Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Structure of γ-conglutin: insight into the quaternary structure of 7S basic globulins from legumes

Jaroslaw Czubinski,[a] Jakub Barciszewski,[b] Miroslaw Gilski,[b,c] Kamil Szpotkowski,[b] Janusz Debski,[d] Eleonora Lampart-Szczapa[a] and Mariusz Jaskolski[b,c]*

[a]Department of Food Biochemistry and Analysis, Poznan University of Life Sciences, Poznan, Poland, [b]Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland, [c]Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland, and [d]Mass Spectrometry Laboratory, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland

Correspondence e-mail: mariuszj@amu.edu.pl

Received 25 September 2014
Accepted 15 November 2014

PDB reference: γ-conglutin, 4pph

γ-Conglutin from lupin seeds is an unusual 7S basic globulin protein. It is capable of reducing glycaemia in mammals, but the structural basis of this activity is not known. γ-Conglutin shares a high level of structural homology with glycoside hydrolase inhibitor proteins, although it lacks any kind of inhibitory activity against plant cell-wall degradation enzymes. In addition, γ-conglutin displays a less pronounced structural similarity to pepsin-like aspartic proteases, but it is proteolytically dysfunctional. Only one structural study of a legume 7S basic globulin, that isolated from soybean, has been reported to date. The quaternary assembly of soybean 7S basic globulin (Bg7S) is arranged as a cruciform-shaped tetramer comprised of two superposed dimers. Here, the crystal structure of γ-conglutin isolated from *Lupinus angustifolius* seeds (LangC) is presented. The polypeptide chain of LangC is post-translationally cleaved into α and β subunits but retains its covalent integrity owing to a disulfide bridge. The protomers of LangC undergo an intricate quaternary assembly, resulting in a ring-like hexamer with noncrystallographic D_3 symmetry. The twofold-related dimers are similar to those in Bg7S but their assembly is different as a consequence of mutations in a β-strand that is involved in intermolecular β-sheet formation in γ-conglutin. Structural elucidation of γ-conglutin will help to explain its physiological role, especially in the evolutionary context, and will guide further research into the hypoglycaemic activity of this protein in humans, with potential consequences for novel antidiabetic therapies.

1. Introduction

The genus *Lupinus*, a member of the legume family (Fabaceae), includes over 450 species, of which only a few have been domesticated (Duranti *et al.*, 2008). The most economically important species are *L. angustifolius* (narrow-leafed lupin), *L. albus* (white lupin) and *L. luteus* (yellow lupin). The increase in the cultivation of lupin seeds is related to the exploration of new and valuable sources of protein for nutritional supplementation. The key nutritional value of lupin seeds is unquestioned, owing to the massive presence of macronutrients and micronutrients, among which proteins play a prominent role (Duranti *et al.*, 2008; Czubinski *et al.*, 2013).

The main proteins present in lupin seeds are storage proteins. They are hydrolyzed during germination and nourish the early stages of seedling growth (Duranti *et al.*, 2008; Foley *et al.*, 2011). The storage proteins in legume seeds have been divided into four main classes, namely 11S globulin, 7S globulin, 7S basic globulin and 2S sulfur-rich albumin (Duranti *et al.*, 2008; Czubinski *et al.*, 2014). In lupins, these protein

classes are referred to as α-conglutin (also known as legumin or glycinin), β-conglutin (also known as vicilin or convicilin), γ-conglutin and δ-conglutin, respectively. The role of γ-conglutin in lupin seeds is of the most interest, but has been the least studied. γ-Conglutin is located in the protein bodies of developing lupin seeds (Shewry *et al.*, 1995). However, this protein is also detected in the cytoplasmic spaces between the protein bodies (Esnault *et al.*, 1996), within the epidermal cotyledonary cells (Duranti *et al.*, 1991) and in the intercellular spaces of germinating cotyledons (Duranti *et al.*, 1994). The stability of γ-conglutin during seed germination and its resistance to proteolysis, as well as the diversity of locations where this protein is found, all suggest a nonstorage role.

Two genes encoding γ-conglutin have been identified, but only one seems to be quantitatively expressed in the developing seed (Foley *et al.*, 2011). γ-Conglutin is synthesized as a single polypeptide chain fused to an N-terminal signal peptide. Based on the amino-acid sequence, the theoretical molecular weight of *L. angustifolius* γ-conglutin (LangC) is 45.4 kDa and its pI is 7.72. A mature monomer of γ-conglutin consists of two subunits derived from a single precursor in a post-translational cleavage event. The two polypeptides are known as the large α subunit, with a molecular weight of $\sim$28 kDa, and the small β subunit, of about 17 kDa, as determined electrophoretically (Duranti *et al.*, 2008). In addition, the γ-conglutin protomer is matured by the formation of six disulfide bridges, one of which links the α and β subunits, and by N-linked glycosylation at a single site. Since the α and β subunits are covalently linked, they are also termed 'domains' to distinguish them from the (α–β) subunits forming the quaternary structure. In an interesting reversible process, the protein forms a pH-dependent association–dissociation equilibrium between the monomer and an oligomeric assembly (Capraro *et al.*, 2010). The oligomeric form changes into the monomeric form with a dimeric transition state when the pH shifts from neutral to slightly acidic.

γ-Conglutin displays several characteristic properties. One of them is the ability to bind divalent metal cations, such as Zn^{2+} and Ni^{2+}, which promote the refolding process after acidic treatment (Duranti *et al.*, 2002). Furthermore, the protein is resistant to pancreatin and trypsin proteolysis (Czubinski *et al.*, 2013, 2014). The ability to bind insulin, which appears to be strongly affected by the ionic strength and pH, is another important property of γ-conglutin (Magni *et al.*, 2004). Moreover, γ-conglutin has the ability to bind to insulin-like growth factors. Various lines of experimental evidence have revealed that γ-conglutin uptake leads to the reduction of blood glucose, which makes this protein pharmacologically similar to the hypoglycaemic drug metformin (Terruzzi *et al.*, 2011). Recent studies have demonstrated unique flavonoid-binding properties of γ-conglutin (Czubinski *et al.*, 2013, 2014).

Homologues of γ-conglutin are widely present in plants. It has been reported that γ-conglutin homologues present in wheat and carrot inhibit the activity of endoglucanases from glycoside hydrolase families GH11 and GH12 (Sansen *et al.*, 2004; Yoshizawa *et al.*, 2012). Because of this property, these homologues were named *Triticum aestivum* xylanase inhibitor I (TAXI-I) and extracellular dermal glycoprotein (EDGP). The implicated glycoside hydrolases cleave xylan and xyloglucan, which are the main components of the plant cell wall. Therefore, it has been suggested that γ-conglutin might play a role in the plant defence system. Surprisingly, however, γ-conglutin from legume plants lacks any inhibitory activity against both of the glycoside hydrolase enzymes (Scarafoni *et al.*, 2010; Yoshizawa *et al.*, 2011).

In this paper, we present the crystal structure of LangC from *L. angustifolius* seeds at 2.0 Å resolution. Until recently, only one crystal structure of a legume protein belonging to the 7S basic globulin class had been determined, which was isolated from soybean and deposited in the PDB as entry 3aup (Yoshizawa *et al.*, 2011). This soybean 7S basic globulin protein (Bg7S) is tetrameric in the crystal and in solution. In contrast, LangC forms a highly intricate hexameric assembly, which makes it significantly different from the soybean homologue. Moreover, both 7S basic globulins (from lupin and from soybean) lack any inhibitory activity against glycoside hydrolases. Therefore, the present structure of γ-conglutin provides additional clues for the understanding of the molecular-evolution changes in this family of legume proteins. The structure of γ-conglutin will also help to explain the intriguing hypoglycaemic properties of this protein.

2. Materials and methods

2.1. Protein purification, crystallization and data collection

γ-Conglutin was extracted from mature lupin seeds (*L. angustifolius* cv. Zeus) as described previously (Czubinski *et al.*, 2014). Briefly, the lupin seeds were milled and sieved to obtain a fraction below 1.6 mm. An automatic Soxhlet system (Büchi Labortechnik AG) with *n*-hexane as the solvent was used to defat the milled seeds. The globulin fraction was extracted with 20 ml 20 m*M* Tris buffer pH 7.5 containing 0.5 *M* NaCl per gram of defatted seeds. The sample was centrifuged at 20 000*g* for 30 min at 277 K. The lupin globulin extract was filtered through a 0.45 µm PVDF syringe filter (Millipore) and applied onto a desalting column filled with Sephadex G-25. Desalted fractions were subjected to a HiTrap Q HP column (GE Healthcare) equilibrated with 20 m*M* Tris buffer pH 7.5, and separation in a linear gradient from 0 to 1 *M* NaCl was carried out. Under these conditions, γ-conglutin is not adsorbed on the column medium. Fractions containing γ-conglutin were collected and loaded onto a HiTrap SP HP column (GE Healthcare) equilibrated with 20 m*M* Tris buffer pH 7.5. Bound proteins were eluted with a linear gradient of NaCl from 0 to 0.5 *M*. Fractions containing γ-conglutin were concentrated using Amicon centrifugal filters (Millipore) and the buffer was exchanged to 20 m*M* Tris pH 7.5. The protein concentration was determined using UV absorption at 280 nm and a calculated extinction coefficient of 33 640 M^{-1} cm^{-1}. The sample at 6 mg ml^{-1} was used for crystallization trials.

Screening for crystallization conditions was performed at 293 K by the sitting-drop vapour-diffusion method using

Table 1
Data-collection and refinement statistics.

Values in parentheses are for the highest resolution shell.

Data collection	
Radiation source	Beamline 14.2, BESSY
Wavelength (Å)	0.91841
Temperature (K)	100
Space group	$P3_1$
Unit-cell parameters (Å)	$a = b = 121.99$, $c = 188.49$
Resolution (Å)	43.76–2.00 (2.12–2.00)
Reflections collected/unique	803093/208256
Completeness (%)	99.7 (99.3)
Multiplicity	3.9 (3.9)
$R_{\rm merge}$† (%)	6.8 (75.5)
$R_{\rm meas}$‡ (%)	7.9 (87.7)
$CC_{1/2}$§ (%)	99.8 (55.6)
$\langle I/\sigma(I)\rangle$	13.2 (1.98)
Refinement	
Unique reflections (work + test)	208256
Test reflections	999
Matthews coefficient (Å^3 Da^{-1})	1.79
Solvent volume (%)	31.1
No. of atoms (non-H)	
Protein	18068
Glycan	108
Solvent	794
$R_{\rm work}/R_{\rm free}$ (%)	14.6/17.4
R.m.s.d. from ideal geometry	
Bond lengths (Å)	0.011
Bond angles (°)	1.4
Ramachandran statistics (%)	
Favoured	95.79
Outliers	0.77
PDB code	4pph

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of observation i of reflection hkl. ‡ $R_{\rm meas} = \sum_{hkl}\{N(hkl)/[N(hkl)-1]\}^{1/2} \times \sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$. § $CC_{1/2}$ is defined as the correlation coefficient between two random half data sets, as described by Karplus & Diederichs (2012).

