Molecular Immunology 131 (2021) 33–43

Contents lists available at ScienceDirect

Molecular Immunology

journal homepage: www.elsevier.com/locate/molimm

Cloning, expression and immunological characterisation of Coc n 1, the first major allergen from Coconut pollen

Bodhisattwa Saha [a,b,*], Bijoya Karmakar [a], Swati Gupta Bhattacharya [a,**]

[a] Division of Plant Biology, Bose Institute, 93/1, Acharya Prafulla Chandra Road, Kolkata 700009, India
[b] Chemistry Research Laboratory, 12 Mansfield Road, OX4 4TG, Oxford, United Kingdom

ARTICLE INFO

Keywords:
Allergen
Vicilin
Immunoglobulin E (IgE)
Sequence similarity networks
Epitope mapping
Coconut pollen

ABSTRACT

Coconut pollen has been documented to be a major contributor to the aeroallergen load in India, causing respiratory allergy in a large cohort of susceptible individuals. Here, we report the identification of the first major allergen from Coconut pollen, Coc n 1. The full-length sequence of the allergen was determined from previously identified peptides and overexpressed in *E. coli*. Recombinant Coc n 1 folded into a trimer and was found to possess allergenicity equivalent to its natural counterpart. Proteolytic processing of Coc n 1 led to the formation of an immunodominant ~20 kDa C-terminal subunit and the site of cleavage was determined by amino acid microsequencing. Five linear IgE binding epitopes were predicted and mapped on the homology modelled structure of Coc n 1. Amongst three immunodominant epitopes, two were present towards the C-terminal end. Coc n 1 was found to belong to the highly diverse cupin superfamily and mimics its structure with known 7S globulin or vicilin allergens but lacks sequence similarity. Using sequence similarity networks, Coc n 1 clustered as a separate group containing unannotated cupin domain proteins and did not include known vicilin allergens except Gly m Bd 28 kDa, a Soybean major allergen. 7S globulins are major storage proteins and food allergens, but presence of such protein in pollen grains is reported for the first time. Further study on Coc n 1 may provide insights into its function in pollen grains and also in the development of immunotherapy to Coconut pollen allergy.

1. Introduction

Allergic diseases are a global health concern that greatly reduces the quality of life. More than 25 % of the world population suffers from some allergic maladies (Pawankar, 2014). Allergenic proteins harboured in various aerobic sources such as pollen, fungal spores and dust mites cause a hypersensitive reaction by triggering the production of IgE antibodies which cross-link together on mast cells. This directs the immune system to release elevated levels of histamine, leukotrienes and recruitment of eosinophils leading to an allergic manifestation (Platts--Mills and Woodfolk, 2011). Pollen grains constitute a major load of aeroallergens and are responsible for rhinitis, allergic asthma and other respiratory ailments in susceptible individuals (Burr et al., 2003).

While Palm trees are common in tropical countries, allergenicity to palm pollen grains has been documented mostly from India and the Middle East (Mansouritorghabeh et al., 2019; Bhattacharya et al., 2018). The Date palm, Betel Nut and Palmyra palm were reported to cause a significant atopic response in an Indian cohort, however, identification of allergens is still lacking (Chowdhury et al., 1998). Since immunotherapy with purified or hypoallergens could be a promising treatment regime, identification of allergens from relevant sources is essential. A bottleneck towards identification and characterisation of allergens is due to the absence of genomic information as majority of the sources are unsequenced. We had previously reported a layered proteomic approach to identify proteins from unsequenced organisms using a manual *de novo* sequencing followed by homology-based search and this led to the identification of multiple allergens from Coconut pollen grains for the first time. Amongst these, a major allergen was identified having similarity to vicilin-like proteins based on three manually sequenced peptides (Saha et al., 2015). However, the peptides from this protein showed very low similarity with commonly reported 7S globulin or vicilin allergens. Instead, they matched with a vicilin-like protein from *Elaeis guineensis*, an unannotated protein. Similar peptides were also identified in Date Palm pollen as an allergen suggesting this protein may

* Corresponding author. Present address: CSIR-Institute of Genomics and Integrative Biology, Mathura Road, New Delhi 110025, India.
** Corresponding author.
E-mail addresses: bodhisattwasaha@gmail.com, bodhisattwa.saha@chem.ox.ac.uk (B. Saha), swati@jcbose.ac.in (S.G. Bhattacharya).

https://doi.org/10.1016/j.molimm.2020.12.026
Received 3 August 2020; Received in revised form 19 November 2020; Accepted 16 December 2020
Available online 21 January 2021
0161-5890/© 2021 Elsevier Ltd. All rights reserved.

belong to a different subgroup of vicilin protein (Saha and Gupta Bhattacharya, 2017). Vicilin proteins are mostly reported as food allergens containing a highly conserved cupin domain and belong to one of the most diverse protein superfamily (Dunwell et al., 2004). Apart from seed storage, proteins of this family include sugar-binding metal-independent proteins as well as metal-dependent enzymes having dioxygenase and decarboxylase activities. It was estimated that there could be more than 18 different functional subclasses of this protein (Galperin and Koonin, 2012). Allergic cupin domain proteins also include 11S legumins which are structurally similar with vicilin proteins and have shown cross-reactivity amongst nut allergens (Bublin and Breiteneder, 2014).

To explore the diversity of cupin domain proteins, studies were done to cluster them based on sequence similarity. Instead of phylogenetic trees, Sequence Similarity Networks (SSNs) have been useful to visualize relationships across diverse superfamilies using large datasets (Atkinson et al., 2009). Ara h 1 core structure was found to be similar to Vicilin and legumin allergens based on global clustering analysis of cupin family proteins (Chruszcz et al., 2011). Vicilins, also known as 7S globulins, are 50−60 kDa monomers, existing as trimers, without having disulfide bonds and known to be proteolytically processed to generate smaller subunits (Kesari et al., 2017). The major allergen from Coconut pollen was found towards the basic pI range at 20 kDa, suggesting it could be a proteolytic product of a bigger protein. A similar observation was also made for Gly m Bd 28 kDa, another vicilin allergen from Soybean that has a 23 kDa C-terminal allergenic subunit (Xiang et al., 2004). To further understand the molecular basis of IgE recognition, a thorough understanding of the antibody binding epitopes is essential. 7S globulins have been studied in detail over the years and with the determination of the crystal structures, epitopes could be predicted on them and mapped. However, the experimental validation of these epitopes has been limited. Determining IgE epitopes can help in the rational designing of hypoallergenic variants and has been used to make successful allergen vaccines (Sircar et al., 2016).

In this study, molecular, immunological and structural characterisation of the major allergen from Coconut pollen was investigated. Using information from *de novo* sequenced peptides, the full-length sequence of the allergen was built using genomic tools and reported to the WHO/IUIS allergen nomenclature. The protein was overexpressed in *E. coli* and immunological characterisation was performed, suggesting that it retains the allergenicity. A homology-based structure was created and IgE binding epitopes were mapped on it, which was further validated. We also investigated its evolutionary relationship with other known cupin domain-containing allergens using a bioinformatics approach, which led to the identification of a novel cluster of proteins with an unknown function.

2. Experimental procedures

2.1. Blood collection

Blood from ten adult patients with a previous record of Coconut pollen allergy was collected for this study. All the patients were suffering from seasonal allergic rhinitis, allergic asthma, and skin allergy and had high total IgE, total histamine. They also had elevated specific IgE responses to crude Coconut pollen extract and >+3 Skin Prick Test report. Smokers and patients suffering from immunodeficiency diseases were not recruited. The study was approved by the hospital from where blood was collected and informed consent was taken from the patients. Demographic details of the patients are described in Supporting Information Table 1.

2.2. Full-length cDNA isolation from Coconut pollen and sequence determination

Pollen grains were collected from the inflorescence by passing through geological scale meshes of decreasing pore size. 100 mg of pure pollen grains at tetrad developmental stage were crushed with liquid nitrogen. Total RNA was extracted by Trizol (Invitrogen, USA) method according to manufacturers' instruction with minor modifications. Due to high polysaccharide content of the pollen grain, 250 μL of high salt solution (0.8 M sodium citrate and 1.2 M NaCl) and 250 μL of isopropanol was used in the RNA precipitation step. cDNA synthesis was done using the RevertAid cDNA synthesis kit (Thermo, USA) using standard Oligo (dT) primers. The full-length gene was amplified by a three step method employing 5′/3′ RACE. Firstly, an internal cDNA fragment was amplified by employing degenerate primers AA(C/T) CCA A(C/G)T GCG ACA GAG TAT GGC and TCC TCG (T/C)TC ACC CCG AAT GC designed based on the *de novo* sequenced peptide SPHVNPSATEYGL and ELAAAFGVSEE respectively from our previous results (Saha et al., 2015). For 3′ RACE, initial cDNA priming was done by Universal adapter primer: AGGAGCACCGACTCACTACAGGTTTTTTTTTTTTTTTT, then amplified by 2 nested gene specific primers and Adapter primer (AP) AGGAGCACCGACTCACTACAGG. 5′ RACE was done according to the protocol described (5′ RACE Invitrogen, ver 2.0). Two antisense gene-specific primers were designed from deduced internal cDNA fragment. The PCR products generated in each step were cloned in TA Cloning vector (Instaclone, Thermo, USA) and sequenced using M13 forward and reverse primers. Finally, two new primers 5′ CTAGCACTACTGTGCTAATTAAGC 3′ (forward) and 5′ TAGACTAGAAAAAGATGGAACATC 3′ (reverse) was used to amplify the full-length gene, which was thereafter sequenced. Signal peptide, N-glycosylation and protein domain were determined using the Signal P4.1 server (Petersen et al., 2011), NetN glyc 1.0 server (Gupta et al., 2004) and Pfam (El-Gebali et al., 2019), respectively.

2.3. Subcloning and protein overexpression

cDNA encoding Coc n 1 was amplified and directionally cloned into pET-21d protein expression vector modified to contain an N-terminal His-Tag and a Tobacco Etch Virus (TEV) protease site (Biobharati Life Science, India) using Forward primer: ATTATGAATTCTACCAAGGGAGAGGAATGGAAGG containing 5′ EcoR1 restriction site and reverse primer TTATAAAGCTTTCATGCAAAGAGCCCCCTCTTCATC containing 3′ HindIII site. The construct containing Coc n 1 was transformed into *E. coli* BL21 (DE3) Rosetta cells and cultured in Luria Bertani (LB) medium containing 100 μg/mL ampicillin and 34 μg/mL chloramphenicol at 37 °C. Once $O.D._{600nm}$ reached 0.4, the expression of Coc n 1 was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and cultured overnight at 16 °C. Proteins were extracted in 25 mM Tris, 150 mM NaCl, and 10 mg/mL of lysozyme, incubated for 30 min. at 4 °C, sonicated and centrifuged at 22,000 *g* for 20 min. The pellet fraction was dissolved in urea buffer: 7 M Urea, 25 mM Tris-HCl, 300 mM NaCl, and 5% glycerol (pH 8). Coc n 1 obtained in the pellet was purified in Ni-NTA column by elution in 250 mM imidazole under denaturing conditions. The purified protein was refolded by reducing the urea concentration by 6 M, 4 M, 2 M through overnight dialysis and finally buffer exchanged into 50 mM Tris, 150 mM NaCl, pH 8. Protein aggregates were removed by centrifugation at 20,000*g* for 10 min. and supernatant was collected. Refolded Coc n 1 at 1 mg/mL concentration was injected into a Superdex 200 10/300 GL column (GE Life Sciences, Uppsala, Sweden) connected to an AKTA purifier, equilibrated and run at 0.3 mL/min in 50 mM Tris, 150 mM NaCl, pH 8.

