Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

PORTLAND PRESS

Research Article

Structure of SPH (self-incompatibility protein homologue) proteins: a widespread family of small, highly stable, secreted proteins

Karthik V. Rajasekar[1], Shuangxi Ji[2,*], Rachel J. Coulthard[3,†], Jon P. Ride[2], Gillian L. Reynolds[4], Peter J. Winn[2], Michael J. Wheeler[5], Eva I. Hyde[2] and Lorna J. Smith[3]

[1]Department of Biochemistry, University of Oxford, Oxford OX1 3QU, U.K.; [2]School of Biosciences, University of Birmingham, Birmingham B15 2TT, U.K.; [3]Department of Chemistry, University of Oxford, Oxford OX1 3QR, U.K.; [4]Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT 59717-3150, U.S.A.; [5]Institute of Science and the Environment, University of Worcester, Worcester WR2 6AJ, U.K.

Correspondence: Eva I. Hyde (e.i.hyde@bham.ac.uk) or Lorna J. Smith (lorna.smith@chem.ox.ac.uk)

SPH (self-incompatibility protein homologue) proteins are a large family of small, disulfide-bonded, secreted proteins, initially found in the self-incompatibility response in the field poppy (*Papaver rhoeas*), but now known to be widely distributed in plants, many containing multiple members of this protein family. Using the Origami strain of *Escherichia coli*, we expressed one member of this family, SPH15 from *Arabidopsis thaliana*, as a folded thioredoxin fusion protein and purified it from the cytosol. The fusion protein was cleaved and characterised by analytical ultracentrifugation, circular dichroism and nuclear magnetic resonance (NMR) spectroscopy. This showed that SPH15 is monomeric and temperature stable, with a β-sandwich structure. The four strands in each sheet have the same topology as the unrelated proteins: human transthyretin, bacterial TssJ and pneumolysin, with no discernible sequence similarity. The NMR-derived structure was compared with a *de novo* model, made using a new deep learning algorithm based on co-evolution/correlated mutations, DeepCDPred, validating the method. The DeepCDPred *de novo* method and homology modelling to SPH15 were then both used to derive models of the 3D structure of the three known PrsS proteins from *P. rhoeas*, which have only 15–18% sequence homology to SPH15. The DeepCDPred method gave models with lower discreet optimised protein energy scores than the homology models. Three loops at one end of the poppy structures are postulated to interact with their respective pollen receptors to instigate programmed cell death in pollen tubes.

*Present address: Shuangxi Ji, 56 Fuhong Road, Chongqing city, Nan'an district 400060, China.
†Present address: Rachel J. Coulthard-Graf, EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany.

Received: 23 October 2018
Revised: 15 February 2019
Accepted: 19 February 2019

Accepted Manuscript online: 19 February 2019
Version of Record published: 12 March 2019

Introduction

Plants express a large number and variety of small, disulfide-bonded, secreted peptides, often acting in defence roles, or in cell signalling. The former group of peptides include plant defensins, thionins, snakins and protease inhibitors, whereas the latter group are important for the control of cell differentiation, such as in the development of seeds, and also for sexual reproduction. While these families of proteins are widespread, their genes have often been missed using conventional genomic algorithms, due to their small open reading frames and high sequence diversity [1,2]. Newer bioinformatics algorithms have found many more of these proteins than identified originally [3,4]. While some members of these families of proteins have been expressed, the presence of multiple disulfide bonds can make them hard to overexpress and purify in soluble form in many host organisms.

One family of proteins that has yet to be structurally characterised is the SPH (self-incompatibility protein homologue) family. The S (self-incompatibility) proteins, now known as PrsS (*Papaver rhoeas* stigma S-determinant) proteins, were discovered in the field poppy, *P. rhoeas* [5]. Self-incompatibility (SI) is used to prevent self-fertilisation in plants. In poppy, this is controlled by a multi-allelic *S*-locus

PORTLAND PRESS

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

which expresses proteins both in pollen and in the stigma. When the *S*-allele of the pollen corresponds to one allele in the stigma, programmed cell death of the pollen tubule is initiated. After the discovery of PrsS proteins in poppy, genes encoding 36 S-protein homologues (SPH proteins) were found from the analysis of the sequence of a large contig from chromosome 4 of *Arabidopsis thaliana* [1], even though *Arabidopsis* is not self-incompatible. The SI systems of other plants, such as the *Brassicaceae* and *Solanaceae*, use different classes of proteins than those used for SI by the *Papaveraceae* (reviewed in refs [6] and [7], respectively), but still contain many SPH family proteins. Currently, 75 members of the SPH family are identified in *A. thaliana* in the Pfam protein domain database (PF05938) [8], with 120 identified using the bioinformatics programme SPADA [3,8]. The proteins appear to be expressed mainly in flowering parts, with some in developing leaves [9]. SPH proteins and domains have now been found in most dicotyledonous plants, with over 1800 homologous sequences identified from 71 species in Pfam [8], and more than 2500 sequences identified in the InterPro protein family database (IPR010264) [10]. While most of the SPH sequences identified to date are from plants, a few similar sequences have been found in fungi (7 sequences) and metazoa (68 sequences) and have been classified in the same protein family [8,10]. The great majority of these proteins contain single domains with signal sequences. The large number of sequences both within a given plant and across a wide range of species, all with signal sequences, suggest that the proteins are involved in many different signalling pathways.

Key features of this family of proteins are their small size with only ∼120 residues, high positive charge, but otherwise quite divergent sequences (Figure 1a). These features are similar to those of the cysteine-rich peptides, but SPH proteins have only 2–6 cysteine residues. Secondary structure predictions suggest that the proteins are composed of 8–9 β-strands and loops but, to date, there is no other structural information about the family. To find out more about this widespread family of proteins, we have overexpressed and purified SPH15, a protein from *A. thaliana* expressed mainly in pollen [11]. We have determined its structure by nuclear magnetic resonance (NMR) spectroscopy, and, from this, predicted the structures of the three known PrsS proteins, the only ones in the SPH protein family with known roles and known receptors.

Experimental

Phylogenetic tree construction

Members of the SPH protein family in *A. thaliana* were predicted via an iterative BLAST search, using the PrsS1 amino acid sequence as the initial search sequence. Model testing and maximum-likelihood tree construction were performed using MEGA 7.0 [12]. Protein sequences were aligned using MUSCLE [13], with default settings. The WAG+G model of evolution was selected based on its BIC (Bayesian information criterion) score and node support values were obtained from 300 bootstrap replicates. No other tree construction parameters were altered.

Sequence and secondary structure prediction

HHblits (version 2.0.16) [14] was used to search for homologous sequences for each query protein sequence from the UniProt sequence database, released in February 2016. The parameter settings used in HHblits were: iteration: 4, *e*-value: 0.001, minimum coverage with the query sequence: 60%, maximum pairwise sequence identity: 90%. Between 900 and 1000 homologous sequences were found for each protein. The 759 sequences that appeared in the hits of all four protein chains (SPH15, SPH1, SPH3 and SPH8) were used to make a multiple sequence alignment (MSA) in HHblits. Based on the MSA, the sequence logo was given by WebLogo [15], and the secondary structure prediction was made using SPIDER2 [16].

Cloning

The signal peptide of SPH15 was predicted using the SignalP web server [17]. The coding sequence of the SPH15 protein, minus the predicted signal peptide, and with a stop codon after the final amino acid, was cloned into pET32b(+) (Novagen), using the *Nco*1 and *Xho*1 sites. Two bases (GC) were inserted after the *Nco*1 site to obtain the correct reading frame. After expression, purification and enterokinase cleavage, this gives SPH15 with three additional amino acids (with sequence AMG) at the N-terminus.

Protein expression and purification

The Origami (λDE3) strain of *Escherichia coli* [18] (Novagen), containing the plasmid pET32b-*SPH15* (expressing a thioredoxin–His_6–SPH15 fusion), was grown at 37°C in LB or, for NMR studies, in M9 media containing

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

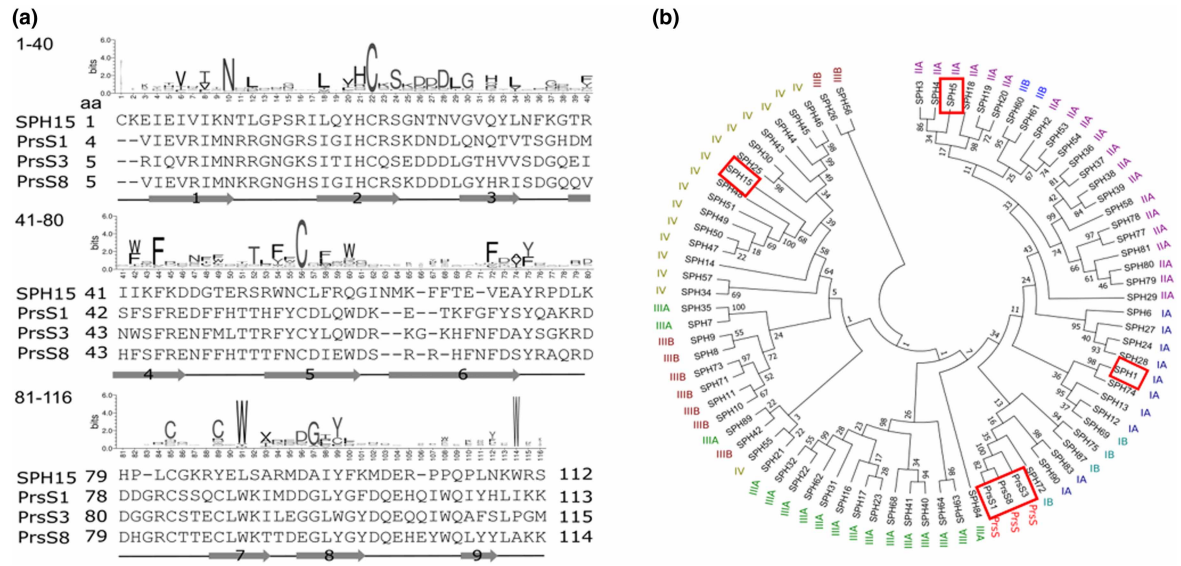

Figure 1. The SPH family of proteins in *A. thaliana* and the PrsS proteins of *P. rhoeas*.

