Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

PAN-modular structure of microneme protein SML-2 from the parasite *Sarcocystis muris* at 1.95 Å resolution and its complex with 1-thio-β-D-galactose

Jürgen J. Müller,[a] Manfred S. Weiss[b] and Udo Heinemann[a,c]*

[a]Kristallographie, Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, 13125 Berlin, Germany, [b]AG Makromolekulare Kristallographie, Helmholtz-Zentrum Berlin für Materialien und Energie, BESSY II, Albert-Einstein-Strasse 15, 12489 Berlin, Germany, and [c]Institut für Chemie und Biochemie, Freie Universität Berlin, Takustrasse 6, 14195 Berlin, Germany

Correspondence e-mail: heinemann@mdc-berlin.de

Received 2 August 2011
Accepted 15 September 2011

PDB References: apo SML-2, 2yil; complex with 1-thio-β-D-galactose, space group $C222_1$, 2yio; space group $P2_12_12_1$, 2yip.

The microneme protein SML-2 is a member of a small family of galactose-specific lectins that play a role during host-cell invasion by the apicomplexan parasite *Sarcocystis muris*. The structures of apo SML-2 and the 1-thio-β-D-galactose–SML-2 complex were determined at 1.95 and 2.1 Å resolution, respectively, by sulfur-SAD phasing. Highly elongated dimers are formed by PAN-domain tandems in the protomer, bearing the galactose-binding cavities at the distal apple-like domains. The detailed structure of the binding site in SML-2 explains the high specificity of galactose-endgroup binding and the broader specificity of the related *Toxoplasma gondii* protein TgMIC4 towards galactose and glucose. A large buried surface of highly hydrophobic character and 24 intersubunit hydrogen bonds stabilize the dimers and half of the 12 disulfides per dimer are shielded from the solvent by the polypeptide chain, thereby enhancing the resistance of the parasite protein towards unfolding and proteolysis that allows it to survive within the intestinal tracts of the intermediate and final hosts.

1. Introduction

In the last decade, considerable progress has been made towards the identification and functional characterization of apical invasion proteins from parasite secretory organelles such as micronemes and rhoptries. Proteins from the parasites *Toxoplasma gondii*, *Eimeria tenella*, *Cryptosporidium* and *Plasmodium falciparum* have been the focus of structural and functional investigations because of their high pathogenicity (Carruthers & Tomley, 2008; Tonkin *et al.*, 2011). These proteins have significant functional versatility that allows them to fulfil diverse biological functions by mediating protein–protein, protein–ligand and especially protein–carbohydrate interactions (Carruthers & Tomley, 2008). Understanding the molecular basis of the cell-surface recognition code established by these interactions is essential in numerous human as well as animal disease processes (Sharon & Lis, 1989) and is the basis for the development of vaccines directed at these pathogens.

Recently, NMR solution structures of the *E. tenella* microneme protein EtMIC5, as well as crystal structures of *T. gondii* and *P. falciparum* AMA1 domains and their complexes with oligosaccharide and peptide ligands, have been analyzed (Brown *et al.*, 2003; Crawford *et al.*, 2010; Bai *et al.*, 2005; Tonkin *et al.*, 2011). Apple-like and partial PAN modules are common structural elements of these proteins, which often lack significant sequence similarity. Modules of this kind are frequently found in proteins from these and related parasites (Brecht *et al.*, 2001).

© 2011 International Union of Crystallography
Printed in Singapore – all rights reserved

Table 1
Data-processing statistics for data sets collected for sulfur phasing.

Values in parentheses are for the outer shell.

	Set 1 (native)	Set 2	Set 3	Set 4	Set 5	Sets 2 + 3	Sets 2 + 3 + 4	Sets 2 + 3 + 4 + 5
Space group	$P2_12_12_1$							
Wavelength (Å)	1.0000	2.0000						
Resolution (Å)	35.0–1.95 (2.00–1.95)	35.0–2.20 (2.26–2.20)	35.0–2.40 (2.46–2.40)	35.0–2.40 (2.46–2.40)	35.0–2.90 (2.98–2.90)	35.0–2.20 (2.26–2.20)	35.0–2.20 (2.26–2.20)	35.0–2.40 (2.46–2.40)
Unit-cell parameters								
a (Å)	53.17	53.26	53.33	53.39	53.42	53.29	53.32	53.33
b (Å)	129.04	129.34	129.55	129.74	129.86	129.43	129.53	129.56
c (Å)	158.10	158.41	158.58	158.68	158.71	158.48	158.54	158.56
Unique reflections	79584	95339	72715	73015	35926	95360	95362	95363
Multiplicity	7.2	7.1	7.1	7.1	5.1	12.5	17.8	19.7
$\langle I/\sigma(I)\rangle$	23.8 (4.2)	28.8 (4.1)	33.7 (4.8)	31.1 (3.8)	24.6 (2.1)	31.8 (3.8)	35.5 (3.5)	35.9 (3.5)
Completeness (%)	99.5 (96.5)	88.9 (47.4)	87.6 (45.3)	87.7 (45.9)	76.0 (23.9)	88.9 (47.4)	88.9 (47.4)	88.9 (47.4)
$R_{\rm merge}$ (%)	5.6 (42.1)	4.5 (29.9)	4.0 (24.6)	4.4 (31.4)	5.1 (36.7)	5.3 (29.9)	6.2 (29.9)	6.7 (29.9)
$R_{\rm r.i.m.}$† (%)	6.1 (45.8)	4.9 (34.4)	4.4 (27.3)	4.7 (36.3)	5.7 (47.6)	5.5 (34.4)	6.4 (34.4)	6.8 (34.4)
Wilson *B* factor (Å^2)	30.8	36.8	39.9	43.0	36.7	36.7	36.7	36.7

† Weiss (2001).

Sarcocystis muris is an intracellular parasite which propagates in mice, rats, voles *etc.* as intermediate hosts and, as known so far, in cats (Mehlhorn & Heydorn, 1978) and ferrets (Rommel, 1979) as final hosts. After adhesion of the parasite to cells of an intermediate host, the microneme *S. muris* lectins (SML) are secreted and attached to the plasma membrane of the infected cell (Entzeroth *et al.*, 1992). Furthermore, moving junctions (Baum & Cowman, 2011) between the parasite and host cell during the invasion process have been identified as SML attachment sites. In general, the SML are thought to contribute in an adhesin-like manner to specific host-cell recognition (Entzeroth *et al.*, 1992), but the direct molecular mechanism of this process is unknown.

Here, we present the high-resolution crystal structure of a galactose-specific *S. muris* lectin: the major micronemal protein SML-2 (Eschenbacher *et al.*, 1993; Müller *et al.*, 2001). A recent very low resolution *ab initio* structure determination of SML-2 provided some insight into the shape and symmetry of the homodimeric protein (Müller *et al.*, 2006). In the absence of a suitable model for molecular-replacement phasing, the presence of six disulfide bridges, as identified by mass spectrometry (Müller *et al.*, 2001), and five methionine residues per SML-2 molecule suggested the use of the sulfur phasing method for structure determination at 1.95 Å resolution. Furthermore, to analyze the structural basis of the specificity of the lectin towards galactose, complexes of SML-2 with 1-thio-β-D-galactose in space groups $C222_1$ and $P2_12_12_1$ were characterized. This study provides insight into the recognition of host cells by *S. muris* and related parasites.

