

# The Crystal Structure of Recombinant proDer p 1, a Major House Dust Mite Proteolytic Allergen<sup>1</sup>

Kåre Meno,<sup>2\*</sup> Peter B. Thorsted,\* Henrik Ipsen,\* Ole Kristensen,<sup>†</sup> Jørgen N. Larsen,\* Michael D. Spangfort,\* Michael Gajhede,<sup>†</sup> and Kaare Lund\*

Allergy to house dust mite is among the most prevalent allergic diseases worldwide. Most house dust mite allergic patients react to Der p 1 from *Dermatophagoides pteronyssinus*, which is a cysteine protease. To avoid heterogeneity in the sample used for crystallization, a modified recombinant molecule was produced. The sequence of the proDer p 1 allergen was modified to reduce glycosylation and to abolish enzymatic activity. The resulting rproDer p 1 preparation was homogenous and stable and yielded crystals diffracting to a resolution of 1.61 Å. The active site is located in a large cleft on the surface of the molecule. The 80-aa pro-peptide adopts a unique fold that interacts with the active site cleft and a substantial adjacent area on the mature region, excluding access to the cleft and the active site. Studies performed using crossed-line immunoelectrophoresis and IgE inhibition experiments indicated that several epitopes are covered by the pro-peptide and that the epitopes on the recombinant mature molecule are indistinguishable from those on the natural one. The structure confirms previous results suggesting a preference for aliphatic residues in the important P2 position in substrates. Sequence variations in related species are concentrated on the surface, which explains the existence of cross-reacting and species-specific antibodies. This study describes the first crystal structure of one of the clinically most important house dust mite allergens, the cysteine protease Der p 1. *The Journal of Immunology*, 2005, 175: 3835–3845.

Inhalation allergy to house dust mite is among the most prevalent allergic diseases worldwide (1, 2). Allergy to house dust mite causes symptoms typical for type 1 immediate hyper-reactivity, such as rhinoconjunctivitis and asthma. The allergic phenotype is characterized by a persistent inflammation in the airway mucosa and the presence of allergen-specific IgE. Allergen-specific IgE binds to high-affinity receptors on the surface of mast cells and basophiles, thereby enabling these cells to become activated by allergen. After activation, mast cells and basophiles release mediators, such as histamine, leukotrienes, enzymes, etc., that are the direct causes of the allergic symptoms. The binding of allergen molecules to receptor-bound IgE and the cross-linking of these on the surface of mast cells and basophiles is therefore the key event in the triggering of a cascade of events leading directly to allergic symptoms. A thorough understanding of the molecular events leading to activation of mast cells and basophiles is thus facilitated by the elucidation of the three-dimensional structure of the allergen molecule. Furthermore, allergen structures facilitate the rational design of allergen variants with reduced IgE binding, which has possible uses in specific allergy vaccination (3).

Several species of house dust mites have been identified contributing to the environmental exposure of human beings (4). The most prevalent species belong to the genus *Dermatophagoides*

(i.e., *D. pteronyssinus* and *D. farinae*). Because the house dust mites are taxonomically related, the different species contain homologous allergens with structural similarities, which causes IgE cross-reactivity (5). Patients sensitized to one species therefore often also react to the other species (6, 7).

Several proteins derived from house dust mites have been characterized as allergens. The most important allergens in terms of prevalence of reactivity are the group 1, 2, 3, and 9 allergens, to which >90% of mite allergic patients have IgE (8). Group 1 and group 2, however, account for most of the IgE on a quantitative basis. Whereas the biological function of the group 2 mite allergen has not yet been identified, the other mentioned allergens are proteases. Groups 3 and 9 are serine proteases, and group 1 is a cysteine protease located in the alimentary canal of the mite (5).

Der p 1 is a major allergen from *D. pteronyssinus* (9). In one study, the prevalence of IgE to Der p 1 was 97% in a population of 35 house dust mite allergic individuals, as measured by radioallergosorbent tests using purified nDer p 1 (8). Der p 1 is a 25-kDa cysteine protease that is present in mite feces in high concentrations. The proteolytic activity of Der p 1 has been proposed to enhance the capacity of the molecule to sensitize human beings (10). The gene encoding Der p 1 has been cloned and sequenced (11) and shown to display isoallergenic variation (12). The open reading frame encodes an 18-aa signal peptide, an 80-aa pro-peptide, and the mature region that comprises an additional 222 aa. The sequence includes four potential N-linked glycosylation sites, three in the mature sequence and one in the pro-peptide. Der p 1 is produced in the mite as an enzymatically inactive pro-enzyme that becomes enzymatically active after cleavage and detachment of the pro-peptide. Apart from inhibiting the activity of the pro-enzyme, the pro-peptide may also act as a folding scaffold for mature Der p 1, as suggested for other proteases (13, 14). When Der p 1 is extracted from mite feces, it is present in the mature form (nDer p 1). Recombinant Der p 1 has been produced in *Escherichia coli*, but the resulting protein had much reduced IgE binding indicating improper folding (15). In *Pichia pastoris*, rproDer p

\*ALK-Abelló A/S, Research Department, Hørsholm, Denmark; and <sup>†</sup>Department of Medicinal Chemistry, Danish University of Pharmaceutical Sciences, Copenhagen, Denmark

Received for publication March 16, 2005. Accepted for publication June 27, 2005.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>1</sup> The coordinates and structure factors of rproDer p 1-CN have been deposited with the Protein Data Bank coordinate file 1XKG and related structure factor file R1XKGSF.