Crystal Screen HT, Index HT and Grid Screen reagents from Hampton Research. The initial crystals grew in 10% PEG 6000 with 0.1 *M* HEPES pH 7.0. The conditions were manually optimized using the hanging-drop technique. Single crystals for diffraction experiments were obtained using 7% PEG 6000 with 0.1 *M* HEPES pH 7.0. For cryoprotection, the crystals were transferred to 30% ethylene glycol and then vitrified in liquid nitrogen for synchrotron-radiation data collection.

X-ray diffraction data were collected on beamline BL14.2 of the BESSY synchrotron, Berlin, Germany with a MAR225 CCD detector. The diffraction data were recorded from a single crystal using the rotation method with an oscillation of 0.5° at 100 K and were processed with *XDS* (Kabsch, 2010), as summarized in Table 1.

2.2. Structure determination and refinement

The crystal structure of LangC was solved by molecular replacement with *Phaser* (McCoy *et al.*, 2007) using as a model a single protomer of Bg7S (PDB entry 3aup, chain *A*; Yoshizawa *et al.*, 2011), with which it shares 64% sequence identity. The molecular-replacement algorithm found six copies of the probe in the asymmetric unit. The final model of LangC was manually built using *Coot* (Emsley *et al.*, 2010) and refined in *phenix.refine* (Afonine *et al.*, 2012). TLS parameters were included in the refinement with three, five, nine, four, five and eight rigid groups for molecules *A*, *B*, *C*, *D*, *E* and *F*, respectively, as suggested by the refinement program. Stereochemical restraints for N-linked glycosylation were generated in *phenix.elbow* (Moriarty *et al.*, 2009) using small-molecule targets from the Cambridge Structural Database (CSD; Allen, 2002). The final model is of high stereochemical quality, as validated by *MolProbity* (Chen *et al.*, 2010). The refinement statistics are reported in Table 1.

2.3. Molecular-weight determination in solution

Static light-scattering (SLS) and dynamic light-scattering (DLS) measurements were carried out using a Zetasizer μV instrument (Malvern Instruments) with the wavelength set to 488 nm and an angle of 90°. All experiments were performed at 293 K in a 2 µl quartz cuvette, with protein concentration in the range 1–7 mg ml^{-1}. A total of 12 scans were accumulated for each sample analyzed. A theoretical radius of gyration corresponding to the crystal structure was calculated using *CRYSOL* (Svergun *et al.*, 1995).

Analytical ultracentrifugation was carried out at 293 K with an Optima XL-I analytical ultracentrifuge (Beckman Coulter). The sedimentation-velocity experiment used LangC at 0.91 mg ml^{-1} concentration in a buffer consisting of 20 m*M* Tris pH 7.5, 250 m*M* NaCl. Absorbance scans at 280 nm were collected during sedimentation at 125 000*g*. The data were analyzed with *SEDFIT* (Schuck, 2000; Schuck *et al.*, 2002) and *SEDNTERP* (Laue *et al.*, 1992).

2.4. Protein deglycosylation

Protein deglycosylation was carried out with endoglycosidase H (EndoH; New England Biolabs, No. P0702S) according to the manufacturer's protocol. Briefly, the protein mixture was dissolved in glycoprotein denaturation buffer and denatured by heating to 373 K for 10 min, followed by addition of EndoH and incubation at 310 K for 60 min.

2.5. Mass-spectrometric analysis

Peptide mixtures were analyzed by LC-MS-MS/MS (liquid chromatography coupled to tandem mass spectrometry) using a Nano-Acquity (Waters) LC system and an Orbitrap Velos mass spectrometer (Thermo Electron Corp.). Prior to analysis, protein samples were subjected to a standard in-solution or in-gel digestion procedure, during which they were reduced with 100 m*M* DTT (30 min at 329 K), alkylated with 0.5 *M* iodoacetamide (45 min at room temperature in the dark) and digested overnight with trypsin at 310 K (sequencing grade modified trypsin; Promega, catalogue No. V5111). The peptide mixture was acidified with 0.1 *M* TFA. For identification of the *N*-glycosylation site, peptides were additionally digested with immobilized pepsin (Thermo Scientific, catalogue No. 20343) for 120 min at 310 K. Following digestion, the resulting peptides were applied onto an RP-18 precolumn (nanoACQUITY Symmetry C18, Waters) using water containing 0.1% TFA as the mobile phase and then transferred to a nano-HPLC RP-18 column (nanoACQUITY BEH C18, Waters) using an acetonitrile gradient (5–35%, 180 min) in the

presence of 0.1% formic acid with a flow rate of 250 nl min^{-1}. The column outlet was directly coupled to the ion source of the spectrometer working in the regime of data-dependent MS to MS/MS switch. A blank run to ensure a lack of cross-contamination from previous samples preceded each analysis. The raw data were processed by *Mascot Distiller* followed by *Mascot Search* (Matrix Science, London, England) against a user database with protein sequences of interest, without a defined sequence specificity of digestion. The search tolerance for precursor and product ion mass was 20 p.p.m. and 0.2 Da, respectively (other parameters: no missed trypsin cleavage sites, variable modification of methionine oxidation, *N*-acetylglucosamine and cysteine methylthio or carbamidomethyl modification). Peptides with a *Mascot* score exceeding a threshold value corresponding to a <5% false-positive rate and with a *Mascot* score above 30 were considered to be positively identified. The standard search was followed by an error-tolerant search, with the parameters set as above. Fragmentation spectra corresponding to glycosylated or N-terminal peptides were validated manually.

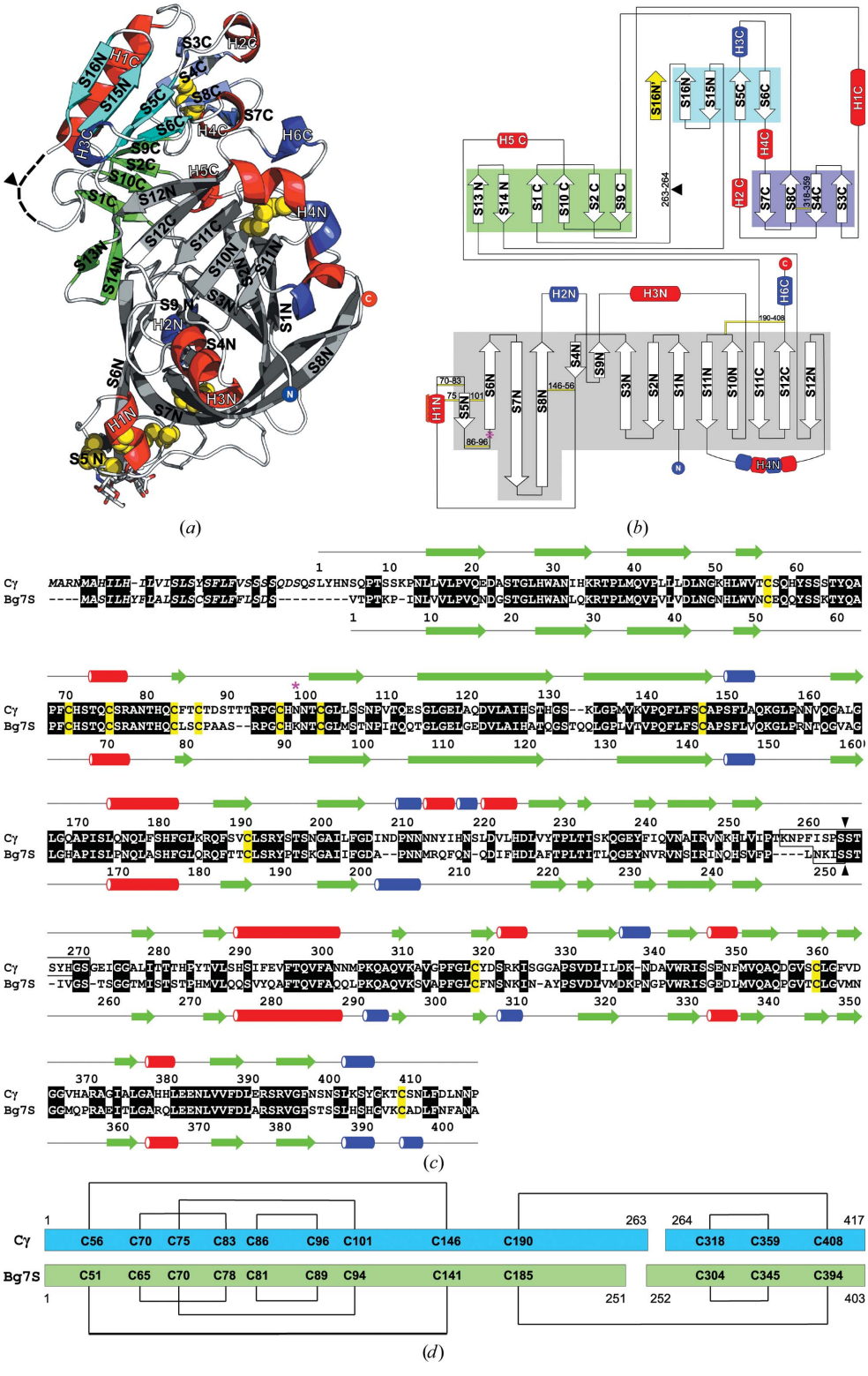


Figure 1
(*a*) Overall structure of molecule *A* of the LangC hexamer. The secondary-structure elements are marked and coloured according to topology. (*b*) Topology diagram of one molecule of the γ-conglutin hexamer. Red/blue cylinders and arrows represent α-helices/3_{10}-helices and β-strands, respectively. The length of the pictographic symbols is not commensurate with the number of amino-acid residues in these elements. The β-strands are organized into four β-sheets highlighted in grey, blue, green and cyan. The post-translational cleavage site is marked by an arrowhead. The glycosylation site is marked by a star. The disulfide bridges are marked in yellow. A β-strand from another γ-conglutin molecule, which is involved in the formation of the intermolecular β-sheet, is indicated in yellow. (*c*) Amino-acid sequence alignment of γ-conglutin (Cγ) and soybean 7S basic globulin (Bg7S) (UniProt accession codes Q42369 and P13917, respectively) calculated using *ClustalW* (http://www.ebi.ac.uk/Tools/msa/clustalw2). Signal peptides are marked in italics. Residues not visible in the electron-density maps of γ-conglutin and Bg7S are boxed. Identical residues are displayed on black/yellow backgrounds. Cysteine residues are displayed on a yellow background. The glycosylation site of γ-conglutin is marked by a star. The post-translational cleavage sites are indicated by arrowheads. Secondary-structure elements are shown as pictograms above the alignment. (*d*) Topology of the disulfide bonds of LangC and Bg7S. The post-translational cleavage sites are shown as gaps.