2. Materials and methods

2.1. Protein isolation, purification and crystallization

SML-2 was isolated and purified from *S. muris* cystozoites from skeletal muscle of infected mice as described by Müller *et al.* (2001). The crystallization of apo SML-2 and the 1-thio-β-D-galactose–SML-2 complex in space groups $P2_12_12_1$ and $C222_1$, respectively, has also been described in detail previously (Müller *et al.*, 2001).

At an early stage of structure determination we decided to use heavy-atom derivatives for phase determination. SML-2 is known to bind galactose and *N*-acetyl-D-galactosamine with high affinity (Klein *et al.*, 1998). As found in a number of proteins, including winged bean lectin (PDB entry 1wbl; Prabu *et al.*, 1998), human galectin-7 (PDB entry 2gal; Leonidas *et al.*, 1998), *Erythrina corallodendron* lectin (PDB entry 1axz; Elgavish & Shaanan, 1998), peanut lectin (PDB entry 1bzw; Ravishankar *et al.*, 1998) and soybean agglutinin (PDB entry 1sbe; Olsen *et al.*, 1997), atoms C^3, C^4, C^5 and C^6 of a bound galactose moiety interact with the protein and an aromatic side chain provides a planar contact area. In these complexes the galactose is oriented with the C^1 hydroxyl group towards the solvent-accessible side of the protein. Therefore, the sugar O^1 was modified by a sulfur-linked Au atom suitable for SIRAS (1-thio-β-D-galactose–gold complex, kindly provided by U. Pfüller, University Witten-Herdecke, Germany). Unfortunately, the metal complex was unstable and dissociated, leaving only the 1-thio-β-D-galactose moiety (GAT) bound to the protein. The diffraction power of the crystals containing the GAT–SML-2 complex in space group $C222_1$ was lower than that of the apoprotein crystals in space group $P2_12_12_1$. Therefore, the cocrystals were tempered by blocking the nitrogen stream of the cryostream cooler for about 5 s. Owing to this tempering, the crystals switched to a slightly tighter packing.

2.2. X-ray diffraction experiments

Three diffraction experiments were conducted independently. The sulfur phasing diffraction experiment was performed on EMBL beamline X12 at DESY, Hamburg. One crystal was mounted at 100 K in a nitrogen stream (Oxford Cryosystems, England). During the experiment five data sets (sets 1–5) were consecutively collected using a MAR Mosaic 225 detector (MAR Research, Hamburg, Germany). The 180

Table 2
Data-processing and refinement statistics for apo SML-2, the SML-2–GAT complex in space group $P2_12_12_1$ and the SML-2–GAT complex in space group $C222_1$.

Values in parentheses are for the outer shell.

	Apo SML-2	SML-2–GAT	SML-2–GAT
Space group	$P2_12_12_1$	$P2_12_12_1$	$C222_1$
Wavelength (Å)	1.0	0.8428	1.542
Resolution (Å)	35.0–1.95 (2.00–1.95)	35.0–2.14 (2.20–2.14)	23.9–2.43 (2.50–2.43)
Unit-cell parameters			
a (Å)	53.17	53.31	74.73
b (Å)	129.04	130.00	81.99
c (Å)	158.10	158.85	130.96
Unique reflections	79584	61346	15317
Multiplicity	7.2 (6.4)	5.0 (4.6)	5.0 (3.9)
$\langle I/\sigma(I)\rangle$	23.8 (4.2)	38.6 (14.1)	29.1 (8.44)
Completeness (%)	99.5 (96.5)	99.5 (95.8)	99.1 (92.6)
$R_{\rm merge}$† (%)	5.6 (42.1)	2.8 (10.3)	4.0 (14.7)
$R_{\rm r.i.m.}$‡ (%)	6.1 (45.8)	3.1 (11.6)	4.5 (17.0)
Wilson *B* factor (Å^2)	30.8	28.6	30.2
Asymmetric unit content (chains/residues)	6/779	6/785	2/259
$V_{\rm M}$ (Å^3 Da^{-1})	3.06	3.06	3.34
Solvent content (%)	59.8	59.8	63.2
$R_{\rm work}/R_{\rm free}$ (%)	16.8/19.7	16.1/20.2	17.0/21.4
R.m.s.d. bonds (Å)	0.016	0.024	0.019
R.m.s.d. angles (°)	1.491	1.905	1.790
Ramachandran plot			
Residues in favoured regions (%)	97.8	97.5	97.6
Residues in allowed regions (%)	99.9	99.7	99.2
No. of outliers	1	2	2
Outliers	*F*91	*B*91, *C*91	*A*91, *B*91
Total No. of atoms	6691	6629	2122
Protein atoms	5948	5985	1976
Waters	661	517	91
GATs	0	6	2
Glycerols	10	7	2
Cl^-	7	13	9
SO_4^{2-}	3	0	2
$B_{\rm average}$			
All protein atoms	26.1	35.4	43.1
Main-chain atoms	24.7	33.9	41.7
Side-chain atoms	27.7	37.0	44.6
Waters	35.5	42.6	43.0
GATs	—	56.3	51.6

† $R_{\rm free}$ calculated using 5% of reflections. ‡ Weiss (2001).

frames ($\Delta\varphi = 1^\circ$) of set 1 were measured at 1.0000 Å wavelength to 1.95 Å resolution. Set 1 is the 'native' apo SML-2 data set because of the low anomalous signal from sulfur at this wavelength. The following sets 2–4 consisted of three complete 360° scans with $\Delta\varphi = 1^\circ$ at a wavelength of 2.0000 Å (starting at φ settings of 0°, 0.333° and 0.666°), optimized for the sulfur signal and the crystal lifetime. The last set, set 5, comprised 270 frames but was discarded during data processing (see Supplementary Material[1]).

The data set for the 1-thio-β-D-galactose–SML-2 complex in space group $P2_12_12_1$ was obtained on EMBL beamline BW7B at DESY using a 345 mm MAR image plate (MAR Research, Hamburg, Germany) from a crystal mounted in a nitrogen stream at 100 K generated by a Cryostream Cooler (Oxford Cryosystems, England). The fixed wavelength was 0.8428 Å.

Diffraction data for the 1-thio-β-D-galactose–SML-2 complex in space group $C222_1$ were collected using a conventional Rigaku Denki RU-H2B X-ray generator with a direct-drive copper anode, a MAR300 imaging-plate detector and crystal cooling in a nitrogen stream at 110 K.

All data sets were processed with the programs *XDS* and *XSCALE* (version 10 May 2010; Kabsch, 2010*a*,*b*). The data-reduction statistics are presented in Table 1. Careful data handling, *e.g.* the removal of dubious images at the beginning and end of scans, the removal of all 'aliens' as annotated by *XDS*, the cutting of possibly erroneous low- and high-resolution bins and the exclusion of set 5, despite the relative low redundancy of 17.8 for the remaining data, permitted substructure determination using the *SHELXC*/*D*/*E* beta suite (Sheldrick, 2010). The exclusion of set 5 was based on slightly increasing unit-cell parameters on comparing sets 1–5 (Table 1) and the reduction of the anomalous signal as detected in SigAno from the *XDS* data output (Supplementary Table S1[1]) owing to radiation damage (as visible in the electron density; see Supplementary Fig. S10[1]). The scaled but unmerged data sets 2–4 as well as the unmerged 'native' apo SML-2 data set 1 were used as input to *SHELXD*. From $\langle I/\sigma(I)\rangle$, $\langle d'/\sigma\rangle$ and the self-anomalous correlation coefficient Self-Anomalous CC calculated in *SHELXC* (Supplementary Figs. S1–S3[1]) the useful resolution range for substructure calculation was determined to be 35–3.1 Å.

