Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The sequence and structure of snake gourd (*Trichosanthes anguina*) seed lectin, a three-chain nontoxic homologue of type II RIPs

Alok Sharma,[a]‡ Gottfried Pohlentz,[b]‡ Kishore Babu Bobbili,[c] A. Arockia Jeyaprakash,[a] Thyageshwar Chandran,[a] Michael Mormann,[b] Musti J. Swamy[c] and M. Vijayan[a]*

[a]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, Karnataka, India, [b]Institute for Hygiene, University of Münster, Robert-Koch-Strasse 41, Münster, Germany, and [c]School of Chemistry, University of Hyderabad, Hyderabad 500 046, Andhra Pradesh, India

‡ These authors contributed equally to this study.

Correspondence e-mail: mv@mbu.iisc.ernet.in

Received 3 December 2012
Accepted 11 April 2013

PDB Reference: snake gourd seed lectin, 4hr6

The sequence and structure of snake gourd seed lectin (SGSL), a nontoxic homologue of type II ribosome-inactivating proteins (RIPs), have been determined by mass spectrometry and X-ray crystallography, respectively. As in type II RIPs, the molecule consists of a lectin chain made up of two β-trefoil domains. The catalytic chain, which is connected through a disulfide bridge to the lectin chain in type II RIPs, is cleaved into two in SGSL. However, the integrity of the three-dimensional structure of the catalytic component of the molecule is preserved. This is the first time that a three-chain RIP or RIP homologue has been observed. A thorough examination of the sequence and structure of the protein and of its interactions with the bound methyl-α-galactose indicate that the nontoxicity of SGSL results from a combination of changes in the catalytic and the carbohydrate-binding sites. Detailed analyses of the sequences of type II RIPs of known structure and their homologues with unknown structure provide valuable insights into the evolution of this class of proteins. They also indicate some variability in carbohydrate-binding sites, which appears to contribute to the different levels of toxicity exhibited by lectins from various sources.

1. Introduction

Protein–glycan interactions are the hallmark of intercellular as well as intracellular communication. Lectins, which are multivalent carbohydrate-binding proteins with stringent selectivity and affinity towards different glycans, glycan linkages and glycoconjugates, are the key players in such communication. They are involved in various biological processes and have also been used in a wide variety of applications for decades (Chandrasekaran *et al.*, 1991; Amara *et al.*, 1992; Lis & Sharon, 1998; Drickamer, 1999; Vijayan & Chandra, 1999; Feizi, 2000; Loris, 2002; Sharon, 2007; Gringhuis *et al.*, 2009). The ubiquitous distribution of lectins across all five kingdoms of life and viruses asserts the importance of lectins *in situ* (http://www.cermav.cnrs.fr/lectines). Plant lectins are among the most thoroughly studied classes of lectins. They have been used extensively in the laboratory of one of us (MV) as a model system to explore the structural variety of proteins, to understand how ligand specificity is generated and to elucidate the structural basis of multivalency (Banerjee *et al.*, 1994; Sankaranarayanan *et al.*, 1996; Chandra *et al.*, 1999; Pratap *et al.*, 2002; Jeyaprakash *et al.*, 2004, 2005; Ramachandraiah *et al.*, 2003; Singh *et al.*, 2005; Natchiar *et al.*, 2006; Kulkarni *et al.*, 2007; Sharma & Vijayan, 2011; Sharma *et al.*, 2007, 2009). Interestingly, of the five structural classes of plant lectins, in three cases the lectin subunits harbour approximate threefold symmetry and each subunit is believed to have evolved through successive gene duplication, fusion and

doi:10.1107/S0907444913010020

divergent evolution (Sankaranarayanan *et al.*, 1996; Ramachandraiah & Chandra, 2000; Singh *et al.*, 2005; Robertus & Ready, 1984; Sharma *et al.*, 2007); β-trefoil lectins are one of these cases. β-Trefoil lectins are also a class of plant lectins whose endogenous role is understood to a substantial extent. Ricin, which is a member of this class, was the first protein to be identified as a lectin in 1888. All β-trefoil plant lectin domains with known structure, except for amaranthin and *Sambucus nigra* agglutinin-II (SNA-II), belong to the type II ribosome-inactivating proteins (RIPs). RIPs are known to terminate protein synthesis in an irreversible and catalytic way by damaging ribosomes. In addition to the lectin chain, type II RIPs contain a catalytic chain with N-glycosidase activity. Such RIPs of known structure include ricin (Montfort *et al.*, 1987), abrin (Tahirov *et al.*, 1995), European mistletoe lectin (Eu-ML; Sweeney *et al.*, 1998; Jiménez *et al.*, 2005; Niwa *et al.*, 2003), *Ricinus communis* agglutinin (RCA; Sweeney *et al.*, 1997; Hegde & Podder, 1998; Sharma *et al.*, 1998), ebulin (Pascal *et al.*, 2001), Himalayan mistletoe lectin (Hm-RIP; Mishra *et al.*, 2004), *Abrus precatorius* agglutinin (APA; Bagaria *et al.*, 2006) and cinnamomin (Azzi *et al.*, 2009). The lectin domain helps the entry of the molecule into the cell. The lectin and the catalytic chains are connected by a disulfide bond. Ricin cleaves the N-glycosidic bond of a single adenine (A4324 in rat liver rRNA or A2660 in prokaryotic rRNA) adjacent to the universally conserved α-sarcin loop of rRNA present on the larger subunit of ribosome (Stirpe *et al.*, 1988). RIPs are also known to remove adenine residues from nucleotides other than rRNA (Barbieri *et al.*, 2001; Stirpe, 2004). Type I RIPs only contain the catalytic chain and are less toxic than type II RIPs. The higher toxicity of type II RIPs is attributed to the lectin component, which probably facilitates entry of the protein into the cell by anchoring onto glycolipids/glycoproteins of the cell membrane. Proteins that are structurally similar to type II RIPs but devoid of toxicity also exist. Ebulin, the toxicity of which is believed to be impaired on account of a defective oligosaccharide-binding site in the lectin chain, is such a nontoxic type II RIP-like protein (Pascal *et al.*, 2001). The structure of another nontoxic RIP-like lectin, that from *Trichosanthes kirilowii* (TKL-1), has been determined at 2.7 Å resolution (Li, Chai *et al.*, 2001). As the sequence of the protein is not available, the information provided by the structure is somewhat limited. The loss of toxicity in this protein is believed to arise from impairment of catalytic activity.

Snake gourd seed lectin (SGSL) isolated from *T. anguina* is a glycosylated galactose-specific nontoxic lectin similar to type II RIPs with a molecular weight of ~62 kDa (Komath *et al.*, 1996, 1998, 2001). Besides carbohydrate ligands containing galactose, the lectin also recognizes noncarbohydrate, predominantly hydrophobic ligands such as porphyrins (Komath *et al.*, 2000, 2006). Preliminary X-ray studies (Manoj *et al.*, 2001) confirmed that the structure of SGSL is similar to those of type II RIPs, although the amino-acid sequence of the protein was not then available. Here, we present the sequence and structure determination of the protein. The new coordinates constitute the most accurate description to date of the structure of a type II RIP-like protein with a partially or wholly impaired catalytic activity, including disulfide links and N-glycosylation. The X-ray results, sequence analysis and modelling studies provide, among other things, valuable insights into the structural basis for the loss of toxicity of the lectin and the evolutionary history of β-trefoil lectins.

2. Materials and methods

2.1. Materials

Snake gourd seeds were purchased from local seed vendors in Hyderabad, India. The guar gum, methyl-α-D-galactose, β-mercaptoethanol, PEG 400 and ammonium sulfate used for crystallization were purchased from Sigma (St Louis, Missouri, USA). Epichlorohydrin, sodium hydroxide and sodium phosphate (both monobasic and dibasic) were purchased from Merck (Mumbai, India). Thermolysin was purchased from Sigma–Aldrich Chemie GmbH (Taufkirchen, Germany). Trifluoroacetic acid was from Roth (Karlsruhe, Germany) and acetic acid was from AppliChem GmbH (Darmstadt, Germany). Trypsin, chymotrypsin and endoproteinase Glu-C were obtained from Roche Diagnostics GmbH (Mannheim, Germany). Formic acid, methanol, 1,4-dithiothreitol and ammonium bicarbonate were purchased from Fluka (Buchs, Switzerland), whereas acetonitrile and water were obtained from Merck (Darmstadt, Germany). All solvents used were of HPLC-grade purity. Micro Bio-Spin P6 Columns were purchased from Bio-Rad (Munich, Germany), ZipTip Pipette Tips C18 were obtained from Millipore (Billerica, USA) and ZIC-HILIC ProteaTips (10–200 µl) and Sample Prep Kit were obtained from Dichrom GmbH (Marl, Germany).

2.2. Purification and N-terminal sequencing

SGSL was purified using affinity chromatography on cross-linked guar gum as the key step, as described previously (Komath *et al.*, 1996, 2001). The purity of the protein was examined on a 12% SDS–PAGE gel (Laemmli, 1970). The catalytic and lectin chains were resolved on an SDS–PAGE gel under denaturing conditions and the protein bands were blotted on a PVDF membrane. The bands were stained with Coomassie Brilliant Blue R-250 and were excised from the membrane for sequencing. The N-terminal sequence was obtained using the sequencing facilities available at the Department of Biochemistry, Indian Institute of Science, Bangalore, India and the Central Food Technological Research Institute, Mysore, India.