2.6. Determination of the post-translational cleavage site

A sample of LangC with the disulfide bonds reduced with β-mercaptoethanol was separated by gel electrophoresis according to Schägger & von Jagow (1987). The separated bands of the α and β subunits were transferred to a 0.22 µm PVDF Immobilon PSQ membrane (Millipore). The single band corresponding to subunit β (16.5 kDa) was cut out from the membrane and subjected to Edman degradation cycles on a fully automated Procise 491 sequencer (Applied Biosystems) in the BioCentrum, Krakow, Poland.

The N-terminal sequence of the β subunit was also identified by LC-MS-MS/MS analysis of peptide mixtures obtained during in-gel digestion of the protein band corresponding to 16.5 kDa, as described in §2.5.

2.7. Other software used

Molecular figures were created with *PyMOL* (v.1.5.0.4; Schrödinger). *ClustalW* (Larkin *et al.*, 2007) was used for sequence alignment. Assignment of secondary-structure elements was based on the *DSSP* algorithm (Kabsch & Sander, 1983). The *PISA* server (Krissinel & Henrick, 2007) was used to calculate the surface area between subunits in the quaternary assembly. Structure alignments based on C^{α} atoms were calculated using *PDBeFold* v.2.55 (Krissinel & Henrick, 2004).

3. Results and discussion

3.1. Purification and crystallization of γ-conglutin

LangC was purified from mature lupin seeds by two-step ion-exchange chromatography as described previously (Czubinski *et al.*, 2014). First, anion exchange was used. Under these conditions, the protein fraction of interest is positively charged and leaves the column in the flowthrough. In the next step, γ-conglutin was separated using a cation-exchange column and collected as a narrow eluting peak. The purity of the protein sample was monitored by SDS–PAGE under nonreducing as well as reducing conditions. The single 44.2 kDa band which was visible on the nonreducing SDS–PAGE gel corresponds to coupled α and β subunits of mature γ-conglutin. When the disulfide bonds were reduced, two bands were detected at 31.4 kDa (α subunit) and 16.5 kDa (β subunit), together with trace amounts of a 46.0 kDa band, indicating that post-translational cleavage was not complete. The SDS–PAGE-determined molecular weight of the α subunit is higher than the theoretical value (28 kDa), and is related to N-linked glycosylation, which may change the electrophoretic mobility of a protein.

Initial screening for crystallization conditions used commercially available screens. Reproducible crystals grew in 10% PEG 6000, 0.1 *M* HEPES pH 7.0 at 293 K but diffracted weakly to ~7 Å resolution. After optimization (PEG 6000 concentration reduced to 7%) the trigonal (space group $P3_1$ or $P3_2$) crystals diffracted X-rays to ~2 Å resolution, although they showed merohedral twinning with a ~3:1 ratio of h, k, l and $-k$, $-h$, $-l$ domains as determined by *phenix.xtriage* (Zwart *et al.*, 2005). According to Matthews analysis (Matthews, 1968), the asymmetric unit could accommodate 5–7 protein molecules with the sequence of γ-conglutin. The molecular-replacement algorithm of *Phaser* (McCoy *et al.*, 2007) identified six copies of the soybean 7S basic globulin (Bg7S) model (PDB entry 3aup; Yoshizawa *et al.*, 2011) in space group $P3_1$. The sequence deposited in UniProt as entry Q42369 (Foley *et al.*, 2011) was used in γ-conglutin model building.

3.2. Overall structure of γ-conglutin

The γ-conglutin fold is very rich in β structures but also contains several α-helices (Fig. 1*a*). The labelling scheme of the secondary-structure elements is illustrated in Fig. 1(*b*). The core of the γ-conglutin fold consists of four β-sheets flanked on the outside by α-helices. The huge 14-stranded antiparallel β-sheet (S5N, S6N, S7N, S8N, S4N, S9N, S3N, S2N, S1N, S11N, S10N, S11C, S12C and S12N) forms the spine of the γ-conglutin molecule and dominates the N-terminal domain (*i.e.* the α subunit). Within this subunit, two elongated strands (S7N and S8N) wrap around the bottom of the molecule (bottom part of Fig. 1*a*). The N-terminal domain has a rich pattern of (four) disulfide bridges. The glycosylation site is also present in this domain and is rigidified by three of these disulfide bonds (Cys70–Cys83, Cys75–Cys101 and Cys86–Cys98). The topology of these disulfide bonds resembles a knot domain, but none of them passes through the macrocycle formed by the other two. Consequently, this cysteine-rich fragment of γ-conglutin can be classified as knot-like.

A unique feature of this domain, which distinctly distinguishes γ-conglutin from other homologues, is a highly curved helix H4N deformed into a banana shape owing to a pattern of alternating α-helical and 3_{10}-helical segments. This banana-shaped helix is connected at its ends to β-strands S11N and S12N of the β-structure of the N-terminal domain.

The C-terminal domain (*i.e.* the β subunit) consists of three β-sheets, two of which are composed of four antiparallel strands each (S7C, S8C, S4C and S3C, and S16C, S15C, S5C and S6C), while the third is composed of six mixed strands (S13N, S14N, S1C, S10C, S2C and S9C). The longest α-helix of the γ-conglutin fold (H1C) is found within the C-terminal domain.

The two domains are covalently linked by the interdomain Cys190–Cys408 disulfide bridge, coupling the loop between strands S10N and S11N from the N-terminal domain with a loop leading to the C-terminal helix H6C. Although the two domains are roughly classified as N-terminal and C-terminal, there is an intricate pattern of chain swaps between these domains. Specifically, a β-hairpin formed by chains S11C and S12C from the C-terminal domain is inserted into the extended N-terminal β-sheet. Conversely, strands S13N–S14N and S15N–S16N are embedded in the mixed and antiparallel β-sheets of the C-terminal domain, respectively.

3.3. Post-translational modifications of γ-conglutin

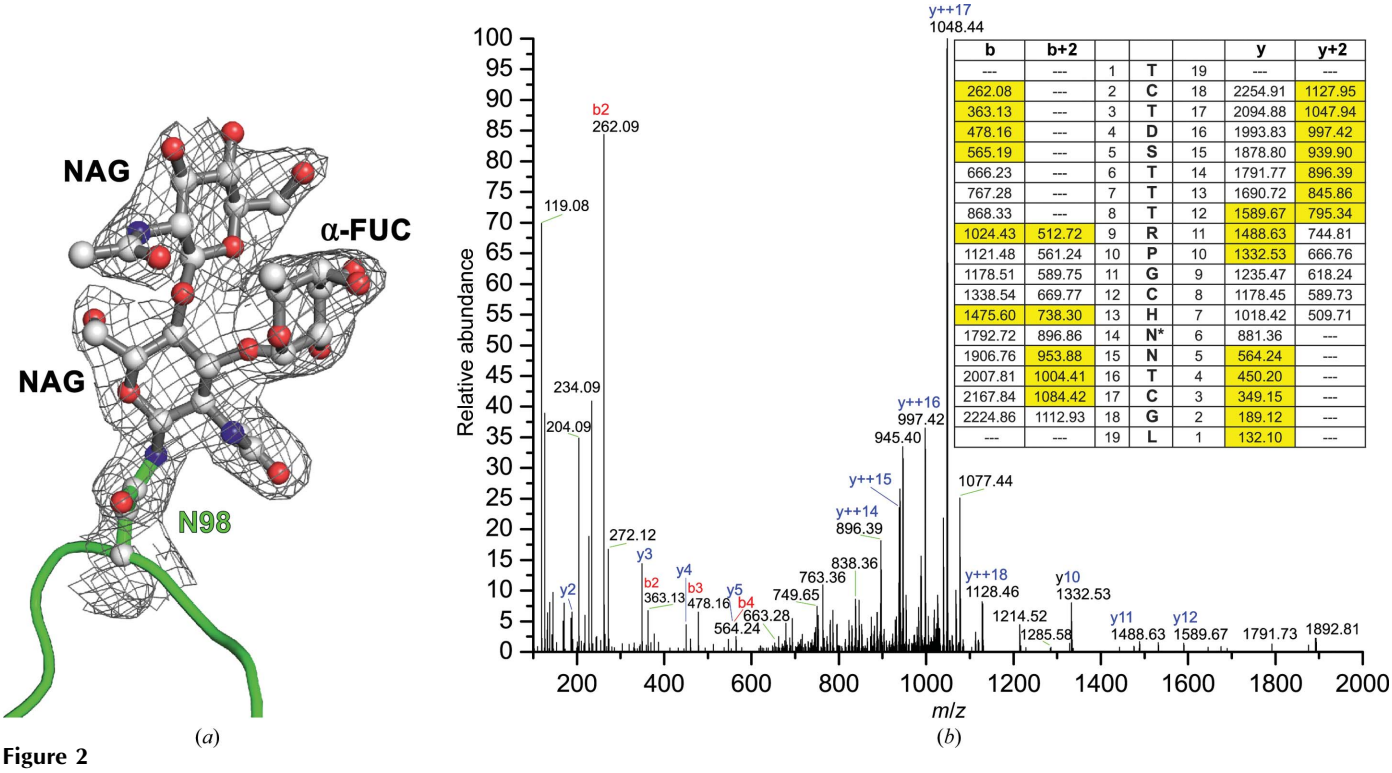
The variations in the post-translation modification of γ-conglutin and of its homologues are of great interest,

although their functional significance is unclear. LangC, with a mature sequence of 449 residues, is synthesized as a single protein chain equipped with a 32-residue signal peptide at the N-terminus. The mature protein is post-translationally cleaved into two subunits commonly termed α and β. The cleavage mechanism is unknown and the precise location of the cleavage site has not yet been determined. We were able to trace only residues Ser10–Thr255 (in subunit α) and Gly271–Pro417 (in subunit β) of mature γ-conglutin in the electron density. The N-terminal, C-terminal and cleavage regions were not visible in the electron-density map.