36 disulfide bridges (treated as super-sulfurs with minimal distance allowed between heavy atoms MIND = 3.5 Å) and 66 heavy atoms of sulfur type including the 30 sulfurs in the methionine residues of six SML-2 monomers are present in the asymmetric unit (Müller *et al.*, 2001, 2006). Cycle 35 165 in *SHELXD* produced a singular solution for the heavy-atom substructure with $CC_{\rm all}$ = 35.5 and $CC_{\rm weak}$ = 10.73 (Supplementary Figs. S4–S6[1]). The next best value, which did not represent a valid solution, was $CC_{\rm all}/CC_{\rm weak}$ = 30.8/7.0. Next, *SHELXE* beta was used for phase determination, density modification and chain tracing. Local scaling of the low-resolution SAD data to the high-resolution 'native' apo SML-2 data set 1 and a high-resolution extrapolation limit of 1 Å were applied. An optimal solvent content of 0.65 was estimated and used instead of the real value of 0.59 calculated from crystal packing, accounting for disordered side chains (Sheldrick, 2010). 105 anomalous scatterers were used for

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: DZ5237). Services for accessing this material are described at the back of the journal.

phasing and tracing (Supplementary Fig. S7). A contrast of 0.743 and a connectivity of 0.8 after density modification and a mean FOM of 0.608 and a pseudo-free CC of 63.4% after three iterations of auto-tracing characterized the quality of the phases.

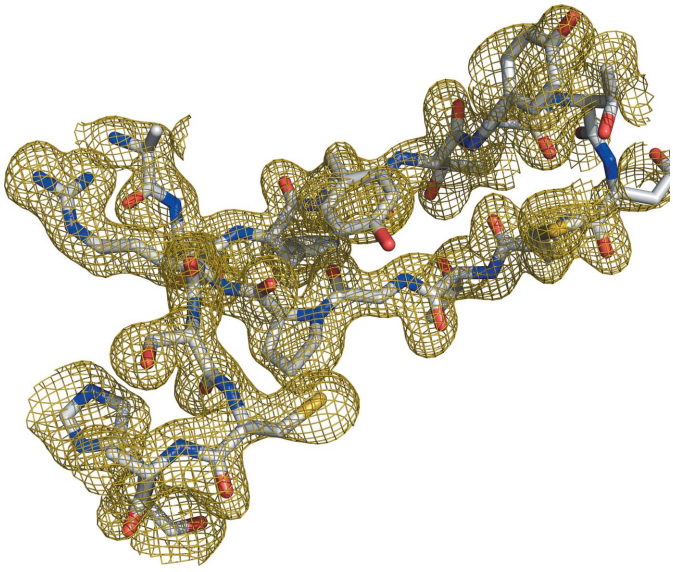


Figure 1
Experimental electron density contoured at 1σ after sulfur phasing with *SHELXC*/*D*/*E* beta. The inserted atomic model is the refined structure of apo SML-2.

The resulting electron density allowed the tracing of poly-alanine chains for 640 residues of the expected 828 (Supplementary Fig. S8). 59 sulfurs were subsequently identified with sulfur positions in the refined structure (*SITCOM* program; Dall'Antonia & Schneider, 2006; Supplementary Fig. S9). The resulting experimental electron density is shown in Fig. 1. Model building with *ARP*/*wARP* v.7.1 (Lamzin & Wilson, 1993) based on the experimental phases resulted in 757 residues (out of 828) in sequence after loop construction, which belonged to six monomers that were arranged as three dimers and related by translational NCS. Several rounds of refinement with *REFMAC* v.5.5.0109 (Murshudov *et al.*, 2011) from *CCP*4 (Winn *et al.*, 2011), interactive corrections with *Coot* v.0.6.1 (Emsley *et al.*, 2010) and the addition of water by *ARP*/*wARP* and of cryoprotectant molecules and ions (Table 2) resulted in a model which fitted the data with $R_{\rm work}$ = 16.8% and $R_{\rm free}$ = 19.7% to 1.95 Å resolution.

The structures of the SML-2–GAT complex in space groups $C222_1$ and $P2_12_12_1$ were solved by molecular replacement in *Phaser* (McCoy *et al.*, 2007) using an apo SML-2 monomer as a search model. For space group $C222_1$ two monomers were found in the asymmetric unit, which were completed to dimers by crystallographic symmetry operations. To reduce the model bias and for completion of the chains and addition of water, *ARP*/*wARP* was used to reconstruct the monomers. Iterative interactive corrections with *Coot* (Emsley *et al.*, 2010) and refinement with *REFMAC*5 (Table 2) resulted in a structure model of the complex with $R_{\rm work}$ = 17.0% and $R_{\rm free}$ = 21.4%.