<sup>2</sup> Address correspondence and reprint requests to Dr. Kåre Meno, ALK-Abelló A/S, Research Department, Bøge Allé 6–8, DK-2970 Hørsholm, Denmark. E-mail address: kme@dk.alk-abello.com

Table I. Primer sequences

| Primer No.        | Primer Sequence 5'-3'                                                 |
|-------------------|-----------------------------------------------------------------------|
| Primer 1, forward | GTCGCTCGAGAAAAGAAGACCATCTTCTATTAAGACTTTCGAAGAATAC                     |
| Primer 2, reverse | GCCTCTAGATCAGTGATGGTGGTGATGGTGACCAGTTTGACCCAAAATGACAACGTATGGATATTCTTC |
| Primer 3, reverse | GTTTCAGCAAGATCCAAATGATTGGTTCACGGTAAGCCAAAATAAGC                       |
| Primer 4, forward | GCTTATTGGCTTACCGTGACCAATCATTGGATCTTGCTGAAC                            |
| Primer 5, reverse | ACCAGAGAAAAGCCCAAGCTGAACACAGCC                                        |
| Primer 6, forward | GGCTGTGGTTCAGCTTGGGCTTCTCTGGT                                         |
| Primer 7, reverse | GTTTCAGCAAGATCCAAATGATTGTTACGGTAAGCCAAAATAAGC                         |
| Primer 8, forward | GCTTATTGGCTTACCGTGAACCAATCATTGGATCTTGCTGAAC                           |

1 can be produced as a hyperglycosylated pro-enzyme with reduced enzymatic activity and IgE binding (16). After in vitro maturation, however, enzymatic activity and IgE binding can be restored independently of glycosylation (17).

The cysteine proteases are divided into five clans, each containing a number of families. Der p 1 belongs to clan CA, family C1, which also includes papain and its relatives. Crystal structures of papain and several closely related proteases of family C1 have been determined, and the catalytic residues have been identified as Cys, His, and Asn. Furthermore, a conserved Gln residue is essential for activity and is believed to help form the oxyanion hole (18), which stabilizes the transition state during catalysis (19). The catalytic Cys and His residues are thought to form a thiolate-imidazolium ion pair stabilized by a direct hydrogen bond between the side chains of the catalytic His and Asn residues (20). Most members of family C1 have pro-peptides homologous to that of papain (115 aa), although the length may differ. The pro-peptide is thought to act by occluding the active site and shielding it from access to substrates, thereby inhibiting the enzymatic activity. Self proteolysis is avoided by binding the pro-peptide in the reverse orientation compared with that of a substrate (18). The shorter Der p 1 pro-peptide has only 17 residues that are identical to propain when the sequences are aligned.

This study provides a detailed description of the crystal structure of the pro form of the major house dust mite allergen Der p 1, its structural relationship with other cysteine proteases, deduced substrate specificities, and the different antibody-binding properties of pro and mature Der p 1. The inferred effect of interspecies sequence variation on IgE binding is furthermore discussed.

## Materials and Methods

### Construction of rproDer p 1 variants

A partially codon-optimized *rproderp1* wild-type (wt)<sup>3</sup> (from now termed *rproderp1*) cDNA containing the complete coding region for rproDer p 1 wt (from now termed rproDer p 1) with the addition of a 10-aa C-terminal His tag was constructed by PCR. The encoded rproDer p 1 protein is equivalent to the proDer p 1 region (aa 19–320) of Swiss-Prot P08176, with the exception of V204A (amino acid numbering throughout this report is from the first amino acid in proDer p 1), which is a naturally occurring variant. Primer 1 and primer 2 (Table I) were used to create a 964-bp fragment with *XhoI* and *XbaI* restriction sites introduced near the 5' and 3' end, respectively. This *rproderp1* gene was cloned into pCR4-TOPO and transformed into TOP10 (Invitrogen), and the sequence was confirmed. Recloning of the *rproderp1* gene into the *E. coli*/P. *pastoris* shuttle vector pGAPZaA was subsequently done by directional cloning with the restriction enzymes *XhoI* and *XbaI*. This resulted in plasmid *rproderp1* with the rproDer p 1 encoded region downstream of the *Saccharomyces cerevisiae*  $\alpha$  factor with a Kex2 cleavage site between the two regions facilitating secretion of the recombinant protein. The plasmid was initially transformed into *E. coli* TOP10 cells.

Construction of site-specific mutations C114A and N132D was done by the overlap extension method (21). Briefly, primer 1 (Table I) was used together with primer 5 in one PCR, primer 3 was used together with primer 6 in another PCR, and primer 2 was used together with primer 4 in a third PCR, all with *rproderp1* as a template. The full fragment containing the mutations was then created by using the three resulting PCR products together with primers 1 and 2 in a final PCR, followed by directional cloning of the 957-bp *XhoI-XbaI* fragment into pGAPZaA as described for *rproderp1*, creating plasmid *rproderp1-C114A/N132D* (from now termed *rproderp1-CN*).

Construction of the single site-specific mutations C114A or N132E were done in a similar manner creating plasmids *rproderp1-C114A* (from now termed *rproderp1-C*) and *rproderp1-N132E* (from now termed *rproderp1-N*), respectively: C114A, primer 1/primer 5 and primer 2/primer 6 (Table I); N132E, primer 1/primer 7 and primer 2/primer 8.

### Expression and purification of rproDer p 1 variants

pGAPZaA-*rproderp1-CN*, -*rproderp1-C*, and -*rproderp1-N* were transformed into *P. pastoris* X33 wt strain (Invitrogen) according to the manufacturer's protocol. Clones were restreaked onto yeast extract/peptone/dextrose/Zeocin (YPDZ)-agar plates and checked for expression. The *P. pastoris* X33::*rproder p 1-CN*, *P. pastoris* X33::*rproder p 1-C*, or *P. pastoris* X33::*rproder p 1-N* clones expressing rproDer p 1-C114A/N132D, rproDer p 1-C114A, or rproDer p 1-N132E (from now termed rproDer p 1-CN, rproDer p 1-C, or rproDer p 1-N, respectively) were grown as described in the protocol supplied by the manufacturer (Invitrogen). Briefly, 5 ml of YPDZ medium (YPD plus 100  $\mu$ g/ml Zeocin) was inoculated with cells from one single colony and grown overnight in a shaker at 220 rpm at 30°C. Two milliliters of culture were then diluted into 1 liter of fresh buffered (pH 6.5) YPDZ medium, and this culture was grown as above for 48–72 h. Culture supernatants were recovered after centrifugation at 4000  $\times$  g for 15 min at 4°C and filtrated through a 0.22- $\mu$ m pore size filter. The protein of interest was precipitated in a two-step procedure.  $(\text{NH}_4)_2\text{SO}_4$  was added to a final concentration of 2 M, followed by incubation overnight at 4°C. The precipitate was discarded after centrifugation at 8000  $\times$  g for 30 min at 4°C.  $(\text{NH}_4)_2\text{SO}_4$  was added to the supernatant to a concentration of 3.2 M, followed by incubation overnight at 4°C. The precipitated proteins were collected by centrifugation at 8000  $\times$  g for 30 min at 4°C, dissolved in 40 ml of wash buffer (50 mM  $\text{NaH}_2\text{PO}_4$ , 300 mM NaCl, and 30 mM imidazole (pH 8)), and dialyzed twice for 8 h at 4°C against 5 liters of wash buffer.