The cleavage site, bracketed by crystallography to within the Thr255–Gly271 segment, was precisely determined by N-terminal sequencing of the β subunit of mature LangC. Two independent experiments revealed that the cleavage site is located at the Ser263–Ser264 peptide. Edman degradation showed that the N-terminal sequence of the β subunit is $_{264}$STSYHG$_{269}$. Mass-spectrometric analysis of the LangC β subunit indicated the presence of several peptides starting with Ser264. A similar cleavage site, Ser251–Ser252, is present in Bg7S (Figs. 1*c* and 1*d*; Yoshizawa *et al.*, 2011).

There are a wide variety of post-translational cleavage sites found within γ-conglutin homologues. The TAXI-I inhibitor exists in at least two molecular forms, both with a total molecular mass of ∼40 kDa. Separation of these two forms under reducing conditions resulted in an unprocessed 40 kDa polypeptide (form *A*), as well as a mixture of 10 and 30 kDa fragments (form *B*). Because the N-terminal sequences of the 30 and 40 kDa polypeptides are identical, it is believed that form *B* is derived from form *A* by post-translational cleavage, possibly by proteolytic enzymes. Both TAXI-I forms are functional and inhibit microbial GH11 enzymes, and thus the cleavage process of this inhibitory protein is not crucial (Sansen *et al.*, 2004). Interestingly, the cleavage points of γ-conglutin and Bg7S are located before the longest α-helix (H1C in the γ-conglutin structure), in contrast to TAXI-I, where the cleavage site follows this α-helix.

γ-Conglutin has 12 cysteine residues, which are paired into six disulfide bridges in a pattern that is a characteristic feature of all legume 7S basic globulins (Sansen *et al.*, 2004; Yoshizawa *et al.*, 2011, 2012). The Cys190–Cys408 bond is particularly important as it links the α and β chains into one covalent γ-conglutin molecule. A similar pattern of cysteine residues and disulfide bridges was found in the structures of Bg7S and EDGP, which share 67 and 44% sequence identity, respectively, with LangC. Despite these similarities, EDGP does not undergo post-translational cleavage. Therefore, the Cys181–Cys406 bond, which corresponds to the disulfide bridge coupling the two subunits of the mature γ-conglutin molecule, is not so crucial for EDGP. The positions of the disulfide bridges in the structure of TAXI-I are different from those in LangC, Bg7S and EDGP. TAXI-I shares 28% sequence identity with LangC but, despite the fact that both proteins consist of two disulfide-coupled subunits, the pattern of the four disulfide bridges within the larger N-terminal subunits is different.



b	b+2				y	y+2
---	---	1	T	19	---	---
262.08	---	2	C	18	2254.91	1127.95
363.13	---	3	T	17	2094.88	1047.94
478.16	---	4	D	16	1993.83	997.42
565.19	---	5	S	15	1878.80	939.90
666.23	---	6	T	14	1791.77	896.39
767.28	---	7	T	13	1690.72	845.86
868.33	---	8	T	12	1589.67	795.34
1024.43	512.72	9	R	11	1488.63	744.81
1121.48	561.24	10	P	10	1332.53	666.76
1178.51	589.75	11	G	9	1235.47	618.24
1338.54	669.77	12	C	8	1178.45	589.73
1475.60	738.30	13	H	7	1018.42	509.71
1792.72	896.86	14	N*	6	881.36	---
1906.76	953.88	15	N	5	564.24	---
2007.81	1004.41	16	T	4	450.20	---
2167.84	1084.42	17	C	3	349.15	---
2224.86	1112.93	18	G	2	189.12	---
---	---	19	L	1	132.10	---

Figure 2
(*a*) The glycosylation site of γ-conglutin, with a fucosylated *N*-acetylglucosamine core. Asn98 and its N-linked sugar moiety, two *N*-acetylglucosamine (NAG) units and one α-fucose (FUC) unit, are shown in $2F_o - F_c$ electron density contoured at the 1.0σ level. (*b*) Annotated fragmentation spectra of the *N*-glycosylated peptide $_{85}$TCTDSTTTRPGCHNNTCGL$_{103}$ obtained by fragmentation of the MH^{3+} ion, m/z = 785.99. A list of theoretically possible fragment ions is presented in a table. The b-series and y-series fragment ions that are present in the spectrum are marked in yellow.

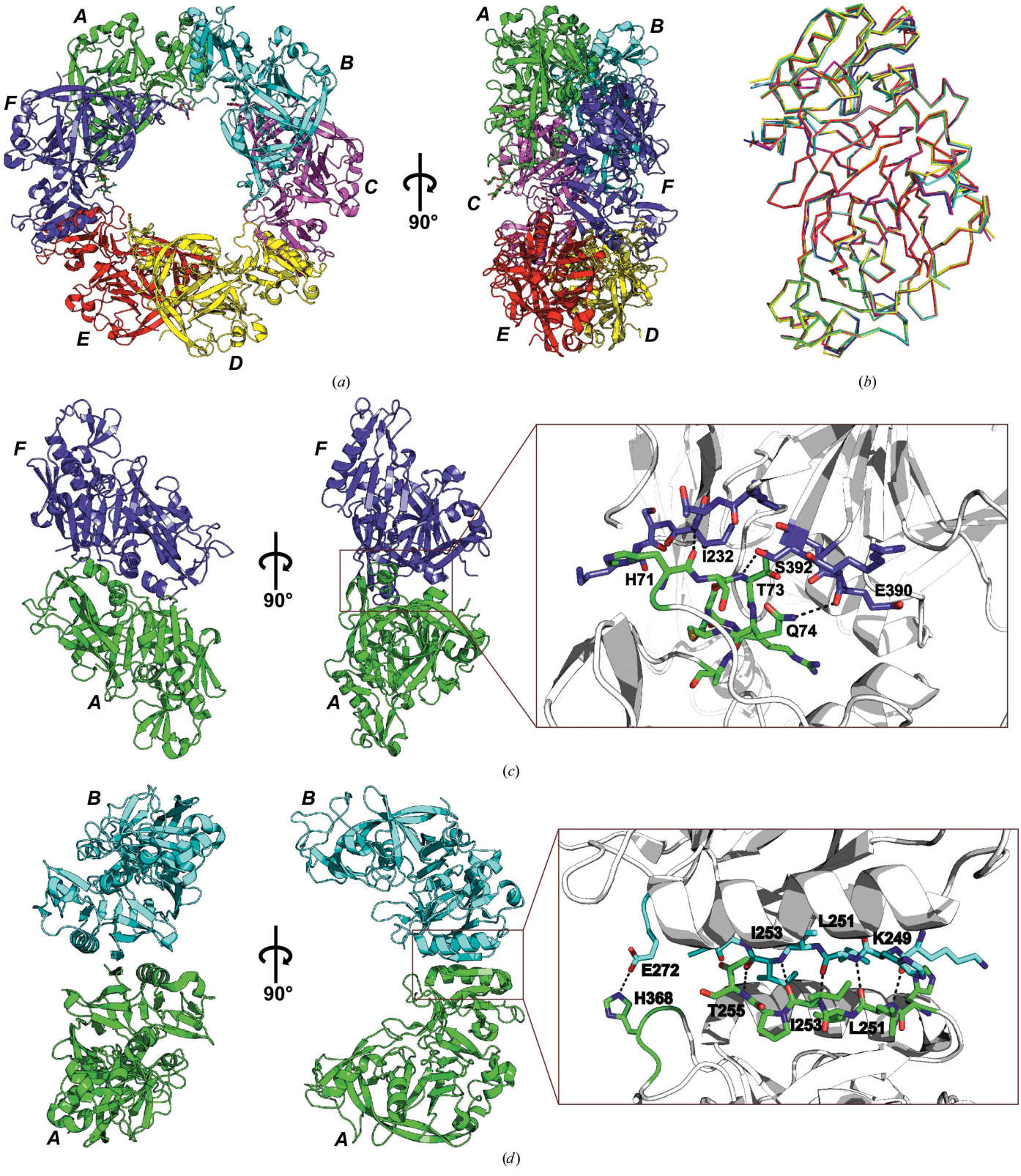


Figure 3
(*a*) Front (left) and side (right) views of the LangC γ-conglutin hexamer. The *A*, *B*, *C*, *D*, *E* and *F* molecules in the asymmetric unit are shown in green, cyan, pink, yellow, red and blue cartoon representation, respectively. (*b*) Superposed C^{α} traces of the six LangC protomers are shown in wire representation using the colour scheme of (*a*). The two different dimerization modes are presented in (*c*) and (*d*) using the colour scheme of (*a*). (*c*) A compact arrangement of two molecules with a helix interaction (motif I). (*d*) A less compact LangC dimer created through interacting β-strands S16N and S16N′ from two molecules, forming an intermolecular β-sheet (motif II). Each of these motifs (I and II) is found in three copies in the γ-conglutin hexamer.

Table 2
R.m.s.d. values for C^{α} atoms in pairwise superpositions of the *A*–*F* components of the γ-conglutin hexamer, listed for subunits $\alpha + \beta$ (first row), α (second row) and β (third row).

The superpositions were calculated with *PDBeFold* v.2.55 (Krissinel & Henrick, 2004).

	R.m.s.d. of C^{α} atoms (Å)				
	A	*B*	*C*	*D*	*E*
B	0.46				
	0.39				
	0.49				
C	0.37	0.50			
	0.29	0.32			
	0.45	0.66			
D	0.58	0.61	0.67		
	0.61	0.58	0.61		
	0.45	0.56	0.61		
E	0.51	0.60	0.59	0.72	
	0.37	0.42	0.34	0.61	
	0.59	0.76	0.84	0.74	
F	0.54	0.55	0.58	0.55	0.59
	0.49	0.47	0.43	0.39	0.51
	0.45	0.59	0.75	0.62	0.85

Table 3
The interaction interface areas between molecules/dimers within the LangC hexamer and the Bg7S tetramer calculated with *PDBePISA* (Krissinel & Henrick, 2007).

Molecule	Surface† (Å^2)	Interface area‡ (Å^2)
LangC		
A–*F*	17160–17447	1235
B–*C*	17128–17354	1224
D–*E*	17541–17186	1297
Average	17303	1252
A–*B*	17160–17128	701
C–*D*	17354–17541	619
E–*F*	17186–17447	664
Average	17303	661
Bg7S		
A–*B*	16565–16392	1389
C–*D*	17288–16736	1643
Average	16745	1516
A–*D*	16565–16736	1436
C–*B*	17288–16392	1409
Average	16745	1422

† Total solvent-accessible surface area per molecule. ‡ The interface area calculated as one-half of the difference in the total accessible surface area of isolated and interacting partners.