2.3. Sample preparation for mass spectrometry

The SGSL solution (250 pmol μl^{-1}, 150 mM NaCl in 20 mM phosphate buffer pH 7.3) was rebuffered to 25 mM ammonium bicarbonate using Micro Bio-Spin P6 columns according to the manufacturer's instructions. The column was equilibrated with 25 mM ammonium bicarbonate buffer and the sample was diluted tenfold with 25 mM ammonium bicarbonate and applied onto the column. The protein was eluted with 25 mM ammonium bicarbonate buffer and

Table 1
X-ray data-collection and structure refinement.

Values in parentheses are for the last resolution shell.

Space group	$P6_122$
Unit-cell parameters	
a (Å)	102.08
c (Å)	271.64
No. of molecules in asymmetric unit	1
Resolution (Å)	30.0–2.25 (2.33–2.25)
No. of observations	340671
No. of unique reflections	40805
Completeness (%)	93.5 (95.7)
$\langle I/\sigma(I)\rangle$	16.7 (3.6)
$R_{\rm merge}$† (%)	14.4 (75.1)
Multiplicity	8.3
No. of protein atoms	3959
No. of sugar atoms	26
No. of solvent atoms	178
R factor‡ (%)	18.7
$R_{\rm free}$‡ (%)	22.9
R.m.s. deviations from ideal values	
Bond lengths (Å)	0.009
Bond angles (°)	1.2
Dihedral angles (°)	6
Ramachandran plot, residues in (%)	
Core region	87.9
Additionally allowed region	12.1
Generously allowed region	0.0
Disallowed region	0.0

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$. ‡ $R = \sum_{hkl} ||F_{\rm obs}| - |F_{\rm calc}|| / \sum_{hkl}|F_{\rm obs}|$. $R_{\rm free}$ is calculated in the same way but for a subset of reflections that were not used in the refinement.

aliquots of the eluate were directly used for proteolysis. For mass-spectrometric (MS) analysis of intact SGSL, the above procedure was performed using 10 mM ammonium acetate pH 7.3 and the eluate was directly subjected to mass-spectrometric analysis.

2.4. In-solution digestion

For thermal unfolding, the samples were dissolved in the appropriate buffer and heated to 368 K for 5 min prior to digestion. Proteolysis was performed in 25 mM ammonium bicarbonate buffer pH 8.6 by incubating SGSL (200 pmol) with thermolysin overnight at 338 K at a substrate:enzyme ratio of 12:1, with trypsin or chymotrypsin overnight at 310 K at a substrate:enzyme ratio of 25:1 or with endoproteinase Glu-C overnight at 298 K at a substrate:enzyme ratio of 50:1. Subsequently, the digest mixtures were dried, redissolved twice in water and dried again.

2.5. Zwitterionic hydrophilic interaction liquid chromatography solid-phase extraction (ZIC-HILIC SPE) of N-glycopeptides

For the enrichment of N-glycopeptides from proteolytic digests, ZIC-HILIC ProteaTips were used as previously described (Neue *et al.*, 2011). The tips were equilibrated with acetonitrile/H_2O + formic acid (80/20 + 2). The digested peptide mixture obtained in the above step was dissolved in acetonitrile/H_2O + formic acid (80/20 + 2) and loaded onto the tips. After washing with the same solvent mixture, the N-glycopeptides were eluted with H_2O/formic acid (98/2) and the solvent was evaporated *in vacuo*.

2.6. Mass spectrometry

The products of in-solution and in-gel proteolysis were analysed by nanoESI Q-ToF MS and MS/MS, and (glyco)-peptide structures were deduced from fragment-ion spectra derived from collision-induced dissociation (CID). NanoESI MS experiments were carried out using a quadrupole time-of-flight (Q-ToF) mass spectrometer (Micromass, Manchester, England) equipped with a Z-spray source in positive-ion mode. The source temperature was kept at 353 K and the desolvation-gas (N_2) flow rate was kept at 75 l h^{-1}. The capillary and cone voltages were adjusted to 1.1 kV and 30 V, respectively. For low-energy CID experiments, the (glyco)-peptide precursor ions were selected in the quadrupole analyser and fragmented in the collision cell using a collision-gas (Ar) pressure of 3.0×10^{-3} Pa and collision energies of 30–60 eV ($E_{\rm lab}$).

2.7. Protein crystallization, data collection and structure solution

Protein crystals were obtained by the hanging-drop method by equilibrating a 10 µl drop of 40 mg ml^{-1} protein in the presence of 10 mM methyl-α-D-galactose, 5 mM β-mercaptoethanol and 1 µl 30% PEG 400 against a reservoir solution consisting of 1 ml of 80% saturated ammonium sulfate in the same buffer (Manoj *et al.*, 2001). Diffraction data were collected at 293 K on the XRD1 beamline at a wavelength of 1.0 Å at the Elettra synchrotron light source, Trieste, Italy using a MAR Research MAR345 imaging plate. The data were processed using *DENZO* and *SCALEPACK* from the *HKL* suite of programs (Otwinowski & Minor, 1997). Intensities were converted to structure factors using *TRUNCATE* from *CCP*4 (French & Wilson, 1978). Solvent content was estimated using the method of Matthews (1968). The structure was solved by the molecular-replacement method using *MOLREP* (Vagin & Teplyakov, 2010) with abrin-a (PDB entry 1abr; Tahirov *et al.*, 1995) as the search model. The solution had a correlation coefficient (CC) of 0.413 and an R factor of 0.586 and led to satisfactory crystal packing with a few short contacts caused by protruding N-terminal residues in the symmetry-related molecules, which were removed during model building and structure refinement. The structure was refined using *REFMAC* from *CCP*4 (Murshudov *et al.*, 2011). Model building was carried out using *Coot* v.0.6 (Emsley & Cowtan, 2004). Addition of sugar ligands and water O atoms using difference maps commenced when R and $R_{\rm free}$ were 25% and 30%, respectively. Water O atoms were located based on peaks with heights greater than 1.0σ in $2F_{\rm o} - F_{\rm c}$ and 3.0σ in $F_{\rm o} - F_{\rm c}$ maps. OMIT maps were used in the course of refinement to check the model. Both the positive and negative densities were carefully examined in such maps. The refined model was validated using *PROCHECK* (Laskowski *et al.*, 1993) and the *MolProbity* web server (Chen *et al.*, 2010). Data-collection and structure-solution statistics are summarized in Table 1.

2.8. Sequence alignment and homology modelling

Sequence homologues of SGSL chain *B* were searched for by iterative *PSI-BLAST* alignment with an *E*-value cutoff of 0.0005 using the nonredundant database available at the NCBI (Altschul *et al.*, 1997; Schäffer *et al.*, 2001). Alignments with less than a 90-residue overlap length were not considered for further study as they cannot form even a single trefoil fold. Sequences thus obtained were made nonredundant using the *CD-HIT* web server (Huang *et al.*, 2010; Li & Godzik, 2006; Li, Jaroszewski *et al.*, 2001; Li *et al.*, 2002). The smaller sequences with more than 90% identity were removed in all-*versus*-all pairwise alignment. Lectin and RIP domains in each sequence were searched for using the *CDD* tool (Marchler-Bauer *et al.*, 2002) available at the NCBI. All pairwise and multiple sequence alignments were carried out using *ALIGN* (Cohen, 1997) and *ClustalW* (Thompson *et al.*, 2002), respectively (both available at http://www.ebi.ac.uk). Sequences with at least one carbohydrate-binding site motif were accepted after analyzing the pairwise alignment of all of the sequences with the SGSL lectin chain sequence. Phylogenetic analyses were carried out using the maximum-parsimony method with 10 000 steps of bootstrapping as implemented in the *MEGA* 5 (Tamura *et al.*, 2007) suite of programs.

Homology modelling was carried out using the program *Modeller* 9v9 (Eswar *et al.*, 2006, 2007). Six two-chain type II RIPs with known structures from different sources were taken together as templates for modelling (PDB entries 2aai, 1abr, 1hwm, 1m2t, 1yf8 and 2vlc; Rutenber *et al.*, 1991; Tahirov *et al.*, 1995; Pascal *et al.*, 2001; Krauspenhaar *et al.*, 2002; Mishra *et al.*, 2005; Azzi *et al.*, 2009). Alignments of the two chains were carried out independently. For each sequence, five different models were generated. Although the five models did not differ much from each other, the model with the best Ramachandran statistics was selected and manually examined using *Coot* v.0.6 (Emsley & Cowtan, 2004). Local restraints were relieved and side-chain and backbone atoms were adjusted wherever necessary. The model was then energy-minimized by *CNS* employing a distance-dependent dielectric constant. The final models were validated using *PROCHECK* (Laskowski *et al.*, 1993).

2.9. Structure analyses

Structure alignments were carried out using *ALIGN* (Cohen, 1997). All pictorial illustrations were generated using *PyMOL* (http://www.pymol.org).

3. Results and discussion

3.1. N-terminal sequencing and MS analysis of SGSL

Type II RIPs can generally be represented as A–S–S–B, where the catalytic chain A is connected by a disulfide bridge to the lectin chain B. Mass-spectrometric analysis of the protein under native conditions in a solution of 10 m*M* ammonium acetate pH 7.3 yielded a molecular weight of 59 028.6 Da (Supplementary Fig. S1[1]), a value comparable to the molecular weight of other well known type II RIPs. However, N-terminal sequencing of the catalytic chain repeatedly yielded a peptide stretch SNRFY which aligned with a stretch of about 46 residues downstream of the N-terminus of the sequences of most other type II RIPs. This discrepancy was resolved by detailed mass-spectrometric analysis of the protein under denaturing conditions in a manner wholly consistent with the electron-density map.