(**a**) Consensus sequence for SPH proteins (top), sequence alignment of SPH15 with PrsS sequences (centre) and (bottom) secondary structure prediction. Top: consensus sequence for 759 proteins homologous to SPH15 and to the three PrsS proteins, from HHblits [14], plotted by WebLogo [15]. Centre: sequence alignment of the SPH domain of SPH15, PrsS1, PrsS3 and PrsS8, from HHblits [14]. For PrsS1, the sequence shown starts at residue 4 and is truncated by seven residues, while for PrsS3 and PrsS8 the sequences start at residue 5 of the mature peptide and are truncated by five and six residues, respectively. Bottom: secondary structure prediction of SPH proteins from SPIDER2 [16], based on the consensus sequence above. (**b**) A maximum-likelihood phylogenetic tree constructed using SPH protein sequences from *A. thaliana* and PrsS protein sequences from *P. rhoeas* (red). Sequences from the BLAST analysis were aligned using MUSCLE [13]. Model selection and tree construction were performed using MEGA7 [12] and the WAG+G model of evolution. Node support values shown on the tree were obtained from 300 bootstrap replicates. Classes are based on the numbers of cysteines (Class I, dark blue and cyan — four cysteines, Class II, purple and blue — five/six cysteines, Class III, green and brown — two cysteines, Class IV, yellow — three cysteines) and the sequence of hydrophilic loop 2 (A — K/RXXD, B — heterogeneous sequence). SPH1, SPH4, SPH8 and SPH15, and the PrsS proteins are outlined with a red box.

50 μg/ml ampicillin, 15 μg/ml kanamycin and 12.5 μg/ml tetracycline. Expression of the protein was induced with 1 mM IPTG at mid-log growth phase. Then, after 30 min, the temperature was reduced to 15°C and the cultures were incubated overnight. The fusion protein was extracted from the cells and purified on a Ni-NTA column using an imidazole gradient in buffer containing 50 mM sodium phosphate (pH 8.0) and 300 mM NaCl, as described recently [19]. It was dialysed into 20 mM Tris-HCl (pH 7.5), 100 mM NaCl, 2 mM $CaCl_2$, followed by cleavage with recombinant enterokinase, at room temperature for 2–3 days.

The cleaved proteins were dialysed into Ni-NTA buffer and re-applied onto the Ni-NTA column. The flow-through from the column, containing SPH15, was kept, and any uncleaved protein and thioredoxin–His_6 fusion eluted from the column using imidazole in Ni-NTA buffer. Alternatively, the cleaved proteins were dialysed into 20 mM sodium phosphate buffer (pH 7.6), 100 mM NaCl, 0.1 mM EDTA and loaded onto a phosphocellulose, P11 column, as described previously [19]. The flow-through samples contained thioredoxin–His_6, while the SPH15 protein was eluted using a 100–600 mM NaCl gradient in the same buffer.

The SPH15 protein was concentrated by ultrafiltration with a 10 kDa cut-off filter and dialysed into the appropriate buffer for subsequent experiments. The concentration of protein was estimated from the absorbance at 280 nm, using a molar absorbance coefficient of 18,700 M^{-1} cm^{-1}, based on its amino acid composition [20].

Analytical ultracentrifugation

AUC (analytical ultracentrifugation) experiments were done using a Beckman ProteomeLab XL-1 analytical ultracentrifuge with an AN50 Titanium rotor at 20°C. Three SPH15 samples at the concentrations, 0.047, 0.028

Biochemical Journal (2019) 476 809–826
https://doi.org/10.1042/BCJ20180828

and 0.005 mM, were used in buffer containing 25 mM sodium phosphate at pH 7.4 and 0.05 mM EDTA. These were spun at 45,000 rpm for 20 h. Absorbance measurements at 280 nm were taken at 10 min intervals and analysed using SEDFIT [21], with υ and ρ set to 0.7327 cm^3 and 1.0017 g/l, respectively.

Circular dichroism spectroscopy

The circular dichroism (CD) spectrum of SPH15 protein at 0.0095 mM, in 25 mM sodium phosphate buffer (pH 7.4), 0.05 mM EDTA, was measured over the range of 190–280 nm, in a 1 mm pathlength cuvette, using a JASCO J-810 CD spectrophotometer at room temperature. The spectrum was subtracted from that of the buffer, taken under the same conditions. The secondary structure was analysed using the CSSTR method [22] in Dichroweb [23].

The CD measurement of a 0.0095 mM SPH15 sample in the buffer above was monitored at 215 nm over a temperature range of 30–90°C at 0.1°C intervals, in a 2 mm cuvette. Far-UV CD spectra of the protein were taken at 5°C intervals. No correction for buffer was applied. The intensity at 215 nm (y) vs. temperature (x) was fitted to a Hill plot with a linear slope, a (eqn 1) and the midpoint (EC_{50}) determined by non-linear regression in Sigmaplot12.

$$y = y_{\min} + ax + \frac{y_{\max} - y_{\min}}{1 + \left(\frac{x}{EC_{50}}\right)^{-\text{Hillslope}}} \quad (1)$$

NMR spectroscopy

The NMR spectra of ^{15}N-, ^{13}C-labelled SPH15 were taken and assigned using triple resonance methods, as described recently [19].

Distance restraints were derived from 3D NOESY-^{15}N HSQC in H_2O and NOESY-^{13}C HSQC in D_2O, both with 100 ms mixing time, acquired on a Varian 800 MHz spectrometer with a room temperature TXI probe using a ~1 mM protein sample in 20 mM sodium phosphate buffer (pH 5.2), 50 mM NaCl, 0.1 mM EDTA. An additional, higher resolution, NOESY-^{13}C HSQC spectrum was acquired using non-linear sampling, on a Bruker 600 MHz spectrometer, with a TXI cryoprobe.

Backbone φ and ψ torsion angles were estimated from Cα, Cβ, C′, N and Hα chemical shifts using the program DANGLE [24]. Hydrogen-bond restraints were based on slowly exchanging amides identified in ^1H-^{15}N HSQC spectra taken at a series of intervals up to several weeks after lyophilisation of the protein and dissolving it in D_2O. Once the basic fold of the protein had been determined, disulfide bond restraints between the two pairs of cysteine residues were added to the structural calculations.

Structures were calculated using the program ARIA 2.3 [25] interfaced to CNS 1.2 [26]. A total of 200 structures were calculated using ambiguous and unambiguous NOE assignments. Assignment ambiguity was reduced during successive iterations in ARIA. Automatic NOE assignments from ARIA were checked manually and the protocol was repeated until there were no validation errors. Finally, the 20 structures with the lowest energies were refined in water. The family of structures was validated using the program PROCHECK [27].

Structure prediction of PrsS proteins

DeepCDPred [28] was used to generate structures for SPH15, PrsS1, PrsS3 and PrsS8. DeepCDPred uses deep learning to predict contacts and distances between residues, based on many inputs including amino acid profile, secondary structure prediction from SPIDER2 [16] and amino acid co-evolution couplings. The couplings are predicted from (i) mutual information [29], (ii) mean-field direct coupling analysis [30], (iii) QUIC [31], (iv) pseudo-likelihood direct coupling analysis [32] and (v) statistical potential [33]. In addition, the programme uses the number of amino acids in the target protein, the number of homologous sequences in the MSA built by HHblits [14] and an estimate of the number of non-redundant sequences in the alignment. DeepCDPred [28] also has a β-sheet prediction algorithm that provides hydrogen-bonding restraints between strands. These restraints, together with restraints to enforce the secondary structure prediction from SPIDER2 [16], are fed into the protein structure modelling program, AbinitioRelax, which is from the Rosetta suite [34]. Three-residue and nine-residue structure fragments required by AbinitioRelax were generated by the Perl script make_fragments.pl from the Rosetta suite. One hundred candidate structures were generated using the protocol. The structure with a lowest Rosetta energy score was chosen as the final model and the TM (template

Biochemical Journal (2019) 476 809–826
https://doi.org/10.1042/BCJ20180828

modelling) score [35] estimated, using a predictor that is part of the DeepCDPred. The best *de novo* (DeepCDPred) model for each protein was further refined using ReFOLD [36].

MODELLER [37] was used to model the poppy proteins, based on the structure of SPH15, with two different methods of sequence alignment. In the first method, HHblits [12] was used to align each of the proteins to SPH15, based on the sequence alone. In the second method, a structural alignment was determined from the structural predictions from DeepCDPred for each protein with the NMR structure of SPH15, using TM alignment [29]. For this second method, MODELLER was used with the SPH15 template either with or without the predicted contact restraints from DeepCDPred. For each of the three proteins, 500 models were calculated by each of these three comparative modelling methods and their DOPE (discreet optimised protein energy) scores [38] were calculated and compared with those from the *de novo* DeepCDPred structure.

Results and discussion

The sequence alignment of the SPH domain of SPH15 from *A. thaliana* and the three PrsS proteins from *P. rhoeas* is shown in Figure 1a, together with the consensus from 759 sequences from HHblits [14] and the secondary structure prediction, showing nine β-strands, from SPIDER2 [16]. The original aim was to determine the structure of one of the PrsS proteins; however, expression of the mature PrsS1 and PrsS3 proteins in *E. coli* gave inclusion bodies, which, after solubilisation, gave very low yields of protein. Our attention, therefore, turned to the *SPH* proteins from *A. thaliana*.

Phylogenetic analysis of SPH proteins in *Arabidopsis thaliana*

An iterative BLAST search of *A. thaliana*, using the PrsS1 amino acid sequence as the initial search sequence, followed by successive BLAST searches with identified SPH amino acid sequences, yielded 92 members of the SPH protein family, more than in the Pfam database [8] but fewer than in SPADA [3]. The 92 proteins were grouped into four classes, based on the number of cysteines. Maximum-likelihood phylogenetic analysis of these sequences with the PrsS proteins of *P. rhoeas*, using MEGA7 [12], shows that all of these proteins are evolutionarily related and can be resolved into the same four classes (Figure 1b). Attempts to include other potential SPH proteins identified using SPADA [3] gave trees that were poorly resolved, with very poor bootstrap results, that displaced members of the original 92 identified proteins that share an obvious identity. This suggests that these additional proteins may have evolved independently; alternatively, the sequence has diverged so far as to render phylogenetic analysis for the entire SPH cohort including these additional proteins impossible. In contrast, *P. rhoeas* PrsS proteins and homologues from *Selaginella moellendorffii* and *Physcomitrella patens* identified using BLAST do not disrupt the phylogenetic tree generation (data not shown). The PrsS proteins of *P. rhoeas* group within the Class I subfamily (Figure 1b), containing four cysteines in strands 2, 5 and 7, and in loop 6 (Figure 1a). Class II proteins have either two additional cysteines, in strands 8 and 9, or are truncated in strand 9 and so have only five cysteines. Class III proteins have only two cysteines in strands 2 and 5, whereas Class IV proteins have three cysteines on strands 2 and 5 and in loop 6. These classes fit the phylogenetic analysis well and it is evident from the analysis that Class II proteins are derived from those in Class I, whereas Class IV proteins (including SPH15) cluster within the Class III proteins. The minor classification (1A, 1B, etc.) refers to the sequence of hydrophilic loop 2 where those in subclass A have the motif K/RXXD, whereas those in subclass B are heterogeneous.