Glycosylation is a frequent post-translational modification of proteins and is often important for functional properties; at the same time, it can protect proteins against hydrolysis. Sequence analysis of LangC indicated the presence of a consensus Asn-Xaa-Ser/Thr motif, where Xaa is any amino acid except Pro, characteristic of N-linked glycosylation. This motif is found in the α-subunit at Asn98-Asn99-Thr100 of the γ-conglutin sequence. The electron-density map at Asn98 revealed a large fragment of the glycan moiety consistent with two *N*-acetylglucosamine (NAG) units and one α-fucose (FUC) unit (Fig. 2*a*). The presence of glycosylation was confirmed by mass spectrometry. Firstly, a γ-conglutin sample was subjected to deglycosylation with EndoH glycosidase, during which the glycan moiety was trimmed to the NAG unit directly coupled to the Asn side chain. This step is necessary because large and heterogeneous glycan structures cannot be analyzed directly together with peptides in MS spectra. Ultimately, a $_{85}$TCTDSTTTRPGCHNNTCGL$_{103}$ peptide carrying a single NAG unit was detected (Fig. 2*b*). The mass of this fragment determined by MS is 2354.97 Da and is the sum of the mass of the peptide (1980.81 Da) with the Cys residues modified to carbamidomethylcysteine ($3 \times 57.02 = 171.07$ Da) and the mass of a single NAG unit (203.08 Da). The presence of *N*-glycosylation was additionally verified by manual inspection of the fragmentation data and unambiguous assignment of the b-series and y-series fragment ions, confirming the peptide modification. Among γ-conglutin homologues, only EDGP is glycosylated at four N-linked glycosylation sites (Asn90, Asn254, Asn299 and Asn410). However, in the EDGP structure only a single NAG moiety linked to each of the above Asn residues could be modelled in the electron-density map (Yoshizawa *et al.*, 2012). In the case of Bg7S, the corresponding glycosylation motif of γ-conglutin is mutated to Lys91-Asn92-Thr93, and thus this glycosylation site is lost.

Since γ-conglutin has the potential to become a component of foodstuffs or therapeutic agents, its glycosylation moiety could become part of an epitope that could elicit the production of human antibodies. Such food-related antibodies are usually specific for the carbohydrate moiety rather than for the N-linked protein. As a consequence, this phenomenon leads to cross-reactivity with other glycoproteins that have a common glycan moiety.

Recently, a glycoproteomic characterization of γ-conglutin form *L. albus* has identified four main micro-heterogeneous variants of a single glycan (Schiarea *et al.*, 2013). Among these glycan forms, two mannosidic-type variants are dominant, while two complex-type N-glycans are less abundant. Unfortunately, it is not possible to determine the complete composition of the glycan moiety of LangC based on the present electron-density maps. The quality of the electron-density map for the six copies of the γ-conglutin molecule within the oligomer is different and thus it can be only concluded that the glycan moiety is probably heterogeneous beyond the visible fragment.

3.4. The quaternary structure of γ-conglutin

A unique feature of legume 7S basic globulin proteins is the formation of quaternary assemblies (Yoshizawa *et al.*, 2011, 2012). In the present case, the asymmetric unit contains six copies (labelled *A*, *B*, *C*, *D*, *E* and *F*) of the γ-conglutin molecule, which form a ring-like hexameric assembly (Fig. 3*a*). In pairwise comparisons, the C^{α} traces of the six molecules superpose quite well, with an average root-mean-square deviation (r.m.s.d.) of 0.5 Å (Table 2; Fig. 3*b*). This indicates that the molecular structure of the six protomers is essentially identical. The hexamer is formed using two types of dimeric interfaces. The interaction interface area between molecules *A*–*F*, *B*–*C* and *D*–*E* is large, on average 1252 Å^2 of the 17 303 Å^2 solvent-accessible area of an isolated monomer,

indicating a strong interaction (Table 3). The dimers formed using this interface (designated type I) must accordingly be very tight. The type I interface is formed *via* a characteristic α-helix motif which is localized within the cysteine-rich region of the molecule (Fig. 3*c*). The key element in this motif is the H1N α-helix, which is additionally stabilized by the Cys75–Cys101 disulfide bridge. Residues Thr73 and Gln74 from this α-helix form hydrogen bonds, respectively, to Ser392 and Glu390 from the loop between β-strands S11C and S12C of the complementary subunit. All of the key residues that participate in this interaction are conserved in the sequence of the soybean homologue. Thus, a similar type I dimerization interface is also found between molecules *A*–*D* and *B*–*C* of the Bg7S tetramer (Table 3). The second type of dimeric interface (II) within the LangC hexamer (*A*–*B*, *C*–*D* and *E*–*F*) is formed through mixed β-sheet interactions between strands S16N and S16N′ from an adjacent molecule. This interaction leads to the extension of one of the molecular β-sheets in an intermolecular context (Fig. 3*d*). Even though the interaction area of motif II is smaller (∼661 Å^2; Table 3), this interaction can also be considered to be quite strong, as it leads to the creation of a large (eight-stranded) intermolecular β-sheet. In contrast to motif I, the dimerization motif II is not present in the soybean homologue. The combination of interactions I and II leads to a hexameric LangC structure comprised of two superposed 'trimers' with threefold pseudosymmetry. In Fig. 3(*a*) the lower triangle is *A*–*C*–*E* and the upper triangle is *B*–*D*–*F*. The dimeric interactions I and II operate alternately

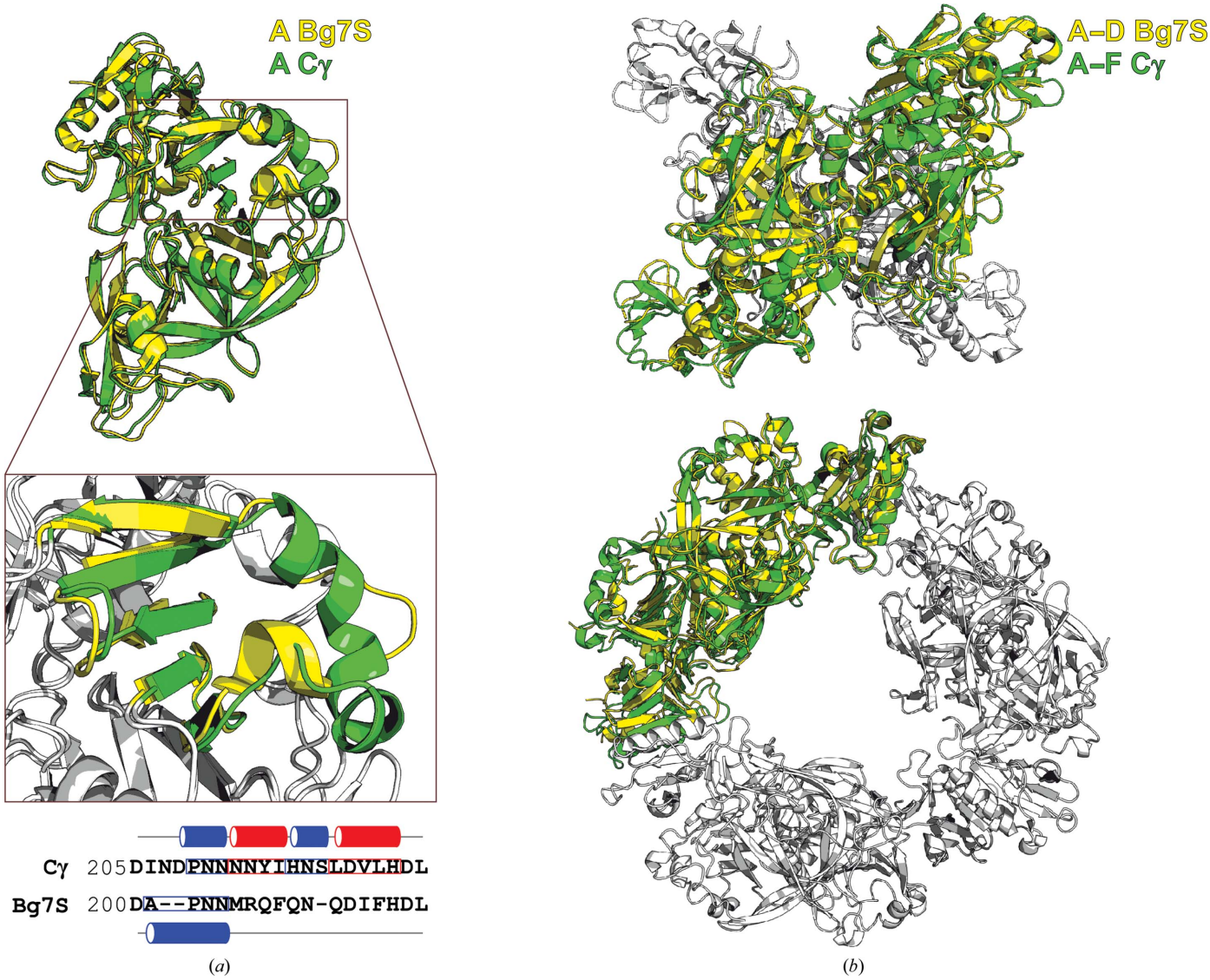


Figure 4
(*a*) Side view of superposed protomers of γ-conglutin (Cγ) and soybean 7S basic globulin (Bg7S; PDB entry 3aup). Chain *A* of the γ-conglutin hexamer and chain *A* of the Bg7S tetramer are shown as green and yellow cartoons, respectively. A close-up view of the γ-conglutin unique curved helix H4N, consisting of alternating α-helical and 3_{10}-helical segments, is presented in a box, together with annotated sequences of the corresponding structural elements. (*b*) Two views of the type I dimer of γ-conglutin and Bg7S, superposed on the Bg7S tetramer (top panel) and on the γ-conglutin hexamer (bottom panel). The *A*–*F* and *A*–*D* dimers of γ-conglutin and Bg7S are shown in green and yellow cartoon representation, respectively. The remaining subunits of each oligomer (Bg7S tetramer, top; γ-conglutin hexamer, bottom) are shown in grey.

Table 4
R.m.s.d. values for C^{α} atoms of chain *A* of LangC superposed on homologous proteins identified by their PDB codes.

The superpositions were calculated with *PDBeFold* v.2.55 (Krissinel & Henrick, 2004).