The mass spectra of the protein in a solution containing 50% acetonitrile and 2% formic acid indicate the presence of a smaller component with a molecular mass of 5209.69 Da and a larger component of mass 53 797.9 Da (Supplementary Fig. S2). This observation, along with the results of *de novo* sequencing and the interpretation of the electron-density map (see below), suggested that SGSL comprises two noncovalently linked components A_α and A_β–S–S–B which dissociate under denaturing conditions. A_α is detected in charge states +5, +6 and +7, while gas-phase ions formed from the larger component appear in charge states +28 to +41 (Supplementary Fig. S2). Thus, the catalytic component (A) of SGSL consists of two chains (A_α and A_β), while it is a single chain in all other well studied type II RIPs. A_α shows some heterogeneity. Besides the full-length species (amino acids 1–46; $\text{aa}^{1–46}$), $\text{aa}^{2–46}$ (loss of the N-terminal asparagine) and $\text{aa}_{2–46}$–H_2O could also be detected. The primary structures of the A_α species were corroborated by N- and C-terminal sequencing based on the CID spectra obtained from the most abundant sixfold-charged precursor ions at *m*/*z* 869.27 (calculated 869.2891), 850.26 (calculated 850.2819) and 847.25 (calculated 847.2802), respectively.

In order to explore whether the A_α and A_β polypeptides are generated owing to proteolysis during the purification procedure, SGSL was purified in the absence as well as in the presence of a cocktail of protease inhibitors. In the latter purification experiment the buffer contained Halt Protease Inhibitor Single Use cocktail (Thermo Scientific, catalogue No. 78430, lot No. MH160898) diluted 100 times as per the instructions provided by the supplier. The cocktail contained 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF), aprotinin, bestatin, E64, leupeptin, pepstatin A and EDTA. The SDS–PAGE of the purified protein obtained from both of the experiments yielded identical patterns involving two major bands corresponding to ~32 and 23 kDa (Supplementary Fig. S3). The same results were obtained when the experiments were performed with Protease Inhibitor Cocktail from Sigma (catalogue No. P9599), indicating that protease inhibitors have no effect on the purification process.

3.2. *De novo* sequencing

Although the fragmentation of precursor ions derived from the in-gel tryptic digests of the protein bands had already

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: BE5225). Services for accessing this material are described at the back of the journal.

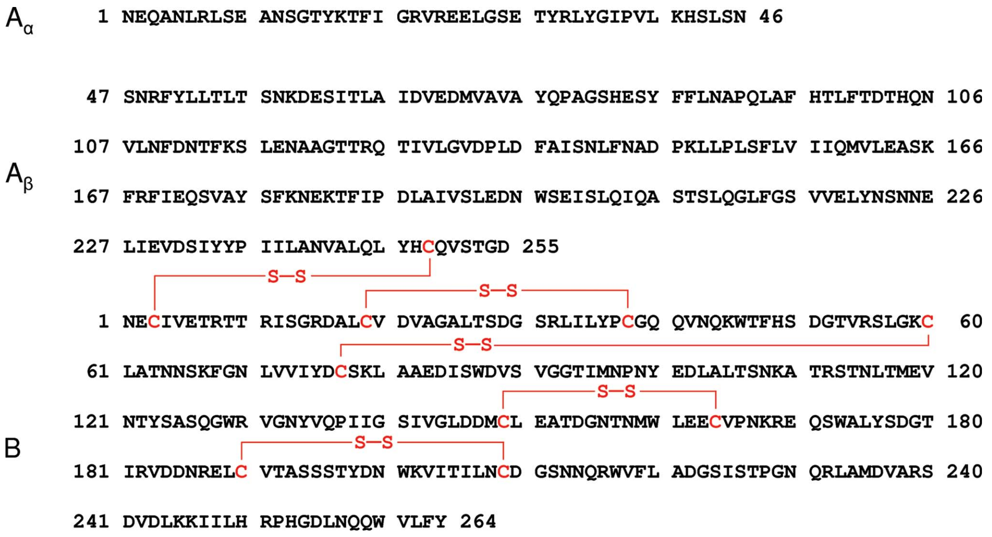

Aα
1 NEQANLRLSE ANSGTYKTFI GRVREELGSE TYRLYGIPVL KHSLSN 46
Aβ
47 SNRFYLLTLT SNKDESITLA IDVEDMVAVA YQPAGSHESY FFLNAPQLAF HTLFTDTHQN 106
107 VLNFDNTFKS LENAAGTTRQ TIVLGVDPLD FAISNLFNAD PKLLPLSFLV IIQMVLEASK 166
167 FRFIEQSVAY SFKNEKTFIP DLAIVSLEDN WSEISLQIQA STSLQGLFGS VVELYNSNNE 226
227 LIEVDSIYYP IILANVALQL YHCQVSTGD 255
S—S
S—S
B
1 NECIVETRTT RISGRDALCV DVAGALTSDG SRLILYPCGQ QVNQKWTFHS DGTVRSLGKC 60
S—S
61 LATNNSKFGN LVVIYDCSKL AAEDISWDVS VGGTIMNPNY EDLALTSNKA TRSTNLTMEV 120
S—S
121 NTYSASQGWR VGNYVQPIIG SIVGLDDMCL EATDGNTNMW LEECVPNKRE QSWALYSDGT 180
S—S
181 IRVDDNRELC VTASSSTYDN WKVITILNCD GSNNQRWVFL ADGSISTPGN QRLAMDVARS 240
241 DVDLKKIILH RPHGDLNQQW VLFY 264

Figure 1
Primary structure of the cleaved A chain (A_α and A_β) and the B chain (B) along with the disulfide links deduced from fragment-ion spectra derived from proteolytic peptides.

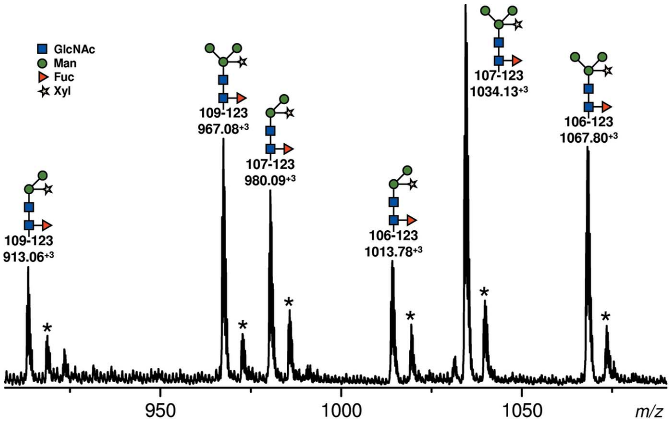


Figure 2
NanoESI Q-ToF mass spectrum of ZIC-HILIC-separated N-glycopeptides derived from a chymotryptic digest of SGSL. The satellite signals labelled by an asterisk (Δm = 16.0 Da) represent the same glycopeptide ions but with oxidized Met118. Monosaccharide symbols are as recommended by the Consortium for Functional Glycomics (http://www.functionalglycomics.org/fg/).

yielded some sequence information, roughly 300 additional CID experiments were performed on precursor peptide ions obtained from in-solution proteolysis by the use of trypsin, chymotrypsin, endoproteinase Glu-C and thermolysin, and the amino-acid sequences were deduced from the resulting fragment-ion spectra. Each amino acid was determined from at least two independent proteolytic peptides and the finally obtained sequences for the three chains are shown in Fig. 1. The residues in A_α and A_β are numbered contiguously for easy comparison with type II RIPs of known structure. Since the amino acids isoleucine (I) and leucine (L) are isobaric, they cannot be discriminated by means of low-energy CID experiments and were assigned on the basis of electron density. Similarly, the masses of glutamine and lysine differ by only 36 mDa and are not unambiguously distinguishable with the Q-ToF instrument used. Thus, in general, amino acids exhibiting an increment mass of 128 are assigned as Gln unless the presence of Lys was proven by tryptic cleavage. Guidance from the electron-density map was also taken wherever necessary.

Intrachain and interchain disulfide bonds were determined using low-energy CID experiments, as described recently (Mormann *et al.*, 2008). Collision-induced dissociation of disulfide-bond-containing peptides obtained from proteolytic peptides resulted in characteristic asymmetric cleavages of the disulfide bridge prior to fragmentation of the amide bonds. The observed fragmentation pattern enabled us to identify an intermolecular S—S bridge joining the A_β and B subunits (Cys249 in subunit A_β to Cys3 in subunit B; *cf.* Supplementary Fig. S4). Furthermore, we identified intramolecular disulfide bonds between Cys19–Cys38, Cys60–Cys77, Cys149–Cys164 and Cys190–Cys209 within the B chain (*cf.* Fig. 1).

The catalytic chain of SGSL shares an average sequence identity of around 33% with those of other type II RIPs of known structure, while the corresponding value for the lectin chain is 38%. The catalytic chain of abrin exhibits the highest relatedness to that of SGSL, with a sequence identity of 37%, whereas the lectin chain of SGSL is closest to that of the nontoxic ebulin, with a sequence identity of 42%.

3.3. Structure determination of the N-linked glycan

The chemical structure of the N-linked glycan of SGSL was determined by proteolysis preceding ZIC-HILIC SPE and MS analysis of the purified glycopeptides (Neue *et al.*, 2011). Fig. 2 shows the MS^1 spectrum obtained from a ZIC-HILIC extract of a chymotryptic digest. Despite heterogeneity in the peptide backbone (owing to miscleavages), SGSL exhibits typical paucimannosidic glycan moieties carrying a 'bisecting' xylose, a core fucose and two or three mannose residues. The glycopeptide structures were confirmed by CID experiments. As an example, the CID spectrum of the triply charged glycopeptide precursor ions at m/z = 1067.78 and the corresponding fragmentation scheme are depicted in Supplementary Fig. S5. Complete sets of B-type and Y-type fragment ions derived from the glycan as well as a few Y-type ions originating from the C-terminus of the peptide moiety allowed the oligosaccharide structure [$GlcNAc_2(Fuc)Man(Xyl)Man_2$] to be deduced as well as the corresponding peptide backbone (amino acids 106–123) of the B chain. This plant-typical paucimannosidic glycosylation occurs at Asn115 of the lectin chain.