SPH15 protein expression and purification

SPH1 in Class 1A, SPH4 in Class IIA, SPH8 in Class IIIB and SPH15 in Class IV were selected for initial trials for expression in *E. coli* as they are relatively distant from each other on phylogenetic analyses and represent the different patterns of conserved cysteine residues observed in sequence alignments (Figure 1a). Initial attempts to express the mature proteins in *E. coli* using the methods used for the poppy proteins also led to inclusion bodies, but SPH15 gave some soluble protein.

Since the cause of the insolubility might be the lack of formation of disulfide bonds in normal strains of *E. coli*, expression trials moved to expression from pET32b in the AD494 strain of *E. coli* which contains a knockout mutation of the thioredoxin reductase (*trxB*) gene, thereby enhancing disulfide bond formation. The section of the *SPH15* gene that encodes the predicted mature protein was cloned between the *Nco*1 and *Xho*1 sites of pET32b(+), thus expressing a thioredoxin–His_6–SPH15 fusion, with an enterokinase cleavage site between the His tag and SPH15. This approach, with both overexpression of thioredoxin, and having the *trxB* mutation,

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

greatly improved the yield of soluble SPH15 protein. The yield was increased further by using the Origami strain of *E. coli*, which contains mutations in both thioredoxin reductase and glutathione reductase [18].

The fusion protein was purified using a nickel affinity column for the His_6 tag (Figure 2a, i), which could also be used for separation of the two proteins produced after cleavage with enterokinase. For large volumes, after enterokinase cleavage, the proteins were separated on a phosphocellulose, cation exchange column. SPH15 contains a large number of basic residues and has a predicted pI 9.7, and so binds to the column at neutral pH, while His-tagged thioredoxin has a predicted pI of 4.7 and flows through the column without binding (Figure 2a, ii). The SPH15 produced has no signal sequence but contains three additional amino acids, AMG, at its N-terminus.

SPH15 protein characterisation and 3D structure

The molecular mass and sedimentation coefficient of SPH15 were determined by AUC (Figure 2b). The sedimentation coefficient calculated, 1.68 S, is slightly greater than expected for a spherical protein of a similar

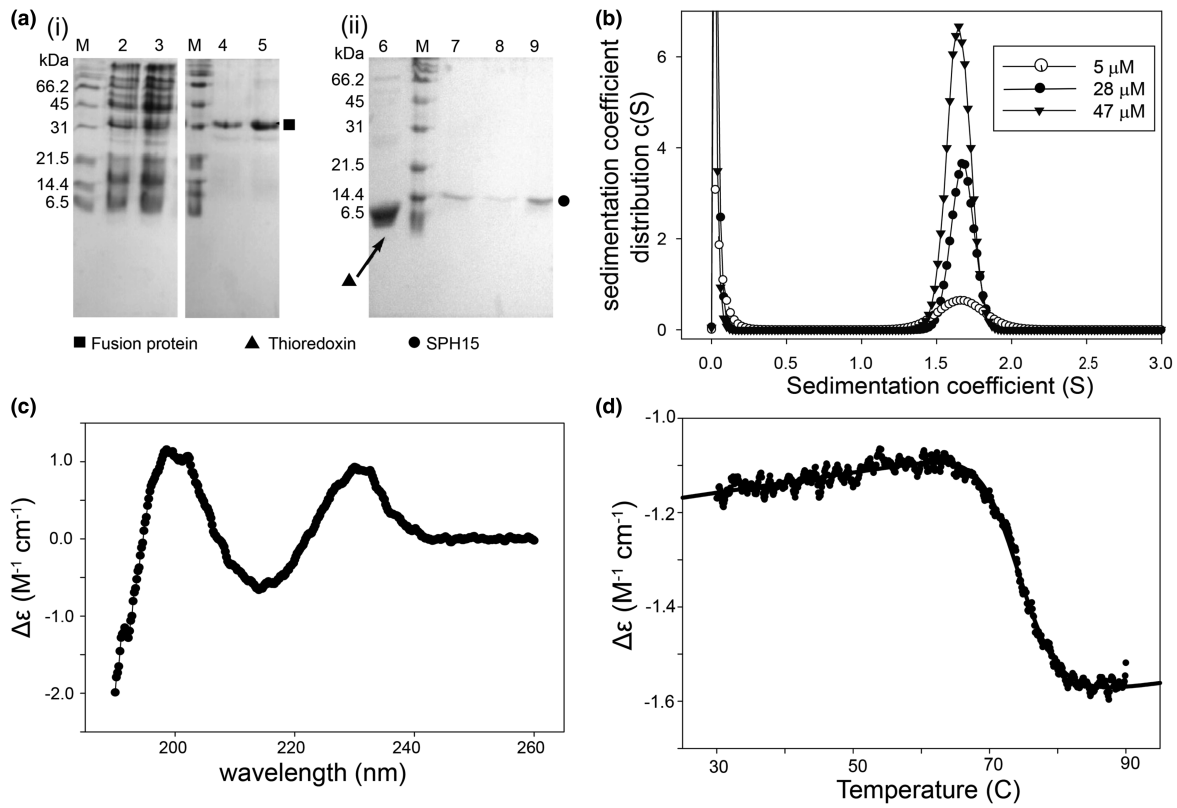

Figure 2. Purification, AUC analysis and CD spectra of SPH15.

(**a**) 15% SDS–PAGE analysis of samples from the purification of SPH15, stained with Coomassie Brilliant Blue. M-protein markers, sizes indicated in the left of figures. (i) Samples from the Ni-NTA column, on the clarified, sonicated, crude protein sample. Lane 2 — protein applied to the column, lane 3 — unbound protein, lanes 4 and 5 — samples from the imidazole gradient. The square symbol indicates the position of the purified thioredoxin–SPH15 fusion protein on the gel. The fusion protein elutes at ~70 mM imidazole. (ii) Samples from the phosphocellulose column, after cleavage of the fusion protein from the Ni-NTA column with enterokinase. Lane 6 — unbound protein, lanes 7–9 — samples from the NaCl gradient. The labels indicate the positions of thioredoxin (triangle) and SPH15 (circle), on the gel. SPH15 elutes at ~350 mM NaCl. (**b**) Sedimentation distribution plot from AUC measurements for SPH15. SPH15 samples were used at concentrations between 0.1 and 1 mg/ml in 25 mM sodium phosphate at pH 7.4 and 0.05 mM EDTA buffer, at 20°C. Absorbance measurements at 280 nm were analysed using SEDFIT [21]. (**c**) Far-UV CD spectrum of SPH15, in 25 mM sodium phosphate (pH 7.4), 0.05 mM EDTA buffer at room temperature. (**d**) The CD signal intensity of SPH15 at 215 nm, in 25 mM sodium phosphate (pH 7.4), 0.05 mM EDTA buffer, over the temperature range 30–90°C. The points show the measurements and the line shows the fit to a Hill curve with a slope (eqn 1).

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

mass, indicating a slightly extended shape ($f/f_0 = 1.26$), and gave an estimated molecular mass of 14.7 kDa, compared with the calculated mass of 13.5 kDa. The sedimentation coefficient was constant for the three concentrations measured, indicating that the protein does not aggregate in the 5–47 μM concentration range and is monomeric.

The CD spectrum of SPH15 in the far-UV shows a minimum at 217.5 nm and a maximum at 198 nm, consistent with it having the β-sheet secondary structure (Figure 2c). Analysis of the secondary structure using the CSSTR method [22] in Dichroweb [23] indicates that it contains 47% β-strand, 24% turn and 28% disordered structure. CD spectra were measured at a series of temperatures. The signal at 215 nm, indicating the extent of secondary structure, showed an initial slow decrease in ellipticity, and then rapid denaturation over ∼10°C, with a midpoint at 75°C (Figure 2d), showing that the protein unfolds co-operatively and is thermally very stable.

The ^{1}H-^{15}N HSQC spectrum of SPH15 (Figure 3a) is well resolved, with a single peak for each amino acid (105 peaks out of 108 expected), showing that the protein is folded and has a single conformation. The NMR spectrum of the protein was assigned by standard triple resonance methods [19]. The secondary ^{13}C, ^{15}N and ^{1}H chemical shifts [39] confirmed that the protein has mainly β-sheet conformation, as predicted [1] and shown by the CD spectrum. The structure was determined from interproton distance restraints estimated from NOESY spectra, complemented by analysis of the H–D exchange rates of the backbone NH groups. Many of the NH signals remain visible for several days after dissolving the protein in D_2O, showing the high persistence