2.4. SDS-PAGE and immunoblot

The uninduced and induced *E. coli* protein extract, along with the purified Coc n 1 protein, was profiled in 12 % SDS-PAGE. Immunoblot

was done according to the method described previously (Dey et al., 2016). In brief, Coc n 1 was transferred electrophoretically to PVDF membrane at 1.2 mA/cm^2 for 1 h. Empty sites were blocked with 3% BSA in TBST and incubated with individual patient sera (1:10 vol/vol) overnight as primary antibody in blocking solution at 4 °C for 2 h. After washing with TBST, incubation was done in monoclonal anti-human IgE alkaline phosphatase tagged produced in mouse (Sigma) at 1:1000 ratio in blocking solution and incubated at room temperature for 2 h. Reactive bands were visualized in NBT-BCIP (Sigma).

2.5. Purification of major allergen

The purification of native Coc n 1 from Coconut pollen was done as described previously (Saha et al., 2015). In brief, total protein extracted in binding buffer (20 mM Tris-HCl, 0.5 M NaCl, 1 mM $MnCl_2$, 1 mM $CaCl_2$, pH 7.4) was loaded onto Con A 4B lectin affinity column (GE Lifesciences), and glycoproteins were eluted by elution buffer (0.5 M alpha-D-glucopyranoside, 20 mM Tris, 0.5 M NaCl, pH 7.4) in ÄKTA Prime (GE Lifesciences). Eluted protein was buffer exchanged in 20 mM sodium acetate pH 5.5, injected into HiTrap SP HP column (GE Lifesciences) and eluted using a 1 M NaCl gradient. Desired fractions were buffer exchanged in 20 mM Tris containing 150 mM NaCl and passed through Superdex 200 10/300 gel filtration column. Eluted fractions were monitored at 280 nm, concentrated using 10 kDa cut-off centrifugal filtration devices (Amicon, Millipore, U.S.A.).

2.6. Stripped basophil histamine release assay

Histamine release assay was performed following the protocol described earlier with minor modifications (Sircar et al., 2015). Briefly, peripheral blood was collected from non-allergic donors, and peripheral blood mononuclear cells were separated by Ficoll Paque (GE Life Sciences). The bound IgE was stripped from basophils by incubation in lactic acid buffer (13.4 mM lactate, 140 mM NaCl and 5 mM KCl at pH 3.5) for 3 min. followed by washing in 4-(2-hydroxyethyl)-1-piperazine ethane sulfonic acid (HEPES) buffer (pH 7.5) and resensitized in 150 µl sera from each patient at 37 °C for 90 min.. Cells were stimulated with 1 µg of Coc n 1 allergen in HEPES buffer containing 1 mM $CaCl_2$ for 1 h at 37 °C, and the reaction was stopped by the addition of ice-cold 0.9 % NaCl (w/v). Following centrifugation, the cell-free supernatant containing histamine was measured by histamine assay kit (Immunotech, Beckman-Coulter). Total histamine was determined by lysing a separate aliquot of mononuclear cells in 3% perchloric acid. Spontaneous histamine release was obtained in unstimulated cell free supernatants. Histamine release was calculated as a percentage of the total histamine content after correcting spontaneous release. The cells sensitized with four non-allergic serum were taken as controls. BSA was used as a negative control.

2.7. Homology modelling

Five templates having the best sequence identity were selected by PSI-BLAST in the PDB database with Coc n 1 amino acid sequence for creating a homology-based 3D model using Modeller 9.15 (Webb and Sali, 2016). The best-matched template was selected for alignment with Coc n 1. Ten models were generated using the alignment file and the final model was selected on the basis of GA341 and Discrete Optimized Protein Energy (DOPE) method assessment functions of MODELLER. The minimum DOPE scored model was selected for the final model and stereochemical analysis. Incorrectly folded loop regions were corrected by the Loop Refinement program in Modeller. The loop refined model was used for predicting IgE binding epitopes. The final structure was visualized in Pymol v 1.8.4.0. Solvent accessible surface area was obtained using "get_area" command in Pymol. The trimer model of Coc n 1 was generated from SWISS-MODEL (Waterhouse et al., 2018) server using 1IPK as a template.

2.8. Circular dichroism

Coc n 1 was dialysed against 10 mM NaH_2PO_4 for Circular Dichroism analysis in a J-815 CD spectropolarimeter (Jasco Corporation, Japan) at 20 °C within a wavelength range of 260–200 nm at 0.5 nm resolution and bandwidth of 1 nm using a protein concentration at 0.25 mg/mL. Measurements were done in a Quartz cell with an optical path length of 0.2 cm with a scan rate of 100 nm/min. The spectral baseline was corrected with a protein free sample in 10 mM NaH_2PO_4. Thermal denaturation of refolded Coc n 1 was checked at temperatures ranging between 20 to 90 °C with intervals of 10 °C. Three scans were taken at each temperature, mean residual ellipticity was calculated and analysed in K2D of Dichroweb server (Whitmore and Wallace, 2004).

2.9. Linear B cell epitope prediction using in-silico tools

ABCPred, BCEPred, and Bepipred epitope prediction servers were used to predict linear B cell epitopes in the amino acid sequence of Coc n 1. The predictions were based on physico-chemical properties such as hydrophilicity, flexibility/mobility, accessibility, polarity, exposed surface and turns (BcePred) (Saha and Raghava, 2004), artificial neural networking (ABCpred) (EL-Manzalawy et al., 2008) and a combination of Hidden Markov Model and propensity scale (BepiPred) (Jespersen et al., 2017). Regions recognised commonly by all the three web servers were selected as probable IgE binding epitopes. A random peptide (RP) N' SCQENVQVMA C' was generated from Expasy random peptide generator with equal composition of amino acids, not having any similarity with Coc n 1 and synthesised for use as a negative control.

2.10. Solid phase peptide synthesis

Peptide synthesis was done by a stepwise procedure in the solid phase using Fmoc chemistry and Rink amide MBHA resin (substitution 0.69 mmol/g; Novabiochem) and Fmoc-protected amino acids (Novabiochem) in Endeavor 90-II peptide synthesizer (AAPTec, Louisville, KY, USA) according to an earlier method (Ganguly et al., 2012). In brief, the Fmoc-amino acid derivatives were coupled with benzotriazol-1-yl-oxytripyrrolidinophosphonium hexafluorophosphate (PyBOP)/hydroxybenzotriazole (HOBT)/diisopropylethylamine (DIPEA) (used as 5:5:10 times excess of the resin, respectively), and Fmoc cleavage was performed with 20 % (v/v) piperidine in dimethylformamide (DMF). DMF was removed by purging N_2 gas and ether. Peptides were cleaved using a cocktail (85 % TFA, 5% H_2O, 5% phenol, 2.5 % anisole, and 2.5 % tri-isopropyl silane) and kept in the cleavage cocktail for 1 h (for Xaa = Arg, the cleavage reaction was carried on for 3 h). TFA was removed by evaporation in a rotary evaporator, and the crude peptides were dissolved in methanol and then purified through reverse-phase HPLC using a Phenomenex C18 column and CH_3OH-H_2O gradients (0–80 % CH_3OH in 40 min) containing 0.1 % TFA. The final products were checked by mass spectrometry for correct molecular weights.

2.11. Specific IgE ELISA with peptides and Coc n 1

Specific IgE ELISA was done according to a previous protocol (Ghosh et al., 2015). Polystyrene plates were coated with 50 ng/well of individual peptides (P1-P5) or Coc n 1. Blocking of wells was done with 1% BSA (wt/vol) and confronted with twelve patient sera at 1:10 dilution in blocking solution overnight at 4 °C. Wells were then incubated in anti-human IgE-alkaline phosphatase tagged, produced in mouse (Sigma) at 1:1000 dilution for 3 h and IgE titers were detected by colorimetric estimation using pNPP substrate (Sigma) in an ELISA Reader (Multiskan-Labsystems, Finland) at 405 nm.

2.12. Competitive ELISA with the peptides

ELISA inhibition assay was performed according to an earlier method (Sircar et al., 2016). Pooled sera from all the twelve patients were pre-incubated with serial dilutions of Coc n 1 or synthetic peptides P1-P5 (0.01−10 μg) overnight at 4 °C. These sera were then exposed to 10 μg (50 ng/μl) plate-bound of Coc n 1 at 1:10 dilution in blocking solution at 4 °C overnight. Plates were then washed, incubated with anti-human IgE-alkaline phosphatase tagged (1:1000 dilution) and IgE titers were recorded similarly as done in ELISA. Inhibition was presented as a percentage decrease in absorbance in plate-bound samples with inhibitor to that of samples without inhibitor.

2.13. N-terminal sequencing

N-terminal protein sequencing of the major allergenic spot from 2D gel identified in the previous study (Saha et al., 2015) was done using Edman degradation. After 2D gel electrophoresis (3−10pI range), the gel was blotted onto PVDF membrane using 200 mMCAPS (Sigma) buffer, pH 11.0. This was stained briefly in 0.1 % (wt/vol) Coomassie Brilliant Blue R-250 in 50 % methanol, destained in 50 % methanol and air-dried. The membrane containing the blotted spot was cut into small pieces and microsequenced using a protein sequencer PPSQ-51A/53A (Shimadzu, Japan).

2.14. Database search and sequence similarity networks

For generating sequence similarity networks (SSNs), the amino acid sequence of Coc n 1 was used as a query in a PSI-BLAST search in the NCBInr and PDB database. Top 100 hits from NCBI and 50 from PDB were selected. Additionally, 20 proteins were selected after searching in the SDAP (Structural database of allergenic proteins) database which had an official WHO/IUIS allergen nomenclature (Ivanciuc et al., 2003). A web tool, EFI-EST was used to generate SSNs (Gerlt et al., 2015). All 170 sequences were submitted in the FASTA format and SSN finalization was done by using an alignment score of 40 after multiple iterations to achieve isofunctional clusters. Networks were visualized in Cytoscape, version 3.8.1 (Shannon et al., 2003). Alignment scores were further refined in Cytoscape and a prefuse force-directed layout was used for better visualization.

2.15. Statistical methods

All statistical calculations were done in Prism 6 Software (GraphPad Software, La Jolla California USA, www.graphpad.com).