Each sample was applied to a 5-ml nickel-NTA agarose column (Qia-gen) equilibrated in wash buffer. The column was then washed in the same buffer. Bound protein was recovered by elution with a 0–100% gradient of elution buffer (50 mM  $\text{NaH}_2\text{PO}_4$ , 300 mM NaCl, and 250 mM imidazole (pH 8)). Fractions containing rproDer p 1-CN, rproDer p 1-C, or rproDer p 1-N were dialyzed twice for 8 h at 4°C against 5 liters of PBS buffer. The samples were finally concentrated in Centricon spin cartridges.

### Maturation of rproDer p 1-N

An overnight culture of *P. pastoris* X33::*rproderp1-N* was grown as described above. This culture was used as inoculum for a 10-ml culture grown in YP medium containing 5% dextrose and 100  $\mu$ g/ml Zeocin. After 72 h, the culture supernatant was recovered, and small-scale purification was performed as described above.

### SDS-PAGE

SDS-PAGE was performed using Nupage gels (10% Bis/Tris; Invitrogen) according to the manufacturer's recommendations (MES buffer plus antioxidant). Gels were silver stained to visualize proteins or glyco stained (Pro-Q Emerald 300 Glycoprotein Gel stain kit, P21857; Molecular Probes) to visualize glycoproteins, and digital images were collected.

<sup>3</sup> Abbreviations used in this paper: wt, wild type; CLIE, crossed-line immunoelectrophoresis; rproDer p 1-C, rproDer p 1-C114A; r(pro)Der p 1-N, r(pro)Der p 1-N132E; rproDer p 1-CN, rproDer p 1-C114A/N132D; YPDZ, yeast extract/peptone/dextrose/Zeocin.

### N-terminal protein sequencing

Five micrograms of purified protein were subjected to SDS-PAGE under reducing conditions and blotted onto a polyvinylidene difluoride membrane. The membrane was Coomassie stained, and protein bands were cut out. Sequencing was performed by Edman degradation at Biocentrum-DTU (Technical University of Denmark).

### Enzymatic deglycosylation of rproDer p 1 preparations

Five micrograms of purified protein were treated with *N*-glycanase under reducing conditions according to the manufacturer's protocol (Prozyme) at 37°C for 18 h.

### Crystallization

The sitting drop vapor diffusion technique with 1 ml of reservoir solution and equilibration at room temperature was used in all crystallization experiments. Initial crystallization conditions were screened using Crystal Screen Cryo (Hampton Research). Clusters of elongated plates were obtained in drops containing 2  $\mu$ l of 10.7 mg/ml rproDer p 1-CN and 2  $\mu$ l of reservoir solution (20% PEG-4000, 80 mM sodium acetate (pH 4.6), 160 mM ammonium sulfate, and 20% glycerol). These conditions were optimized using Additive Screen 1 (Hampton Research). Diffraction quality single crystals were obtained with drops containing 2  $\mu$ l of 4 mg/ml rproDer p 1-CN, 0.44  $\mu$ l of 0.1 M YCl<sub>3</sub>, and 2  $\mu$ l of reservoir solution (20% PEG-4000, 40 mM sodium acetate (pH 4.6), 160 mM ammonium sulfate, and 20% glycerol).

### X-ray data collection and processing

Crystals were mounted in a cryo loop directly from the crystallization drop and flash frozen in liquid nitrogen. X-ray diffraction data were collected at 120 K on a Rigaku RU300 rotating anode generator operating at 90 mA and 50 kV, equipped with a MAR345 image plate detector and Osmic mirrors. High- and low-resolution data sets were collected on the same crystal to optimize data quality. The high-resolution data set consisted of 140 successive frames (1° oscillation per frame) from which data in the resolution range 1.61–11 Å were used. The low-resolution data set consisted of 40 successive frames (3° oscillation per frame) from which data in the resolution range 3.85–40 Å were used. The data were processed using Denzo and Scalepack (22). Statistics from the data processing are listed in Table II. The crystal contained one molecule in the asymmetric unit and had a solvent content of 45%.

### Molecular replacement and model building

Molecular replacement was performed with data extending to a resolution of 3.5 Å using the CCP4 (23) program Molrep (24) and model coordinates

from structures of cysteine proteases with homologous sequences. The Protein Data Bank (PDB) entries 1PPO (papaya proteinase  $\Omega$ ), 1AEC (actinidin), 1CVZ (papain), 1MEM (cathepsin K), 1CS8 (procathepsin L), and 1YAL (chymopapain) were all tested. 1YAL (25) returned a solution with a convincing contrast to the rest of the solutions and the highest correlation coefficient (0.21). The initial molecular replacement model was improved using the CCP4 implementation of the ARP/wARP program (26). The automatic tracing generated a nearly complete model including the pro-peptide.

### Structure refinement

Remaining interpretable electron density was model built manually using the O program (27). Restrained refinement of the structure was performed using the CCP4 (23) program Refmac5 (28). Five percent of the data was omitted from the refinement and used as a test set (29). Iterative cycles of model building in O were followed by refinement in Refmac5. Water molecules were automatically picked using the CCP4 implementation of ARP/Waters (30) and verified manually. Strong  $F_o - F_c$  electron density was observed in a depression on the surface of rproDer p 1-CN. A  $Y^{3+}$  ion complex-bound to the side chain carboxylic acid groups of D136, E139, and E171, the backbone carbonyl oxygen of L137, and three water molecules was placed at this position, and it returned nice electron density with a final B-factor of 16.7 Å<sup>2</sup>. Two of the complex-bound water molecules hydrogen bond to a symmetry-related rproDer p 1-CN molecule through a hydrogen bonding network involving several other water molecules and a well ordered sulfate ion. This could explain the major improvement in crystal quality observed when YCl<sub>3</sub> was included in the crystallization. An additional sulfate ion and two glycerol molecules were furthermore included in the final model. Statistics from the structure refinement are listed in Table II.