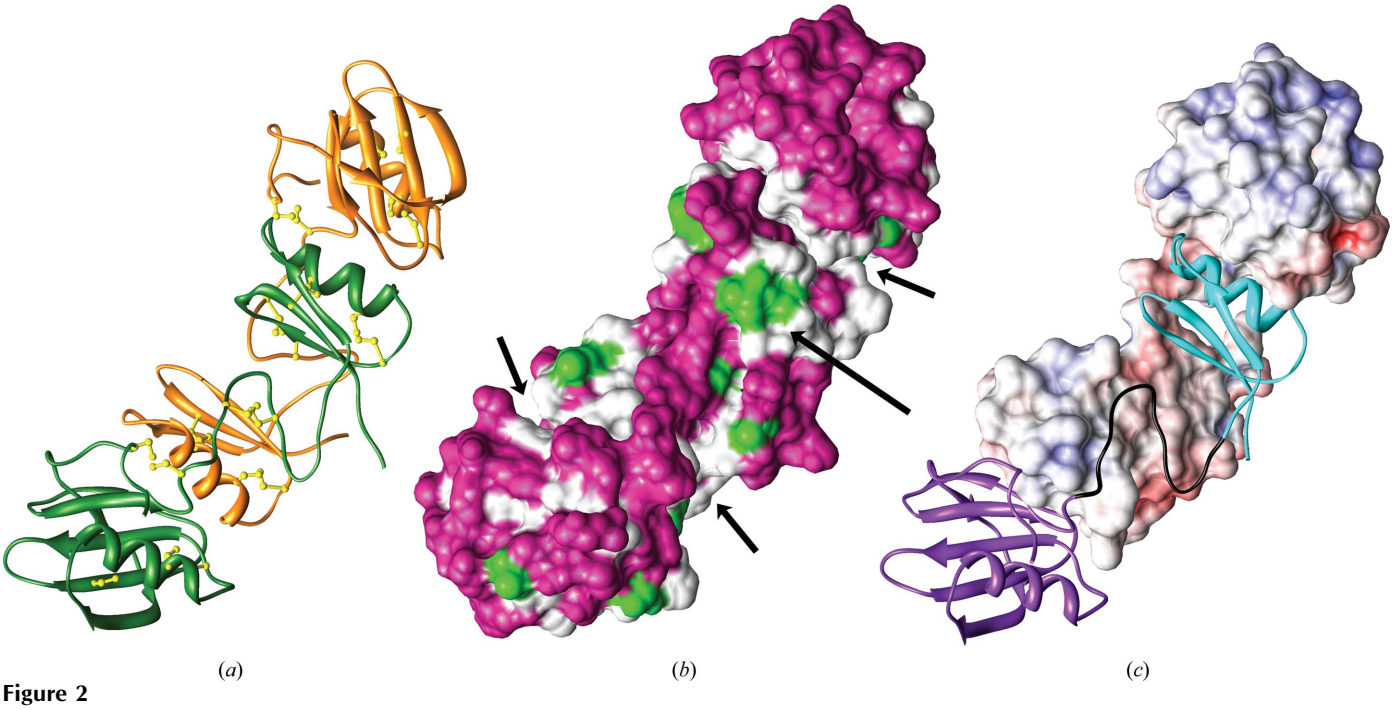


Figure 2
Asymmetric unit of the apo SML-2 crystal. Chains *A*/*B* (*a*), *C*/*D* (*b*) and *E*/*F* (*c*) form homodimers. (*a*) Chains *A*/*B* (coloured green and orange) are shown with their secondary structure emphasized and with disulfide bridges (cysteines in ball-and-stick representation). (*b*) The surface of chains *C*/*D* is coloured white where it covers hydrophobic residues and deep pink around hydrophilic residues. Hydrophobic residues differing from homologous structures from the SML family are coloured green. (*c*) Chain *E* is shown as a secondary-structure plot with its apple-like domain (PAN_AP) coloured magenta, the C-terminal PAN_1 domain coloured cyan and the connecting peptide stretch between them in black. The surface of chain *F* is coloured by electrostatic potential (±6 kcal mol^{-1} for deepest blue and red) as calculated by *PyMOL* (DeLano, 2002). The secondary-structure plots were generated by *CHIMERA* (Pettersen *et al.*, 2004).

The procedure was identical for the SML-2–GAT complex in crystals belonging to space group $P2_12_12_1$, for which the refinement converged with R_{work} = 16.1% and R_{free} = 20.2% at 2.14 Å resolution (Table 2).

3. Structure analysis

3.1. Apo SML-2 in space group $P2_12_12_1$

The composition of the asymmetric unit as described recently at 16 Å resolution (Müller *et al.*, 2006) was confirmed. Six chains assemble as three tightly packed dimers with a total of 779 residues and are positioned nearly equidistant (52.0 Å) along a vector parallel to the *c* axis within the asymmetric unit (Fig. 2). This translational NCS was expected from Patterson maps and was indicated by relatively large structure-factor amplitudes for *h*, *k*, 3*n* reflections at very low resolution (Müller *et al.*, 2006). The dimer has the overall shape of a slightly curved cylinder with a maximal diameter of 80 Å and a minimal diameter of 25 Å. Roughly half of the total surface area of 12 100 Å^2 of the dimer is polar (5200 Å^2). Hydrophobic areas on the dimer surface are quite extended and comprise two large patches, one between the chains and one atop the C-terminal domain of each chain (marked by arrows in Fig. 2*b*).

Approximately 130 residues per polypeptide chain on average were located in the electron density, lacking residues 1–3 and 135–138 at the chain termini. The dimer subunits are related by noncrystallographic twofold axes and matched with r.m.s.d. values for C$^\alpha$ atoms of 0.83, 0.18, and 1.20 Å, respectively.

Each SML-2 molecule consists of two domains of the PAN superfamily, as defined by PROSITE (PS50948), extending from Gln4 to Cys78 and from Cys82 to Ser134. The N-terminal PAN domain is furthermore grouped into the PAN_AP subfamily of apple domains (apple-like) by *SMART* (SM00473), where the apple-domain fold consists of a central four-stranded β-sheet with antiparallel strands a–d (Fig. 3; Gln22–His27, His46–Asn50, Lys55–Lys60 and Asp71–Arg76) that wraps around the ten-residue α-helix H1 (Ala31–Ala40). The central sheet is flanked by a two-stranded β-sheet formed by the antiparallel strands α and β (Ile16–Ser18 and Phe65–Thr67). The conserved disulfide bonds Cys9–Cys78, Cys34–Cys56 and Cys38–Cys44 stabilize the PAN_AP module, where the latter two link helix H1 to the β-sheet. The disulfides Cys82–Cys86, Cys107–Cys127 and Cys111–Cys117 belong to the C-terminal domain and the latter two attach H2 to the three-stranded antiparallel β-sheet e–g (Ala95–Asp98, Leu119–Thr121 and Thr126–Tyr130). Residues Asn92–Gln132 of the second domain can be superimposed onto Glu21–Gly61 of the first domain with an r.m.s.d. of 1.0 Å for 39 C$^\alpha$ positions with 23% sequence identity. Both smaller modules belong to

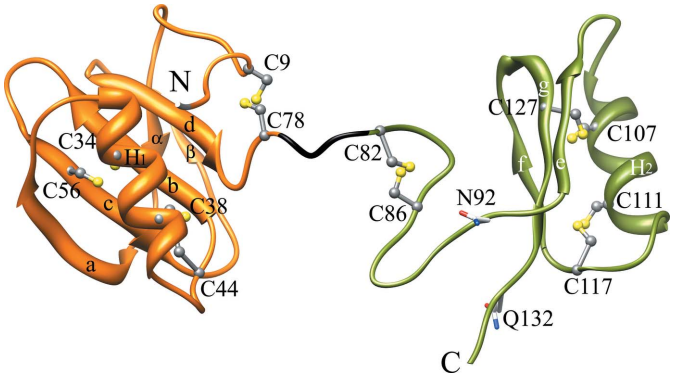


Figure 3
C$^\alpha$ trace of SML-2 monomer *A*. The PAN domains (PROSITE PS50948) are coloured gold and green and the linker is coloured black. Residues Leu5–Cys78 belong to an apple-like PAN_AP motif (*SMART* SM00473) and residues Asn92–Gln132 belong to a PAN_1 motif (PFAM PF00024). Disulfide bridges are shown in ball-and-stick representation. The strands are denoted a, b, c and d in the PAN_AP sheet I and α and β in sheet II. Strands e, f and g and α-helix H2 belong to the PAN_1 motif in the C-terminal domain (residues Asn92–Gln132).