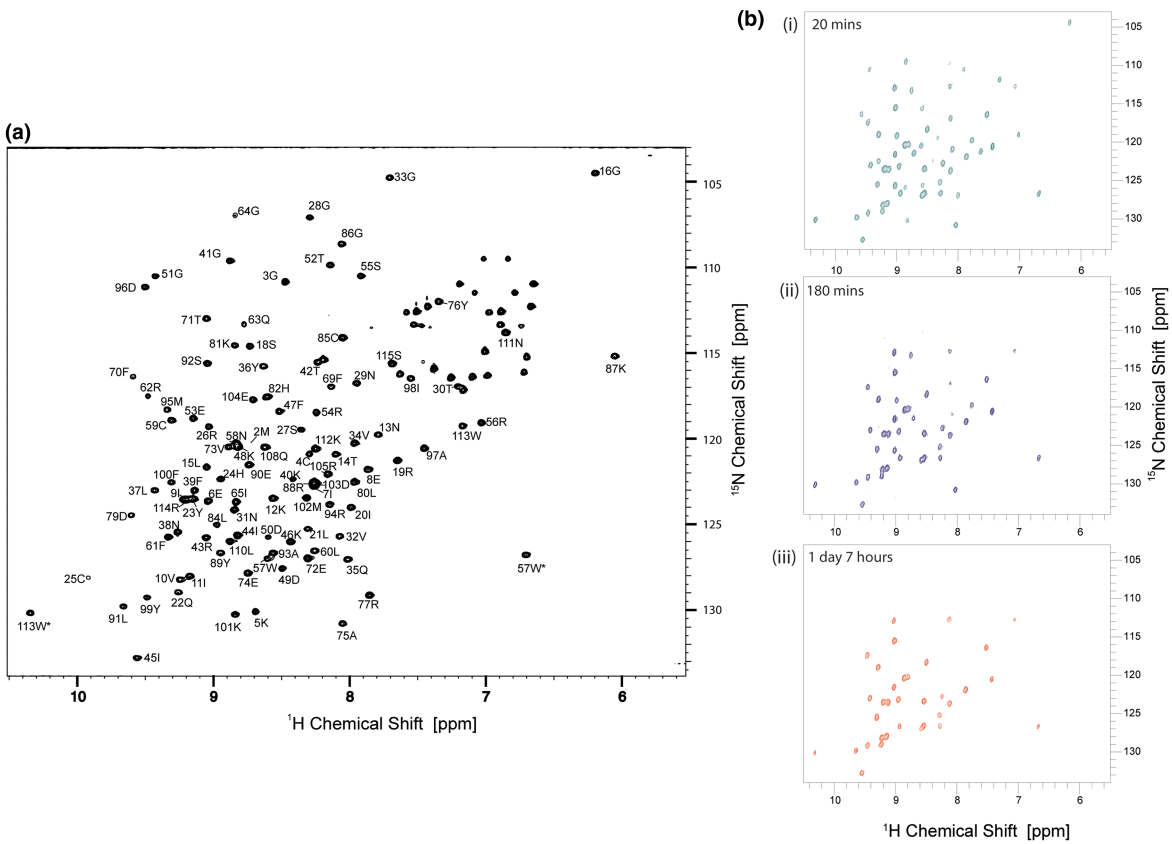

Figure 3. ^{1}H-^{15}N HSQC spectra of SPH15.

(**a**) The ^{1}H-^{15}N HSQC NMR spectrum of ∼1 mM SPH15 in 20 mM sodium phosphate buffer (pH 5.2), 50 mM NaCl, 0.1 mM EDTA, at 25°C, acquired on a Varian 800 MHz spectrometer. Assignments are shown with the residues numbered from the cloned sequence, i.e., including the three N-terminal residues from the vector [19]. (**b**) ^{1}H-^{15}N HSQC NMR spectra of SPH15 taken at intervals after lyophilisation and dissolving the sample in D_2O, under the conditions in (**a**), but acquired on a Bruker 500 MHz spectrometer. (i) Turquoise-spectrum taken after 20 min (0.33 h); (ii) dark blue-spectrum taken after 180 min (3 h) and (iii) orange-spectrum taken after 1 day, 7 h (31 h).

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

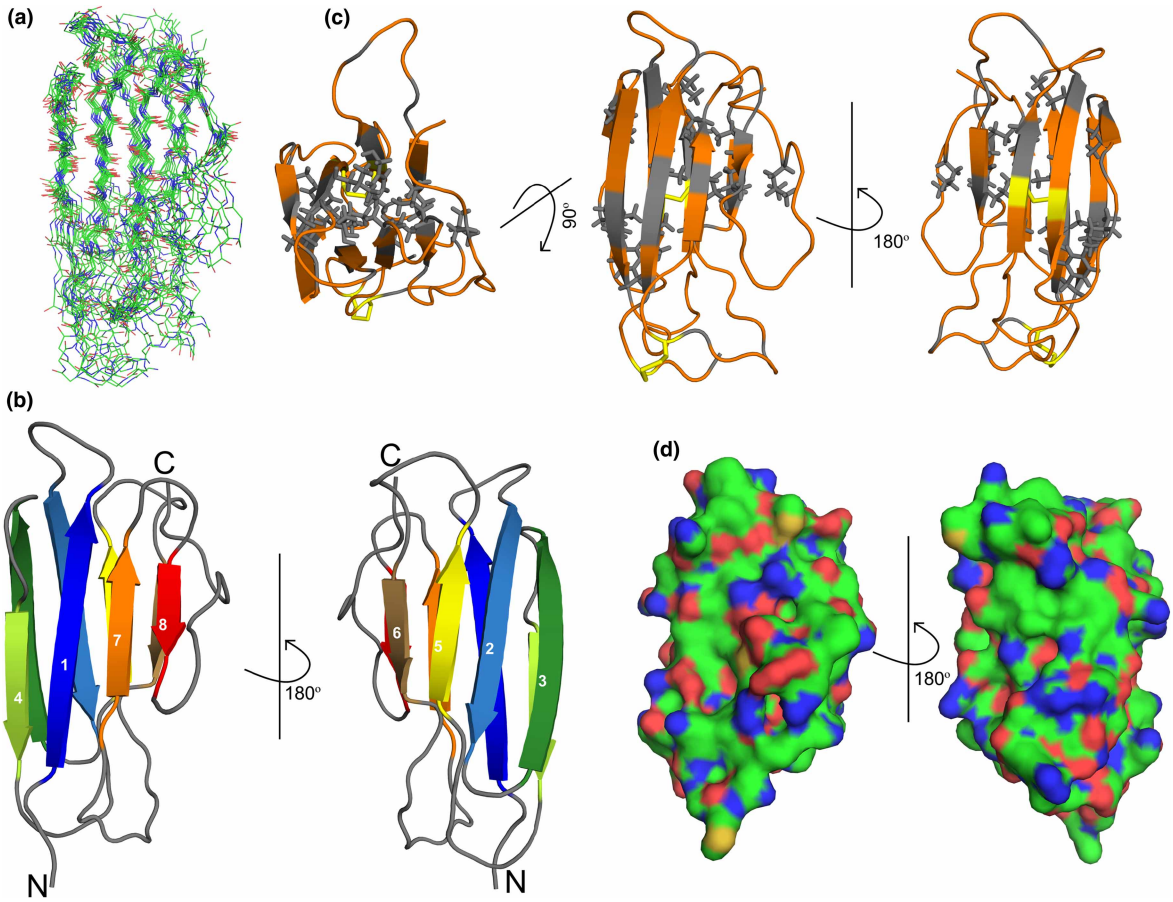

Figure 4. 3D structure of SPH15.

(**a**) Overlay of the backbones of the 20 SPH15 structures with the lowest energies calculated from the NMR restraints. The protein backbone is coloured green for Cα, blue for NH and red for CO. (**b**) Ribbon representation of two orientations of the lowest energy structure of SPH15. The termini are labelled and the strands are numbered and coloured in rainbow colours, from blue to red, from N- to C-terminus. (**c**) Ribbon representation of the lowest energy structure in three orientations, showing the hydrophobic residues and hydrophobic side chains in grey and the disulfide bonds in yellow. (**d**) Surface representation of the protein, in the same two orientations as in **b**, with hydrophobic residues as green, basic residues as blue, acidic residues as red and sulfur as yellow.

of the NH bonds (Figure 3b). The disulfide-bonding pattern predicted previously from sequence alignments [1] was confirmed by NOEs between the Cβ protons of the bonded cysteine residues. Figure 4a shows an overlay of the backbone structures of the 20 lowest energy structures calculated from the NMR data, with a ribbon diagram of the lowest energy structure in two orientations in Figure 4b. The structure is generally well determined, with the backbone RMSD of the top 20 structures from the mean structure of 1.6 Å for the β-strands; however, some of the loops are less well-defined, leading to an overall RMSD of the backbone of 2.3 Å (Table 1). In particular, residues 48–51 and 75–80, at the tips of the longer loops, loop 4 and loop 6, respectively, have backbone RMSDs >4 Å, possibly due to flexibility.

SPH15 contains eight β-strands arranged in two, four-stranded sheets, as a sandwich. The strands are mainly in agreement with the secondary structure prediction from the MSA, shown in Figure 1a. The last strand predicted in Figure 1 contains only three residues, with a central proline, and is irregular, but shown in some of the models. In the 3D structure, two of the strands in one sheet, namely strands 1 and 7, are parallel, while the other strands are all antiparallel (Figure 4b,c). There is a large hydrophobic core containing residues from each strand (Figure 4c), many of which are conserved as hydrophobic residues in the consensus sequence (Figure 1a). In SPH15, Cys 1, at the N-terminus, is bonded to Cys 81, in loop 6, keeping strand 1 close to

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

Table 1 NMR restraints and refinement statistics for top 20 NMR models of SPH15

NMR distance and dihedral restraints	
Distance restraints	
Total NOEs:	
Unambiguous	2278
Ambiguous	339
Intra-residue NOEs	1255
Inter-residue NOEs:	
Sequential ($\|i-j\|=1$)	511
Short ($\|i-j\|=2$–3)	44
Medium-range ($\|i-j\|=4$–5)	22
Long-range ($\|i-j\|>5$)	446
Hydrogen bonds	52
Total dihedral angle restraints	
φ	63
ψ	63
Structure statistics for top 20 structures	
Number of dihedral angle violations (>5°)	0
Number of distance restraint violations (>0.5 Å)	0
Deviations from idealised geometry	
Bond lengths (Å)	0.00808 ± 0.00049
Bond angles (°)	0.891 ± 0.058
Impropers (°)	2.7 ± 0.3
Average RMSD to the mean structure for top 20 structures (Å)	
β-strand residues, heavy atoms *	2.52 ± 0.35
β-strand residues, backbone (Cα, N, C′)*	1.55 ± 0.27
All residues, heavy atoms	3.01 ± 0.38
All residues, backbone	2.26 ± 0.39
Ramachandran analysis of top structure (%)	
Residues in most favoured regions	63.6
Residues in additional allowed regions	30.3
Residues in generously allowed regions	6.1
Residues in disallowed regions	0.0

*Statistics applied to residues 4–11, 17–24, 31–35, 37–44, 54–59, 66–71, 86–91 and 94–99 (numbered from Sph15 Cysteine 1, as residue 1).

strand 7, while Cys 21, on strand 2, is bonded to Cys 55, on strand 5, these being neighbouring strands within the same sheet (Figure 4c). While the protein is overall highly basic, with 21 positively charged residues and 12 acidic residues, the charges are distributed over the whole of the surface of the protein (Figure 4d). The overall shape of SPH15 is similar to that of protein domains with a Greek key fold, such as the immunoglobulin constant domain; however, the topology of the strands differs as, in the latter, all strands are antiparallel, and the single disulfide bond is between the two sheets.