### In situ crossed-line immunoelectrophoresis (CLIE)

CLIE was performed in agarose gels cast on 5 × 7-cm glass plates as described by Krøll (31) using Tris Veronal (73 mM Tris, 24 mM 5,5-diethylbarbituric acid, and 0.35 mM calcium lactate, pH 8.6) as buffer. The rabbit anti-nDer p 1 Abs were produced and purified as described previously (9, 32). First-dimension electrophoresis was as follows: 10  $\mu$ g of Der p extract (ALK-Abelló) was subjected to zone electrophoresis in a 1 × 5-cm agarose gel (Litex agarose type HSA) for 25 min at 10 V/cm. Second-dimension electrophoresis (perpendicular to the first dimension) was as follows: two agarose gels were cast in the anodic direction, one agarose gel (1 × 5 cm) containing either 10  $\mu$ g of rDer p 1-N or 10  $\mu$ g of rproDer p 1-CN and one agarose gel containing the anti-nDer p 1 Abs (0.4  $\mu$ l/cm<sup>2</sup>, 5 × 5 cm). Second-dimension electrophoresis was performed overnight at 2 V/cm. After electrophoresis, the gels were washed, dried, and stained with Coomassie brilliant blue R250 as described previously (33).

### IgE inhibition

The IgE inhibition experiments were performed on an ADVIA Centaur system (Bayer Diagnostics) (34), as described previously (35). Briefly, the binding of ~100 ng of biotinylated nDer p 1 to human IgE was inhibited by serial dilutions of nDer p 1, rproDer p 1, rproDer p 1-C, rproDer p 1-N, or rproDer p 1-CN in the concentration range 1–100,000 ng/ml. The IgE in each sample originated from 25  $\mu$ l of a serum pool that had been absorbed to paramagnetic particles covalently coated with anti-human IgE Abs. Non-IgE Abs were removed by washing the paramagnetic particles before the incubation with the inhibitor/biotinylated nDer p 1 mixtures. The amount of biotinylated nDer p 1 bound to the captured IgE was estimated from the relative light units detected after incubation with acridiniumester-conjugated streptavidin. The biotinylated nDer p 1 was kindly provided by N. Johansen (ALK-Abelló In Vitro Diagnostics Business Unit). The serum pool was prepared by mixing equal volumes of serum from 10 house dust mite allergic individuals.

### Miscellaneous computational methods

All structure images were made with PyMol (36). Solvent exposure of individual amino acids was calculated using the GETAREA program (37) available at [www.scsb.utmb.edu/cgi-bin/get\\_a\\_form.tcl](http://www.scsb.utmb.edu/cgi-bin/get_a_form.tcl). The program designates residues as being buried if the ratio of side chain surface area to a random coil reference value per residue is <20% and solvent exposed if the ratio exceeds 50%. The random coil value of a residue Xxx is the average solvent accessible surface area of Xxx in the tripeptide Gly-Xxx-Gly in an ensemble of 30 random conformations. The solvent-accessible surface area on mature proteins covered by the pro-peptide was calculated using the CCP4 (23) program ArealMol (38).

Table II. Crystallographic statistics

|                                                            |                                               |
|------------------------------------------------------------|-----------------------------------------------|
| Data collection                                            |                                               |
| Space group                                                | P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> |
| Unit cell dimensions (Å)                                   | a = 64.7, b = 66.6, c = 73.4                  |
|                                                            | $\alpha = \beta = \gamma = 90^\circ$          |
| Resolution range (Å)                                       | 40–1.61 (1.67–1.61) <sup>a</sup>              |
| Reflections                                                | 463,368                                       |
| Unique reflections                                         | 41,503                                        |
| Completeness (%)                                           | 99.7 (96.9)                                   |
| R <sub>sym</sub> (%) <sup>b</sup>                          | 5.5 (40.3)                                    |
| Redundancy                                                 | 5.6 (3.5)                                     |
| Average I/ $\sigma$ (I)                                    | 27.9 (2.9)                                    |
| Refinement                                                 |                                               |
| Unique reflections <sup>c</sup>                            | 41,444/2,083                                  |
| R <sub>work</sub> /R <sub>free</sub> (%) <sup>d</sup>      | 15.3/18.6                                     |
| Number of atoms                                            |                                               |
| Protein/hetero <sup>e</sup> /water                         | 2,396/23/312                                  |
| RMSD from ideal values                                     |                                               |
| Bond lengths (Å)                                           | 0.014                                         |
| Bond angles (°)                                            | 1.5                                           |
| Average B-factors (Å <sup>2</sup> )                        |                                               |
| Protein/hetero <sup>e</sup> /water                         | 22.7/33.3/34.8                                |
| Ramachandran plot (non Gly or Pro), residues in region (%) |                                               |
| Most favored/additional                                    | 89.4/9.4                                      |
| Generous/disallowed                                        | 1.1/0.0                                       |

<sup>a</sup> Values in parentheses refer to the highest resolution shell.

<sup>b</sup>  $R_{sym} = \sum_i \sum_j |I_i(h) - I_j(h)| / \sum_i \sum_j I_i(h)$ .

<sup>c</sup> All reflections with  $F > 0$  were included (working set/test set).

<sup>d</sup>  $R_{work} = \sum_h ||F_o(h)| - k|F_c(h)|| / \sum_h |F_o(h)|$ .  $R_{free}$  is calculated as for  $R_{work}$  but on a 5% test set of the reflections.

<sup>e</sup> Y<sup>3+</sup>, sulfate and glycerol.