Protein structure	PDB code	R.m.s.d. of C^{α} atoms† (Å)	No. of aligned residues	Sequence identity (%)	*Q*-score‡
Bg7S	3aup	1.04/1.24	365	67	0.71
EDGP	3vlb	1.45/1.46	330	44	0.56
TAXI-IA	1t6g	1.98/2.03	301	28	0.44
Porcine pepsin	4pep	2.34	258	20	0.33
Cockroach allergen Bla g 2	1yg9	2.59	249	14	0.28
Phytepsin	1qdm	2.86/2.91	259	17	0.21

† If more than one protein chain was present in a crystal structure, the lowest/highest r.m.s.d. values are provided. ‡ *Q*-score represents the quality function of C^{α} alignment. It reduces the effect of the r.m.s.d./$N_{\rm algn}$ (the number of aligned residues) balance on the estimation of alignments: $Q = (N_{\rm algn}N_{\rm algn})/\{[1 + ({\rm r.m.s.d.}/R_0)^2]N_{\rm res1}N_{\rm res2}\}$, where $N_{\rm res1}$ and $N_{\rm res2}$ represent the number of residues in the aligned proteins and the empirical parameter R_0 is set to 3 Å.

between molecules from the lower and upper triangles across pseudo-dyads. In consequence, the hexameric assembly has local D_3 symmetry.

3.5. pH-dependent assembly of γ-conglutin

It has been reported before that quaternary assemblies of legume 7S basic globulins are strongly pH-dependent (Capraro *et al.*, 2010; Yoshizawa *et al.*, 2011). Capraro *et al.* (2010) studied the pH-dependent structural dynamics of γ-conglutin using size-exclusion chromatography coupled with light scattering. The authors noted that at pH 4.5, which was the most acidic pH tested, γ-conglutin appeared as a single peak with a molecular mass of ∼50 kDa corresponding to the monomeric state. At pH 6.5 and 7.5 the monomeric form completely disappeared, while oligomeric forms with molecular masses of about 100, 255 and 480 kDa were observed. Among these forms, the dominating form (90%) was that with a molecular mass of ∼255 kDa. Owing to poor consistency of the results from chromatographic separations, the final information about the number of γ-conglutin molecules forming the quaternary assembly was not clearly confirmed. To overcome difficulties with equilibrium shifts among the aggregates during the chromatographic procedures, the authors used cross-linking of the γ-conglutin molecules. In the end, they were able to obtain a clear separation of aggregates with homogeneous size and concluded that γ-conglutin coexisted as a dimer and tetramer of the 50 kDa molecule. Those conclusions were partially confirmed by structural analysis of Bg7S (Yoshizawa *et al.*, 2011). This homologue of γ-conglutin forms a tetrameric assembly both in the crystal and in solution at pH 7.0 or above. Under weakly basic and weakly acidic conditions Bg7S has dimeric transitions. By analyzing the electrostatic potential of this protein, Yoshizawa *et al.* (2011) noted that the two types of hypothetical dimers utilized acidic or basic surfaces of the molecule in order to form quaternary assemblies.

The structural changes in the quaternary assembly of lupin γ-conglutin take place at pH values of about 4.5 and 6.5 (Capraro *et al.*, 2010). These pH conditions correspond to the p$K_{\rm a}$ values of the glutamate and histidine side chains, respectively. Glu and His residues can be found in the structural elements forming the dimeric interfaces of both motif types (I and II) within LangC in the crystal structure. For example, as part of the type II interface, His368 located in the loop between S9C and S10C from molecule *A* forms a salt bridge to Glu272 located in a loop near the cleavage site of molecule *B*. The formation of this salt bridge seems to be of particular importance for the quaternary assembly of LangC.

Light-scattering measurements of LangC in solution confirmed that the protein forms a hexameric assembly at pH 7.5. The γ-conglutin sample was prepared in a similar way as for crystallization, *i.e.* by purification using two-step ion-exchange chromatography and buffer exchange to 20 m*M* Tris pH 7.5. The radius of gyration ($R_{\rm g}$) calculated for the monomer and hexamer corresponding to the crystal structure was 2.34 and 4.81 nm, respectively. The experimental value of the hydrodynamic radius extrapolated to zero concentration determined from dynamic light scattering is 5.72 nm. For spherical particles, the hydrodynamic radius should be ∼15% higher than the radius of gyration (García de la Torre *et al.*, 2000), and this relation is satisfied in the case of the γ-conglutin hexamer. The molecular weight of LangC determined by static light scattering is ∼280 kDa. This result was additionally confirmed by analytical ultracentrifugation (AUC). The experiment, based on sedimentation velocity, revealed a major peak (68% based on *SEDFIT* calculations) at ∼250 kDa corresponding to a γ-conglutin hexamer, and some minor peaks. The molecular weight of LangC determined by AUC is ∼250 kDa (sedimentation coefficient of ∼10.6 S). All of the molecular-mass determination experiments confirm that LangC is hexameric not only in the crystal but also in solution.

3.6. Comparison of γ-conglutin with soybean 7S basic globulin

A search for structural homologues of γ-conglutin was performed using *PDBeFold* (*SSM*; Krissinel & Henrick, 2004; http://www.ebi.ac.uk/msd-srv/ssm/). The server detected similarity to (i) Bg7S (soybean 7S basic globulin; PDB entry 3aup; *Q*-score 0.71; Yoshizawa *et al.*, 2011), (ii) EDGP (extracellular dermal glycoprotein; PDB entry 3vlb; *Q*-score 0.56; Yoshizawa *et al.*, 2012), (iii) TAXI-IA (*T. aestivum* xylanase inhibitor IA; PDB entry 1t6g; *Q*-score 0.44; Sansen *et al.*, 2004), (iv) pepsin-like aspartic proteases, *e.g.* porcine pepsin (PDB entry 4pep; *Q*-score 0.33; Sielecki *et al.*, 1990) and plant phytepsin (PDB entry 1qdm; *Q*-score 0.21; Kervinen *et al.*, 1999), and (v) cockroach allergen (PDB entry 1yg9; *Q*-score 0.28; Gustchina *et al.*, 2005).

Structure-based sequence alignment indicated that the secondary-structural elements are well conserved between

LangC and Bg7S, although deletions and insertions in some helices and loops could be observed (Fig. 1*c*). One protomer of γ-conglutin superimposes relatively well on the Bg7S molecule (Fig. 4*a*), and the r.m.s.d. of 1.04 Å for 365 superposed C^{α} atoms indicates that these proteins are structurally related (Table 4). Despite the fact that the general fold is preserved, the structure of γ-conglutin is distinctly different because of the presence of the unique curved helix H4N (Fig. 4*a*). Moreover, the protomers of these two proteins undergo a completely different quaternary assembly. In the case of Bg7S the molecules form a pseudo-D_2-symmetric tetramer (dimer of dimers), while LangC exists as a cyclic D_3 hexamer with noncrystallographic symmetry (trimer of dimers) (Fig. 4*b*). The same type I dimerization principle is found in the γ-conglutin and Bg7S structures. However, the dimers within the oligomers are organized differently, as the type II dimer found in γ-conglutin is not present in the Bg7S structure. Bg7S exists as a tetramer with a cruciform shape formed by two superposed type I dimers (upper part of Fig. 4*b*), while γ-conglutin is arranged in a circular hexameric

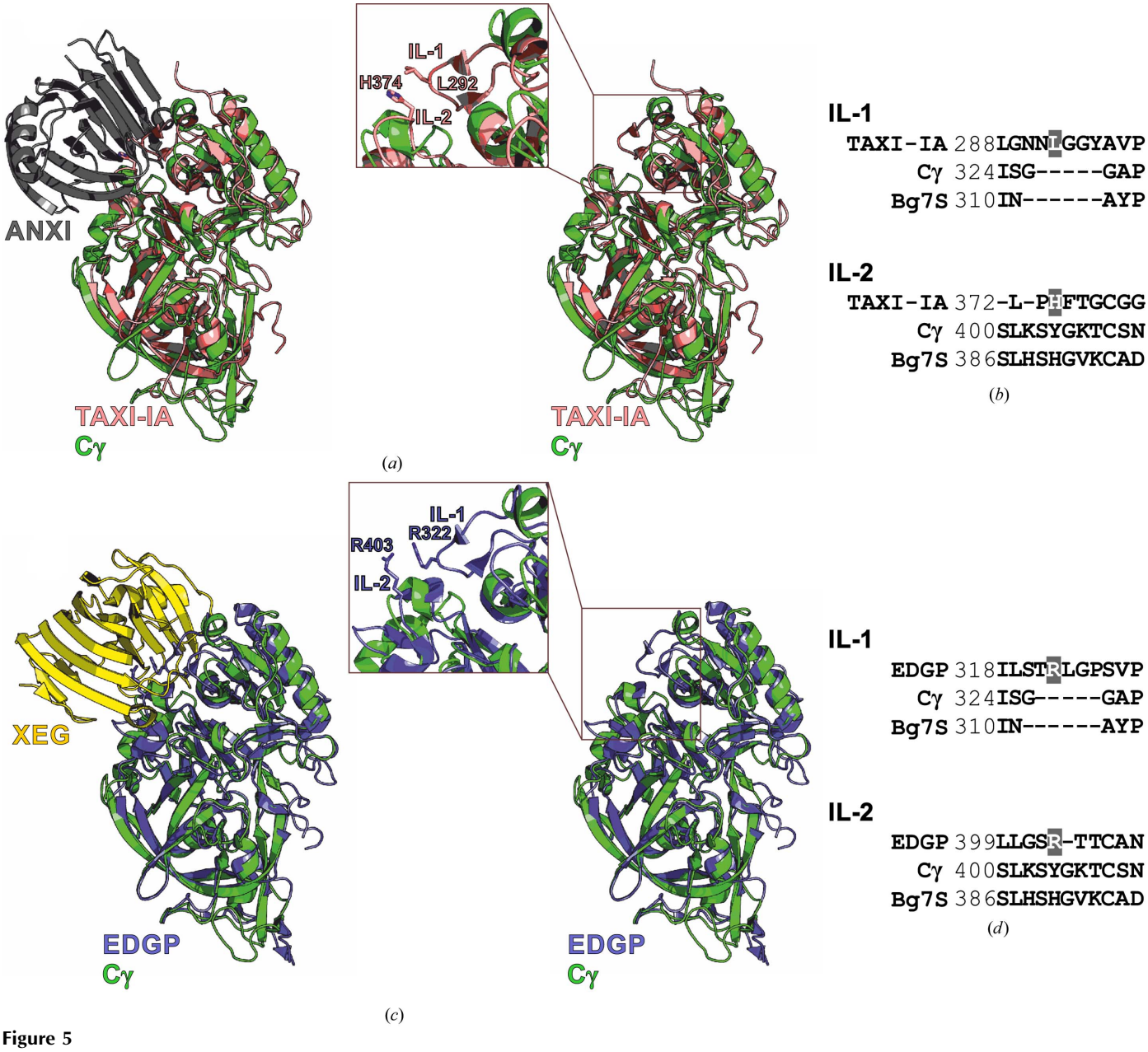


Figure 5
(*a*) Structural superposition of molecule *A* of LangC (green) on *T. aestivum* xylanase inhibitor-I (TAXI-IA, pink) in complex with *A. niger* xylanase (ANXI, grey) (PDB entry 1t6g). A close-up view of the inhibitory loops 1 (IL-1) and 2 (IL-2) of TAXI-IA, with residues involved in the inhibition of ANXI shown in stick representation and labelled. (*b*) Sequence alignment of IL-1 and IL-2 of TAXI-IA with γ-conglutin and Bg7S. Leu292 and His374 of TAXA-IA are highlighted in grey. (*c*) Structural superposition of molecule *A* of LangC on carrot extracellular dermal glycoprotein (EDGP, blue) in complex with xyloglucan-specific endo-β-1,4-glucanase (XEG, yellow) (PDB entry 3vlb). A close-up view of the inhibitory loops IL-1 and IL-2 of EDGP, with residues involved in the inhibition of XEG shown in stick representation and labelled. (*d*) Sequence alignment of IL-1 and IL-2 of EGDP with γ-conglutin and Bg7S. Arg322 and Arg403 of EDGP are highlighted in grey.