3. Results

3.1. Determination of full-length sequence of Coc n 1

The first step towards cloning Coc n 1 was to amplify the region

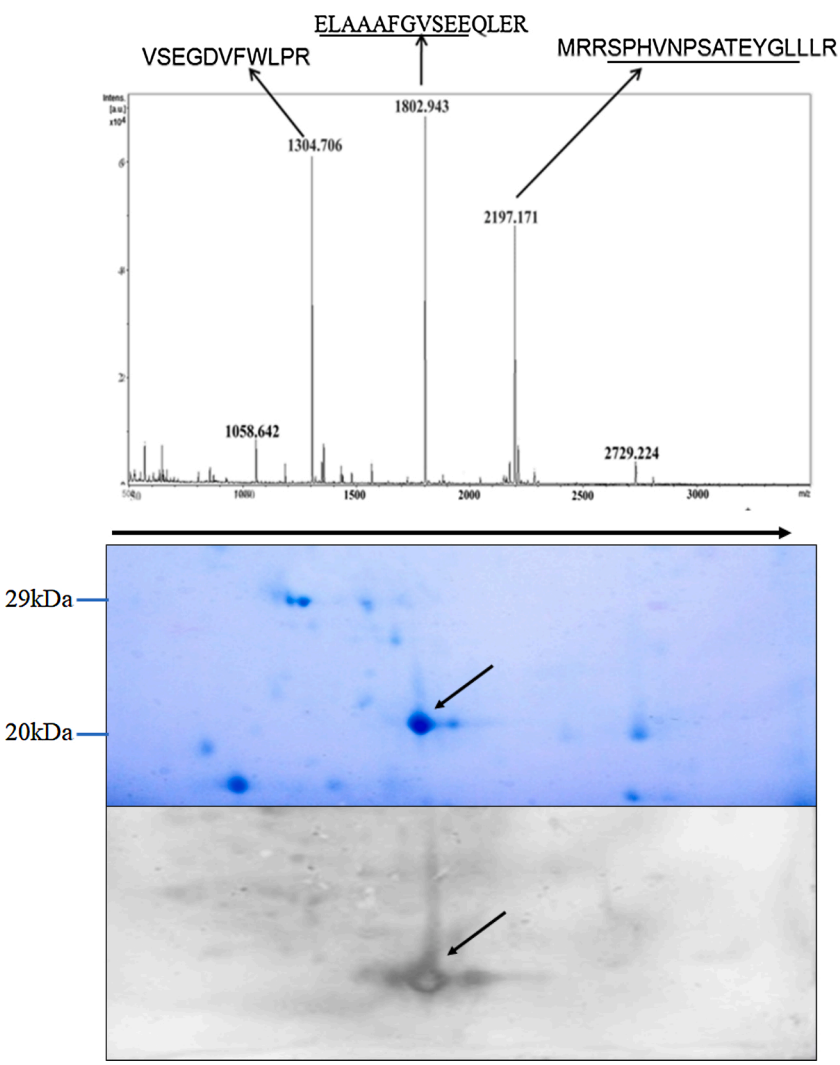


Fig. 1. Schematic showing peptides of major allergen in Coconut pollen proteome for determination of full-length sequence. The upper panel shows sequence of three peptides after MALDI-TOF identified by *de novo* sequencing (Saha et al., 2015). Underlined amino acid sequences are the regions from where the degenerate primers were designed to clone the full-length gene. The middle figure shows the position of the major allergen in the 2D gel (3-10pI) and lower figure shows the immunoblot with pooled patients' sera.

flanking between two extreme peptides identified by mass spectrometry using *de novo* sequencing from our previous study, as shown in the schematic (Fig. 1). Degenerate primers designed based on the deduced peptides amplified a partial gene sequence of 290bp corresponding to 96 amino acids from Coconut pollen RNA. 3′ RACE with Universal adapter primer linked to oligo dT amplified a 621bp cDNA amplicon showing 97 % identity with predicted "vicilin-like antimicrobial peptides" from *Elaeis guineensis,* an oil palm. Nested gene-specific primer was used to confirm amplification of the correct gene. 5′ RACE with Universal adapter primer amplified a 1047bp amplicon towards the 5′ end of the gene, which was further checked by a nested primer downstream (Supporting Information Fig S1). Upon combining the 3′ and 5′ RACE sequences, a 1738bp mRNA sequence was generated in which 1473bp coded for a protein having 490 amino acids (Fig. 2). The resulting protein sequence has a 29 amino acid (aa) signal peptide directing the protein towards the secretory pathway. N-linked glycosylation site was found at 205 aa position. Analysis in Pfam indicated the presence of two cupin domains separated by 88 aa. The first cupin domain existed between 56–211 and the second one between 299–448 aa position. The sequence was submitted to NCBI and given an accession number of KT880469. This sequence was also submitted to the WHO/IUIS allergen nomenclature subcommittee and was named as Coc n 1, the first allergen from Coconut pollen.

3.2. Overexpression of Coc n 1 in E. coli

The intact gene was cloned into pET-21d vector and expressed in *E. coli* BL21 DE3 (Rosetta) strain. Both the soluble extract in Tris buffer and the cell pellet dissolved in urea buffer were run in SDS-PAGE. No expression was seen in the soluble fraction of the lysate while high expression was seen in the pellet at 52 kDa (Fig. 3A). This was further purified by passing through affinity chromatography using Ni-NTA column. Coc n 1 was thereafter refolded by dialysis and was observed at 52 kDa as a single band without any visible contaminations.

3.3. Recombinant Coc n 1 retains allergenicity

Immunogenicity of Coc n 1 was monitored by western blotting, specific IgE ELISA and the ability to release histamine. Western blotting with the sera from ten allergic patients susceptible to Coconut pollen on a PVDF membrane blotted with Coc n 1 showed an IgE reactive band at 52 kDa region, suggesting that the recombinant allergen retains its allergenic potency. Immunoblot of purified native Coc n 1 showed immunoreactivity to two ~20 kDa bands using pooled sera. No allergenicity was seen with pooled control sera (Fig. 3B). Coc n 1 was used to assess the allergenic potency by specific IgE ELISA with individual sera from ten patients. All of the patients showed more than 3 fold higher IgE titers compared to healthy normal (Fig. 3C). Coc n 1 was analysed, whether it causes IgE to cross-link and release histamine by stripped basophil histamine release assay. More than 45 % of total histamine was released due to incubation with Coc n 1. There was less than 10 % histamine release in healthy controls and 6% after incubating stripped basophils with BSA as a negative control instead of the allergen (Fig. 3D).

3.4. Secondary and tertiary structural characteristic of Coc n 1

The secondary structural characteristic of Coc n 1 was assessed by Circular Dichroism. Upon analysing the spectra in K2D program of Dichroweb, the protein was inferred to have 7% alpha-helix, 45 % beta-sheet, 48 % random coils (Fig. 4A). Thermal stability and denaturation were checked by recording spectra at temperatures ranging from 25 to 90 °C. On observing the total spectra between 200 nm–260 nm, the negative ellipticity values increased after 45 °C. The spectrum was monitored at 222 nm and 218 nm to determine the stability of alpha-helix and beta-sheet, respectively. Alpha helix was found to have a melting temperature of 40 °C and beta-she et at 45 °C. This suggests that the protein is not temperature stable and denatures completely above 50 °C (Fig. 4B). Upon heating to 90 °C, the protein was found to precipitate in the cuvette. Cooling to 25 °C did not regain its original conformation suggesting that the heating has an irreversible effect on its proper folding. Using Size exclusion chromatography, Coc n 1 was found to exist as a trimer with ~150 kDa molecular weight (Fig. 4C).

3.5. Homology modelling

A 3D model of the protein was developed by homology modelling using five templates having maximum identity (Fig. 5A). Best homology

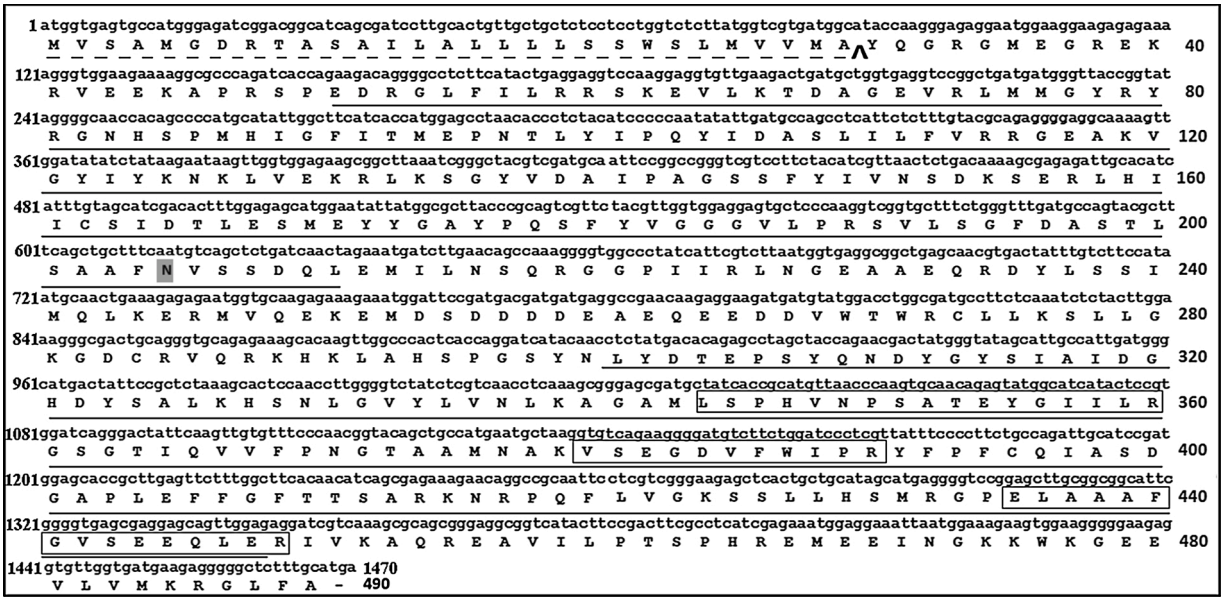


Fig. 2. Nucleotide and deduced amino acid sequence of Coc n 1. Numbers to the left show nucleotide residues and numbers to right show the amino acid residues. Potential N and O-glycosylation site is shaded. Boxed peptides were identified by MALDI-TOF during identification from crude extract. Underlined regions are part of the cupin domain. Dotted lines indicate the signal peptide.