Comparison of SPH15 protein structure with other proteins of known structure

Comparison of the structure of SPH15 with other proteins in the protein structure database using DALI [40] gave, as hits, many proteins with a Greek key topology. The top hits containing domains with the same topology

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

as SPH15 were the membrane-binding domains (domain 4) of the cytotoxic proteins pneumolysin from *Streptococcus pneumoniae* [41,42] and perfringolysin from *Clostridium perfringens* [43], with *Z*-scores >6.5 and small RMSDs (2.8 Å) to SPH15, despite negligible sequence identity (~8–12%) (Figure 5a). A bacterial-type VI secretion protein, TssJ, from a pathogenic strain of *E. coli* [44] was also identified as containing a domain with the same topology, with a *Z*-score of 6.4 and an RMSD of 3.1 Å (Figure 5b). One entire protein identified with the same topology as SPH15 was human transthyretin [45], with a *Z*-score of ~6 and RMSD ~3.5 Å (Figure 5c). Again, neither TssJ nor transthyretin shows any discernible sequence similarity to SPH15.

The proteins identified have very different functions, oligomeric states, and interact with their partners using different surfaces in each case. Perfringolysin and pneumolysin belong to a group of toxins from Gram-negative organisms that kill cells by forming large circular aggregates that make holes in eukaryotic membranes [46,47] (Figure 5a). TssJ is an outer membrane lipoprotein that forms part of the outer shell of the type VI secretion system from *E. coli* and hence also an important virulence factor [44]. The syringe-like, type VI secretion complex, contains 10 copies of TssJ and its partners [48]. In contrast, transthyretin is a homotetramer that is involved in the transport of thyroid hormone and, separately, using a different interface, transport of retinol-binding protein bound to retinoic acid [49,50] (Figure 5c). Given the differences in oligomeric state and interaction surfaces in each case, little can be deduced from these proteins about how S-proteins may function.

De novo structure prediction of SPH15

The only proteins in the SPH family with known functions and interaction partners are three poppy PrsS proteins, PrsS1, PrsS3 and PrsS8, involved in the SI response. The corresponding receptor proteins, PrpS1, PrpS3 and PrpS8 (*P. rhoeas*, pollen S-determinant) [51], are ~20 kDa transmembrane proteins. The three secreted

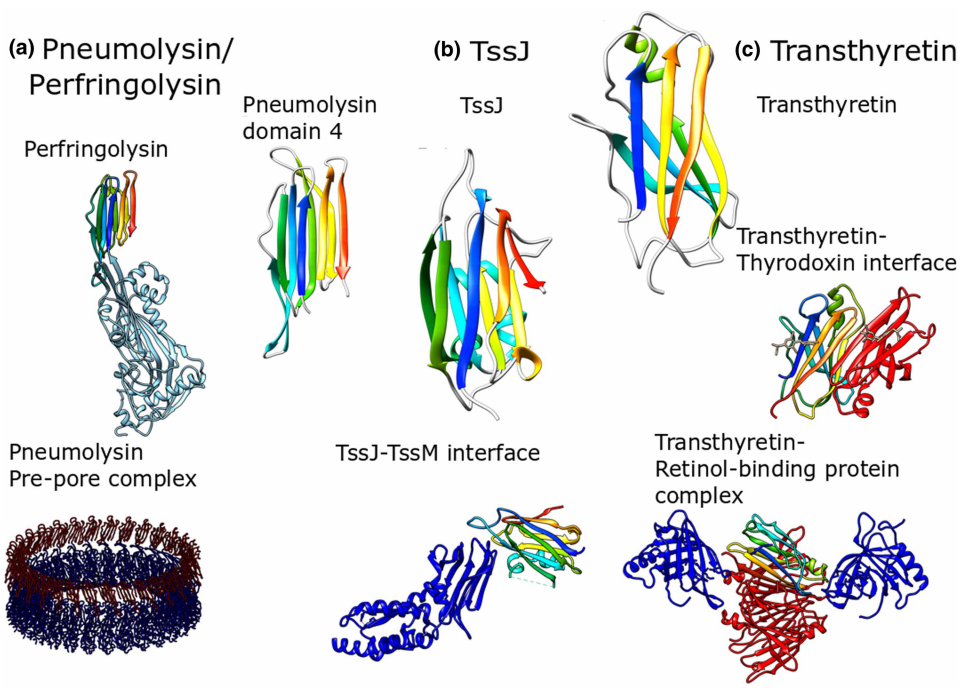

Figure 5. 3D structures of proteins with the same fold as SPH15 and some of their complexes.

(**a**) Pneumolysin domain 4, from PBD 5CR8 [42], perfringolysin full monomer, from 1PFO [46], pneumolysin prepore complex, from 2BK2 (from cryoEM) [47]. (**b**) TssJ from PDB 4Y7O [48] and the monomer of the TssJ–TssM complex from PDB 3RX9 [44]. The dotted line connects residues in TssJ that were not observed in the structure of the complex. (**c**) Transthyretin monomer and the transthyretin–thyrodoxin dimer interface, both from PDB 2ROX [49], and the transthyretin homotetramer complexed with retinol-binding protein from PDB 1RLB [50]. Each monomer/domain is shown as a ribbon structure, coloured in rainbow colours from blue to red, from N- to C-terminus, with strand 1 in a similar orientation to SPH15 in Figure 4a. In the complexes, one monomer is coloured in rainbow colours and, where relevant, other monomers of the same protein/domain are coloured red, the partner proteins are coloured blue.

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

PrsS proteins show ~60% sequence identity and ~75% sequence similarity to each other, with a similar level of sequence identity between the receptor proteins, across the length of the sequences. While SPH15 and the PrsS proteins clearly belong to the same protein family, they are in different classes within the family and there is only 15–18% sequence identity between them, rather small for reliable homology modelling (Figure 1a,b). However, the relatively large number of proteins in this family enabled us to test a newly developed *de novo* structure prediction method, DeepCDPred [28], based on co-evolution/correlated mutation to model SPH15. We then used this method to predict structures of the three poppy proteins.

The *de novo* prediction of the structure of SPH15 using the DeepCDPred method was made solely from the MSA of the protein and homologous sequences, independently of the NMR measurements. The method uses deep learning to predict contacts and distances between residues and then produces a series of structural models. Figure 6a shows the *de novo* structure of SPH15 with the lowest DOPE score [38], in a similar orientation to that of the NMR-derived structure in Figure 4a. The two structures are very similar with a TM score of 0.62 and an RMSD of 3.34 Å for all Cα atoms and 2.91 Å for the Cα atoms of residues in the β-strands. One major difference between the two structures is that in the *de novo* structure the β-strands are slightly longer, with more regular geometry, hence showing the short ninth strand, antiparallel to strand 8. The *de novo* structure also shows one turn of an α helix in loop 4, not seen in the NMR measurements.

To examine the precision of the *de novo* structure prediction, the five models with the lowest Rosetta energies were aligned and the differences between their Cα positions determined compared with the restraint strength at each position derived from the co-evolution map (Figure 6b). The differences in structures are below 3 Å, apart from the N-terminal three amino acids, and the regions 75–85 and 100–104, where there are no restraints. For the top 10 predicted SPH15 structures, the pairwise RMSD of the Cα atoms for residues in the β-strands is 3.1 Å, and it is 5.3 Å for all Cα atoms (Table 2); somewhat larger than the RMSDs between the top 20 NMR structures (Table 1). The lack of restraints between residues 75–85 could arise if this region of the protein was conserved across homologues, or if it were flexible so that amino acid substitutions in one part of the loop do not allow prediction of substitutions in other parts of the loop. The latter is consistent with the NMR structure, where the residues at the tip of loop 6 are poorly constrained.

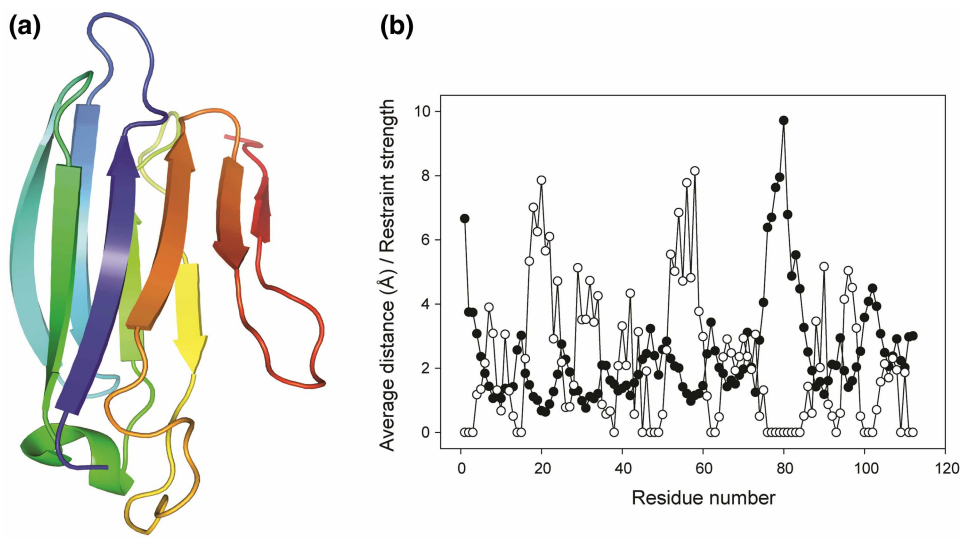

Figure 6. *De novo* structure predictions of SPH15.

(**a**) Ribbon representation of the *de novo* structure prediction of SPH15 from DeepCDPred with the lowest DOPE score in the same orientation as that of the NMR-derived structure in Figure 4a. The structure is coloured in rainbow colours, from blue to red, from N- to C-terminus. (**b**) Differences in Cα positions of the top five models of SPH15 from DeepCDPred, compared with the restraint strength. The five *de novo* SPH15 structures with the lowest energies were aligned and the average distance between the corresponding Cα atoms was calculated (black circles, solid lines). The restraint strength is the sum of the predictions of contacts at each position (white circles, dashed lines). The regions with more restraints are more converged (lower average distance).