## Results

### Characterization of the recombinant pro and mature Der p 1 variants

When expressed in *P. pastoris*, rproDer p 1 displays a heterogenic smear in SDS-PAGE with a larger estimated molecular mass than the calculated 35.4 kDa, indicating a heterogeneous glycosylation (Fig. 1A, lane 2). Accordingly, removal of *N*-linked glycosylations by *N*-glycanase treatment reduced the heterogeneity and yielded a protein band with the expected mobility (38 kDa; Fig. 1A, lane 3). Because both glycosylation and spontaneous maturation of rproDer p 1 could impede crystallization, sequence modifications were introduced. A N132E mutation rendered the protein insensitive to glycosylation at this site, yielding rproDer p 1-N. The active site Cys, inferred by sequence comparisons with other cysteine proteases to be C114 in the proDer p 1 sequence, was additionally mutated to an Ala residue in rproDer p 1-CN, preventing the formation of an enzymatically active protease. Purified rproDer p 1-N migrated as two distinct bands with an estimated molecular mass of 38 and 40 kDa, respectively, and a third much less intense band with an estimated molecular mass of 36 kDa (Fig. 1A, lane 4). The absence of a large molecular mass smear confirms the presence of a large glycosylation on N132 in rproDer p 1. Purified rproDer p 1-CN migrated as two distinct bands with an estimated molecular mass of 38 and 40 kDa, and nDer p 1 migrated as a uniform 26-kDa band as shown in Fig. 1A, lanes 6 and 8, respectively. When rproDer p 1-N is processed during fermentation to rDer p 1-N, only one band with an estimated molecular mass of 28 kDa is observed (Fig. 1B, lane 2), which is close to the calculated value of 26.1 kDa for the mature His-tagged protein. Treatment with *N*-glycanase caused the 40-kDa band to disappear in the rproDer p 1-N and rproDer p 1-CN preparations, whereas no detectable change in mobility was observed for nDer p 1 (Fig. 1A, lanes 5, 7, and 9, respectively).

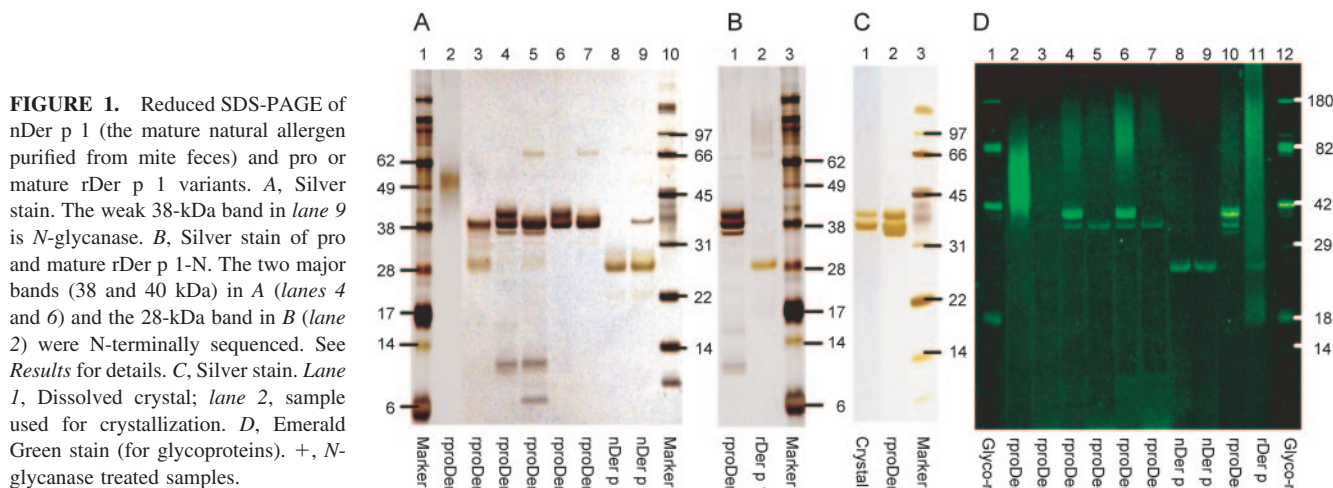
To elucidate the glycosylation pattern in the preparations of pro and mature Der p 1, their ability to bind a glyco-binding dye (Emerald Green) before and after treatment with *N*-glycanase was investigated after separation by SDS-PAGE. As expected, rproDer p 1 stained heavily with the dye. However, the 38-kDa band resulting from *N*-glycanase treatment still bound the dye (Fig. 1D, lanes 2 and 3). All three rproDer p 1-N bands were recognized by the dye, and *N*-glycanase treatment did not alter the 38- and 36-kDa bands (Fig. 1D, lanes 4 and 5). Similarly, the 40- and 38-kDa rproDer p 1-CN bands both bound the dye, and *N*-glycanase treatment eliminated the 40-kDa band (Fig. 1D, lanes 6 and 7). *N*-glycanase had no effect on nDer p 1, which bound the dye both before and after

treatment (Fig. 1D, lanes 8 and 9). The mature recombinant protein rDer p 1-N also bound the dye (Fig. 1D, lane 11).

N-terminal sequencing of the two major protein species (40/38-kDa) in the rproDer p 1-N and rproDer p 1-CN preparations showed that the sequences were identical to the expected N-terminal sequence for proDer p 1 (RPSSI). Similarly, the 28-kDa rDer p 1-N band had the authentic mature Der p 1 N terminus (TNACSING). The rproDer p 1-CN preparation characterized above was used for crystallization trials.

### The overall structure of rproDer p 1-CN

Like in other papain-like members of family C1 with known structure, the mature region of rproDer p 1-CN is folded to form a globular protein with two interacting domains that delimit a cleft on the surface where substrates bind (Fig. 2). The left domain in Fig. 2A consists of residues 101–196, which adopt a predominantly  $\alpha$ -helical conformation. The N-terminal residues (82–100) of the mature region cross over and embrace the  $\beta$ -sheet-dominated right domain consisting of residues 197–299. The three C-terminal residues (300–302) are placed roughly between the two domains pointing back into the left domain. Residues 77–81, which constitute the border between the pro-peptide (aa 1–80) and the mature region, were poorly defined by the electron density, and these residues are consequently omitted in the final structure. The N-terminal part of the pro region forms a distinct domain containing two  $\alpha$ -helices ( $\alpha$ 1,  $\alpha$ 2) at one end of the substrate-binding cleft, as shown in Fig. 2B. The remainder of the pro region contains two additional  $\alpha$ -helices ( $\alpha$ 3,  $\alpha$ 4) that span the entire cleft with helix  $\alpha$ 3 wedged into the S' subsites and the C-terminal helix ( $\alpha$ 4) covering the S subsites. The placement of the pro region is similar to that found for two other papain-like members of family C1 that are available in the PDB in their pro forms, procathepsin L (1CS8) and procaricain (1PCI). In these two cases, however, the pro-peptides adopt a fold containing short N-terminal helix  $\alpha$ 1, followed by long helix  $\alpha$ 2, strand  $\beta$ 1, short helix  $\alpha$ 3, and an extended C-terminal part with a short strand  $\beta$ 2 (Fig. 3A). Two of the three disulfide bridges found in papain are conserved in Der p 1 (C111–C151, C145–C183; Fig. 3B). A third disulfide bridge unique to Der p 1 is observed between residues C84 and C197, as predicted previously (39). C84 is placed in the N terminus of mature Der p 1, in a region that is part of the pro-peptide in other papain-like cysteine proteases. Consequently, if a homolog of this disulfide bridge