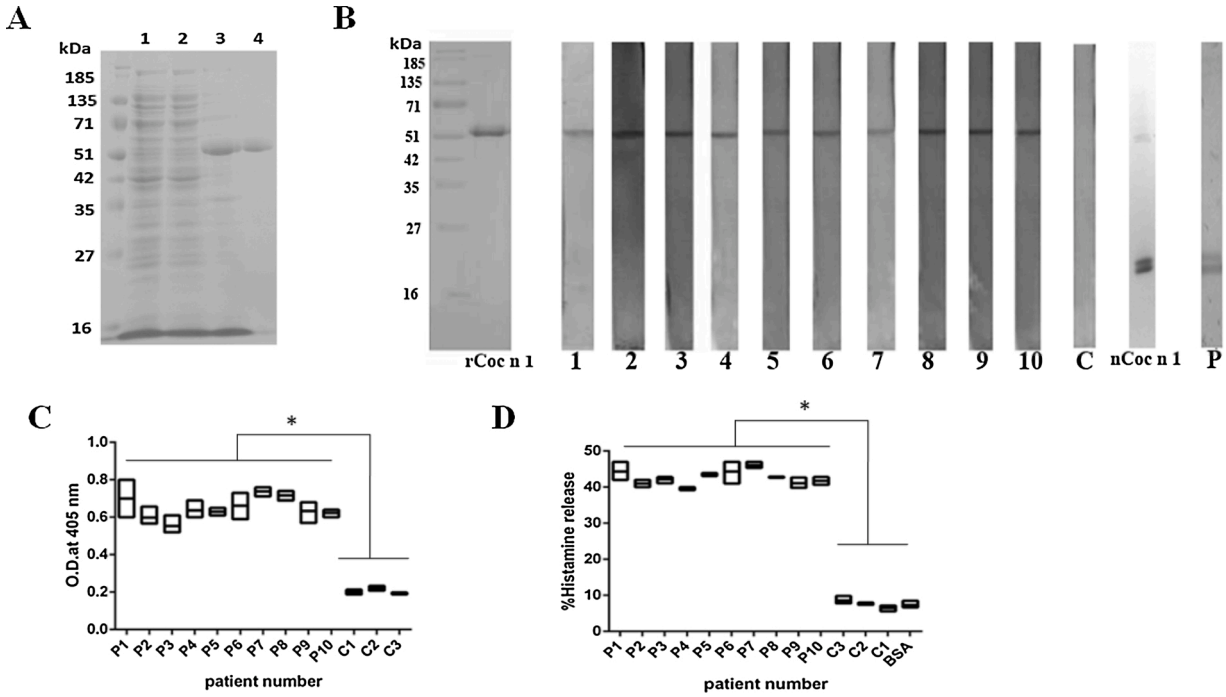


Fig. 3. Generation of Coc n 1 and validation of allergenic potency (A) Purification of Coc n 1 (A-1), uninduced cell lysate (A-2) supernatant after induction with 0.5 mM IPTG (A-3). Pellet fraction dissolved in denaturation buffer containing 7 M Urea (A-4). Purification of Coc n 1 in Ni-NTA Column. (B) Immunoblot of Coc n 1. 1-10 represent immunoblot with individual patient sera. C: Immunoblot with pooled control sera. nCoc n 1: SDS-PAGE of natively purified Coc n 1 from Coconut pollen grains. P: Immunoblot of nCoc n 1 with 10 pooled patient sera. (C) Specific IgE ELISA with 10 individual patients, C1-C3 represents IgE titers in control patients. (D) Release of histamine represented as a percentage of total histamine after deducting spontaneous release upon stimulating PBMCs with Coc n 1. C1-C3: healthy control. BSA is a negative control. * represents $p < 0.001$.

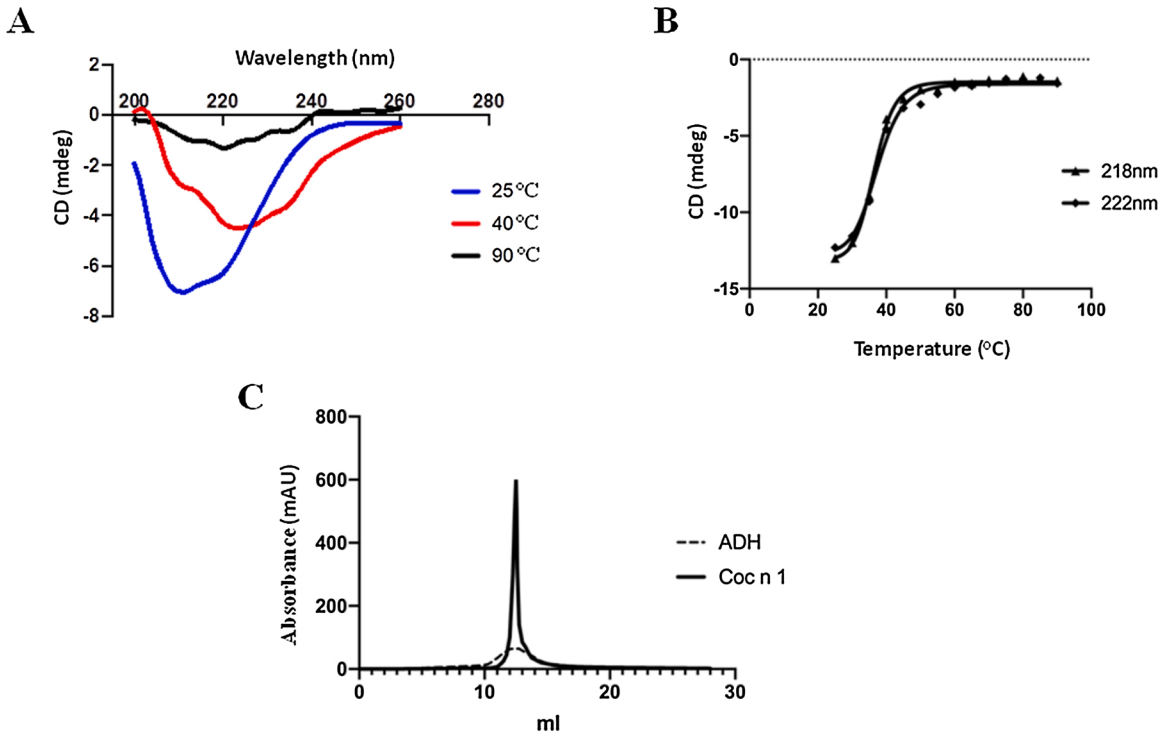


Fig. 4. Biophysical characterisation Coc n 1. (A) Far UV Circular Dichroism spectra of Coc n 1 at 25 °C, 40 °C and 90 °C. (B) Change in CD spectra at 222 nm and 218 nm with increasing temperature from 25 °C to 90 °C. Sigmoidal curve fitting done to represent data. (C) Size-exclusion chromatogram of Coc n 1 in 24 mL Superdex 200 Increase column. Alcohol Dehydrogenase 150 kDa (ADH) used as a standard.

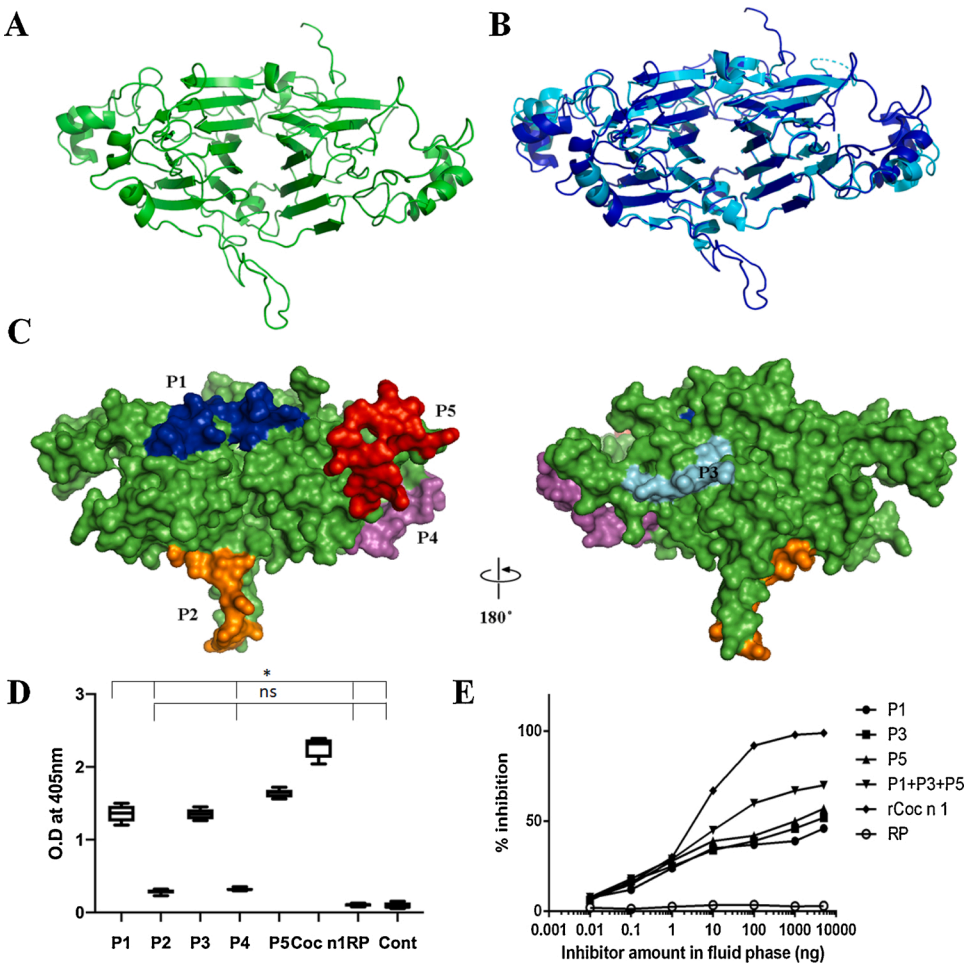


Fig. 5. Epitope mapping of Coc n 1. (A) Homology modelled structure of Coc n 1. (B) Superimposition of Coc n 1 with its template: 1IPK_A, Beta conglycinin from Soybean. (C) Mapping of predicted IgE binding epitopes on the surface filled structure of Coc n 1. (D) Specific IgE ELISA with the synthesized peptides using sera pooled from ten Coconut pollen sensitive patients. Experiments were done in triplicates and O.D. values are represented in a Box and Whisker plot where whiskers are the maximum and minimum values. The middle line in the box represents the median value. Statistical analysis was done by One way ANOVA using Tukey's Multiple comparisons test. * represents $p < 0.0001$. ns: Non significant. (E) Competitive ELISA where plate-bound Coc n 1 was confronted with increasing doses of peptides as inhibitors (P1, P2, P3, P4, P5) or combination of peptides (Equimolar mix of P1+P3+P5) or Coc n 1 (auto inhibitor) or control sera (Cont). Inhibition is represented as percentage of IgE binding inhibition to Coc n 1. RP: Random Peptide.

Table 1
IgE binding studies on predicted linear epitopes of Coc n 1 using pooled Coconut pollen sensitive sera (n = 10). Coc n 1 was used as positive control, while negative control sera were taken from healthy individuals. Solvent accessible surface area was calculated from Pymol using "get_area" command. $O.D._{pos}$ represents O.D. at 405 nm of positive control (the full-length protein), $O.D._{pep}$ represents O.D. value of the peptides, $O.D._{neg}$ represents O.D. value of the negative control.