PORTLAND PRESS

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

Table 2 RMSD between pairs of top 10 *de novo* calculated structures of SPH15, PrsS1, PrsS3 and PrsS8

Pairwise RMSDs between top 10 *de novo* DeepCDPred models, with the lowest DOPE scores.
Lower half, bold, RMSD between all Cα atoms.
Upper half, italics, RMSD for residues in β-strands only (as defined by the multiple sequence alignment in Figure 1).
The diagonal (grey boxes) shows the RMSD for 10 models within the same protein, whereas off-diagonal numbers show the RMSDs between 10 models of each protein.

	SPH15	PrsS1	PrsS3	PrsS8
SPH15	*3.1 ± 0.3 Å* **5.3 ± 0.8 Å**	*4.9 ± 1.2 Å*	*4.7 ± 1.6 Å*	*4.0 ± 0.7 Å*
PrsS1	**6.6 ± 1.0 Å**	*5.2 ± 1.6 Å* **6.6 ± 1.4 Å**	*5.4 ± 1.6 Å*	*4.7 ± 1.2 Å*
PrsS3	**6.2 ± 1.3 Å**	**6.4 ± 1.4 Å**	*4.8 ± 2.1 Å* **6.0 ± 1.9 Å**	*4.9 ± 1.5 Å*
PrsS8	**5.9 ± 0.8 Å**	**6.0 ± 1.3 Å**	**6.2 ± 1.5 Å**	*3.5 ± 0.9 Å* **4.9 ± 1.1 Å**

Structure prediction of the poppy SI proteins, PrsS1, PrsS3 and PrsS8

The similarity of the fold of the *de novo* prediction to the NMR-determined structure, and the ability to estimate errors in the structure predictions, gave us confidence to use this method to model each of the related proteins, PrsS1, PrsS3 and PrsS8, separately. To improve the structural predictions, 3–7 amino acids at the N- and C-terminals of each of the proteins were not included in the calculations. These amino acids give no restraints and are likely to be highly flexible. This left the core of ∼110 amino acids for each protein shown in Figure 1a, for which the structure was calculated, without disulfide bond constraints. As each protein has a different length and sequence, each has a slightly different set of homologous proteins within the SPH family. For each, over 900 homologous sequences were found using HHblits [14], with 759 sequences found in common in homology searches of the three proteins and SPH15. The different set of homologous proteins gives a slightly different co-evolution contact map for each PrsS protein, but, given the large number of common sequences, the contact maps and the overall predicted secondary structures are similar.

As a comparison of the DeepCDPred method, we also used MODELLER [37] to model the poppy proteins based on the NMR structure of SPH15. Two different methods of the sequence alignment to the NMR structure of SPH15 were used; either solely based on the HHblits sequence alignment [14], as in Figure 1a, or alignment using the structure prediction of each of the proteins calculated from DeepCDPred, with the program TM-align [35]. The TM alignment was then used in MODELLER with or without the predicted contact restraints from DeepCDPred. For each of the three PrsS proteins, 500 models were calculated in MODELLER by each of these three comparative modelling methods. The DOPE score [38] of each model was calculated and compared with the DOPE scores of 100 models from the *de novo* DeepCDPred calculations (Figure 7). For each of the three comparative modelling methods, MODELLER gave similar structures with a narrow range of DOPE scores. In contrast, the *de novo* DeepCDPred models had a much wider range of DOPE scores, a few of the structures had poor DOPE scores, but the mean DOPE scores were all more negative (better) than those from MODELLER. In all three proteins, the lowest DOPE scores for the DeepCDPred models were very much lower than the lowest energy models from MODELLER. For PrsS1 and PrsS8, the entire interquartile range of scores for the DeepCDPred models was lower than the lower quartile of all the other methods, suggesting more stable structures. The *de novo* DeepCDPred model with the lowest DOPE score was taken to be the most representative structure for each protein and is shown in Figure 8.

The topology of all three PrsS proteins predicted from DeepCDPred (Figure 8) is the same as that of SPH15; the hydrophobic core is maintained, but the exact lengths and regularity of the β-strands and the orientation of the loops vary. In particular, as for the SPH15 structures, loop 6, between strands 6 and 7, is poorly constrained and varies in the models, as does the length and orientation of the C-terminal region, which has a longer final β-strand, strand 9, particularly in PrsS1, than in SPH15. Table 2 shows the pairwise distance distribution

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

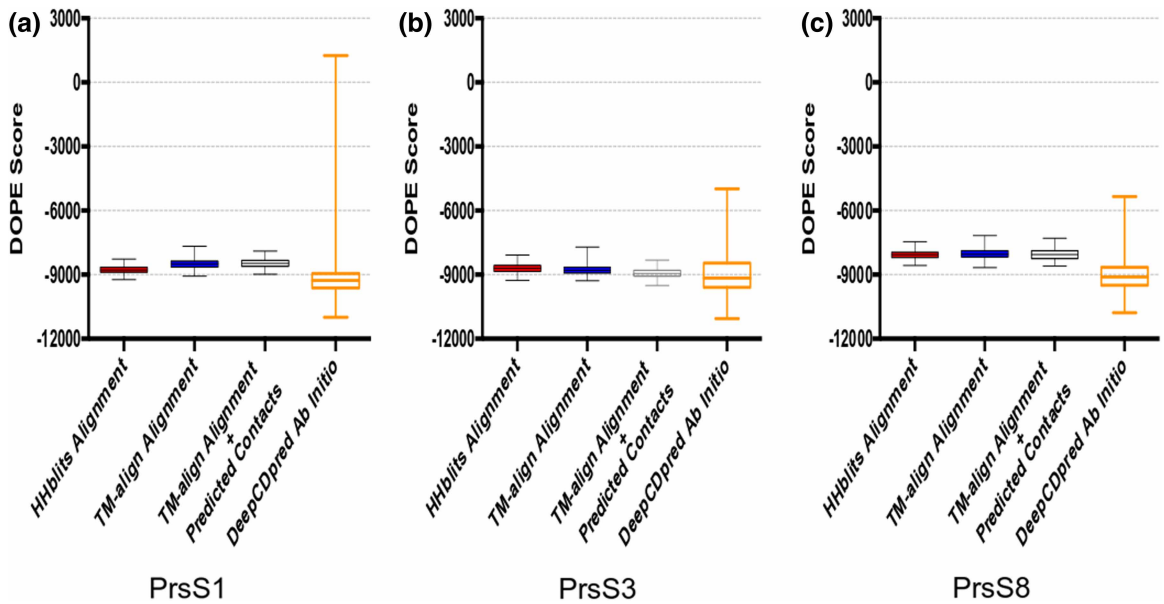

Figure 7. Comparison of DOPE scores for three different comparative modelling methods and *de novo* DeepCDPred modelling of the PrsS proteins.

(**a**) PrsS1 (left), (**b**) PrsS3 (centre) and (**c**) PrsS8 (right). For the comparative modelling (black outlines), MODELLER [37] was used, based on the structure of SPH15, and DOPE scores [38] for 500 structures were calculated. For each protein, left histogram (red): HHblits [14] alignment of protein sequence with SPH15 followed by MODELLER. Centre histograms: structural alignment of the NMR template structure with the structural predictions from DeepCDPred [28], using TM-align [35], either without (blue) or with (white) the predicted contact restraints from DeepCDPred, followed by MODELLER. Right histogram (yellow outline): DOPE scores of 100 models from the *de novo* DeepCDPred calculation.

between the models with the 10 lowest DOPE scores, both within predictions of the same protein and comparing the models of the different poppy proteins and SPH15. The average pairwise RMSDs between models of the same protein (4.9–6.6 Å for all atoms) and between models of the different proteins (5.9–6.6 Å) are similar. This suggests that the four proteins are very similar in structure to each other, within error, despite the disulfide-bonding pattern in SPH15 (Class IV) being different from that in the poppy proteins (Class I) (Figure 1). One of the disulfide bonds in SPH15, that between Cys 22 in strand 2 and Cys56 in strand 5, is conserved throughout all the SPH proteins and the two cysteines are close in all four structures. Cys 82 in loop 6 is conserved in all four proteins, but, in the poppy proteins, and other Class I proteins, it is disulfide bonded to a residue in strand 7, rather than to Cys 1, as found in SPH15. These disulfide bonds were added to the *de novo* calculations of the PrsS proteins as further constraints and shown to be compatible with the overall structures. The addition of the disulfide bonds had no effect on the DOPE score of the best model for PrsS1 and PrsS8, but gave a small 4% decrease in the DOPE score for PrsS3, with an equivalent improvement in the DOPE score for SPH15.

Interactions of PrsS proteins with their receptors

In a previous study [52], the predicted surface loops of PrsS1 were mutated at one or more sites, the proteins purified and refolded from inclusion bodies, and their activity to inhibit germination of pollen assayed. The mutation with the largest effect was D79 at the tip of loop 6 (Figure 8a,b). When this was mutated to either G or H (as found in PrsS3 and PrsS8, respectively), both totally abrogated activity. Mutation of the other aspartate residues in the loop (D77, D78), which are conserved in the other alleles, each to His, also removed PrsS1 activity, but the double mutation D77E/D78E was as active as wild type, suggesting that the negative charge on this loop is important. The only other single mutation tested that showed any effect was D27H, in loop 2, a conserved residue in all three poppy sequences. This mutation showed only 78% wild-type activity when assayed at low protein concentrations (25 μg/ml), but was as active as wild type at 75 μg/ml. Loop 2 is on the