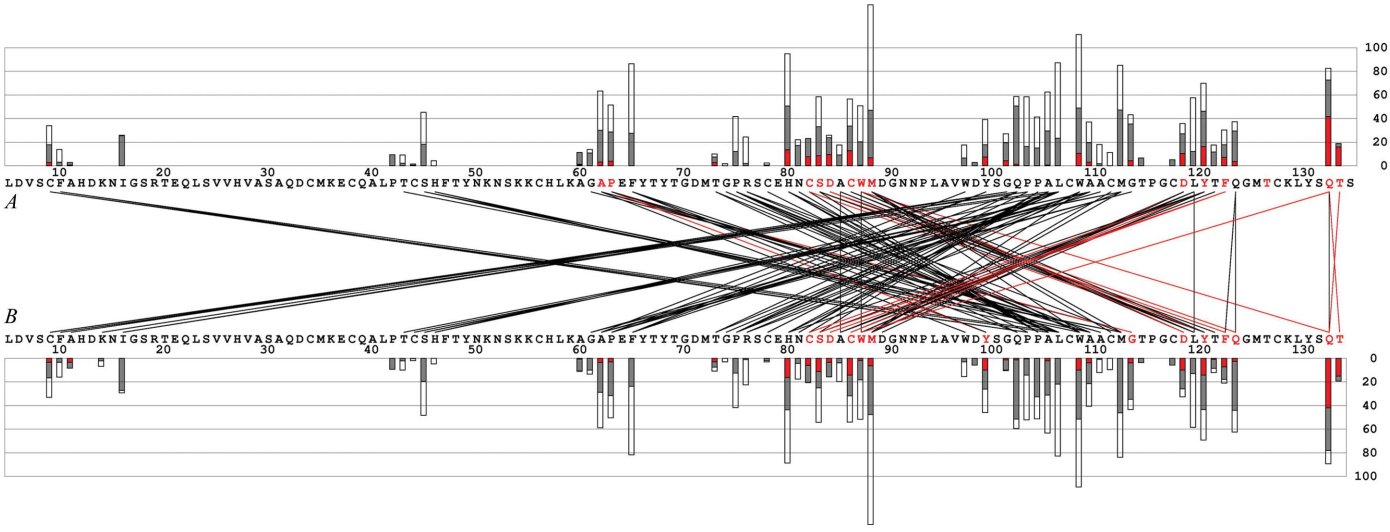


Figure 4
Buried area plot of the interaction site in dimer *A*/*B* of apo SML-2 (*XSAE*; Broger, 2011). The buried solvent-accessible surface per residue is partitioned into polar (red bars), apolar (white bars) and mixed type (grey bars). The scale on the right is in absolute Å^2. Hydrogen-bond partners in the interface are connected by red lines and hydrogen-bond donors or acceptors are shown in red capitals; van der Waals contacts are marked by black lines.

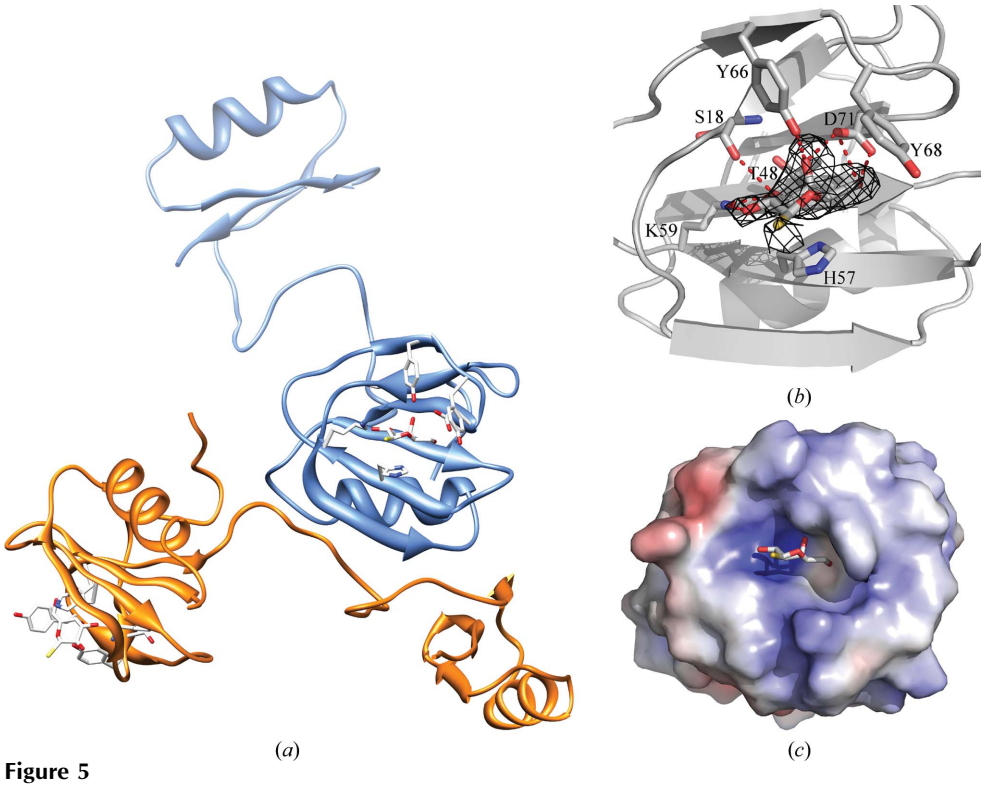


Figure 5
1-Thio-β-D-galactose–SML-2 complex in space group $C222_1$. (*a*) Two monomers in the asymmetric unit. Galactose moieties and interacting residues are shown in stick representation. (*b*) 1-Thio-β-D-galactose in the binding cleft with OMIT electron density contoured at 3σ (black). Hydrogen bonds are shown in red. (*c*) Solvent-accessible surface of the apple-like domain binding cavity coloured by potential (*PyMOL* graphics).

the PAN_1 members of the PAN superfamily as defined by PFAM (PF00024), which are characterized by a three-stranded β-sheet opposite a nine-residue α-helix, and only two conserved disulfide bridges are located in these modules. The buried disulfides Cys34–Cys56, Cys38–Cys44 and Cys107–Cys127 stabilize the PAN globules and are the most resistant to radiation damage (see Supplementary Material).

In the monomer nearly all residues outside β-strands b and d are solvent-exposed, with a solvent-accessible surface area of about 8100 Å^2. The biologically active unit is a dimer of SML-2 (Montag *et al.*, 1997). The dimers *A*/*B*, *C*/*D* and *E*/*F*, with a total dimer surface of 11 900 Å^2 each, interact over an area of 2000 Å^2 per monomer, a value comparable to the most stable protein complexes (LoConte *et al.*, 1999). The dimer is stabilized by a polar interface area of about 220 Å^2 per monomer with 23 hydrogen bonds, but also by 960 Å^2 of buried hydrophobic surface per monomer. The remainder is of mixed type (Fig. 4). Whereas the apple-like domain Cys9–Cys78 contributes 453 Å^2 per chain (polar, 14 Å^2; apolar, 243 Å^2; mixed, 196 Å^2), the linker and the PAN_1 motif Glu79–Ser134 contribute the main contact area of 1548 Å^2 (polar, 187 Å^2; apolar, 708 Å^2; mixed, 654 Å^2) per chain.

3.2. Complex of 1-thio-β-D-galactose with SML-2 in space group $C222_1$

SML-2 binds sugars noncovalently but with high affinity and displays specificity towards galactose and *N*-acetyl-D-galactosamine (Klein *et al.*, 1998). 1-Thio-β-D-galactose derived from a 1-thio-β-D-galactose–gold complex synthesized for phasing was cocrystallized with the lectin. Fig. 5(*a*) shows the asymmetric unit of the 1-thio-β-D-galactose–SML-2 complex. Two monomers present in the asymmetric unit form dimers, as in apo SML-2, related by crystallographic dyad symmetry. The secondary interface of about 415 Å^2 per monomer is much smaller than the inner dimer interface, but may account for the trend of the dimers to oligomerize (Montag *et al.*, 1997). This interaction site contains the only Ramachandran outlier, Asn91, in the SML-2 structures.