3.4. Overall three-dimensional structure

SGSL adopts essentially the same fold as observed for other type II RIPs (Fig. 3). Chain A can be divided into three domains. As in other similar proteins (Montfort *et al.*, 1987), domain I spanning residues 4–109 (the first three residues are not observed in the electron-density map) is made up of a six-stranded mixed β-sheet with a 3_{10}-helix between an α-helix and a small two-stranded β-sheet. There is a break in the chain. A_α forms its N-terminal stretch. The last two residues of A_α (45 and 46) and the first residue of A_β (47) are also not observed in the electron-density map. The 45–47 segment connects two strands in a β-sheet. Despite the break in the chain, the small chain forms an integral part of the three-dimensional structure, such as, for example, in pea lectin and jacalin. Protein-sequence comparison with other type II RIPs of known structure appears to suggest that A_α and A_β are from the same gene product. The biological significance of the cleavage is still unclear. Domain II, made up of residues 110–192, consists of five helices. The third and last domain is the smallest and it consists of one 11-residue α-helix and a two-stranded antiparallel β-sheet. Chain B or the lectin component is made up of two β-trefoils. Each is believed to have evolved through successive gene duplication and fusion of a 40-residue galactose-binding stretch (Robertus & Ready, 1984). Residues 1–136 fold into one trefoil, whereas residues 137–264 form the second trefoil. The electron density unambiguously corroborates the presence of the highly conserved disulfide bond between A_β and B as well as the four disulfide bonds in chain B, which had already been demonstrated by the MS analysis.

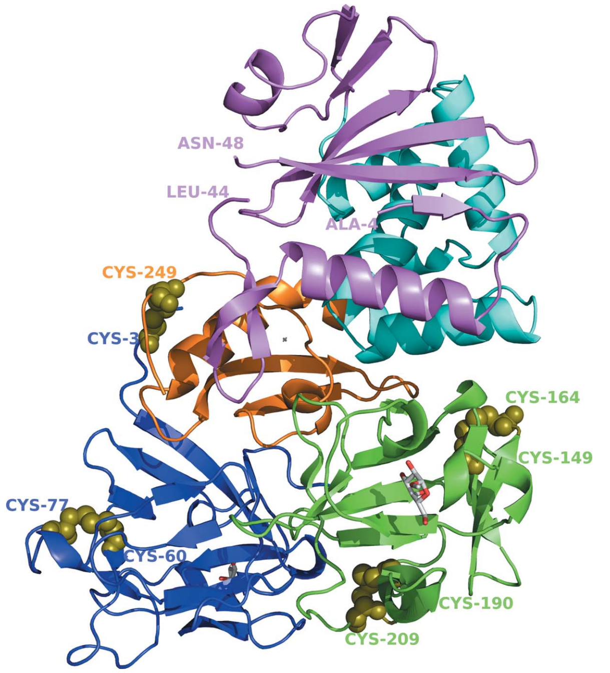


Figure 3
Three-dimensional structure of SGSL. The three domains of the catalytic chain are shown in different colours (domain I, pink; domain II, cyan; domain III, orange). Domain I of the lectin chain is shown in blue and domain II is shown in green. Cysteine residues involved in disulfide bonds are labelled and shown as spheres. Sugar atoms are shown in stick representation.

3.5. Lectin–sugar interactions

The B chain of type II RIPs, which is involved in carbohydrate binding, consists of two β-trefoil domains. Each domain has three subdomains, designated α, β and γ, that are related to each other by an approximate threefold axis. However, only one subdomain in each domain carries a carbohydrate-binding site. The site in domain 1 is on the α subdomain and is designated 1α. Similarly, that in domain 2 is designated 2γ (Fig. 4). Each carbohydrate-binding site is characterized by the presence of an aromatic ring which stacks against the *b* face of the galactose ring at the primary site, as indeed happens in all galactose-binding plant lectins. A three-residue kink occurs on the other side of the ring. The side chains of an Asp and an Asn and a main-chain amido N atom interact through hydrogen bonds with galactose in all cases. In ebulin, a defect in the 2γ site alone presumably leads to nontoxicity (Pascal *et al.*, 2001). Therefore, the 2γ site is believed to be more important than the 1α site. In the protein from Himalayan mistletoe only, a third carbohydrate-binding site, 1β, has been identified (Mikeska *et al.*, 2005).

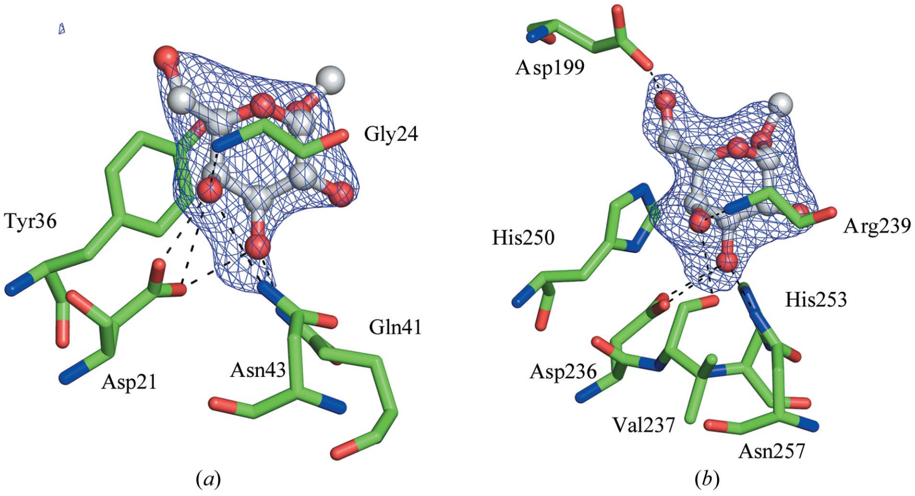
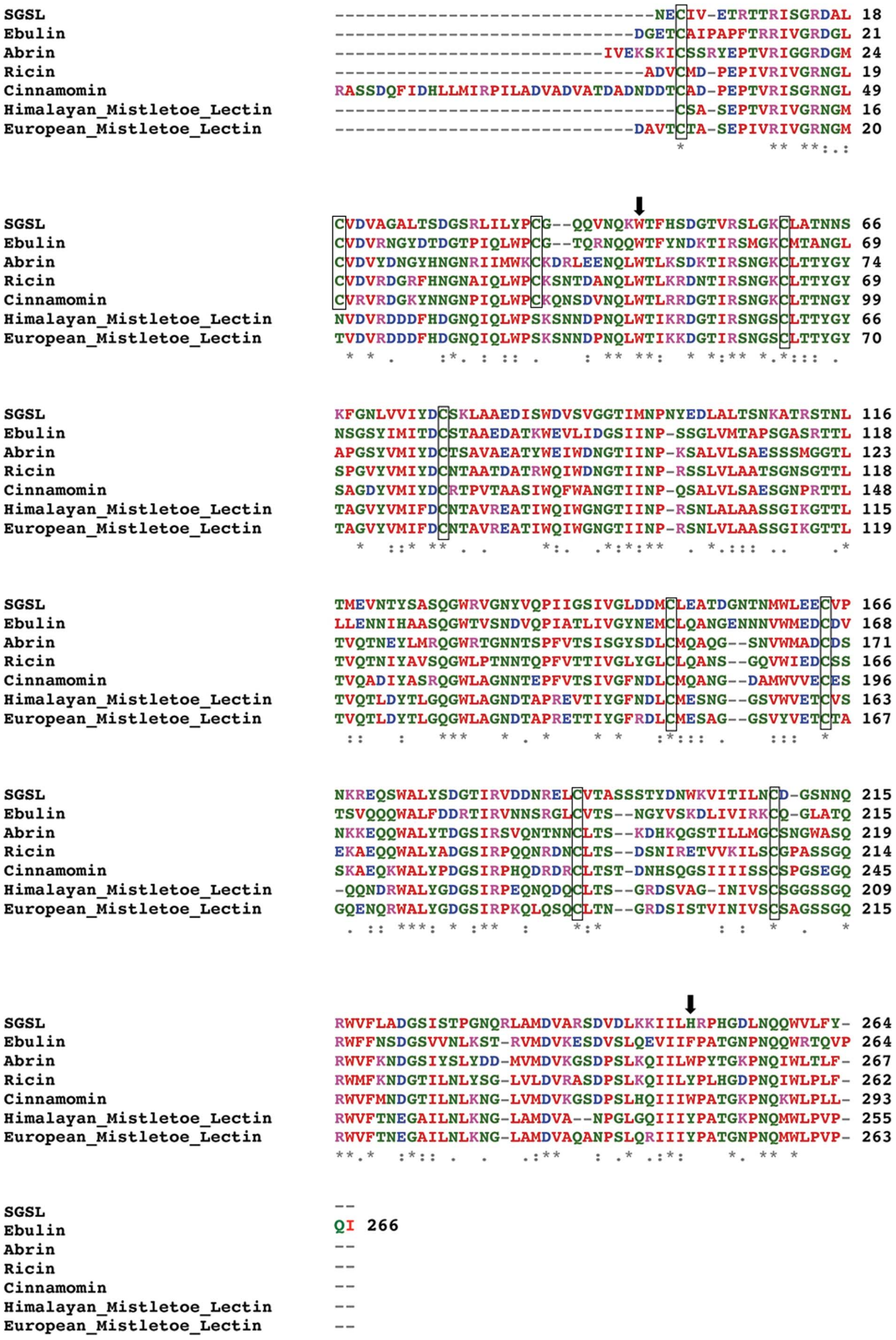


Figure 4
Simulated-annealing OMIT maps of carbohydrate ligands at (*a*) the 1α site and (*b*) the 2γ site in the SGSL structure, along with lectin–carbohydrate interactions. Electron density is contoured at the 3σ level.