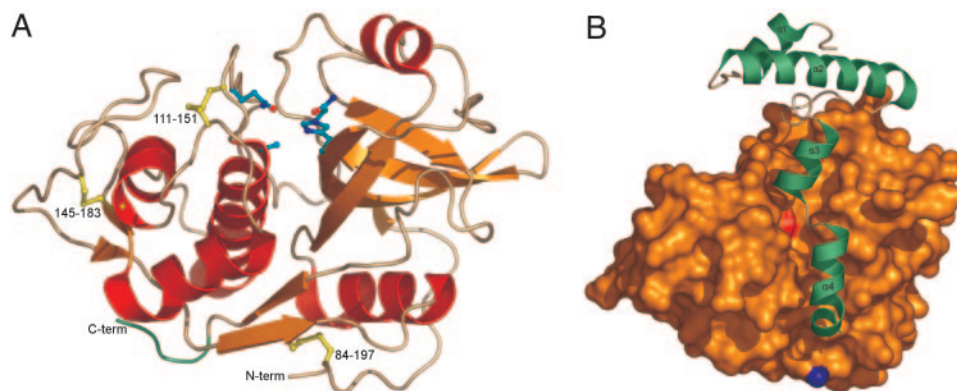


**FIGURE 2.** Overall structure of rproDer p 1-CN. *A*, Schematic representation of the mature region, including the visible part of the His-tag (colored green).  $\alpha$ -Helices are shown as red coils,  $\beta$ -strands are shown as orange arrows, and disulfide bridges are shown as yellow sticks with the residue numbers indicated. The side chains of the active site residues Q108, A114, H250, and N270 are shown as cyan sticks. *B*, The folding of the pro region on the surface of the mature region. The residues from the pro-peptide that are visible in the electron density are shown schematically in green, and the part of the structure that corresponds to the mature protein (i.e., without pro-peptide and His-tag) is represented by the molecular surface. The N-terminal nitrogen atom of the first visible residue in the mature region (N82) is colored blue. The  $C_{\beta}$  atom of A114 corresponding to the active site Cys residue has been colored red.

was present in the latter enzymes, the pro and mature regions would be covalently bound after proteolytic activation. The structurally equivalent residue of C84 is F99p (the residues of the pro regions of procaricain and procathepsin L are indicated by the suffix p) in procaricain, and this side chain binds to a conserved hydrophobic pocket on the surface of the mature region, thus forming a similar albeit much weaker interaction. Overall, 197 of the 221 observed residues in the mature region of rproDer p 1-CN can be structurally aligned with the structure of mature papain (PDB ID 9PAP) resulting in a root mean square deviation of 1.38 Å and a sequence identity of 28.9% (Fig. 3*B*).

Both the rproDer p 1-CN preparation and a dissolved crystal show two bands at ~38 and 40 kDa in SDS-PAGE (Fig. 1*C*, lanes 2 and 1, respectively) with the 38-kDa band being the dominating one. The electron density map indicated the presence of a glycosylation on N195, although the density was too weak to allow insertion of any sugar residues.

#### *The folding of the pro-peptide*

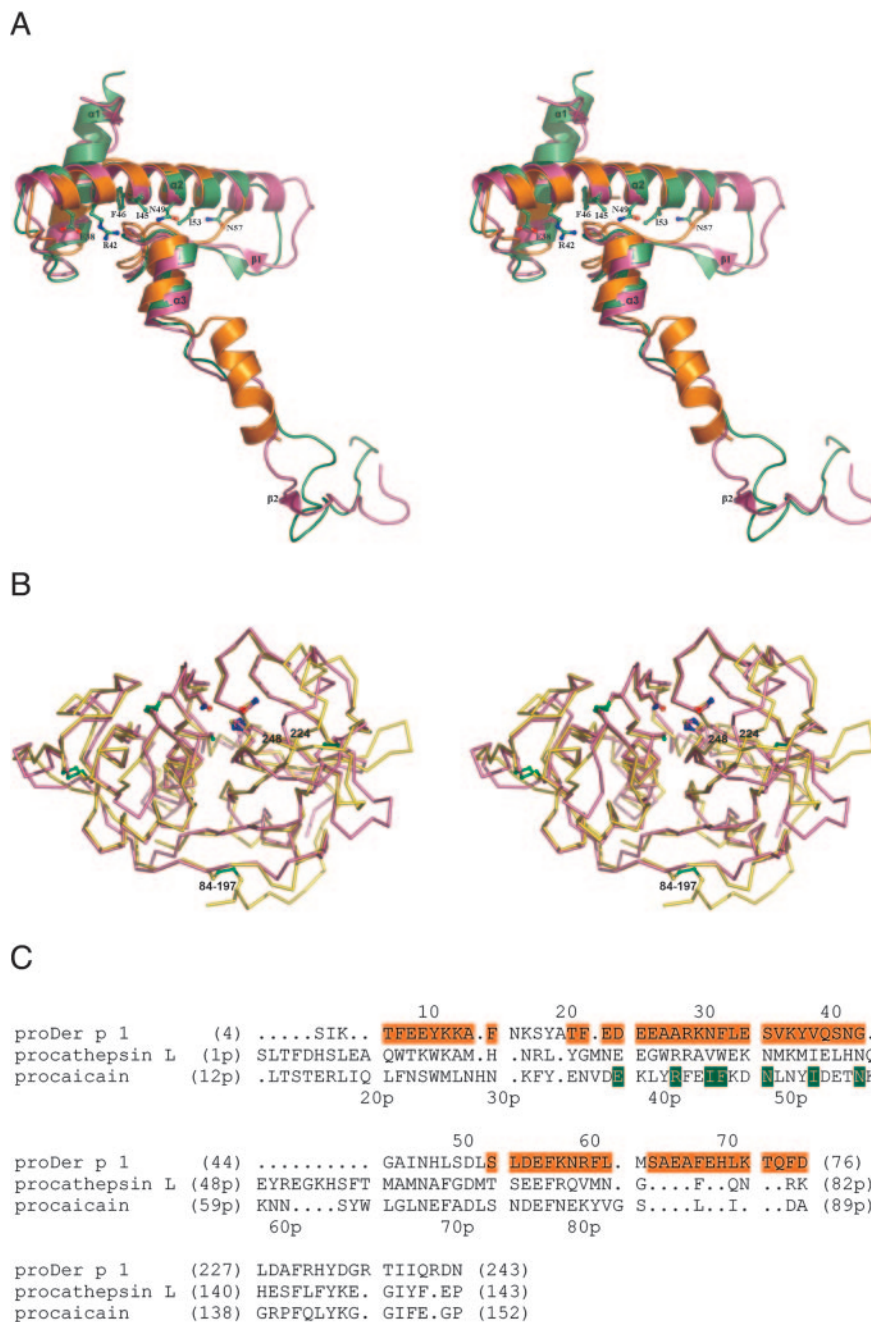
Karrer et al. (40) have defined two distinct subfamilies within the cysteine protease family on the basis of conserved amino acids in the pro-peptides and the length of these. Although not completely conserved, the ERFNIN subfamily contains the  $E^{38p}X_3RX_2(IV)FX_2NX_3IX_3N^{57p}$  motif separated by a stretch of variable length from another conserved block of amino acids containing residues F70p, D72p, and E77p (1PCI numbering<sup>4</sup>). The proteins in this family typically have pro-peptides of ~100 aa. The other subfamily contains the cathepsin B-like enzymes, which have much shorter pro-peptides (62 aa in human cathepsin B) that do not contain the ERFNIN motif. The pro-peptides of these two subfamilies do not show any notable sequence homology, and proDer p 1 has not been assigned to either of them. No structure of propapain has been published, but a search for structures similar to rproDer p 1-CN performed using Secondary Structure Matching (41) revealed that the only published homologous structures of papain-like cysteine proteases in their pro form are human procathepsin L and K and procaricain from papaya (PDB ID 1CS8, 1BY8, and 1PCI, respectively). Furthermore, the structures of pro-