The galactose-binding cleft is localized at the distal side of the PAN_1 disulfide motif of the PAN_AP domain. In apo SML-2, five water molecules on average are found in the binding cavity for the galactose molecule (not shown). The vicinity of the 1-thio-β-D-galactose is shown in Fig. 5(*b*) and the electrostatic potential is shown in Fig. 5(*c*). Seven hydrogen bonds shorter than 3.3 Å and nine hydrophobic contacts (not shown) with distances of less than 4.0 Å fix the ligand in the cavity. The presence of an aromatic residue (His57) facing the nonpolar side of galactose is a common feature of galactose-specific lectins (Sujatha *et al.*, 2005). The O^4 of the galactose moiety forms two hydrogen bridges to Tyr66 and Asp71, respectively. In a putative complex with β-D-glucose these bonds could not form and, on the other side of the sugar ring, the stacking interaction with His57 would be distorted by the glucose O^4, explaining the specificity for galactose. The structures of apo SML-2 and holo SML-2 around the binding site are identical within coordinate error (the r.m.s.d. of C^α of Ser18, Thr48, His57, Lys59, Tyr66, Tyr68 and Asp71 is 0.08 Å). Neither water molecules nor ions are involved in the bonding network. 1-Thio-β-D-galactose fills the active site almost completely and a modelled *N*-acetyl-D-galactosamine, which binds SML-2 with highest affinity (Klein *et al.*, 1998), fits into the pocket like a key into a lock (not shown).

3.3. Complex of 1-thio-β-D-galactose with SML-2 in space group $P2_12_12_1$

Tempering crystals of space group $C222_1$ on the beamline by stopping the cryocooler stream induces a switch to a tighter crystal packing in space group $P2_12_12_1$ with a solvent content

of 59.8% (Table 2). By this reorganization of the dimers, the crystal contact interface is enlarged by about 140 Å^2 and nine instead of six inter-chain hydrogen bonds are formed.

The *ab initio* phasing method of Lunin *et al.* (2002) was previously used to derive a very low resolution structure of SML-2 at 16–18 Å. The data presented above confirm the validity of the low-resolution map (Müller *et al.*, 2006). Fig. 6 provides a comparison of the current structure with the low-resolution model.

3.4. Polymorphism of the SML family

The SML sequence variations discussed previously (Müller *et al.*, 2001; Klein *et al.*, 1998) may represent an immune-evasion strategy of the parasite as discussed for AMA1 of the malaria parasite *P. falciparum* (Bai *et al.*, 2005). SML-1 and SML-2 are expressed in cyst merozoites, while SML-3 has not been isolated to date. The core of the PAN_AP module is highly conserved within the family (Fig. 7), but 31 of 51 non-conserved residues are localized in the dimerization domain. Residues Ser18, Thr48, His57, Lys59, Tyr66, Tyr68 and Asp71 of the galactose-binding site are conserved in all three family members, explaining the galactose-binding activity found for all three (Klein *et al.*, 1998). The hydrophobic residues Ala29, Ala40, Ala60, Ala85, Gly90, Ala95, Val96, Met112 and Met125 are exchanged in the SML family through nonconservative substitutions (marked in green on the surface in Fig. 2*b*). The hydrophobic cleft at the central surface of the PAN_AP module remains conserved within the SML family, but no functionality, as proposed for the hydrophobic patches in *P. falciparum* AMA1 (Bai *et al.*, 2005), has been attributed to the SMLs so far.

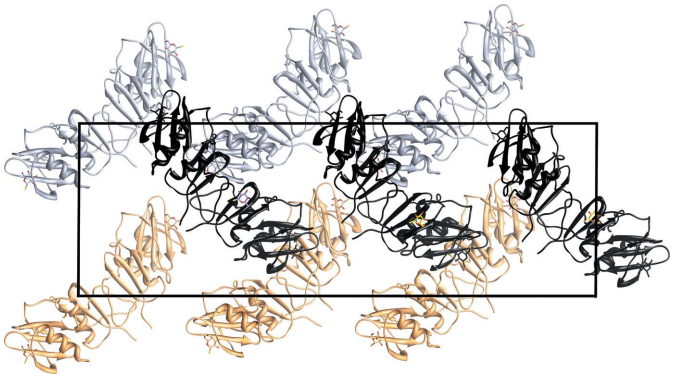


Figure 6
Crystal packing of complete SML-2 holodimers in a slice of $a \times c \times b/2$ in space group $P2_12_12_1$. The projection is along b. The view is the same as that in Fig. 2 of the recent publication by Müller *et al.* (2006).

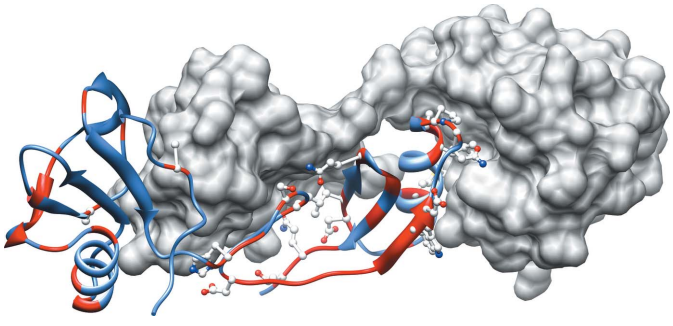


Figure 7
Monomer interaction site in dimers of the SML family. Conserved residues are coloured blue and nonconserved residues are coloured red. All residues in direct contact with the other protomer are drawn in ball-and-stick representation.

4. Conclusions

PAN modules are structural motifs that are found in about 1000 proteins and are widely spread over higher eukaryotes as plasminogen/hepatocyte growth-factor-related proteins, plasma prekallikrein/coagulation-factor-XI-type molecules, nematode and protozoan proteins (Tordai *et al.*, 1999). Generally, their structure is highly conserved despite low sequence identity, *e.g.* the r.m.s.d. values to the SML-2 PAN_AP domain are about 2.0 Å for sequence identities far below 30% (coagulation factor XI, PDB entry 2f83 chain *A*, 69 aligned residues, r.m.s.d. 2.3 Å, 29% identity, Papagrigoriou *et al.*, 2006; human hepatocyte growth factor, PDB entry 1gp9, 69 aligned residues, r.m.s.d. 2.2 Å, 14% identity, Watanabe *et al.*, 2002; leech *Haementeria officinali* anti-platelet protein, PDB entry 1i8n chain *A*, 69 aligned residues, r.m.s.d. 2.9 Å, 10% identity, Huizinga *et al.*, 2001) (obtained from the *DaliLite* server at EBI; http://www.ebi.ac.uk/Tools/dalilite). An increasing number of parasitic and more specialized microneme protein (MICs) structures have been analyzed by the Parasite Genome Projects (*e.g.* proteins from *E. tenella*, *P. falciparum* and *T. gondii*; see http://www.tigr.org/parasiteProjects.shtml). PAN modules with an even higher structural similarity to PAN_AP of SML-2 [*e.g. P. falciparum* apical membrane antigen 1, AMA1, PDB entry 1z40, 69 aligned residues, r.m.s.d. 1.7 Å, 13% identity, Bai *et al.*, 2005; *P. vivax* AMA1, PDB entry 1w81, 69 aligned residues, r.m.s.d. 1.8 Å, 10% identity, Pizarro *et al.*, 2005; *E. tenella* microneme protein 5 precursor EtMIC5 (apple domain 9), PDB entry 1hky, 68 aligned residues, r.m.s.d. 2.2 Å, 26% identity, Brown *et al.*, 2003; *T. gondii* AMA1, PDB entry 2x2z, 66 aligned residues, 2.3 and 2.5 Å r.m.s.d. for two modules per chain *A*, 12% identity] have been described.