In the structure of SGSL, methyl-α-Gal is well defined at both the 1α and the 2γ sites (Fig. 4), as are the residues involved in carbohydrate binding. Hydrogen bonds involving an Asp, an Asn and a main-chain N atom to methyl-α-Gal exist in both the 1α and the 2γ sites (Fig. 4) as in other type II RIPs. Additional interactions involving a Gln also occur. These additional interactions have been observed previously in some instances. However, there are some crucial differences at both binding sites. The stacking residue at the 1α site is tryptophan in all other cases, whereas this position is occupied by a tyrosine in the case of SGSL. Perhaps the most important difference at this site is the substitution of an Asp/Asn residue at position 24 by a glycine in SGSL. The Asp/Asn residues in all of the other homologues are appropriately positioned to interact with the second sugar ring in the case of oligosaccharides. This interaction is likely to be absent in the case of SGSL. Also, the stacking of an aromatic residue on methyl-α-Gal does not occur in the 2γ site in SGSL. The aromatic residue at this site is replaced by a histidine (Supplementary Fig. S6). This observation is in consonance with chemical modification studies (Komath *et al.*, 1998), which suggest the involvement of a histidine residue in carbohydrate binding. Interestingly, although the carbohydrate-binding site at the 2γ position is fairly conserved, the stretch which precedes the site has an insertion of two serines in SGSL (Fig. 5). This could lead to the protrusion of Tyr198 and Asp199 in the carbohydrate-binding site, which might interfere with the binding of oligosaccharides at this position in a fashion similar to that in toxic lectins. Such variations may possibly lead to different orientations of the carbohydrate at the two carbohydrate-binding sites in toxic and nontoxic lectins (Fig. 6). Thus, it would appear that although the primary binding sites of SGSL are similar to those of the toxic homologues, some crucial changes in the length and the amino-acid composition of specific segments are likely to perturb or modulate the binding of the lectin domain to the glycan receptors on the cell surface.



```
SGSL                            -------------------------------NECIV-ETRTTRISGRDAL 18
Ebulin                          -----------------------------DGETCAIPAPFTRRIVGRDGL 21
Abrin                           --------------------------IVEKSKICSSRYEPTVRIGGRDGM 24
Ricin                           ------------------------------ADVCMD-PEPIVRIVGRNGL 19
Cinnamomin                      RASSDQFIDHLLMIRPILADVADVATDADNDDTCAD-PEPTVRISGRNGL 49
Himalayan_Mistletoe_Lectin      ---------------------------------CSA-SEPTVRIVGRNGM 16
European_Mistletoe_Lectin       -----------------------------DAVTCTA-SEPIVRIVGRNGM 20
                                                                 *        ** **:.:

                                                             ⬇
SGSL                            CVDVAGALTSDGSRLILYPCG--QQVNQKWTFHSDGTVRSLGKCLATNNS 66
Ebulin                          CVDVRNGYDTDGTPIQLWPCG--TQRNQQWTFYNDKTIRSMGKCMTANGL 69
Abrin                           CVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGKCLTTYGY 74
Ricin                           CVDVRDGRFHNGNAIQLWPCKSNTDANQLWTLKRDNTIRSNGKCLTTYGY 69
Cinnamomin                      CVRVRDGKYNNGNPIQLWPCKQNSDVNQLWTLRRDGTIRSNGKCLTTNGY 99
Himalayan_Mistletoe_Lectin      NVDVRDDDFHDGNQIQLWPSKSNNDPNQLWTIKRDGTIRSNGSCLTTYGY 66
European_Mistletoe_Lectin       TVDVRDDDFHDGNQIQLWPSKSNNDPNQLWTIKKDGTIRSNGSCLTTYGY 70
                                 * * .    :*. : :: .    : ** **:  * *:** *.*::: .

SGSL                            KFGNLVVIYDCSKLAAEDISWDVSVGGTIMNPNYEDLALTSNKATRSTNL 116
Ebulin                          NSGSYIMITDCSTAAEDATKWEVLIDGSIINP-SSGLVMTAPSGASRTTL 118
Abrin                           APGSYVMIYDCTSAVAEATYWEIWDNGTIINP-KSALVLSAESSSMGGTL 123
Ricin                           SPGVYVMIYDCNTAATDATRWQIWDNGTIINP-RSSLVLAATSGNSGTTL 118
Cinnamomin                      SAGDYVMIYDCRTPVTAASIWQFWANGTIINP-QSALVLSAESGNPRTTL 148
Himalayan_Mistletoe_Lectin      TAGVYVMIFDCNTAVREATIWQIWGNGTIINP-RSNLALAASSGIKGTTL 115
European_Mistletoe_Lectin       TAGVYVMIFDCNTAVREATIWQIWGNGTIINP-RSNLVLAASSGIKGTTL 119
                                  *  ::* ** . .     *:.  .*:*:**  . *.::: ..    .*

SGSL                            TMEVNTYSASQGWRVGNYVQPIIGSIVGLDDMCLEATDGNTNMWLEECVP 166
Ebulin                          LLENNIHAASQGWTVSNDVQPIATLIVGYNEMCLQANGENNNVWMEDCDV 168
Abrin                           TVQTNEYLMRQGWRTGNNTSPFVTSISGYSDLCMQAQG--SNVWMADCDS 171
Ricin                           TVQTNIYAVSQGWLPTNNTQPFVTTIVGLYGLCLQANS--GQVWIEDCSS 166
Cinnamomin                      TVQADIYASRQGWLAGNNTEPFVTSIVGFNDLCMQANG--DAMWVVECES 196
Himalayan_Mistletoe_Lectin      TVQTLDYTLGQGWLAGNDTAPREVTIYGFNDLCMESNG--GSVWVETCVS 163
European_Mistletoe_Lectin       TVQTLDYTLGQGWLAGNDTAPRETTIYGFRDLCMESAG--GSVYVETCTA 167
                                 ::   :   ***   * . *    * *   :*::: .    :::  *

SGSL                            NKREQSWALYSDGTIRVDDNRELCVTASSSTYDNWKVITILNCD-GSNNQ 215
Ebulin                          TSVQQQWALFDDRTIRVNNSRGLCVTS--NGYVSKDLIVIRKCQ-GLATQ 215
Abrin                           NKKEQQWALYTDGSIRSVQNTNNCLTS--KDHKQGSTILLMGCSNGWASQ 219
Ricin                           EKAEQQWALYADGSIRPQQNRDNCLTS--DSNIRETVVKILSCGPASSGQ 214
Cinnamomin                      SKAEQKWALYPDGSIRPHQDRDRCLTST-DNHSQGSIIIISSCSPGSEGQ 245
Himalayan_Mistletoe_Lectin      -QQNDRWALYGDGSIRPEQNQDQCLTS--GRDSVAG-INIVSCSGGSSGQ 209
European_Mistletoe_Lectin       GQENQRWALYGDGSIRPKQLQSQCLTN--GRDSISTVINIVSCSAGSSGQ 215
                                 . :: ***: * :**  :    *:*           : :  *  .   *

                                                                  ⬇
SGSL                            RWVFLADGSISTPGNQRLAMDVARSDVDLKKIILHRPHGDLNQQWVLFY- 264
Ebulin                          RWFFNSDGSVVNLKST-RVMDVKESDVSLQEVIIFPATGNPNQQWRTQVP 264
Abrin                           RWVFKNDGSIYSLYDD-MVMDVKGSDPSLKQIILWPYTGKPNQIWLTLF- 267
Ricin                           RWMFKNDGTILNLYSG-LVLDVRASDPSLKQIILYPLHGDPNQIWLPLF- 262
Cinnamomin                      RWVFMNDGTILNLKNG-LVMDVKGSDPSLHQIIIWPATGKPNQKWLPLL- 293
Himalayan_Mistletoe_Lectin      RWVFTNEGAILNLKNG-LAMDVA--NPGLGQIIIYPATGKPNQMWLPVP- 255
European_Mistletoe_Lectin       RWVFTNEGAILNLKNG-LAMDVAQANPSLQRIIIYPATGNPNQMWLPVP- 263
                                **.*  :*:: .  .   .:**   : .* .:*:    *. ** *

SGSL                            --
Ebulin                          QI 266
Abrin                           --
Ricin                           --
Cinnamomin                      --
Himalayan_Mistletoe_Lectin      --
European_Mistletoe_Lectin       --
```

Figure 5
Multiple sequence alignment of the lectin chain of SGSL with corresponding chains of type II RIPs of known three-dimensional structure.

Lectins are known to use different strategies to generate specificity for sugars and glycoconjugates (Vijayan & Chandra, 1999). These include water bridges (Ravishankar *et al.*, 1997), loop length (Jeyaprakash *et al.*, 2004), post-translational modification (Sankaranarayanan *et al.*, 1996) and oligomerization (Chandra *et al.*, 1999; Wright & Hester, 1996). Type II RIPs appear to present a case in which subtle variations in essentially the same binding site are used to introduce variability in sugar binding.