cathepsin B from both human and rat and human procathepsin X are known (PDB ID 3PBH, 1MIR, and 1DEU, respectively), but these were not identified by Secondary Structure Matching using standard settings, presumably because of their shorter pro-peptides of 62, 62, and 38 aa, respectively. The first three enzymes belong to the ERFNIN subfamily according to the definition above. For comparisons with rproDer p 1-CN, we will use procaricain (116-aa pro-peptide) (42) and procathepsin L (96-aa pro-peptide) (43), which are 72/69% and 19/39% (pro/mature region) identical to papain (115-aa pro-peptide) in sequence alignments, respectively. The corresponding numbers for proDer p 1 are 17/25%, which with its 80-aa pro-peptide has a considerably smaller pro domain than the other three proteins. The Der p 1 pro-peptide thus appears to be intermediate in length compared with the ERFNIN and the cathepsin B subfamilies.

In the ERFNIN subfamily, the residues in the ERFNIN motif are located on the core face of helix  $\alpha 2$  where they are involved in a series of interactions (Fig. 3*A*). The primary role of this motif is thought to be stabilization of the structural arrangement of helices  $\alpha 1$ – $\alpha 3$  and strand  $\beta 1$  of the pro-peptide into a discrete globular domain. In procaricain/procathepsin L, the packing of helices  $\alpha 1$  and  $\alpha 2$  is determined by the interactions of the three aromatic residues F/W22p, W25p, and F/W46p (1PCI numbering), among which the last residue is part of the ERFNIN motif (44). The residues that occupy the corresponding positions in the rproDer p 1-CN structure are F8, Y11, and F32 (Fig. 3*C*), implying that these interactions are conserved. The packing of the pro region is further stabilized by intra-pro-domain salt bridges and hydrogen bonds between the highly conserved residues K31p-D72p-Y33p and E38p-R42p-E77p. The former hydrophilic interactions are conserved in rproDer p 1-CN by means of K17, D51, and Y19. However, neither E38p nor R42p, which are part of the ERFNIN motif, are structurally conserved in rproDer p 1-CN, and this part of the network is thus not present in the mite enzyme. The relative orientation of helix  $\alpha 2$  and strand  $\beta 1$  is similar in procaricain and procathepsin L, and the two residues that stabilize this fold are I53p and N57p through side chain hydrophobic contacts and hydrogen bonds, respectively (44). Both residues are part of the ERFNIN motif and are again not conserved in rproDer p 1-CN because they are located C terminally in helix  $\alpha 2$ , which is truncated in rproDer p 1-CN. N49p, also a member of the ERFNIN motif, forms hydrogen bonds with the backbone of residues in the

<sup>4</sup> There is a discrepancy between the numbering of the pro-peptide of procaricain in the 1PCI PDB entry, 106 aa, and the P10056 Swiss-Prot entry, 116 aa, because the 10 N-terminal residues are excluded in the former. The pro-peptide numbering used in this report refers to the PDB entry.

**FIGURE 3.** Comparison of rproDer p 1-CN with other structures. *A*, Schematic stereo image showing a superposition of the pro-peptides of rproDer p 1-CN (orange), procaricain (green), and procathepsin L (purple). The secondary structure elements in procathepsin L are labeled. The residues from the ERFNIN motif in procaricain are shown as sticks. *B*, Stereo image showing a superposition of C $\alpha$  traces of the mature regions of rproDer p 1-CN (yellow) and papain (violet, PDB ID 9PAP). The side chains of the active site residues Q108, A114, H250, and N270 (proDer p 1 numbering) together with cysteine residues involved in disulfide bridges are shown as sticks. The disulfide bridge that is unique to Der p 1 is labeled. Nitrogen, oxygen, and sulfur atoms are colored blue, red, and green, respectively. The first and last residues in the helix/loop region that have different conformations in the two structures are labeled in rproDer p 1-CN. The coordinates were superimposed using the protein structure comparison service Secondary Structure Matching (SSM) at the European Bioinformatics Institute ([www.ebi.ac.uk/msd-srv/ssm](http://www.ebi.ac.uk/msd-srv/ssm)) (41). *C*, Structure-based sequence alignment of the pro-peptides and the pro region binding loops of proDer p 1-CN, procathepsin L, and procaricain generated using SSM. Sequence numbers are shown for proDer p 1-CN and procaricain above and below the sequences, respectively. Helices  $\alpha$ 1– $\alpha$ 4 are highlighted in orange in the proDer p 1 sequence. The ERFNIN motif residues are highlighted in green in the procaricain sequence.



loop preceding helix  $\alpha$ 3 via its side chain in procaricain and procathepsin L. N49p is not conserved in rproDer p 1-CN. S35, however, performs a similar function. The final member of the ERFNIN motif, which can be either an Ile or a Val residue, is I45p in procaricain, and it is part of the hydrophobic core. This residue is not conserved in rproDer p 1-CN where the hydrophilic side chain of N31 occupies the corresponding position.