form composed of two three-membered rings (bottom part of Fig. 4*b*). In the Bg7S tetramer, a different 'type II' interface (*A*–*B* and *C*–*D*) is formed between loop elements of two molecules in a head-to-head arrangement. Interestingly, the amino-acid residues that are directly involved in the formation of the intermolecular β-sheet in the type II dimer of the γ-conglutin hexamer differ from the corresponding residues in Bg7S. The differences involve three residues, Leu251, Ile253 and Thr255, which form parallel-type β-sheet hydrogen bonds, as well as Glu272 and His368, which form a salt bridge.

3.7. Comparison of γ-conglutin with glycoside hydrolase inhibitor proteins

Plant microbial pathogens secrete enzymes to digest the polysaccharide chains of plant cell walls. These enzymes are classified into glycoside hydrolase families in the CAZY carbohydrate-active enzymes database (Lombard *et al.*, 2014). The degradation of cell walls, which act as a physical barrier against microorganisms, is particularly important during pathogen invasion. To form a defence against this attack, plants synthesize specific protein inhibitors of glycoside hydrolases.

In the cases of EDGP and TAXI-IA, which are monomeric, the level of structural similarity with γ-conglutin is low, despite the fact that the structural cores are well preserved. The C^{α} r.m.s.d. of molecule *A* of γ-conglutin superposed on EDGP and TAXI-IA is 1.45 and 1.98 Å, respectively (Table 4, Figs. 5*a* and 5*c*). Despite detectable structural homology with these glycoside hydrolase inhibitor proteins, γ-conglutin does not inhibit GH enzymes. This phenomenon is common to all legume 7S basic globulins (Scarafoni *et al.*, 2010; Yoshizawa *et al.*, 2011). The lack of inhibitory activity against both GH11 and GH12 hydrolases is specifically related to structural changes in two inhibition loops, namely IL-1 and IL-2 (Figs. 5*a* and 5*c*). The GH11 and GH12 enzymes have a similar β-jelly-roll fold with a substrate-binding groove formed by a concave β-sheet, suggesting that similar elements of inhibitor proteins should be involved in the inhibition mechanism (Sansen *et al.*, 2004).

The active site of GH11 consists of two conserved glutamates located on either side of an extended open cleft. Hydrolysis of xylosidic substrates by a member of the GH11 family, *Aspergillus niger* xylanase I (ANXI), is believed to proceed *via* a double-displacement mechanism, in which the nucleophilic catalyst Glu79 forms a covalent glycosyl-enzyme intermediate which is subsequently hydrolyzed by Glu170, which acts as a general acid/base (Sansen *et al.*, 2004). In the ANXI–TAXI-IA complex structure (PDB entry 1t6g), the imidazole ring of His374 of TAXI-IA (from IL-2) is located between the two catalytic glutamates of ANXI. Additionally, Leu292 of TAXI-IA (from IL-1) forms a hydrophobic interaction with Tyr10 of ANXI (Sansen *et al.*, 2004).

The mechanism of hydrolysis catalyzed by the GH12 enzymes is similar, *i.e.* it involves two glutamates, where one acts as the nucleophile and the other as an acid/base. The putative nucleophile and acid/base of xyloglucan-specific endo-β-1,4-glucanase (XEG), a representative of the GH12 family, are Glu119 and Glu205, respectively. The crystal structure of an XEG–EDGP complex (PDB entry 3vlb) revealed that Arg322 and Arg403 located in IL-1 and IL-2, respectively, of EDGP penetrate the active-site cleft of XEG and interact with the catalytic glutamates of the enzyme *via* salt bridges. Additionally, hydrophobic interactions are created by the aliphatic moiety of Arg403 of EDGP and Trp28 of XEG, as well as by Leu202 and Pro203 of EDGP and Trp13 of XEG (Yoshizawa *et al.*, 2012).

Although overall the ANXI–TAXI-IA and XEG–EDGP complexes are structurally comparable, the individual differences in the architecture of the active sites of GH enzymes indicate that specific proteins are required to inhibit them. Interestingly, superposition of the C^{α} atoms of γ-conglutin on TAXI-IA (an inhibitor of GH11) and EDGP (an inhibitor of GH12) shows the absence of the IL-1 loop in LangC (Fig. 5*b* and 5*d*). The IL-2 loop of TAXI-IA and EDGP is marked by the His374 and Arg403 residues, respectively. In γ-conglutin, Tyr404 is present at this position of IL-2. These structural observations explain why γ-conglutin does not interact with the active site of either the GH11 or GH12 enzymes.

3.8. Comparison of γ-conglutin with pepsin-like aspartic proteases

Plants are equipped with a complex proteolytic machinery, which is particularly involved in protein turnover. At the same time, plant proteases play key roles in the defence system against pathogen invasion (Hoorn & Jones, 2004). Based on the key residues used by the enzymes to cleave peptide bonds, the superfamily of proteolytic enzymes is subdivided into five classes, containing serine proteases, aspartic proteases, cysteine proteases, metalloproteases and threonine proteases. According to structural homology analysis, the fold of γ-conglutin bears resemblance to several pepsin-like proteins from the aspartic protease family, despite a very low level of sequence identity (below 20%; Table 4). Pepsin-like aspartic proteases are built from two similarly folded domains, each contributing an Asp-Thr-Gly triad to the active site. They cleave peptide bonds using a nucleophilic water molecule activated by the two catalytic Asp residues located at the bottom of the catalytic cleft. A superposition of porcine pepsin (PDB entry 4pep) on γ-conglutin locates the active-site of the enzyme in the cleft between the α and β subunits of γ-conglutin (Fig. 6*a*). In a detailed comparison, the catalytic residues of porcine pepsin, Asp32-Thr33-Gly34 and Asp215-Thr216-Gly217, correspond to the Asp46-Leu47-Asn48 and Thr279-Thr280-Thr281 residues of γ-conglutin, respectively (Fig. 6*b*). These mutations of the catalytic residues explain why γ-conglutin is catalytically dysfunctional and does not have any protease activity.

Cockroach allergen, Bla g 2 (PDB entry 1yg9), is a 36 kDa glycoprotein that also shares homology with aspartic proteases (Table 4). However, this protein is inactive owing to critical amino-acid substitutions around the catalytic residues. The Gly34 and Gly217 residues in the two Asp-Thr-Gly triads of

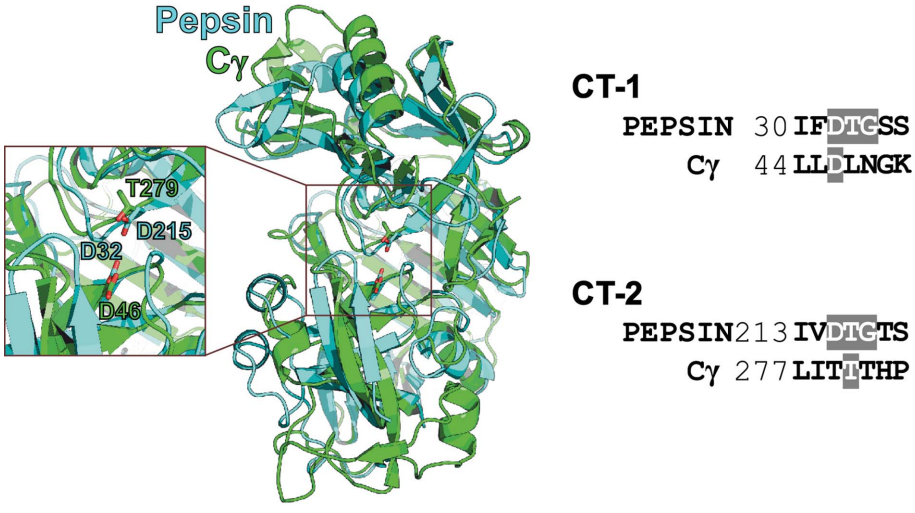
Bla g 2 are substituted by Thr and Ser, respectively. These mutations lead to a disruption of the critical hydrogen-bond network involving the catalytic Asp residues (Gustchina *et al.*, 2005). Two other features also differentiate Bla g 2 from classical aspartic proteases: the presence of five disulfide bridges stabilizing the protein fold and the ability to bind metal ions. The large number of disulfide bridges (six) found in γ-conglutin and its zinc-binding ability are interesting parallels between these two apparently unrelated proteins.