Peptide	Sequence	Position in Coc n 1	Size(aa)	IgE binding $O.D._{pos}$:$O.D._{pep}$	IgE binding $O.D._{pep}$:$O.D._{neg}$	Surface area
P1	MEGREKRVEEKAPRSPEDR	35–54	19	1.5	12.0	738.0 $Å^2$
P2	KEMDSDDDDEAEQEE	251–266	15	4.2	5.0	1270.4 $Å^2$
P3	FTTSARKNRPQF	410–422	12	1.3	13.5	685.5 $Å^2$
P4	SEEQLERIVKAQREAV	443–459	16	6.0	3.0	1297.3 $Å^2$
P5	EEINGKKWKGEE	469–481	12	1.2	15.0	1554.3 $Å^2$

was observed with the beta chain of Beta Conglycinin from Soybean (PDB ID: 1IPK_A) as determined through DOPE score in Modeller 9.15. The main chain conformation of the modelled protein was found to be present in the acceptable region of Ramachandran plot. As observed in the plot, 89 % of the residues were in the most favoured region and 6% were in the allowed region. The loop-refine command in Modeller helped to refine the areas where there was no homology. The model can be divided into two very similar modules, the N- and C-terminal modules, separated by a pseudo-dyad axis. Each module contains a conserved beta-barrel core domain also known as the cupin domain and an extended loop domain containing several helices (Supporting Information Fig S2). Upon superimposing the homology the modelled Coc n 1 and Beta conglycinin from Soybean, it was observed that while most parts of the core structure match with an RMSD value of 0.4 Å, a disordered loop region between the two cupin domains in Coc n 1 is unusually longer (Fig. 5B). The C-terminal also showed a 23 amino acid disordered region due to a lack of homology with known structures.

3.6. Epitope mapping and peptide synthesis

B cell epitopes were predicted on the mature full-length sequence of Coc n 1 using *in-silico* tools. Five regions commonly identified by all the three tools showing a high probability of antibody binding were selected and synthesized by FMOC chemistry. The peptides are P1: 35–54, P2: 251–266, P3: 410–422, P4: 443–459, P5: 469–481 (Table 1). The peptides were mapped on the surface filled model of Coc n 1 (Fig. 5C). P1 was located near the N terminus and P2 was present in the loop between the two cupin domains. P3 was present towards the beginning of the C-terminal cupin domain, while P4 and P5 were located towards the C-terminal end. P2, P3 and P5 were found mostly in the random coil region. P4 was present in the helix region and P1 was found in the junction of a helix and random coil.

3.7. Specific IgE ELISA with the epitope containing peptides

Specific IgE ELISA with the 5 peptides were done to assess the ability of the synthesised peptides to bind IgE antibodies. Pooled human blood sera allergic to Coconut pollen grain were added as a primary antibody. P1, P3 and P5 showed markedly higher IgE reactivity compared to the other two (Fig. 5D). The highest IgE binding was observed in P5 followed by P3 and P1. The IgE reactivity was more than 10 times in the immunodominant epitopes compared to a Random Peptide (RP) used as a negative control. P2 and P4 showed more than four times less IgE reactivity compared to the positive control, Coc n 1 (Table 1). There was no IgE binding to all peptides combined with pooled sera from control patients.

3.8. Specific IgE inhibition with 3 immunodominant epitopes

Using a competitive ELISA, the immunogenic strength of the peptides (liquid phase) was quantified as a measure to inhibit IgE binding to intact Coc n 1 (solid phase). Graded amounts of immunodominant peptides: P1, P3 and P5 (0.01 ng to 10 μg) were incubated with pooled Coconut pollen sensitive sera (n = 10) and a dose-dependent inhibition of IgE reactivity was monitored. While P3 and P5 showed 50 % IgE inhibition (IC_{50}) at 5 μg and 1 μg, respectively. P1 did not reach an IC_{50} in the desired range of inhibition. Using a combination of all three peptides, the IC_{50} value reduced to 20 ng. Coc n 1 was used as a self-inhibitor and the IC_{50} reached at 5 ng (Fig. 5E).

3.9. N-terminal microsequencing

The total proteome of Coconut pollen was profiled in a 3−10pI range 2D gel and the spot containing Coc n 1 protein at ~20 kDa was microsequenced by Edman degradation. A ten amino acid N-terminal sequence of HSPGSYNLYD was determined. This sequence was searched in the full length protein of Coc n 1 and was found to exist between the 294th and 303rd amino acid position. The 294th amino acid position was found at the extended loop region between the cupin domains, suggesting that the protein breaks into two parts at this point, each part containing one cupin domain and the C-terminal cupin domain gives rise to this ~20 kDa allergenic subunit.

3.10. Sequence similarity network generation and multiple sequence alignment

Coc n 1 was queried for similarity with other proteins in the allergen (SDAP), NCBInr and the RCSB PDB database. 170 related protein sequences were collated and clustered based on sequence similarity. Using EFI-EST tool, sequence similarity networks (SSNs) were generated by grouping proteins having similar protein sequences. Proteins are represented by nodes, while edges represent sequence similarity between two nodes. After further optimizing parameters and filtering alignment scores to display network that have higher stringency, three well-defined clusters could be identified (Fig. 6). Coc n 1 was present in Vicilin 2 cluster which includes proteins that are uncharacterised or unannotated. This cluster is very compact, suggesting a high degree of sequence similarity. Vicilin 1 cluster included all 7S globulin proteins and allergens. 7S globulins having crystal structure also fell in this cluster. The proteins in these two clusters are highly similar in structure; however, they had 25–29 % sequence similarity. Between the two clusters, Korean Pine vicilin (PDB ID: 4LEJ) seemed to be evolutionarily closer to the vicilin 2 cluster and could be a close link between the two groups. The legumin cluster included few 11S globulin allergens that were obtained after a similarity search in the SDAP allergen database. Vicilin-like proteins from *Phoenix dactylifera* and *Elaeis guineensis* were found to be the closest homologs with more than 99 % sequence similarity. A multiple sequence alignment was done with 13 proteins representing both the clusters showed structural conservation in the cupin region despite having a low degree of sequence similarity (Supporting Information Fig S3).

4. Discussion

Amongst the abundant seed proteins, 7S globulin and 11S legumins are considered as most potent allergenic proteins. However, vicilin-like proteins were never reported from pollen grains. In this study, we report the full-length sequence and IgE reactivity of an unannotated vicilin-like protein, Coc n 1 from Coconut pollen grains.

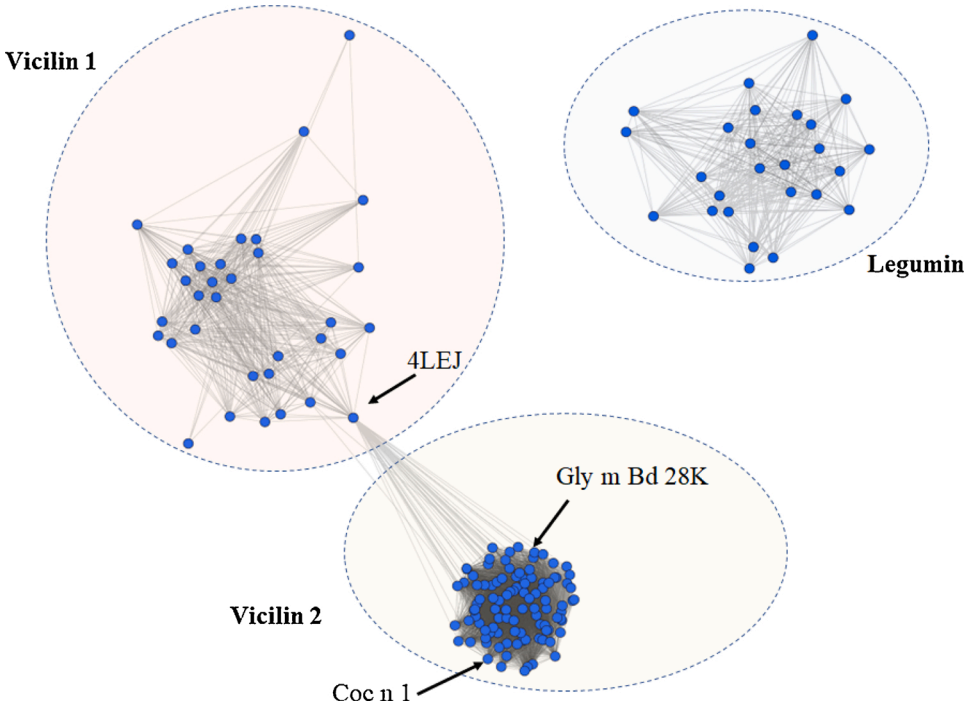


Fig. 6. Sequence similarity network. 170 proteins collected by PSI-BLAST with Coc n 1 full-length sequence in NCBInr, PDB and allergen database (SDAP). SSNs generated by EFI-EST enzyme similarity tools. Proteins are represented by dots and edges represent similarity between two proteins beyond threshold (above alignment score 40). Clusters visualized in Cytoscape using Prefuse Force Directed Layout method. Three clusters represent Legumin, Vicilin 1 and Vicilin 2. Vicilin 1 contains known 7S globulin allergens, while Vicilin 2 contains structural similar vicilin-like proteins with unknown functions.

Coconut being unsequenced, the *de novo* sequenced peptides identified previously, were used as the only information to determine an internal sequence and thereafter the full-length sequence of Coc n 1. As many allergens arise from unsequenced plants, peptides have been used to unravel full-length genes. For instance, peptides from Pectin Methylesterase, an allergen from olive, assisted in cloning the full-length gene using RACE (Salamanca et al., 2010). Coc n 1 shared the highest sequence identity with a family of proteins named "vicilin-like antimicrobial peptides" which were generated from automated genome annotation. Thereafter, the identity varies between 55–60 % with hypothetical proteins, 25–30 % with established 7S globulin allergens whose structures are known or have been experimentally validated and finally 20 % similarity with 11S legumins. Upon superimposing Coc n 1 with 1IPK_A, a 7S globulin from Soybean, the core structure overlapped exceptionally well except for a considerably longer linker region in between the two cupin domains. Such extended linker region is also present in three other cupin family proteins, such as soybean allergen Gly m Bd 28 kDa, MP27/MP32 precursor protein from pumpkin and Gea8 globulin like protein from carrot. They were predicted to form a separate cluster in the evolutionary tree of seed globulins (Shutov et al., 1998).

IgE reactivity was observed in a 20 kDa spot having peptides from Coc n 1 in the 2D of the total proteome of Coconut pollen in our previous study (Saha et al., 2015). This was also confirmed by purifying the protein from Coconut pollen which showed IgE reactivity in two bands both identified as vicilin protein near 20 kDa. Since the full-length protein is now determined to be 52 kDa, this led us to hypothesize that the allergenic spot could be a proteolytic product of the mature form. Using N-terminal sequencing, Coc n 1 was inferred to be processed into a 197 amino acid (21.7 kDa), 6.66pI C-terminal subunit. The cleavage occurs between Alanine (293) and Histidine (294) in the linker region between two cupin domains but presence of a specific protease cleavage site could not be ascertained. This linker region is supposed to be a hypervariable site in legume seed storage proteins, which contains specific sites for proteolytic cleavage *in vivo* resulting in acidic and basic subunits. As vicilins are food allergens, this region is also predicted to be the site of protease activity in the digestive tract (Xiang et al., 2015). A similar observation was seen in Gly m Bd 28 kDa, which gets proteolytically processed to 212 and 240 amino acid subunits. Both the N and C-terminal subunits of Gly m Bd 28 kDa were described to be allergenic (Tsuji et al., 2001). Proteolytic cleavage plays an important part in storage protein deposition and reactivation in seeds. Vicilin-like proteins have signal peptides which guide into the ER-lumen, entering the secretory pathway through the endomembrane system. The internal degradation happens inside the vacuolar membrane (Müntz, 1996). In the recombinant protein, no such activity was not found and an intact full-length protein band was observed. In the recombinant form, Coc n 1 retained its stoichiometry as a trimer similar to the native protein (Saha et al., 2015). Hence a model was suggested where the Coc n 1 trimer is processed *in vivo* to generate a C-terminal subunit that is recognised as a major allergen in Coconut pollen (Fig. 7 A, B).