The apical membrane antigen 1 (AMA1) from *P. falciparum* is a leading malaria vaccine candidate. This molecule consists of a tandem repeat of two complete PAN_AP modules (Bai *et al.*, 2005). The same is found for AMA1 from *P. vivax* (Pizarro *et al.*, 2005). In contrast to hepatocyte growth factor (HGF) and SML-2, they belong to a single back-folded chain and diverse long loops extend from their PAN_AP cores (Bai *et al.*, 2005), protecting a potentially functionally essential hydrophobic patch of the molecule. This site is different from the position of the galactose-binding cavity in SML-2 and from hydrophobic patches on the SML-2 surface (see Fig. 2*b*).

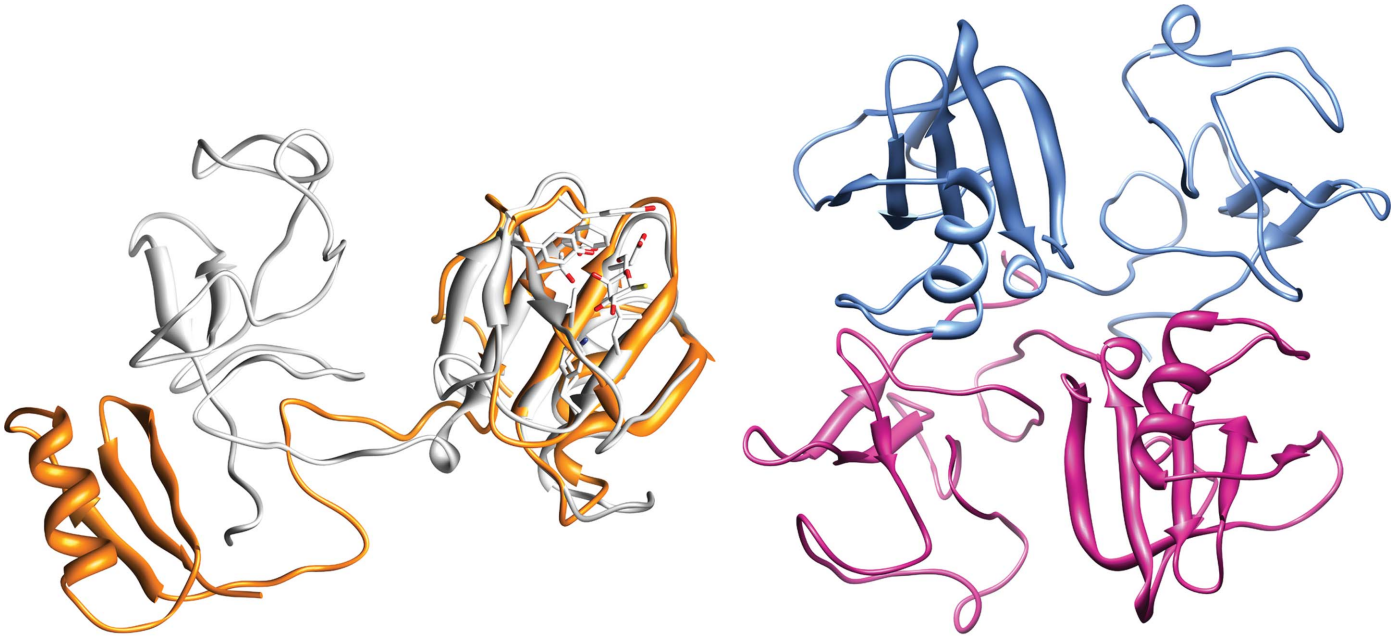
The relatively short chains of 138 residues in SML-2 form a dimer in solution with a tendency to oligomerize (Montag *et al.*, 1997). The open shape of a monomer with a large hydrophobic surface is not an energetically favourable conformation. Two different possibilities may exist to bury large

hydrophobic patches in the interior of the native protein: (i) both domains may form a separate independent module or (ii) after domain swapping the PAN_AP and PAN_1 motifs of one protomer form a more stable, rigid, elongated dimer by tight packing with the PAN_1 and PAN_AP domains of the second molecule. This domain swapping has also been discussed previously for NK1 fragments of HGF/SF in the context of growth-factor dimerization and receptor binding (Chirgadze *et al.*, 1999; Watanabe *et al.*, 2002) and HGF and SML-2 show secondary, tertiary and quaternary structure similarity (Fig. 8).

Potentially, this dimer spans a longer distance and may show enhanced activity during host-cell occupation because it presents two exposed sites for recognition and interaction with carbohydrate moieties at cell surfaces exhibiting galactose endgroups. The bridging distance may be further enlarged by oligomerization of dimers as shown above and found by Montag *et al.* (1997). Furthermore, AMA1 from *P. falciparum* and *P. vivax* (Pizarro *et al.*, 2005) and EtMIC5 PAN domains stack upon one another, also forming elongated structures that possibly project away from the recognized surface and are engaged in forming solid connections between receptors and parasite components (Carruthers & Tomley, 2008). Dimerization and oligomerization have also been discussed as strategies for lectins (Rini, 1995) to improve their specificity and enhance affinity. The dimeric structure of SML-2 may also hint at a bridging function, in which one SML-2 dimer may link the parasitic vesicle to an apposing cell.

Several protein–protein complexes and protein–ligand interactions have been structurally characterized to date for parasite micronemal proteins, *e.g. P. falciparum* AMA1 (Coley *et al.*, 2007, PDB entry 2q8a; Henderson *et al.*, 2007, PDB entry 2z8v) and *T. gondii* AMA1 (Tonkin *et al.*, 2011, PDB entry 2y8t), as well as for PAN domains in HGF (Lietha *et al.*, 2001; PDB entry 1gmn). All interaction sites include the PAN-domain helix at the opposite side to the galactose-binding cavity in SML-2. Thus, either the same region of the SML-2 PAN_AP domain of the same protomer or its counterpart on the dimer partner could be engaged in protein–cell-surface attachment. This must be regarded as speculative in the absence of detailed information about such processes.

The second known activity of PAN modules concerns their carbohydrate recognition and discrimination. For the molecules mentioned above, sequence similarity to the galactose-binding region of SML-2 does not exist, but it does to another micronemal protein, the TgMIC4 apple domain 5. These residues are highly conserved (Fig. 9), in agreement with the recently described specificity of the *T. gondii* protein for galactose, *N*-acetylgalactosamine and *N*-acetylglucosamine



(*a*) (*b*)

Figure 8
(*a*) Superposition of the PAN_AP module from SML-2 (coloured orange) onto the PAN_AP module of human HGF (PDB entry 1gp9; grey). The 1-thio-β-D-galactose is drawn in ball-and-stick representation. The r.m.s.d. value for 48 residues is 1.0 Å, with a sequence identity of 19%. (*b*) Dimer of human HGF.

```
          10        20        30        40        50        60        70
SML-2     CFAHDKNIGSR.TEQL..SVVHVASAQDCMKECQALPTCSHFTYNKNSKKCHLKAGAPEFYTYT..GDMTGPRSCE
EtMIC5    CYQN..GVSFTGGKAI..SEAKAASSQACQELCEKDAKCRFFTLAS..GKCSLFADDAALRPTKSDGAVSGNKRCI
TgMIC4    CV.HTGNIGSKA..QTIKEVKRASSLSECRARCQAEKECSHYTYNVKSGLCYPKRGKPQFYKYL..GDMTGSRTCD
                    o                   o         o o       o o     o
```

Figure 9
Structure-based sequence alignment of SML-2 PAN_AP, EtMIC5 (PDB entry 1hyk; 23% identity) and TgMIC4 apple domain 5 (46% identity). Residues coloured cyan are conserved compared with SML-2. The residues marked with circles belong to the galactose-binding site.

