
- Kotani E, Yamakawa M, Iwamoto S, Tashiro M, Mori H, Sumida M, Matsubara F, Taniai K, Kadono-Okuda K, Kato Y, Mori H. 1995. Cloning and expression of the gene of hemocytin, an insect humoral lectin which is homologous with the mammalian von Willebrand factor. *Biochem Biophys Acta* 1260:245–258.
- Ogura K, Nara K, Watanabe Y, Kohno K, Tai T, Sanai Y. 1996. Cloning and expression of cDNA for O-acetylation of GD3 ganglioside. *Biochem Bio*phys Res Commun 225:932–938.
- Ohno I, Hashimoto J, Shimizu K, Takaoka K, Ochi T, Matsubara K, Okubo K. 1996. A cDNA cloning of human AEBP1 from primary cultured osteoblasts and its expression in a differentiatin osteoblastic cell line. *Biochem Biophys Res Commun.* 228:411–414.
- Piendl W, Eichenseer C, Viell P, Altenbucher J, Cullum J. 1994. Analysis of putative DNA amplification genes in the element AUD1 of *Streptomyces lividans*. Mol Gen Genet 244:439–443.
- Poole S, Firtel R, Lamar E. 1981. Sequence and expression of the discoidin 1 family in *Dictyostelium discoideum*. J Mol Biol 153:273–289.
- Rost B. 1996. PHD: Predicting one-dimensional protein structure by profilebased neural networks. *Methods Enzymol* 266:525–539.
- Sauer C, Gehrig A, Warneke-Wittstock R, Marquard A, Ewing CC, Gibson G, Lorenz B, Jurklies B, Weber BHF. 1997. Position cloning of the gene associated with X-linked juvenile retinoschisis. *Nature Gen* 17:164–170.
- Stubbs JD, Lekutis C, Singer KL, Bui A, Yukuzi D, Srinivasan U, Parry G. 1990. cDNA cloning of a mouse mammary epithelial cell surface protein reveals the existence of epidermal growth factor-like domains linked to factor VIIIlike sequences. *Proc Natl Acad Sci USA 87*:8417–8421.
- Takagi S, Hirata T, Agata K, Mochii M, Eguchi G, Fujisawa H. 1991. The A5 antigen, a candidate for the neuronal recognition molecule, has homologies to complement components and coagulation factors. *Neuron* 7:295–307.
- Wood WI, Capon DJ, Simonson CC, Eaton DL, Gitschier J, Keyt B, Seeburg PH, Smith DH, Hollingshead P, Wion KL, Delwart E, Tuddenham EGD, Vehar GA, Lawn RM. 1984. Expression of active human factor VIII from recombinant DNA clones. *Nature* 312:330–337.