SSN is an alternative method to group closely related proteins based on sequence similarity and ease to represent on a larger dataset. SSNs

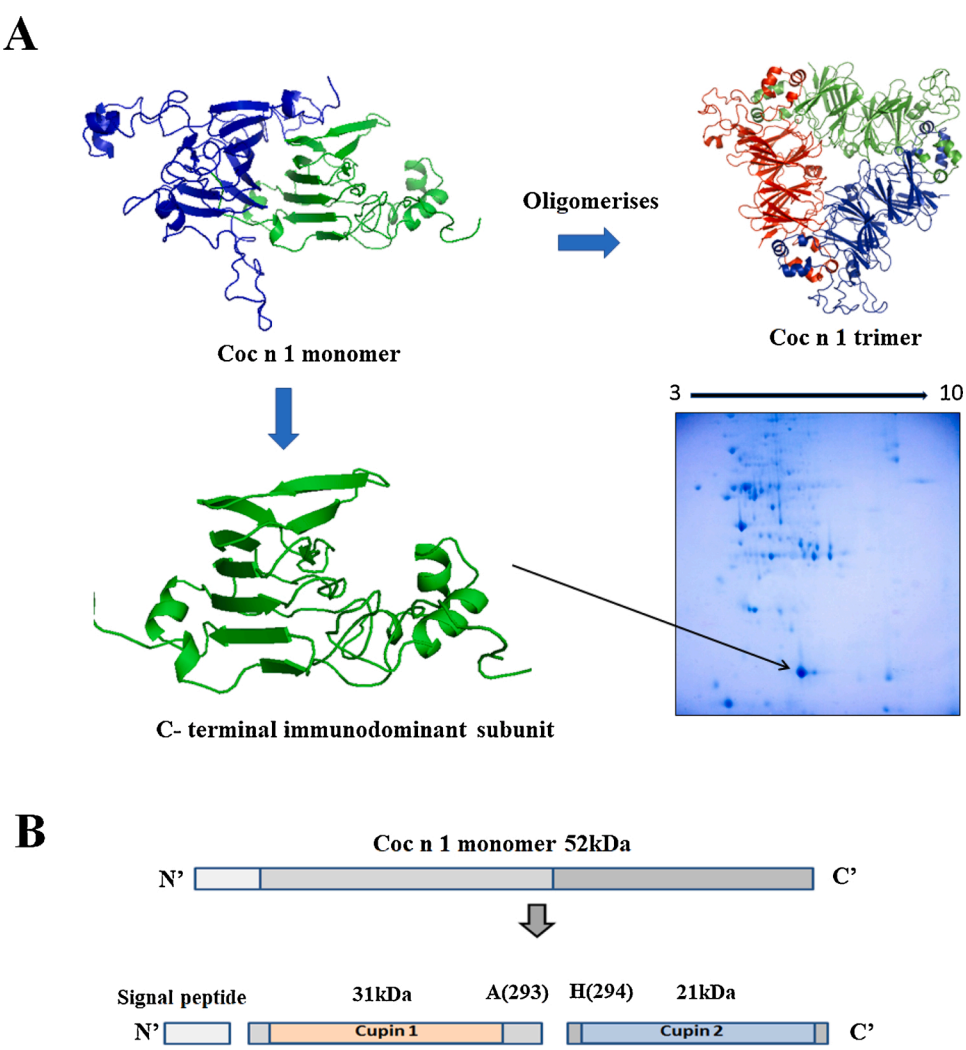


Fig. 7. Model of proteolytic processing of mature Coc n 1 to generate immunodominant allergen. (A) Coc n 1 oligomerises to a trimer and its monomer also gets cleaved *in vivo* to give two subunits. The C-terminal immunodominant subunit identified to be the major allergen from Coconut pollen grains and its position is shown in the 2D gel. (B) Schematic representation of precursor Coc n 1 showing site of cleavage of signal peptide, and a proteinase K site to produce the immunodominant subunit. The position of the cupin domains in each subunit is also shown.

were used to infer functions and evolutionary relationships to proteins in the cupin superfamily (Uberto and Moomaw, 2013). SSNs were used to analyse Coc n 1 with related sequences, which resulted in three different protein clusters. Legumins are hexameric cupin domain-containing proteins that formed a separate cluster (Vicilin 1). They are also abundant seed storage proteins and major allergen in nuts such as Ara h 3 in peanut (Jin et al., 2009). While all of the proteins in the Coc n 1 cluster (Vicilin2) were unannotated, only Gly m Bd 28 kDa allergen was found to be studied in considerable detail. Proteins designated as "Vicilin-like antimicrobial peptides" also belonged to this cluster. One such protein was reported to be processed into peptides in Capsicum seeds that were responsible for antifungal activity (Vieira Bard et al., 2014). However, there has been no study on the antimicrobial property of pollen grains. It can be hypothesised that pollen grains are viable for long periods and being a rich source of proteins, carbohydrates may generate such peptides to prevent a microbial attack. Further study may shed light on the function of such proteins in pollen grains, which could unearth some exciting biology.

The low percentage of helix deduced from Circular Dichroism is evident from the homology-based structure where helices are only found in the extended loop region towards the N and C terminus. CD spectra of Coc n 1 showed a minima at 212 nm. Jug r 6, another vicilin allergen from Walnut, showed a predominantly beta-sheet with minima at 218 nm (Dubiela et al., 2018). With increasing temperature, the spectral shape changed, indicating unfolding and precipitation. Overall, Coc n 1 started to denature above 40 °C and was completely denatured at 50 °C. Jug r 6 showed an almost similar folding characteristic where it started denaturing at 45 °C and became completely unfolded at 75 °C. However, another study revealed that Korean Pine vicilin is highly heat-stable and does not denature until 95 °C (Jin et al., 2008). Though 7S globulins are reported to have high temperature stability, Coc n 1 did not show such characteristics. As Coc n 1 does not fall in cluster of established 7S globulin allergens, structural stability may not be as comparable. Hence biophysical characterisation of proteins in the Vicilin 2 cluster are also need to be studied to validate further and understand the heat-induced structural changes.

Strong IgE binding in western blotting suggested that the recombinant protein retains its IgE binding. Since after heat denaturation, Coc n 1 is incapable of refolding, it can be presumed that the signals in immunoblot are entirely due to linear epitopes. Specific IgE ELISA also showed markedly higher IgE titers compared to control. Conformational epitopes are difficult to assess as it requires a thorough understanding of the crystal structure in conjunction with the antibody and has not been investigated here. The linear nature of continuous B-cell epitopes has given rise to a number of prediction approaches that are based on sequence analysis methods. Using such prediction tools, five linear epitopes were mapped on the Coc n 1 surface. P1, P3 and P5, which showed elevated IgE binding, were present in the coiled regions where P5 occupied the maximum solvent accessible surface area amongst the three. In Ara h 1, a well characterised vicilin allergen, the epitopes were mainly found in the extended loop region between the cupin domains (Chruszcz et al., 2011). Peptide P2 was seen in a similar region but did not show IgE reactivity. The low IgE binding ability of peptide P2 and P4 could be due to combined effect of surface accessibility, low antigenic propensity and hydrophobicity. The peptides were able to inhibit a maximum of 42 % (P1), 48 % (P3) and 51 % (P5) of IgE binding to Coc n 1 while using a combination of the three peptides, maximum inhibition reached 60 %. This suggests that P3 and P5 are more potent in IgE binding. The equivalent region of P3 in Gly m Bd 28 kDa was also reported to bind IgE, suggesting that they could display cross-reactivity (Xiang et al., 2004). Both P3 and P5 peptides reside in the C-terminal subunit of Coc n 1, supporting its immunodominant nature. Partial inhibition of the peptides to prevent IgE binding to intact Coc n 1 suggests that these peptides might not represent explicitly the library of linear epitopes or that additional conformational epitopes are important for IgE antibody binding. However, it is also possible that some IgE antibodies may prefer topological conformations on the Coc n 1 along with its neighbouring amino acids.

5. Conclusion

The present study has delineated the identification and characterisation of Coc n 1, a vicilin-like protein from Coconut pollen. This is the first report of a cupin domain protein from any pollen grain. As a novel allergen in the cupin superfamily, cross-reactivity with known 7S globulin allergens could be investigated to unearth presence of shared epitopes. Identification of dominant linear IgE binding epitopes on Coc n 1 may assist in the rational designing of hypoallergens that can be used for immunotherapy in sensitive patients.

Author contributions

The entire study was conceived and designed by SGB and BS. BS and BK performed all the experiments. BS and SGB wrote the manuscript.

Funding sources

BS is a supported by a fellowship from University Grants CommissionF.2-8/2002(SA-I). The entire study was supported by Department of Science and Technology, Government of India, Institutional Plan Project-1, Bose Institute.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

The authors are thankful to Chanchal Chakraborty, Jadab Ghosh, Swarup Biswas, Smriti Ranjan Maji, Amarendra Nath Biswas, Tanmoy Debnath, Kaberi Ghosh and Soumyo Shubhro Gupta for their technical help and support. Dr. Sudipto Saha is acknowledged for helping in epitope mapping and bioinformatics.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.molimm.2020.12.026.

References

Atkinson, H.J., Morris, J.H., Ferrin, T.E., Babbitt, P.C., 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One 4, e4345. https://doi.org/10.1371/journal.pone.0004345.

Bhattacharya, K., Sircar, G., Dasgupta, A., Gupta Bhattacharya, S., 2018. Spectrum of allergens and allergen biology in India. Int. Arch. Allergy Immunol. 177, 219–237. https://doi.org/10.1159/000490805.

Bublin, M., Breiteneder, H., 2014. Cross-reactivity of peanut allergens. Curr. Allergy Asthma Rep. 14, 426. https://doi.org/10.1007/s11882-014-0426-8.

Burr, M.L., Emberlin, J.C., Treu, R., Cheng, S., Pearce, N.E., ISAAC Phase One Study Group, 2003. Pollen counts in relation to the prevalence of allergic rhinoconjunctivitis, asthma and atopic eczema in the International Study of Asthma and Allergies in Childhood (ISAAC). Clin. Exp. Allergy 33, 1675–1680. https://doi.org/10.1111/j.1365-2222.2003.01816.x.

Chowdhury, I., Chakraborty, P., Gupta Bhattacharya, S., Chanda, S., 1998. Allergenic relationship among four common and dominant airborne palm pollen grains from Eastern India. Clin. Exp. Allergy 28, 977–983. https://doi.org/10.1046/j.1365-2222.1998.00364.x.

Chruszcz, M., Maleki, S.J., Majorek, K.A., Demas, M., Bublin, M., Solberg, R., Hurlburt, B. K., Ruan, S., Mattisohn, C.P., Breiteneder, H., Minor, W., 2011. Structural and immunologic characterization of Ara h 1, a major peanut allergen. J. Biol. Chem. 286, 39318–39327. https://doi.org/10.1074/jbc.M111.270132.