3.6. Nucleotide-binding site

All of the toxic type II RIPs of known structure, including ebulin, possess a buried and substantially conserved nucleotide-binding/processing site made up of two tyrosines, one arginine and one glutamate. In European mistletoe lectin (PDB entry 1m2t), for instance, Tyr76 and Tyr115 appear to be involved in orienting the nucleotide in the proper position for enzymatic activity by sandwiching the adenine base between them. The conserved Glu165 and Arg168 are believed to be involved in transition-state stabilization (Krauspenhaar *et al.*, 2002). In an attempt to understand the possible structural basis of the nontoxicity of SGSL, the adenine from mistletoe lectin was docked into SGSL by superposing the mistletoe lectin–adenine complex on SGSL. Fig. 7 shows a comparison of the location of adenine in the mistletoe lectin and SGSL. Except for a couple of changes in the side chains, the surroundings of adenine are the same in the two proteins. In fact, as shown in Fig. 7, water molecules are found in SGSL at locations corresponding to the two N atoms of the adenine attached to the mistletoe structure. The major difference at the adenine-binding site is the replacement of Tyr76 in mistletoe lectin by an aliphatic residue (Val73) in SGSL (Supplementary Fig. S7). Tyrosine at this position has been shown to be crucial for toxicity in the case of ricin by site-directed mutagenesis experiments (Lord *et al.*, 1994). In another difference, Tyr115 in mistletoe lectin is replaced by Phe114. Both Tyr and Phe can stack on the adenine base. Furthermore, interaction with adenine at this position involves a hydrogen bond between a main-chain carbonyl O atom and N1 of the adenine base. Therefore, this Tyr-to-Phe substitution is unlikely to affect binding. The substitution of Arg168 by a lysine residue is also likely to play a role in the hampered catalytic activity of SGSL.

3.7. Internal symmetry in sequence and structure: a possible clue to evolutionary history

As mentioned earlier, it has been suggested that the lectin chain of type II RIPs evolved through gene duplication and fusion (Robertus & Ready, 1984). Each domain in the chain is believed to have originated through successive gene duplication, fusion and divergent evolution of a primitive sugar-binding stretch of nearly 40 residues in length as in β-prism fold lectins. A further duplication and fusion presumably resulted in the modern lectin chain. Among proteins of known structure and known sequence, the sequence identity between pairs of domains varies between 17 and 28%. The sequence identity among the subdomains in each domain is still lower. Certainly, the relatedness of sequences among subdomains is not as pronounced as is observed in the case of β-prism II lectins or β-prism I lectins from monocots or algae (Sharma *et al.*, 2007). When a phylogenetic tree was constructed by treating the α, β and γ domains of type II RIPs of known structure as individual sequences, equivalent subdomains were clustered together, irrespective of their source (Supplementary Fig. S8). Interestingly, three classes of plant lectins which are thought to be the product of successive gene duplication, fusion and divergent evolution, namely, β-prism I fold lectins, β-prism II fold lectins and β-trefoil fold lectins, behave differently from each other. β-Prism II fold lectins, in which the sequence similarities among the three sheets are highest, show a high degree of intermixing in a phylogenetic evaluation. Such intermixing has previously been observed in β-prism I fold lectins (Sharma *et al.*, 2007) from monocots and algae. The clustering pattern can be interpreted as suggesting that in the case of type II RIPs the duplication of the lectin domain is a much more recent event than the triplication of the subdomains. The divergent evolution of the two-domain chain then continued.

3.8. Sequence homologues and evolutionary implications

With the X-ray analysis of SGSL, the three-dimensional structures of nine type II RIPs with an uncleaved or cleaved catalytic chain and one lectin chain are now available. Six of them, namely ricin, abrin-a, *Abrus* agglutinin, cinnamomin and those from European and Himalayan mistletoe, are cytotoxic, while the remainder are not. While the lack of toxicity of ebulin has been attributed to a defect in sugar binding, the loss of toxicity of TKL-1 has been suggested to arise from a defect in substrate binding at the A chain. TKL-1 and SGSL originate from the same family. Our work suggests that the loss of toxicity in SGSL results from a combination of changes in the active site in the catalytic chain and alterations in the carbohydrate-binding sites. However, the variations in structure and function are well within the common framework of the type II RIP architecture. Thus, it appeared worthwhile to explore the situation in other homologous proteins of unknown structure but of known sequence. This has been

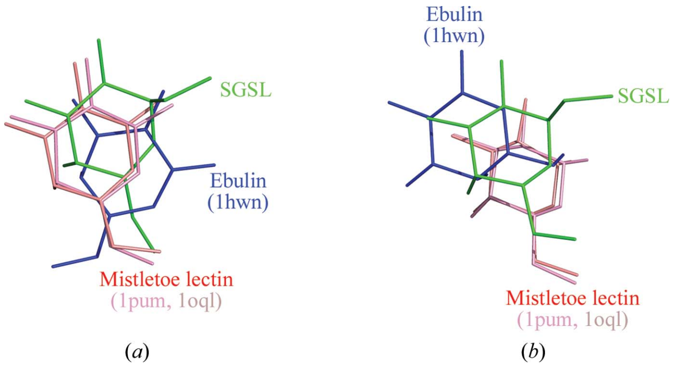


Figure 6
Orientation of galactose in the sugar complexes of type II RIPs at (*a*) the 1α site and (*b*) the 2γ site when the lectin chains are superimposed. Carbohydrates from the toxic mistletoe lectin (PDB entries 1oql and 1pum) are shown in pink, those from the nontoxic ebulin (PDB entry 1hwn) are in blue and those from SGSL are shown as green sticks.

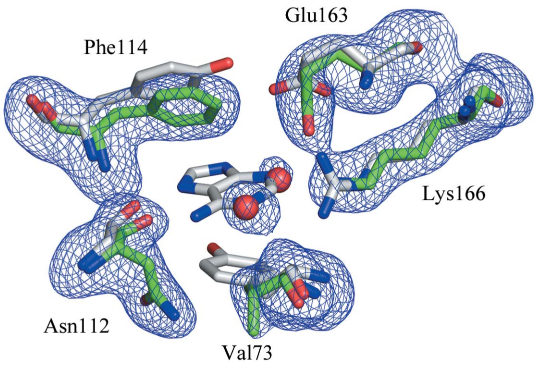


Figure 7
Structural superposition of the nucleotide-binding site residues of SGSL (green), along with the simulated-annealing OMIT map at the 3σ level, and European mistletoe lectin (PDB entry 1m2t; white). The adenine molecule observed in mistletoe lectin is shown in green sticks. The water molecules observed in SGSL are shown as red spheres.

performed through a comparative analysis of sequences and homology modelling.

A global search of amino-acid sequences using the sequence of the lectin subunit of SGSL, employing the criteria mentioned in §2, resulted in the identification of 160 proteins with at least one carbohydrate-binding site. The β-trefoil fold is known to exhibit substantial functional diversity (Murzin *et al.*, 1995). Therefore, the 160 sequences were further searched for the presence of the catalytic domain using the *CDD* web server available at the NCBI. This resulted in the identification of 30 proteins containing the lectin as well as the catalytic chains.

A phylogenetic tree based on the 160 lectin homologues obtained from a search using the sequence of the lectin chain of mistletoe type II RIP was constructed. Those belonging to RIPs cluster in one branch. The other branches belong to sequences of glycosyl hydrolases, metalloproteases, endotoxins *etc.* The type II RIP branch of the phylogenetic tree referred to above is shown in Fig. 8. Interestingly, all of the proteins in the branch belong to plants. Thus, unlike β-prism fold lectins (Sharma *et al.*, 2007), type II RIPs appear to occur only in plants.

The RIPs shown in Fig. 8 belong to 11 taxonomic families. Of these, including SGSL, crystal structures of proteins with known sequence representing six families exist. There are five plant families containing type II RIPs with known sequences but no crystal structure. Type II RIPs from *Camellia sinensis* (ADF45510), *Iris hollandica* (AF256085), *Ximenia americana* (CAJ38823), *Polygonatum multiflorum* (AF213984) and *Adenia volkensii* (CAD61022) were chosen as representatives of the five families for homology modelling. Models were constructed in an identical manner for all five, employing the techniques outlined in §2. The catalytic subunits of all five models superpose well on those of toxic type II RIPs of known structure and sequence. The sequence and structure of the binding sites of the five models and the six crystal structures are very similar. Therefore, the subunits are likely to be catalytically active to nearly the same extent. However, the carbohydrate-binding sites in the models exhibit some differences.

3.9. Variability in carbohydrate-binding sites

Carbohydrates bound to type II RIPs have been shown to exhibit different orientations in toxic and nontoxic proteins even when there are only minor differences in the binding-site residues. With the exception of Himalayan mistletoe lectin (PDB entry 1yf8; Mishra *et al.*, 2004), lactose occupies the carbohydrate-binding site in a nearly identical fashion in all toxic type II RIPs. The lactose molecules from complexes with ricin, Himalayan mistletoe lectin and ebulin were docked on the carbohydrate-binding sites of all of these models and examined for the most feasible orientation. Local restrained refinement of side chains was carried out using *Coot* v.0.6 (Emsley & Cowtan, 2004). Carbohydrate positions were also manually adjusted wherever required.