Structurally, apart from the main pepsin-like domain, plant aspartic proteases (phytepsin; PDB entry 1qdm) also have an additional plant-specific insert of ~100 residues near the C-terminus that is not present in animal or microbial homologues (Kervinen *et al.*, 1999). In detailed topology (PDB entry 3rfi), the plant-specific insert consists of four amphipathic α-helices, which are folded in a boomerang-shaped subdomain with a hydrophobic internal surface. This unique domain is additionally stabilized by three disulfide bridges (Bryksa *et al.*, 2011). Sequence analysis indicates that the plant-specific insert is similar to saposins and saposin-like proteins. Therefore, a physiological function related to membrane fusion was proposed for the insert (Egas *et al.*, 2000). The pH dependence of the helix content of the plant-specific insert indicates that membrane–protein interactions would require acidic conditions. The plant-specific insert of plant aspartic proteases is crucial for the molecular distribution of the enzymes in the cell. After translocation of the enzyme to a vacuole, final maturation takes place, during which the plant-specific domain near the C-terminus of the enzyme is removed. Although the amino-acid sequence of γ-conglutin is also ~100 residues longer than the sequences of typical animal or microbial aspartic proteases, the additional residues are distributed all over the sequence and do not form a plant-specific insertion domain. The main structural differences between γ-conglutin and pepsin-like proteins are found in the bottom part of the N-terminal subunit in Fig. 6(*a*), where the cysteine-rich domain of the lupin protein is present. In particular, the H1N α-helix, which is involved in the formation of the type I dimer characteristic of legume 7S basic globulins, is present in this differentiating region. Somewhat analogously to the plant-specific insert of aspartic proteases, three disulfide bridges are present in this region of γ-conglutin. Furthermore, in contrast to plant pepsin-like proteins, γ-conglutin is equipped with the unique banana-shaped helix H4N. The character of the H4N structure may suggest membrane-interaction potential of this element. Taken together, the above observations suggest that γ-conglutin probably evolved from pepsin-like protease ancestors. In the course of evolution, γ-conglutin lost the proteolytic activity and the identity of the plant-specific insert, although the general folding pattern has been preserved. In particular, the primordial pattern of two ψ-loops (Andreeva *et al.*, 1984) is still distinctly recognizable in the S3N–S4N/S9N and S1C–S2C/S10C motifs (Fig. 1*b*).

4. Conclusions

The crystal structure of the lupin seed γ-conglutin LangC, a member of the legume 7S basic globulins, determined at 2.0 Å resolution reveals a fold rich in disulfide bridges and in β-structures, including a huge 14-stranded β-sheet at the molecular core. The protein is maturated by (i) the formation of six disulfide bridges, (ii) post-translational cleavage at the Ser263-Ser264 peptide into subunits α and β, which remain linked *via* one of the S—S bonds (Cys190–Cys408), (iii) N-linked glycosylation and (iv) the removal of an N-terminal signal peptide (32 residues). The presence of multiple disulfide bridges creates a knot-like domain in subunit α. The single Asn-linked glycosylation of γ-conglutin at Asn98 is clearly visible in the electron-density maps as a branched structure consisting of two NAG units and one FUC unit, up to the point where the glycan moiety becomes nonhomogeneous. The glycosylation site was additionally confirmed by advanced enzymatic/MS analysis. The γ-conglutin fold includes a unique highly curved helix H4N comprised of alternating α-helical and 3_{10}-helical segments, suggestive of specific anchoring at target cellular structures, such as membranes.

γ-Conglutin undergoes quaternary assembly in a pH-dependent manner. To date, only the quaternary structure of Bg7S, a soybean homologue of γ-conglutin, has been reported. Bg7S forms D_2-symmetric tetramers assembled from two dimers



(*a*) (*b*)

Figure 6
(*a*) Structural superposition of molecule *A* of LangC (green) on porcine pepsin (PDB entry 4pep; blue). A close-up view of the catalytic triad 1 (CT-1) and 2 (CT-2) of pepsin, with residues involved in the proteolysis shown in stick representation and labelled. (*b*) Sequence alignment of CT-1 and CT-2 of porcine pepsin with LangC. The DTG triads are highlighted in grey.

arranged in a cruciform shape. Surprisingly, LangC forms a D_3-symmetric ring-like hexamer in which a Bg7S type I dimer undergoes a completely novel higher-order arrangement. One of the interfaces (type II) of the hexameric structure involves β-sheet interactions that extend the intramolecular β-structure of γ-conglutin into an intermolecular context. The crystallographic observation of hexamerization of γ-conglutin has been confirmed by dynamic and static light-scattering experiments in solution, as well as by analytical ultracentrifugation.

In structural comparisons, γ-conglutin shows homology to plant inhibitors of glycoside hydrolases GH11 and GH12. However, detailed analysis reveals that the γ-conglutin lineage has accumulated mutations in the key binding loops IL-1 and IL-2 and has lost any inhibitory activity. The fold of γ-conglutin is also reminiscent of pepsin-like aspartic proteases. However, in this case the key catalytic triads (Asp-Thr-Gly) have also been mutated out, leading to a proteolytically nonfunctional lupin protein. The plant-specific insert of phytepsins (plant aspartic proteases) has lost its identity, despite the preservation of the overall length of the protein sequence. In addition, LangC shows interesting similarities to the cockroach antigen Bla g 2, which is also evolutionarily related to aspartic proteases.

All of the above observations indicate that γ-conglutin is a protein with an intriguing evolutionary history, possibly providing cornerstone ancestry information for protein groups that have not been considered to be related. The fact that this abundant seed protein has no identified physiological function makes it even more intriguing. Its resistance to proteolysis strongly suggests a nonstorage role, despite the provisional classification of γ-conglutin together with seed storage proteins. Finally, the ability of γ-conglutin to bind insulin and to reduce blood glucose make it an attractive candidate as an antidiabetic drug. The accurate structural information presented in this paper will guide further research into the structure and function of γ-conglutin.

The research leading to these results received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under BioStruct-X (grant agreement No. 283570). We thank HZB for the allocation of synchrotron-radiation beamtime. This work was supported by the Ministry of Science and Higher Education (project No. N N312 493340) and National Science Center (DEC-2013/10/M/NZ1/00251). We also thank Ms Katarzyna Dabrowska for technical assistance with the MS analysis of glycosylation.

References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Andreeva, N. S., Zdanov, A. S., Gustchina, A. E. & Fedorov, A. A. (1984). *J. Biol. Chem.* **259**, 11353–11365.
Bryksa, B. C., Bhaumik, P., Magracheva, E., De Moura, D. C., Kurylowicz, M., Zdanov, A., Dutcher, J. R., Wlodawer, A. & Yada, R. Y. (2011). *J. Biol. Chem.* **286**, 28265–28275.
Capraro, J., Spotti, P., Magni, C., Scarafoni, A. & Duranti, M. (2010). *Int. J. Biol. Macromol.* **47**, 502–507.
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.
Czubinski, J., Dwiecki, K., Siger, A., Kachlicki, P., Neunert, G., Lampart-Szczapa, E. & Nogala-Kalucka, M. (2013). *J. Agric. Food Chem.* **60**, 1830–1836.
Czubinski, J., Dwiecki, K., Siger, A., Neunert, G. & Lampart-Szczapa, E. (2014). *Food Chem.* **143**, 418–426.
Duranti, M., Consonni, A., Magni, C., Sessa, F. & Scarafoni, A. (2008). *Trends Food Sci. Technol.* **19**, 624–633.
Duranti, M., Di Cataldo, A., Sessa, F., Scarafoni, A. & Ceciliani, F. (2002). *J. Agric. Food Chem.* **50**, 2029–2033.
Duranti, M., Faoro, F. & Harris, N. (1991). *Protoplasma*, **161**, 104–110.
Duranti, M., Scarafoni, A., Gius, C., Negri, A. & Faoro, F. (1994). *Eur. J. Biochem.* **222**, 387–393.
Egas, C., Lavoura, N., Resende, R., Brito, R. M., Pires, E., de Lima, M. C. & Faro, C. (2000). *J. Biol. Chem.* **275**, 38190–38196.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Esnault, M. A., Citharel, J., Thomas, D., Guegan, P. & Cavalier, A. (1996). *Plant Physiol. Biochem.* **34**, 101–109.
Foley, R. C., Gao, L.-L., Spriggs, A., Soo, L. Y. C., Goggin, D. E., Smith, P. M. C., Atkins, C. A. & Singh, K. B. (2011). *BMC Plant Biol.* **11**, 59.
García de la Torre, J., Huertas, M. L. & Carrasco, B. (2000). *Biophys. J.* **78**, 719–730.
Gustchina, A., Li, M., Wünschmann, S., Chapman, M. D., Pomés, A. & Wlodawer, A. (2005). *J. Mol. Biol.* **348**, 433–444.
Hoorn, R. A. L. van der & Jones, J. (2004). *Curr. Opin. Plant Biol.* **7**, 400–407.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.
Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.
Kervinen, J., Tobin, G. J., Costa, J., Waugh, D. S., Wlodawer, A. & Zdanov, A. (1999). *EMBO J.* **18**, 3947–3955.
Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). *Bioinformatics*, **23**, 2947–2948.
Laue, D. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. (1992). *Analytical Ultracentrifugation in Biochemistry and Polymer Science*, pp. 90–125. Cambridge: Royal Society of Chemistry.
Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. (2014). *Nucleic Acids Res.* **42**, D490–D495.
Magni, C., Sessa, F., Accardo, E., Vanoni, M., Morazzoni, P., Scarafoni, A. & Duranti, M. (2004). *J. Nutr. Biochem.* **15**, 646–650.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
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.
Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. (2009). *Acta Cryst.* D**65**, 1074–1080.
Sansen, S., De Ranter, C., Gebruers, K., Brijs, K., Courtin, C., Delcour, J. & Rabijns, A. (2004). *J. Biol. Chem.* **279**, 36022–36028.
Scarafoni, A., Ronchi, A. & Duranti, M. (2010). *Pytochemistry*, **71**, 142–148.
Schägger, H. & von Jagow, G. (1987). *Anal. Biochem.* **166**, 368–379.
Schiarea, S., Arnoldi, L., Fanelli, R., De Combarieu, E. & Chiabrando, C. (2013). *PLoS One*, **8**, e73906.
Schuck, P. (2000). *Biophys. J.* **78**, 1606–1619.
Schuck, P., Perugini, M. A., Gonzales, N. R., Howlett, G. J. & Schubert, D. (2002). *Biophys. J.* **82**, 1096–1111.
Shewry, P. R., Napier, J. A. & Tatham, A. S. (1995). *Plant Cell*, **7**, 945–956.

Sielecki, A. R., Fedorov, A. A., Boodhoo, A., Andreeva, N. S. & James, M. N. G. (1990). *J. Mol. Biol.* **214**, 143–170.
Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Terruzzi, I., Senesi, P., Magni, P., Montesano, A., Scarafoni, A., Luzi, L. & Duranti, M. (2011). *Nutr. Metab. Cardiovasc. Dis.* **21**, 197–205.
Yoshizawa, T., Shimizu, T., Hirano, H., Sato, M. & Hashimoto, H. (2012). *J. Biol. Chem.* **287**, 18710–18716.
Yoshizawa, T., Shimizu, T., Yamabe, M., Taichi, M., Nishiuchi, Y., Shichijo, N., Unzai, S., Hirano, H., Sato, M. & Hashimoto, H. (2011). *FEBS J.* **278**, 1944–1954.
Zwart, P. H., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *CCP4 Newsl. Protein Crystallogr.* **42**, contribution 8.