Dey, D., Saha, B., Sircar, G., Ghosal, K., Gupta Bhattacharya, S., 2016. Mass spectrometry-based identification of allergens from Curvularia pallescens, a prevalent aerospore in India. Biochim. Biophys. Acta (BBA)-Proteins Proteom. 1864, 869–879. https://doi.org/10.1016/j.bbapap.2016.03.012.

Dubiela, P., Kabasser, S., Smargiasso, N., Geiselhart, S., Bublin, M., Hafner, C., Mazzucchelli, G., Hoffmann-Sommergruber, K., 2018. Jug r 6 is the allergenic vicilin

present in walnut responsible for IgE cross-reactivities to other tree nuts and seeds. Sci. Rep. 8, 1–14. https://doi.org/10.1038/s41598-018-29656-4.

Dunwell, J.M., Purvis, A., Khuri, S., 2004. Cupins: the most functionally diverse protein superfamily? Phytochemistry 65, 7–17. https://doi.org/10.1016/j.phytochem.2003.08.016.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47, D427–D432. https://doi.org/10.1093/nar/gky995.

EL-Manzalawy, Y., Dobbs, D., Honavar, V., 2008. Predicting linear B-cell epitopes using string kernels. J. Mol. Recognit.: Interdiscipl. J. 21, 243–255. https://doi.org/10.1002/jmr.893.

Galperin, M.Y., Koonin, E.V., 2012. Divergence and convergence in enzyme evolution. J. Biol. Chem. 287, 21–28. https://doi.org/10.1074/jbc.R111.241976.

Ganguly, H.K., Majumder, B., Chattopadhyay, S., Chakrabarti, P., Basu, G., 2012. Direct evidence for CH··· π interaction mediated stabilization of pro-cis pro bond in peptides with pro-pro-aromatic motifs. J. Am. Chem. Soc. 134, 4661–4669. https://doi.org/10.1021/ja209334v.

Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., Whalen, K.L., 2015. Enzyme function initiative-enzyme similarity tool (EFI-EST): a web tool for generating protein sequence similarity networks. Biochim. Biophys. Acta (BBA)-Proteins Proteom. 1854, 1019–1037. https://doi.org/10.1016/j.bbapap.2015.04.015.

Ghosh, N., Sircar, G., Saha, B., Pandey, N., Gupta Bhattacharya, S., 2015. Search for allergens from the pollen proteome of sunflower (Helianthus annuus L.): a major sensitizer for respiratory allergy patients. PLoS One 10, e0138992. https://doi.org/10.1371/journal.pone.0138992.

Gupta, R., Jung, E., Brunak, S., 2004. Prediction of N-Glycosylation Sites in Human Proteins. NetNGlyc 1.0 Server.

Ivanciuc, O., Schein, C.H., Braun, W., 2003. SDAP: database and computational tools for allergenic proteins. Nucleic Acids Res. 31, 359–362. https://doi.org/10.1093/nar/gkg010.

Jespersen, M.C., Peters, B., Nielsen, M., Marcatili, P., 2017. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 45, W24–W29. https://doi.org/10.1093/nar/gkx346.

Jin, T., Albillos, S.M., Chen, Y.W., Kothary, M.H., Fu, T.J., Zhang, Y.Z., 2008. Purification and characterization of the 7S vicilin from Korean pine (Pinus koraiensis). J. Agric. Food Chem. 56, 8159–8165. https://doi.org/10.1021/jf801138q.

Jin, T., Guo, F., Chen, Y.W., Howard, A., Zhang, Y.Z., 2009. Crystal structure of Ara h 3, a major allergen in peanut. Mol. Immunol. 46, 1796–1804. https://doi.org/10.1016/j.molimm.2009.01.023.

Kesari, P., Sharma, A., Katiki, M., Kumar, P., R Gurjar, B., Tomar, S., K Sharma, A., Kumar, P., 2017. Structural, functional and evolutionary aspects of seed globulins. Protein Pept. Lett. 24, 267–277. https://doi.org/10.2174/0929866523666161220112641.

Mansouritorghabeh, H., Jabbari-Azad, F., Sankian, M., Varasteh, A., Farid-Hosseini, R., 2019. The most common allergenic tree pollen grains in the Middle East: a narrative review. Iran. J. Med. Sci. 44, 87.

Müntz, K., 1996. Proteases and proteolytic cleavage of storage proteins in developing and germinating dicotyledonous seeds. J. Exp. Bot. 47, 605–622. https://doi.org/10.1093/jxb/47.5.605.

Pawankar, R., 2014. Allergic diseases and asthma: a global public health concern and a call to action. World Allergy Organ. J. 7, 12. https://doi.org/10.1186/1939-4551-7-12.

Petersen, T.N., Brunak, S., Von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786. https://doi.org/10.1038/nmeth.1701.

Platts-Mills, T.A., Woodfolk, J.A., 2011. Allergens and their role in the allergic immune response. Immunol. Rev. 242, 51–68. https://doi.org/10.1111/j.1600-065X.2011.01021.x.

Saha, B., Gupta Bhattacharya, S., 2017. Charting novel allergens from date palm pollen (Phoenix sylvestris) using homology driven proteomics. J. Proteomics 165, 1–10. https://doi.org/10.1016/j.jprot.2017.05.021.

Saha, S., Raghava, G.P.S., 2004. BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In: International Conference on Artificial Immune Systems. Springer, Berlin, Heidelberg, pp. 197–204.

Saha, B., Sircar, G., Pandey, N., Gupta Bhattacharya, S., 2015. Mining novel allergens from coconut pollen employing manual de novo sequencing and homology-driven proteomics. J. Proteome Res. 14, 4823–4833. https://doi.org/10.1021/acs.jproteome.5b00657.

Salamanca, G., Rodríguez, R., Quiralte, J., Moreno, C., Pascual, C.Y., Barber, D., Villalba, M., 2010. Pectin methylesterases of pollen tissue, a major allergen in olive tree. FEBS J. 277, 2729–2739. https://doi.org/10.1111/j.1742-4658.2010.07689.x.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504. https://doi.org/10.1101/gr.1239303.

Shutov, A.D., Braun, H., Chesnokov, Y.V., Bäumlein, H., 1998. A gene encoding a vicilin-like protein is specifically expressed in fern spores: evolutionary pathway of seed storage globulins. Eur. J. Biochem. 252, 79–89. https://doi.org/10.1046/j.1432-1327.1998.2520079.x.

Sircar, G., Saha, B., Mandal, R.S., Pandey, N., Saha, S., Gupta Bhattacharya, S., 2015. Purification, cloning and immuno-biochemical characterization of a fungal aspartic protease allergen Rhi o 1 from the airborne mold Rhizopus oryzae. PLoS One 10, e0144547. https://doi.org/10.1371/journal.pone.0144547.

Sircar, G., Jana, K., Dasgupta, A., Saha, S., Gupta Bhattacharya, S., 2016. Epitope mapping of Rhi o 1 and generation of a hypoallergenic variant A CANDIDATE MOLECULE FOR FUNGAL ALLERGY VACCINES. J. Biol. Chem. 291, 18016–18029. https://doi.org/10.1074/jbc.M116.732032.

Tsuji, H., Hiemori, M., Kimoto, M., Yamashita, H., Kobatake, R., Adachi, M., Fukuda, T., Bando, N., Okita, M., Utsumi, S., 2001. Cloning of cDNA encoding a soybean allergen, Gly m Bd 28K. Biochim. Biophys. Acta (BBA)–Gene Struct. Expression 1518, 178–182. https://doi.org/10.1016/S0167-4781(00)00310-9.

Uberto, R., Moomaw, E.W., 2013. Protein similarity networks reveal relationships among sequence, structure, and function within the cupin superfamily. PLoS One 8, e74477. https://doi.org/10.1371/journal.pone.0074477.

Vieira Bard, G.C., Nascimento, V.V., Oliveira, A.E.A., Rodrigues, R., Da Cunha, M., Dias, G.B., Vasconcelos, I.M., Carvalho, A.O., Gomes, V.M., 2014. Vicilin-like peptides from Capsicum baccatum L. seeds are α-amylase inhibitors and exhibit antifungal activity against important yeasts in medical mycology. J. Pept. Sci. 102, 335–343. https://doi.org/10.1002/bip.22504.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296–W303. https://doi.org/10.1093/nar/gky427.

Webb, B., Sali, A., 2016. Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinformatics 54, 5–6. https://doi.org/10.1002/0471250953.bi0506s15.

Whitmore, L., Wallace, B.A., 2004. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. Nucleic Acids Res. 32, W668–W673. https://doi.org/10.1093/nar/gkh371.

Xiang, P., Haas, E.J., Zeece, M.G., Markwell, J., Sarath, G., 2004. C-Terminal 23 kDa polypeptide of soybean Gly m Bd 28 K is a potential allergen. Planta 220, 56–63. https://doi.org/10.1007/s00425-004-1313-7.

Supporting Information

Cloning, expression and immunological characterisation of Coc n 1, the first major allergen from Coconut pollen

Bodhisattwa Saha[1,2*], Bijoya Karmakar[1], Swati Gupta Bhattacharya[1*]

[1] Division of Plant Biology, Bose Institute, 93/1, Acharya Prafulla Chandra Road, Kolkata 700009, India.
[2] Present address: CSIR-Institute of Genomics and Integrative Biology, Mathura Road, New Delhi 110025, India.
Running title: Novel allergen from Coconut pollen.

*To whom correspondence should be addressed:
Prof. Swati Gupta Bhattacharya, Division of Plant Biology, Bose Institute, 93/1, Acharya Prafulla Chandra Road, Kolkata 700009, India. Telephone: (+91) 3323031131, Fax: (+91) 3323506790. Email: swati@jcbose.ac.in
*Bodhisattwa Saha, Institute of Genomics and Integrative Biology, Mathura Road, New Delhi 110025, India. Telephone: +91 11-29879487. Email: bodhisattwasaha@gmail.com

Conflict of interest:

The authors declare no conflicts of interest.

This document contains:

Supplementary Table 1

Supplementary Fig 1-3

Index

Table 1: Demographic data of allergic patients recruited in the study.

Fig S1: Generation of full-length Coc n 1 by 3' and 5' RACE

Fig S2: Homology modelling of Coc n 1.

Fig S3: Multiple sequence alignment of 14 proteins having cupin domains using T-Coffee server

Patient	Age	Sex	Symptoms	Skin Test	Specific IgE (Patient: Normal)	Total IgE (IU/ml)	Other pollen allergy
1	20	M	Allergic rhinitis, rhino conjunctivitis, allergic asthma	+3	3.5	4243.28	PS, BF, AC
2	35	M	Allergic rhinitis, urticaria, allergic asthma	+3	4.2	2736.33	PS, HA, AS
3	39	M	Allergic rhinitis, rhino conjunctivitis	+4	5.5	2179.56	PS, AC, BF
4	42	F	Allergic rhinitis	+3	3.8	4011.23	PS, AC
5	58	M	Allergic rhinitis, allergic asthma	+2	3.9	3982.81	PS,BF
6	43	F	Allergic rhinitis	+3	4.5	3536.77	PS, AC, AS
7	38	F	Allergic rhinitis	+4	4.7	3421.54	BF, HA
8	55	M	Allergic rhinitis, allergic asthma	+4	3.6	2899.63	AC, BF, HA
9	52	F	Allergic rhinitis, urticaria, allergic asthma	+4	4.0	3423.88	BF, HA
10	39	M	Allergic rhinitis, rhino conjunctivitis	+3	3.6	4545.11	PS, BF
C1	26	M	No symptom	---	---	102.77	---
C2	29	M	No symptom	---	---	69.88	---
C3	36	M	No symptom	---	---	73.77	---

Table S1: Demographic data of allergic patients recruited in the study.