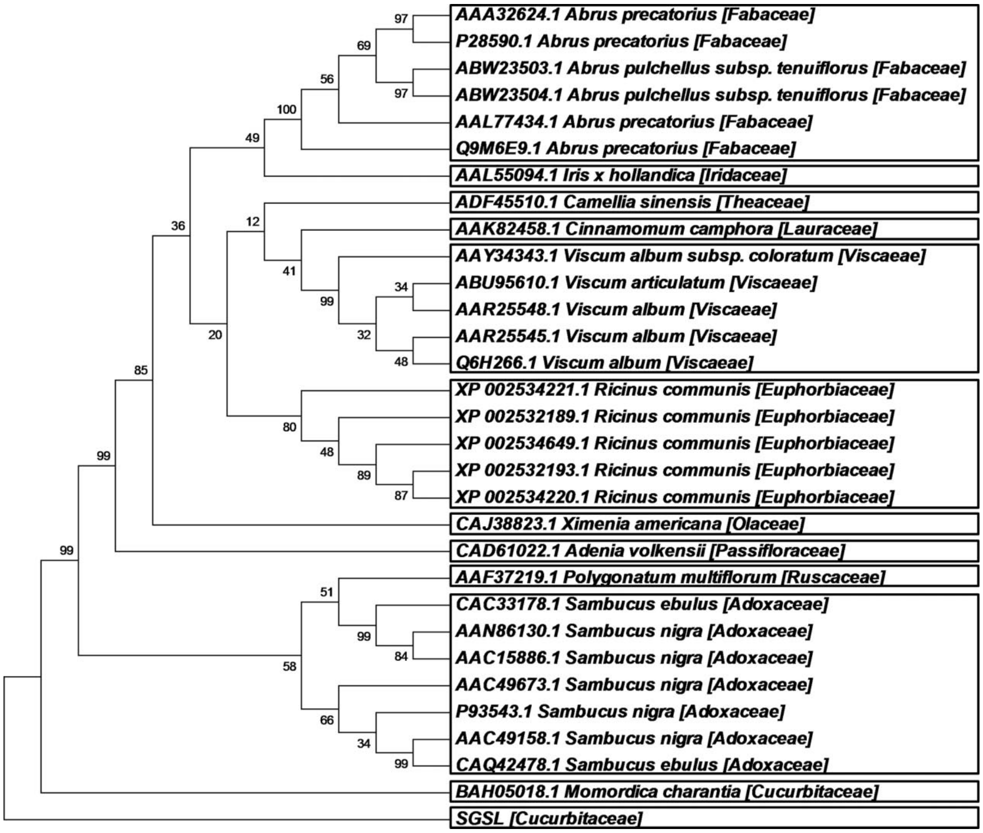

97
69
56
97
100
49
AAA32624.1 Abrus precatorius [Fabaceae]
P28590.1 Abrus precatorius [Fabaceae]
ABW23503.1 Abrus pulchellus subsp. tenuiflorus [Fabaceae]
ABW23504.1 Abrus pulchellus subsp. tenuiflorus [Fabaceae]
AAL77434.1 Abrus precatorius [Fabaceae]
Q9M6E9.1 Abrus precatorius [Fabaceae]
AAL55094.1 Iris x hollandica [Iridaceae]
ADF45510.1 Camellia sinensis [Theaceae]
AAK82458.1 Cinnamomum camphora [Lauraceae]
AAY34343.1 Viscum album subsp. coloratum [Viscaeae]
ABU95610.1 Viscum articulatum [Viscaeae]
AAR25548.1 Viscum album [Viscaeae]
AAR25545.1 Viscum album [Viscaeae]
Q6H266.1 Viscum album [Viscaeae]
36
12
41
34
99
32
48
20
85
80
48
89
87
XP 002534221.1 Ricinus communis [Euphorbiaceae]
XP 002532189.1 Ricinus communis [Euphorbiaceae]
XP 002534649.1 Ricinus communis [Euphorbiaceae]
XP 002532193.1 Ricinus communis [Euphorbiaceae]
XP 002534220.1 Ricinus communis [Euphorbiaceae]
CAJ38823.1 Ximenia americana [Olaceae]
CAD61022.1 Adenia volkensii [Passifloraceae]
99
99
AAF37219.1 Polygonatum multiflorum [Ruscaceae]
CAC33178.1 Sambucus ebulus [Adoxaceae]
AAN86130.1 Sambucus nigra [Adoxaceae]
AAC15886.1 Sambucus nigra [Adoxaceae]
AAC49673.1 Sambucus nigra [Adoxaceae]
P93543.1 Sambucus nigra [Adoxaceae]
AAC49158.1 Sambucus nigra [Adoxaceae]
CAQ42478.1 Sambucus ebulus [Adoxaceae]
51
99
84
58
66
34
99
BAH05018.1 Momordica charantia [Cucurbitaceae]
SGSL [Cucurbitaceae]

Figure 8

Phylogenetic tree constructed from the sequence alignment of lectin domains of type II RIP homologues. Clusters belonging to sequences from the same genera are separately boxed. Family names are included in each branch.

The carbohydrate-binding sites of the model ADF45510 from *C. sinensis* (a dicot plant) are similar to those of toxic type II RIPs, but the four other models exhibited some differences. The model of CAD61022 from *A. volkensii*, a dicot plant, indicates that the mode of carbohydrate binding at the 1α site is likely to be similar to that in ricin, while the 2γ site could prefer a carbohydrate orientation similar to that in Himalayan mistletoe lectin. The protein from *X. americana*, a dicot, with sequence CAJ38823 is likely to have a weaker 1α site as the modelled carbohydrate seems to have lost three hydrogen bonds in comparison to the situation in ricin. The two monocot sequences studied here, corresponding to sequences AF213984 and AF256085 from the plants *P. multiflorum* and *I. hollandica*, respectively, yield interesting observations. The AF213984 model suggests that the mode of carbohydrate binding at both sites is most likely to be similar to that in ebulin and hence the protein would probably behave in a nontoxic manner *in situ*. In the model of AF256085, although the most preferred carbohydrate orientation is similar to that in ricin, the replacement of the stacking residue Trp37 by serine and the Gln35Thr substitution, which results in the loss of one hydrogen bond, could together lead to weak or no binding at the 1α site. Thus, homology models of relevant proteins with known sequences but unknown structures also suggest variation in carbohydrate-binding sites with the same overall structure of type II RIPs.

In another modelling study, the possibility of carbohydrate-binding sites at other equivalent positions, such as 1β, 1γ, 2α and 2β, were examined. Interestingly, in the homology model of sequence AF256085 the residues and their locations at the 1β site have very high similarity to those in Himalayan mistletoe lectin and ricin (Lee *et al.*, 1994; Frankel *et al.*, 1996; Steeves *et al.*, 1999; Mikeska *et al.*, 2005). Thus, it seems that the protein could also have a functional carbohydrate-binding site at the 1β position. It is important to note here that although the 1α site of this model seems to be either very weak or nonfunctional, the protein is toxic. The 1β site along with the 2γ site in this protein presumably compensates for the loss of binding at the 1α site.

4. Conclusions

SGSL is the first three-chain RIP or RIP homologue to be reported. Although its catalytic chain is cleaved in two, the integrity of the three-dimensional structure of the protein is maintained. Most of the type II RIPs of known structure are toxic. The impaired toxicity of ebulin has been attributed to a ‘defective’ carbohydrate-binding site. Detailed crystallographic, sequencing and modelling studies indicate that the loss of toxicity in SGSL is caused by the cumulative effect of changes in the nucleotide and carbohydrate-binding sites. A careful study of the sequences of lectin chains of known structure suggests that the fusion of the two domains was preceded by divergent evolution of the domain resulting from the fusion of the three subdomains. A comprehensive analysis of the sequences of homologous proteins shed considerable light on the evolution of this class of lectins. The carbohydrate-binding sites exhibit some variability within the framework of the overall common structure of type II RIPs. This variability appears to contribute to different levels of observed or suggested toxicity of proteins from different sources.

The protein sequence has been deposited in the UniProt database with accession number B3EWX5 and the coordinates of SGSL–methyl-α-galactose have been deposited in the PDB as entry 4hr6.

Financial assistance to MV from the Department of Science and Technology and to MJS from the Department of Biotechnology (DBT) is acknowledged. MM and MJS gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG; IRTG-MCGS, GRK 1549) and UGC. Computations were performed at the Bioinformatics Centre and the Interactive Graphics facility, supported by the DBT, Government of India. MV is a DAE Homi Bhabha Professor. TC and KBB are Senior Research Fellows of the CSIR (India). The authors gratefully acknowledge Dr Lalitha Gowda, CFTRI, Mysore, India for her generous help in the amino-terminal sequencing of the protein and thank Professor H. S. Savithri and Mr Kunduri Govind, Department of Biochemistry, IISc, India for their valuable suggestions during this study.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

Amara, J. F., Cheng, S. H. & Smith, A. E. (1992). *Trends Cell Biol.* **2**, 145–149.

Arockia Jeyaprakash, A., Jayashree, G., Mahanta, S. K., Swaminathan, C. P., Sekar, K., Surolia, A. & Vijayan, M. (2005). *J. Mol. Biol.* **347**, 181–188.

Azzi, A., Wang, T., Zhu, D.-W., Zou, Y.-S., Liu, W.-Y. & Lin, S.-X. (2009). *Proteins*, **74**, 250–255.

Bagaria, A., Surendranath, K., Ramagopal, U. A., Ramakumar, S. & Karande, A. A. (2006). *J. Biol. Chem.* **281**, 34465–34474.

Banerjee, R., Mande, S. C., Ganesh, V., Das, K., Dhanaraj, V., Mahanta, S. K., Suguna, K., Surolia, A. & Vijayan, M. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 227–231.