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

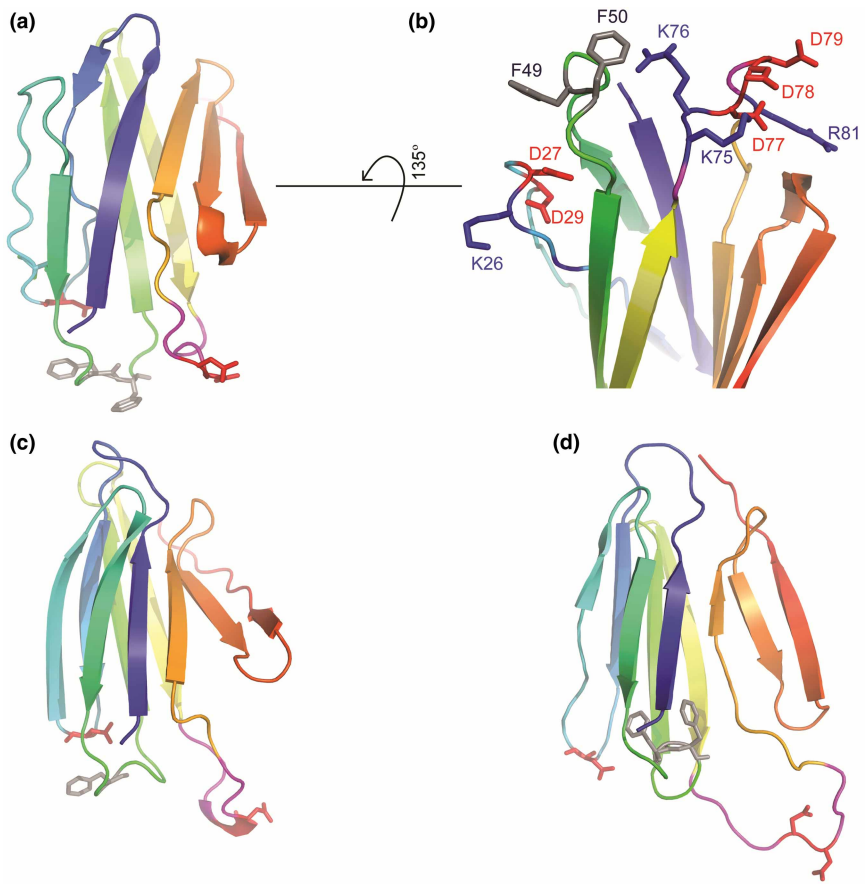

Figure 8. *De novo* structure predictions of the three poppy proteins PrsS1, PrsS3 and PrsS8.
(**a**) Ribbon representation of the *de novo* structure prediction of the PrsS1, with the lowest DOPE score. The backbone structure is coloured from blue to red, from N- to C-terminus. Side chains of selected amino acids mutated in ref. [52] are shown as sticks. (**b**) Part of the *de novo* predicted structure of PrsS1 in (**a**), showing the side chains of charged and hydrophobic amino acids in loops 2, 4 and 6, thought to interact with the PrpS1 receptor, as coloured sticks, labelled with an amino acid number. Red: acidic, grey: hydrophobic, blue: basic. (**c**) Ribbon representation of the *de novo* structure prediction of the PrsS3, with the lowest DOPE score, coloured as in (**a**). (**d**) Ribbon representation of the *de novo* structure prediction of the PrsS8, with the lowest DOPE score, coloured as in (**b**).

same side of the protein as loop 6 and both contain several aspartate and charged residues at their tips in all alleles (Figures 1a and 8a,b). These loops are separated by loop 4, which contains mainly charged residues, but with two central phenylalanine residues at its tip in PrsS1 (sequence REDFFH). Similar hydrophobic residues are found at the tips of the loop in the alleles (Figure 8). The mutation F49M in PrsS1 to the residue found in PrsS3 had little effect on activity, suggesting that the aromatic ring here is not essential for activity, but both F and M are hydrophobic, so the importance of hydrophobicity for activity was not tested.

A single mutation in loop 1 (N15H) in PrsS1, two double conservative mutations in loop 3 (T36S/S37D and H39Q/D40E) and a single mutation in loop 8 (D99H) showed no effect. A single conservative mutation in loop 5, K63R, also showed no effect, but a triple mutant K63R/E64K/T65G showed only 58% of the activity of wild-type protein at 25 μg/ml, less than the single mutation in loop 2, D27H, but much higher than D79G. However, our structural model shows that T65 is at the beginning of strand 6 of PrsS1, rather than in a loop, so this triple mutation may affect protein folding rather than simply receptor binding.

Overall, these mutagenesis studies, together with the structural models, suggest that loops 1, 3, 5 and 8 have little effect on the activity of PrsS1, while loop 6 at the tip of the protein interacts with the receptor, probably alongside loops 2 and 4 (Figure 8b). Given the length and flexibility of the loop 6, this is the loop most likely to interact with the receptor and confer specificity. It is also highly divergent in sequence across the SPH

protein family. Loops 2 and 6 are largely charged, and aspartate residues in these loops have been shown to affect activity. In contrast, the tip of loop 4 contains two hydrophobic Phe groups, with charged groups on either side. The exposure of such hydrophobic residues on the outside of proteins is unusual, and they are good candidates for intermolecular interactions with hydrophobic partners.

The receptor proteins PrpS1, S3 and S8 are highly hydrophobic, 20 kDa proteins, that have been shown to localise to the plasma membrane of the pollen tube. Interestingly, no homologues to these sequences were found using HHblits; however, secondary structure predictions suggest that they contain six transmembrane helices, with an extracellular, extended, domain of ∼35 amino acids. The central, 15 amino acid segment of this extracellular loop of the PrpS1 protein, namely DQKWVVAFGTAAICD, has been synthesised, and shown to interact with purified PrsS1, in a slot blot, whereas a randomised peptide of the same composition did not [51]. The same peptide could block the inhibition of germination of pollen by PrsS1 protein, and, surprisingly, was allele specific, despite this part of the PrpS1 receptor protein being similar to the other alleles. Hence, this peptide is likely to bind to the PrsS1 protein. This peptide is largely hydrophobic in the central region (underlined residues). We postulate that the central hydrophobic amino acids of the receptor interact with the exposed F49 and F50 residues in loop 4, while the charged residues within the receptor may interact with the charged residues on this loop and loops 2 and 6.

Conclusion

The SPH family of proteins are extremely widespread in dicotyledonous plants and are thought to be involved in a large variety of signalling pathways. *Arabidopsis thaliana* contains 92 core members of this family that are evolutionarily related and have sequence resemblance to *Papaver* PrsS proteins. Other core members of the SPH family are found only in dicotyledonous plants and the lower plants, Lycopodiopsida (*Selaginella*, spike moss) and Bryophyta (*Physcomitrella*, spreading earthmoss), and have not been identified in monocotyledonous plants, despite numerous attempts using BLAST. This specific phylogenetic distribution suggests that, like the additional SPADA-identified homologues in *Arabidopsis*, proteins identified in fungi and animals and placed in Pfam group PF05938 may be the result of independent evolution or divergence to the point at which little sequence homology remains.

The SPH proteins are highly stable and have a β-sandwich structure, with 8–9 β-sheets in a topology distinct from that found in most other proteins to date. The different classes of SPH proteins are evolutionarily related and have distinct disulfide-bonding patterns that can be readily accommodated within the proposed structure. All the proteins have a disulfide bond between the neighbouring strands 2 and 5. The PrsS proteins and those in Class I have four cysteines, with an additional disulfide bond between strand 7 and loop 6. Class II proteins have an additional disulfide bond between the adjacent strands 8 and 9, whereas Class III proteins only have the one conserved disulfide bond. SPH15 is a Class IV protein, with three cysteines within the conserved sequence and one, additional one, at residue 1, outside of this, giving rise to a disulfide bond between residue 1 and loop 6. The classes have been subclassified on the sequence of the hydrophilic loop 2, proteins in subclass A have the motif K/RXXD while those in subclass B are heterogeneous. PrsS1 and PrsS8 have KXXD in this loop, whereas PrsS3 contains E as the first amino acid of the motif but still contains the conserved D.

From the limited mutagenesis data, involving this conserved D residue in loop 2 and the larger effects of mutation in loop 6 [52], together with the structural analysis presented here, we speculate that loops 2, 4 and 6, on one face of the protein, interact with the receptor PrpS1 to mediate programmed cell death. However, other unrelated proteins with the same strand topology, such as transthyretin and TssJ, show that a range of interaction interfaces is possible (Figure 5). Thus, this family of proteins may have evolved to act as a versatile and stable scaffold to display a variety of peptides in the loops, each interacting with a different receptor. As such, in addition to their broad roles in cell signalling, the SPH family may be a useful scaffold for synthetic biology applications.

Database Deposition

The NMR-derived 3D co-ordinates of SPH15 have been deposited in the Protein Data Bank under PDB ID code 6G7G.

Abbreviations

AUC, analytical ultracentrifugation; AMG, additional amino acids; CD, circular dichroism; DOPE, discreet optimised protein energy; MSA, multiple sequence alignment; NMR, nuclear magnetic resonance; PrsS, *Papaver*

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

rhoeas stigma S-determinant; S-protein, self-incompatibility protein; SI, self-incompatibility; SPH, self-incompatibility protein homologue; TM, template modelling.

Author Contribution

J.P.R. cloned and overexpressed SPH15 and did the initial protein purification. G.L.R. and M.J.W. did the phylogenetic analysis of the SPH proteins. E.I.H. did the initial protein characterisation. E.I.H. and K.V.R. collected NMR data. R.J.C., L.J.S. and E.I.H. assigned the backbone NMR spectrum, K.V.R. completed the assignments and determined the NMR structure of the protein. S.J. and P.J.W. modelled the structure of the poppy proteins and SPH15 using DeepCDPred and MODELLER. M.J.W., P.J.W., L.J.S. and E.I.H. wrote the paper.

Funding

S.J. was funded by an Elite Scholarship by the University of Birmingham. NMR spectra were collected mainly at the HWB-NMR centre in Birmingham under Wellcome Trust grant 099185/Z/12/Z, spectra were also collected at the MRC Biomedical NMR centre at NIMR, London and at the NMR Facility, Department of Biochemistry, University of Oxford.

Acknowledgements

We are grateful to the NMR staff, S. Whittaker and C. Ludwig (HWB-NMR, Birmingham); T. Frenkiel and G. Kelly (NIMR, London); and C. Redfield (University of Oxford), for NMR time and assistance.

Competing Interests

The Authors declare that there are no competing interests associated with the manuscript.