PS: *Phoenix sylvestris*; BF: *Borassus flabellifer*; AC*: Areca* catechu; HA: *Helianthus annuus*; AS: *Alstonia scholaris*. C1-C4 are negative control patients who are healthy and do not have demonstrated allergic syndromes. All clinical data of the human subjects were retrieved from medical records maintained at Mediland Diagnostics, Kolkata from where blood sample was also collected. Total IgE was determined from Total IgE Elisa Kit (Omega Diagnostics).

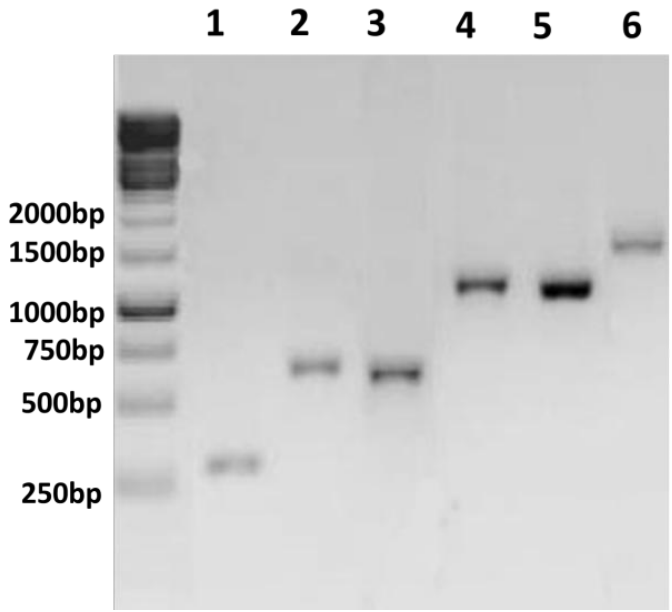
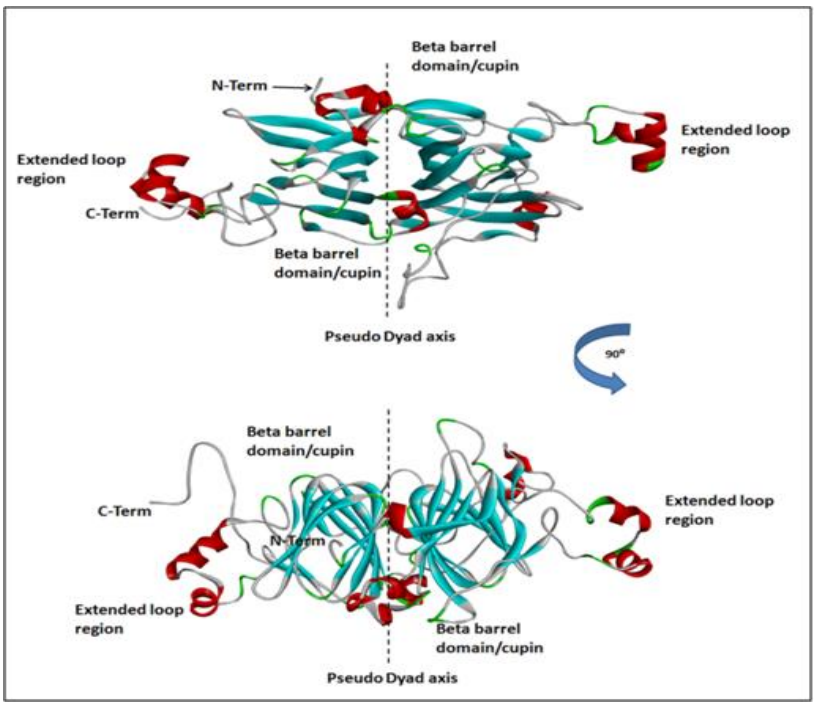


Fig S1: Generation of full length Coc n 1 by 3' and 5' RACE.

(Lane 1): Amplicon using degenerate primers from *de novo* sequenced peptides, (Lane 2): 3'RACE, (Lane 3): 3'RACE with nested primer, (Lane 4): 5'RACE, (Lane 5): 5'RACE with nested primer, (Lane 6): Full-length cDNA.

A



B

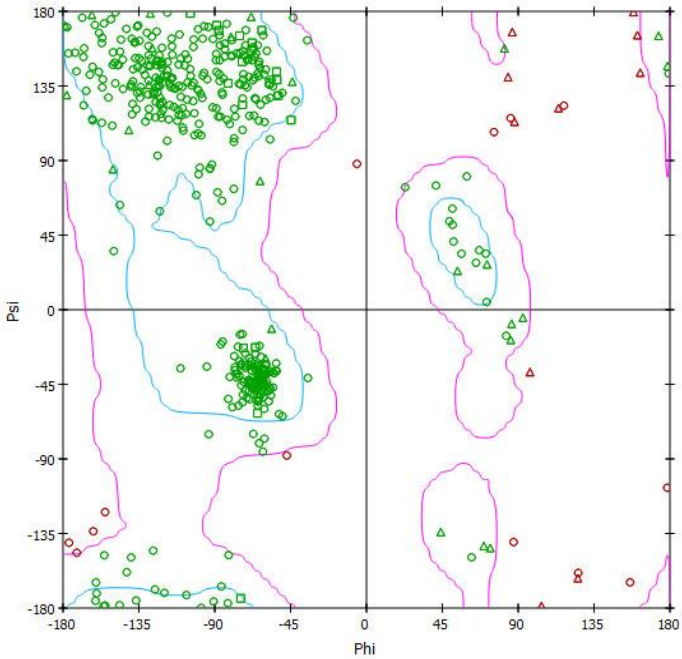


Fig S2: Homology modelling of Coc n 1.

A: Homology modelled structure of Coc n 1 with characteristic features common in 7S globulins. Image generated using Discovery studio. (https://discover.3ds.com/discovery-studio-visualizer-download). B: Ramachandran Plot of homology based model of Coc n 1 using SAVES server (https://servicesn.mbi.ucla.edu/SAVES/).

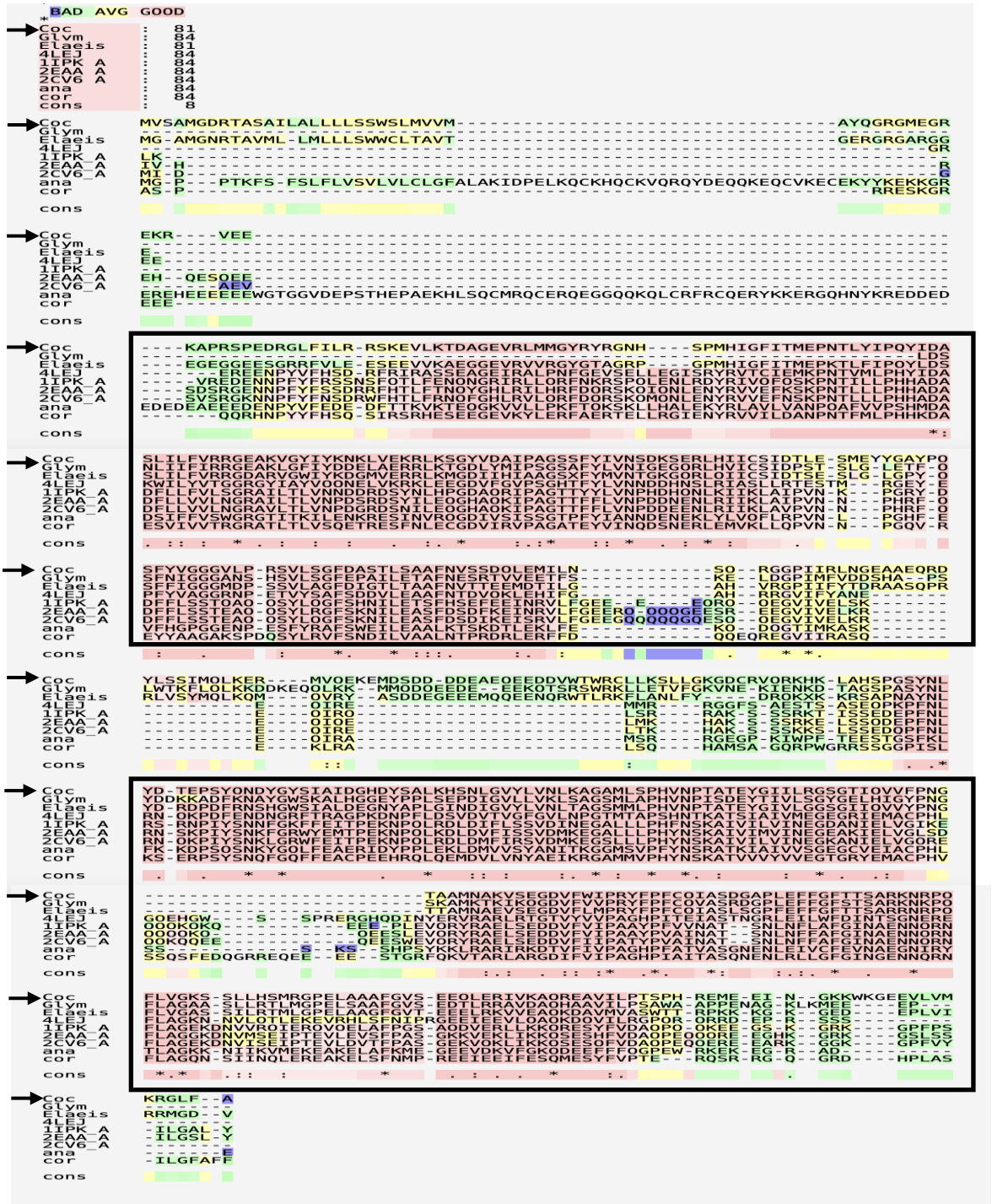


Fig S3: Multiple sequence alignment of 14 proteins having cupin domains using T-Coffee server (Di Tomasso et al., 2011).

1IPK, 2CV6, 3SMH, 2EAA, 4LEJ are known allergens having defined PDB structures. Cor a 11 and Ana o 1 are also known cupin domain containing allergens. Coc, Glym and Elaeis represent Coc n1, Gly m bd 29kD allergen and vicilin-like protein from *Elaeis guineensis* respectively. All others are predicted vicilin-like proteins having high sequence similarity with Coc n 1. Boxes show region of Cupin domains which are highly conserved. Arrow indicates the Coc n 1 protein sequence.

Reference:

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.M., Taly, J.F., Notredame, C., 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 39, W13-W17. https://doi.org/10.1093/nar/gkr245.