Barbieri, L., Valbonesi, P., Bondioli, M., Alvarez, M. L., Dal Monte, P., Landini, M. P. & Stirpe, F. (2001). *FEBS Lett.* **505**, 196–197.

Chandra, N. R., Ramachandraiah, G., Bachhawat, K., Dam, T. K., Surolia, A. & Vijayan, M. (1999). *J. Mol. Biol.* **285**, 1157–1168.

Chandrasekaran, S., Dean, J. W. III, Giniger, M. S. & Tanzer, M. L. (1991). *J. Cell. Biochem.* **46**, 115–124.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.

Cohen, G. H. (1997). *J. Appl. Cryst.* **30**, 1160–1161.

Drickamer, K. (1999). *Curr. Opin. Struct. Biol.* **9**, 585–590.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U. & Sali, A. (2006). *Curr. Protoc. Bioinformatics*, Unit 5.6. doi:10.1002/0471250953.bi0506s15.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U. & Sali, A. (2007). *Curr. Protoc. Protein. Sci.*, Unit 2.9. doi:10.1002/0471140864.ps0209s50.

Feizi, T. (2000). *Immunol. Rev.* **173**, 79–88.

Frankel, A., Tagge, E., Chandler, J., Burbage, C. & Willingham, M. (1996). *Protein Eng.* **9**, 371–379.

French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.

Gringhuis, S. I., den Dunnen, J., Litjens, M., van der Vlist, M. & Geijtenbeek, T. B. (2009). *Nature Immunol.* **10**, 1081–1088.

Hegde, R. & Podder, S. K. (1998). *Eur. J. Biochem.* **254**, 596–601.

Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. (2010). *Bioinformatics*, **26**, 680–682.

Jeyaprakash, A. A., Srivastav, A., Surolia, A. & Vijayan, M. (2004). *J. Mol. Biol.* **338**, 757–770.

Jiménez, M., Sáiz, J. L., André, S., Gabius, H.-J. & Solís, D. (2005). *Glycobiology*, **15**, 1386–1395.

Komath, S. S., Kavitha, M. & Swamy, M. J. (2006). *Org. Biomol. Chem.* **4**, 973–988.

Komath, S. S., Kenoth, R., Giribabu, L., Maiya, B. G. & Swamy, M. J. (2000). *J. Photochem. Photobiol. B*, **55**, 49–55.

Komath, S. S., Kenoth, R. & Swamy, M. J. (2001). *Eur. J. Biochem.* **268**, 111–119.

Komath, S. S., Nadimpalli, S. K. & Swamy, M. J. (1996). *Biochem. Mol. Biol. Int.* **39**, 243–252.

Komath, S. S., Nadimpalli, S. K. & Swamy, M. J. (1998). *Biochem. Mol. Biol. Int.* **44**, 107–116.

Krauspenhaar, R., Rypniewski, W., Kalkura, N., Moore, K., DeLucas, L., Stoeva, S., Mikhailov, A., Voelter, W. & Betzel, C. (2002). *Acta Cryst.* D**58**, 1704–1707.

Kulkarni, K. A., Katiyar, S., Surolia, A., Vijayan, M. & Suguna, K. (2007). *Proteins*, **68**, 762–769.

Laemmli, U. K. (1970). *Nature (London)*, **227**, 680–685.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Lee, R. T., Gabius, H. J. & Lee, Y. C. (1994). *Carbohydr. Res.* **254**, 269–276.

Li, M., Chai, J., Wang, Y., Wang, K. & Bi, R. (2001). *Protein Pept. Lett.* **8**, 81–87.

Li, W. & Godzik, A. (2006). *Bioinformatics*, **22**, 1658–1659.

Li, W., Jaroszewski, L. & Godzik, A. (2001). *Bioinformatics*, **17**, 282–283.

Li, W., Jaroszewski, L. & Godzik, A. (2002). *Bioinformatics*, **18**, 77–82.

Lis, H. & Sharon, N. (1998). *Chem. Rev.* **98**, 637–674.

Lord, J. M., Roberts, L. M. & Robertus, J. D. (1994). *FASEB J.* **8**, 201–208.

Loris, R. (2002). *Biochim. Biophys. Acta*, **1572**, 198–208.

Manoj, N., Jeyaprakash, A. A., Pratap, J. V., Komath, S. S., Kenoth, R., Swamy, M. J. & Vijayan, M. (2001). *Acta Cryst.* D**57**, 912–914.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. & Bryant, S. H. (2002). *Nucleic Acids Res.* **30**, 281–283.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Mikeska, R., Wacker, R., Arni, R., Singh, T. P., Mikhailov, A., Gabdoulkhakov, A., Voelter, W. & Betzel, C. (2005). *Acta Cryst.* F**61**, 17–25.

Mishra, V., Ethayathulla, A. S., Sharma, R. S., Yadav, S., Krauspenhaar, R., Betzel, C., Babu, C. R. & Singh, T. P. (2004). *Acta Cryst.* D**60**, 2295–2304.

Montfort, W., Villafranca, J. E., Monzingo, A. F., Ernst, S. R., Katzin, B., Rutenber, E., Xuong, N. H., Hamlin, R. & Robertus, J. D. (1987). *J. Biol. Chem.* **262**, 5398–5403.

Mormann, M., Eble, J., Schwöppe, C., Mesters, R. M., Berdel, W. E., Peter-Katalinić, J. & Pohlentz, G. (2008). *Anal. Bioanal. Chem.* **392**, 831–838.

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.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Natchiar, S. K., Srinivas, O., Mitra, N., Surolia, A., Jayaraman, N. & Vijayan, M. (2006). *Acta Cryst.* D**62**, 1413–1421.

Neue, K., Mormann, M., Peter-Katalinić, J. & Pohlentz, G. (2011). *J. Proteome Res.* **10**, 2248–2260.

Niwa, H., Tonevitsky, A. G., Agapov, I. I., Saward, S., Pfüller, U. & Palmer, R. A. (2003). *Eur. J. Biochem.* **270**, 2739–2749.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Pascal, J. M., Day, P. J., Monzingo, A. F., Ernst, S. R., Robertus, J. D., Iglesias, R., Pérez, Y., Férreras, J. M., Citores, L. & Girbés, T. (2001). *Proteins*, **43**, 319–326.

Pratap, J. V., Jeyaprakash, A. A., Rani, P. G., Sekar, K., Surolia, A. & Vijayan, M. (2002). *J. Mol. Biol.* **317**, 237–247.

Ramachandraiah, G. & Chandra, N. R. (2000). *Proteins*, **39**, 358–364.

Ramachandraiah, G., Chandra, N. R., Surolia, A. & Vijayan, M. (2003). *Glycobiology*, **13**, 765–775.

Ravishankar, R., Ravindran, M., Suguna, K., Surolia, A. & Vijayan, M. (1997). *Curr. Sci.* **72**, 855–861.

Robertus, J. D. & Ready, M. P. (1984). *J. Biol. Chem.* **259**, 13953–13956.

Rutenber, E., Katzin, B. J., Ernst, S., Collins, E. J., Mlsna, D., Ready, M. P. & Robertus, J. D. (1991). *Proteins*, **10**, 240–250.

Sankaranarayanan, R., Sekar, K., Banerjee, R., Sharma, V., Surolia, A. & Vijayan, M. (1996). *Nature Struct. Biol.* **3**, 596–603.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V. & Altschul, S. F. (2001). *Nucleic Acids Res.* **29**, 2994–3005.

Sharma, S., Bharadwaj, S., Surolia, A. & Podder, S. K. (1998). *Biochem. J.* **333**, 539–542.

Sharma, A., Chandran, D., Singh, D. D. & Vijayan, M. (2007). *J. Biosci.* **32**, 1089–1110.

Sharma, A., Sekar, K. & Vijayan, M. (2009). *Proteins*, **77**, 760–777.

Sharma, A. & Vijayan, M. (2011). *Glycobiology*, **21**, 23–33.

Sharon, N. (2007). *J. Biol. Chem.* **282**, 2753–2764.

Singh, D. D., Saikrishnan, K., Kumar, P., Surolia, A., Sekar, K. & Vijayan, M. (2005). *Glycobiology*, **15**, 1025–1032.

Steeves, R. M., Denton, M. E., Barnard, F. C., Henry, A. & Lambert, J. M. (1999). *Biochemistry*, **38**, 11677–11685.

Stirpe, F. (2004). *Toxicon*, **44**, 371–383.

Stirpe, F., Bailey, S., Miller, S. P. & Bodley, J. W. (1988). *Nucleic Acids Res.* **16**, 1349–1357.

Sweeney, E. C., Tonevitsky, A. G., Palmer, R. A., Niwa, H., Pfueller, U., Eck, J., Lentzen, H., Agapov, I. I. & Kirpichnikov, M. P. (1998). *FEBS Lett.* **431**, 367–370.

Sweeney, E. C., Tonevitsky, A. G., Temiakov, D. E., Agapov, I. I., Saward, S. & Palmer, R. A. (1997). *Proteins*, **28**, 586–589.

Tahirov, T. H., Lu, T.-H., Liaw, Y.-C., Chen, Y.-L. & Lin, J.-Y. (1995). *J. Mol. Biol.* **250**, 354–367.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). *Mol. Biol. Evol.* **24**, 1596–1599.

Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2002). *Curr. Protoc. Bioinformatics*, Unit 2.3. doi:10.1002/0471250953.bi0203s00.

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.

Vijayan, M. & Chandra, N. (1999). *Curr. Opin. Struct. Biol.* **9**, 707–714.

Wright, C. S. & Hester, G. (1996). *Structure*, **4**, 1339–1352.