References

1 Ride, J.P., Davies, E.M., Franklin, F.C.H. and Marshall, D.F. (1999) Analysis of Arabidopsis genome sequence reveals a large new gene family in plants. *Plant Mol. Biol.* **39**, 927–932 https://doi.org/10.1023/A:1006178511787
2 Silverstein, K.A.T., Moskal, W.A., Wu, H.C., Underwood, B.A., Graham, M.A., Town, C.D. et al. (2007) Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *Plant J.* **51**, 262–280 https://doi.org/10.1111/j.1365-313X.2007.03136.x
3 Zhou, P., Silverstein, K.A.T., Gao, L.L., Walton, J.D., Nallu, S., Guhlin, J. et al. (2013) Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). *BMC Bioinformatics* **14**, 335 https://doi.org/10.1186/1471-2105-14-335
4 Joly, V. and Matton, D.P. (2015) KAPPA, a simple algorithm for discovery and clustering of proteins defined by a key amino acid pattern: a case study of the cysteine-rich proteins. *Bioinformatics* **31**, 1716–1723 https://doi.org/10.1093/bioinformatics/btv047
5 Foote, H.C.C., Ride, J.P., Franklin-Tong, V.E., Walker, E.A., Lawrence, M.J. and Franklin, F.C. (1994) Cloning and expression of a distinctive class of self-incompatibility (S) gene from *Papaver rhoeas* L. *Proc. Natl Acad. Sci. U.S.A.* **91**, 2265–2269 https://doi.org/10.1073/pnas.91.6.2265
6 Tantikanjana, T., Nasrallah, M.E. and Nasrallah, J.B. (2010) Complex networks of self-incompatibility signaling in the Brassicaceae. *Curr. Opin. Plant Biol.* **13**, 520–526 https://doi.org/10.1016/j.pbi.2010.06.004
7 McClure, B., Cruz-García, F. and Romero, C. (2011) Compatibility and incompatibility in S-RNase-based systems. *Ann. Bot.* **108**, 647–658 https://doi.org/10.1093/aob/mcr179
8 Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L. et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 https://doi.org/10.1093/nar/gkv1344
9 Wheeler, M.J., Vatovec, S. and Franklin-Tong, V.E. (2010) The pollen S-determinant in Papaver: comparisons with known plant receptors and protein ligand partners. *J. Exp. Bot.* **61**, 2015–2025 https://doi.org/10.1093/jxb/erp383
10 Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J. et al. (2017) Interpro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 https://doi.org/10.1093/nar/gkw1107
11 Jordan, N.D., Ride, J.P., Rudd, J.J., Davies, E.M., Franklin-Tong, V.E. and Franklin, F.C.H. (2000) Inhibition of self-incompatible pollen in *Papaver rhoeas* involves a complex series of cellular events. *Ann. Bot.* **85**, 197–202 https://doi.org/10.1006/anbo.1999.1034
12 Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 https://doi.org/10.1093/molbev/msw054
13 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 https://doi.org/10.1093/nar/gkh340
14 Remmert, M., Biegert, A., Hauser, A. and Söding, J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 https://doi.org/10.1038/nmeth.1818
15 Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 https://doi.org/10.1101/gr.849004
16 Yang, Y.D., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A. et al. (2017) SPIDER2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. *Methods Mol. Biol.* **1484**, 55–63 https://doi.org/10.1007/978-1-4939-6406-2_6

17 Nielsen, L., Frokjaer, S., Brange, J., Uversky, V.N. and Fink, A.L. (2001) Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry* **40**, 8397–8409 https://doi.org/10.1021/bi0105983
18 Bessette, P.H., Aslund, F., Beckwith, J. and Georgiou, G. (1999) Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proc. Natl Acad. Sci. U.S.A.* **96**, 13703–13708 https://doi.org/10.1073/pnas.96.24.13703
19 Coulthard, R.J., Rajasekar, K.V., Ride, J.P., Hyde, E.I. and Smith, L.J. (2018) ^{1}H, ^{13}c and ^{15}N NMR assignments of self-incompatibility protein homologue 15 from *Arabidopsis thaliana*. *Biomol. NMR Assign.* https://doi.org/10.1007/s12104-018-9853-0
20 Pace, C.N., Vajdos, F., Fee, L., Grimsley, G. and Gray, T. (1995) How to measure and predict the molar absorption-coefficient of a protein. *Protein Sci.* **4**, 2411–2423 https://doi.org/10.1002/pro.5560041120
21 Schuck, P. (2000) Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. *Biophys. J.* **78**, 1606–1619 https://doi.org/10.1016/S0006-3495(00)76713-0
22 Sreerama, N. and Woody, R.W. (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal. Biochem.* **287**, 252–260 https://doi.org/10.1006/abio.2000.4880
23 Whitmore, L. and Wallace, B.A. (2008) Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers* **89**, 392–400 https://doi.org/10.1002/bip.20853
24 Cheung, M.S., Maguire, M.L., Stevens, T.J. and Broadhurst, R.W. (2010) DANGLE: a Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. *J. Magn. Reson.* **202**, 223–233 https://doi.org/10.1016/j.jmr.2009.11.008
25 Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E. and Nilges, M. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* **23**, 381–382 https://doi.org/10.1093/bioinformatics/btl589
26 Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W. et al. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 https://doi.org/10.1107/S0907444998003254
27 Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–486 https://doi.org/10.1007/BF00228148
28 Ji, S., Oruç, T., Mead, L., Rehman, M.F., Thomas, C.M., Butterworth, S. et al. (2019) DeepCDpred: inter-residue distance and contact prediction for improved prediction of protein structure. *PLoS ONE* **14**, e0205214 https://doi.org/10.1371/journal.pone.0205214
29 Dunn, S.D., Wahl, L.M. and Gloor, G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 https://doi.org/10.1093/bioinformatics/btm604
30 Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S. and Rost, B. (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15**, 85 https://doi.org/10.1186/1471-2105-15-85
31 Hsieh, C.J., Sustik, M.A., Dhillon, I.S. and Ravikumar, P. (2014) QUIC: quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.* **15**, 2911–2947 http://www.jmlr.org/papers/v15/hsieh14a.html
32 Seemayer, S., Gruber, M. and Söding, J. (2014) CCMpred—fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128–3130 https://doi.org/10.1093/bioinformatics/btu500
33 Betancourt, M.R. and Thirumalai, D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **8**, 361–369 https://doi.org/10.1110/ps.8.2.361
34 Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R. et al. (2011) Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 https://doi.org/10.1016/B978-0-12-381270-4.00019-6
35 Zhang, Y. and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 https://doi.org/10.1093/nar/gki524
36 Shuid, A.N., Kempster, R. and McGuffin, L.J. (2017) ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. *Nucleic Acids Res.* **45**, W422–W428 https://doi.org/10.1093/nar/gkx249
37 Webb, B. and Sali, A. (2014) Protein structure modeling with MODELLER. *Methods Mol. Biol.* **1137**, 1–15 https://doi.org/10.1007/978-1-4939-0366-5_1
38 Shen, M.Y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 https://doi.org/10.1110/ps.062416606
39 Cornilescu, G., Delaglio, F. and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289–302 https://doi.org/10.1023/A:1008392405740
40 Holm, L. and Laakso, L.M. (2016) Dali server update. *Nucleic Acids Res.* **44**, W351–W355 https://doi.org/10.1093/nar/gkw357
41 Lawrence, S.L., Feil, S.C., Morton, C.J., Farrand, A.J., Mulhern, T.D., Gorman, M.A. et al. (2015) Crystal structure of *Streptococcus pneumoniae* pneumolysin provides key insights into early steps of pore formation. *Sci. Rep.* **5**, 14352 https://doi.org/10.1038/srep14352
42 Marshall, J.E., Faraj, B.H.A., Gingras, A.R., Lonnen, R., Sheikh, M.A., El-Mezgueldi, M. et al. (2015) The crystal structure of pneumolysin at 2.0 Å resolution reveals the molecular packing of the pre-pore complex. *Sci. Rep.* **5**, 13293 https://doi.org/10.1038/srep13293
43 Rossjohn, J., Feil, S.C., McKinstry, W.J., Tweten, R.K. and Parker, M.W. (1997) Structure of a cholesterol-binding, thiol-activated cytolysin and a model of its membrane form. *Cell* **89**, 685–692 https://doi.org/10.1016/S0092-8674(00)80251-2
44 Felisberto-Rodrigues, C., Durand, E., Aschtgen, M.S., Blangy, S., Ortiz-Lombardia, M., Douzi, B. et al. (2011) Towards a structural comprehension of bacterial type VI secretion systems: characterization of the TssJ–TssM complex of an *Escherichia coli* pathovar. *PLoS Pathog.* **7**, e1002386 https://doi.org/10.1371/journal.ppat.1002386
45 Peterson, S.A., Klabunde, T., Lashuel, H.A., Purkey, H., Sacchettini, J.C. and Kelly, J.W. (1998) Inhibiting transthyretin conformational changes that lead to amyloid fibril formation. *Proc. Natl Acad. Sci. U.S.A.* **95**, 12956–12960 https://doi.org/10.1073/pnas.95.22.12956
46 Rossjohn, J., Polekhina, G., Feil, S.C., Morton, C.J., Tweten, R.K. and Parker, M.W. (2007) Structures of perfringolysin O suggest a pathway for activation of cholesterol-dependent cytolysins. *J. Mol. Biol.* **367**, 1227–1236 https://doi.org/10.1016/j.jmb.2007.01.042
47 Tilley, S.J., Orlova, E.V., Gilbert, R.J.C., Andrew, P.W. and Saibil, H.R. (2005) Structural basis of pore formation by the bacterial toxin pneumolysin. *Cell* **121**, 247–256 https://doi.org/10.1016/j.cell.2005.02.033
48 Durand, E., Nguyen, V.S., Zoued, A., Logger, L., Péhau-Arnaudet, G., Aschtgen, M.S. et al. (2015) Biogenesis and structure of a type VI secretion membrane core complex. *Nature* **523**, 555–560 https://doi.org/10.1038/nature14667

Biochemical Journal (2019) **476** 809–826
https://doi.org/10.1042/BCJ20180828

49 Wojtczak, A., Cody, V., Luft, J.R. and Pangborn, W. (1996) Structures of human transthyretin complexed with thyroxine at 2.0 Å resolution and 3′,5′-dinitro-*N*-acetyl-L-thyronine at 2.2 Å resolution. *Acta Crystallogr. D Biol. Crystallogr.* **52**, 758–765 https://doi.org/10.1107/S0907444996003046

50 Monaco, H.L., Rizzi, M. and Coda, A. (1995) Structure of a complex of 2 plasma-proteins: transthyretin and retinol-binding protein. *Science* **268**, 1039–1041 https://doi.org/10.1126/science.7754382

51 Wheeler, M.J., de Graaf, B.H.J., Hadjiosif, N., Perry, R.M., Poulter, N.S., Osman, K. et al. (2009) Identification of the pollen self-incompatibility determinant in *Papaver rhoeas*. *Nature* **459**, 992–995 https://doi.org/10.1038/nature08027

52 Kakeda, K., Jordan, N.D., Conner, A., Ride, J.P., Franklin-Tong, V.E. and Franklin, F.C.H. (1998) Identification of residues in a hydrophilic loop of the *Papaver rhoeas* S protein that play a crucial role in recognition of incompatible pollen. *Plant Cell* **10**, 1723–1731 https://doi.org/10.1105/tpc.10.10.1723