(Brown *et al.*, 2003). The only difference between the binding sites of the two lectins is the exchange of His57 to tyrosine, which is accompanied by a loss of specificity towards galactose and a change in geometry that allows *N*-acetylglucosamine to enter the cleft and possibly form a hydrogen bond from O^4 to tyrosine O^{η}. To date, no three-dimensional structure of TgMIC4 is known and site prediction may promote further experiments. Also, for both molecules the cognate galactose-terminated carbohydrate ligand remains to be detected.

We express our gratitude to T. Montag, H. Klein, N. Zyto and B. Löschner from the Paul-Ehrlich-Institut, Bundesinstitut für Impfstoffe und biomedizinische Arzneimittel, Langen, Germany, who provided the original lectin SML-2. The modified 1-thio-β-D-galactose was a gift from U. Pfüller, University of Witten-Herdecke, Germany. This work was supported by the Fonds der Chemischen Industrie.

References

Bai, T., Becker, M., Gupta, A., Strike, P., Murphy, V. J., Anders, R. F. & Batchelor, A. H. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 12736–12741.
Baum, J. & Cowman, A. F. (2011). *Science*, **333**, 410–411.
Brecht, S., Carruthers, V. B., Ferguson, D. J., Giddings, O. K., Wang, G., Jakle, U., Harper, J. M., Sibley, L. D. & Soldati, D. (2001). *J. Biol. Chem.* **276**, 4119–4127.
Broger, C. (2011). *XSAE* v.1.6.2. F. Hoffmann–La Roche AG, Basel, Switzerland.
Brown, P. J., Mulvey, D., Potts, J. R., Tomley, F. M. & Campbell, I. D. (2003). *J. Struct. Funct. Genomics*, **4**, 227–234.
Carruthers, V. B. & Tomley, F. M. (2008). *Subcell. Biochem.* **47**, 33–45.
Chirgadze, D. Y., Hepple, J. P., Zhou, H., Byrd, R. A., Blundell, T. L. & Gherardi, E. (1999). *Nature Struct. Biol.* **6**, 72–79.
Coley, A. M., Gupta, A., Murphy, V. J., Bai, T., Kim, H., Foley, M., Anders, R. F. & Batchelor, A. H. (2007). *PLoS Pathog.* **3**, 1308–1319.
Crawford, J., Tonkin, M. L., Grujic, O. & Boulanger, M. J. (2010). *J. Biol. Chem.* **285**, 15644–15652.
Dall'Antonia, F. & Schneider, T. (2006). *J. Appl. Cryst.* **39**, 618–619.
DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.
Elgavish, S. & Shaanan, B. (1998). *J. Mol. Biol.* **277**, 917–932.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Entzeroth, R., Kerckhoff, H. & König, A. (1992). *Eur. J. Cell Biol.* **59**, 405–413.
Eschenbacher, K. H., Klein, H., Sommer, I., Meyer, H. E., Entzeroth, R., Mehlhorn, H. & Rüger, W. (1993). *Mol. Biochem. Parasitol.* **62**, 27–36.
Henderson, K. A., Streltsov, V. A., Coley, A. M., Dolezal, O., Hudson, P. J., Batchelor, A. H., Gupta, A., Bai, T., Murphy, V. J., Anders, R. F., Foley, M. & Nuttall, S. D. (2007). *Structure*, **15**, 1452–1466.
Huizinga, E. G., Schouten, A., Connolly, T. M., Kroon, J., Sixma, J. J. & Gros, P. (2001). *Acta Cryst.* D**57**, 1071–1078.
Kabsch, W. (2010*a*). *Acta Cryst.* D**66**, 125–132.
Kabsch, W. (2010*b*). *Acta Cryst.* D**66**, 133–144.
Klein, H., Löschner, B., Zyto, N., Pörtner, M. & Montag, T. (1998). *Glycoconj. J.* **15**, 147–153.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
Leonidas, D. D., Vatzaki, E. H., Vorum, H., Celis, J. E., Madsen, P. & Acharya, K. R. (1998). *Biochemistry*, **37**, 13930–13940.
Lietha, D., Chirgadze, D. Y., Mulloy, B., Blundell, T. L. & Gherardi, E. (2001). *EMBO J.* **20**, 5543–5555.
Lo Conte, L., Chothia, C. & Janin, J. (1999). *J. Mol. Biol.* **285**, 2177–2198.
Lunin, V. Y., Lunina, N. L., Podjarny, A., Bockmayr, A. & Urzhumtsev, A. (2002). *Z. Kristallogr.* **217**, 668–685.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Mehlhorn, H. & Heydorn, A. O. (1978). *Adv. Parasitol.* **16**, 43–91.
Montag, T., Bornhak, M., Loeschner, B., Klein, H., Otto, A., Zyto, N. & Entzeroth, R. (1997). *Eur. J. Cell Biol.* **74**, Suppl. 46, 21.
Müller, J. J., Lunina, N. L., Urzhumtsev, A., Weckert, E., Heinemann, U. & Lunin, V. Y. (2006). *Acta Cryst.* D**62**, 533–540.
Müller, J. J., Müller, E.-C., Montag, T., Zyto, N., Löschner, B., Klein, H., Heineman, U. & Otto, A. (2001). *Acta Cryst.* D**57**, 1042–1045.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Olsen, L. R., Dessen, A., Gupta, D., Sabesan, S., Sacchettini, J. C. & Brewer, C. F. (1997). *Biochemistry*, **36**, 15073–15080.
Papagrigoriou, E., McEwan, P. A., Walsh, P. N. & Emsley, J. (2006). *Nature Struct. Mol. Biol.* **13**, 557–558.
Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.
Pizarro, J. C., Vulliez-Le Normand, B., Chesne-Seck, M. L., Collins, C. R., Withers-Martinez, C., Hackett, F., Blackman, M. J., Faber, B. W., Remarque, E. J., Kocken, C. H., Thomas, A. W. & Bentley, G. A. (2005). *Science*, **308**, 408–411.
Prabu, M. M., Sankaranarayanan, R., Puri, K. D., Sharma, V., Surolia, A., Vijayan, M. & Suguna, K. (1998). *J. Mol. Biol.* **276**, 787–796.
Ravishankar, R., Surolia, A., Vijayan, M., Lim, S. & Kishi, Y. (1998). *J. Am. Chem. Soc.* **120**, 11297–11303,
Rini, J. M. (1995). *Annu. Rev. Biophys. Biochem. Struct.* **24**, 551–577.
Rommel, M. (1979). *Parasitol. Res.* **58**, 187–188.
Sharon, N. & Lis, H. (1989). *Science*, **246**, 227–234.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Sujatha, M. S., Sasidhar, Y. U. & Balaji, P. V. (2005). *Biochemistry*, **44**, 8554–8562.
Tonkin, M. L., Roques, M., Lamarque, M. H., Pugnière, M., Douguet, D., Crawford, J., Lebrun, M. & Boulanger, M. J. (2011). *Science*, **333**, 463–467.
Tordai, H., Bányai, L. & Patthy, L. (1999). *FEBS Lett.* **461**, 63–67.
Watanabe, K., Chirgadze, D. Y., Lietha, D., de Jonge, H., Blundell, T. L. & Gherardi, E. (2002). *J. Mol. Biol.* **319**, 283–288.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
Winn, M. D. et al. (2011). *Acta Cryst.* D**67**, 235–242.