#### Interactions between the pro-peptide and the mature region

In procaricain and procathepsin L, two main areas of contact between the pro-peptide and the mature region have been defined (44). The first of these is the loop containing residues 138–152 (1PCI numbering) in the mature region, known as the pro region binding loop. Positions 139, 140, 145, and 146 are often occupied by charged residues in this enzyme family. This is also the case in proDer p 1-CN, where the structurally equivalent residues are D228, A229, D234, and G235 (Fig. 3C). The shorter  $\alpha$ 2 helix in

rproDer p 1-CN prevents the interaction between the pro-peptide and the C-terminal part of the loop, so only the first three of these residues interact with the pro-peptide. The second area of contact in procaricain and procathepsin L is in the substrate binding site along the S2- to S2' subsites. Of particular importance is the hydrophobic surface within the S' subsites that contain W181 and W185, which the pro-peptide packs onto. These two tryptophan residues form part of a highly conserved aromatic cluster, which also contains F141, Y144, and F149 (Y in cathepsin L), located in the pro region binding loop. Except for F149, these five residues are structurally conserved in rproDer p 1-CN with the following numbers: W272, W276, F230, and Y233. F149 is located beneath  $\beta$ 1, which is present in the pro-peptides in procaricain and procathepsin L but absent in rproDer p 1-CN (Fig. 3A).

Helix  $\alpha$ 3 that spans from S53 to L62 in rproDer p 1-CN is conserved in procaricain (S74p–Y82p) and procathepsin L

(T67p-N76p). Two residues within the helix are completely conserved in the three proteins, E56/E77p/E70p and F57/F78p/F71p (proDer p 1/1PCI/procathepsin L numbering; Fig. 3C). Moreover, a large hydrophobic amino acid that presumably enters the S2' subsite is conserved in position F61/Y82p/M75p. The first of the conserved helix  $\alpha 3$  residues (E77p in 1PCI) was mentioned above as one of the conserved residues in the ERFNIN subfamily. The second (F78p) and third (Y82p) are located in the interface between the pro-peptide and the mature region where they interact with the conserved aromatic cluster containing W272 and W276 (proDer p 1 numbering) described above. The length and position of helix  $\alpha 3$  is thus well conserved. However, the backbone conformation differs dramatically immediately on the C-terminal side of this helix. The presence of helix  $\alpha 4$  in rproDer p 1-CN (S64-D76) places the residue bridging helices  $\alpha 3$  and  $\alpha 4$ , M63, much further away from the catalytic residues compared with the corresponding residue in procaricain (S85p) and procathepsin L (G77p) (Fig. 3A). The side chain of M63 points toward but not into S2, in contrast to the two other enzymes in which S2 is occupied by a bulky hydrophobic side chain (L86p and F78p, respectively). This leaves room for a glycerol molecule in the active site of rproDer p 1-CN underneath the pro-peptide, which is hydrogen bonded to the backbone nitrogen and carbonyl atoms of W115 and D154, respectively. Subsite S2 is therefore not occupied by any well ordered atoms in rproDer p 1-CN, but the glycerol molecule and the

F75 side chain from helix  $\alpha 4$  line it closely. In procaricain and procathepsin L, the pro-peptides are anchored near the active site by the two backbone hydrogen bonds G66-S85p/187p and G68-G77p/Q79p, respectively, and this interaction is consequently not conserved in rproDer p 1-CN. However, several interactions between helix  $\alpha 4$  and the mature region anchor the pro-peptide in the mite protein. These include the hydrogen bonds Q74-R158 and F75-Y298 and the hydrophobic interactions F68-Y249 and F75-T155/I221/Y296. Furthermore, like in procathepsin L, the carbonyl oxygen of F61 in rproDer p 1-CN points into the oxyanion hole defined by the backbone nitrogen of A114 and the side chain amide of Q108 to which it forms a hydrogen bond.

Groves et al. (44) suggested that the charge complementarity between residue D48p/K37p in helix  $\alpha 2$  and R139/E141 in the pro region binding loop observed in procaricain/procathepsin L could be a determinant in the binding of a pro-peptide to its cognate mature region. However, this charge complementarity is not conserved in rproDer p 1-CN where the corresponding residues are E34 and D228 (Fig. 3C).

#### Active site and substrate-binding cleft

The active site on rproDer p 1-CN and other papain-like members of family C1 is located in a prominent cleft on the surface of the mature protein. The residues lining the cleft determine the substrate specificity of the protease. The substrate-binding cleft on

**FIGURE 4.** Substrate-binding cleft of rproDer p 1-CN compared with papain. *A*, Stereo image showing the  $C_{\alpha}$  trace in and around the substrate-binding cleft of the mature region of rproDer p 1-CN with the side chains of the residues lining the cleft in a stick representation. The disulfide bridge between residues 111 and 151 is included for orientation. *B*, Same as *A* but with an inhibitor-bound papain structure (violet) superimposed. The inhibitor molecule covalently bound to the active site C25 (acetyl-Ala-Ala-Phe-methylenyl-Ala) in papain is shown in yellow in a stick representation. The side chain oxygen and nitrogen atoms of Asn and Gln residues are designated as the same atom type in 1PAD, and these atoms are therefore shown with the same color, wheat. Nitrogen, oxygen, and sulfur atoms are colored blue, red, and green, respectively. Gly residues are shown with a ball at the  $C_{\alpha}$  position. The structures were superimposed as described in Fig. 3.

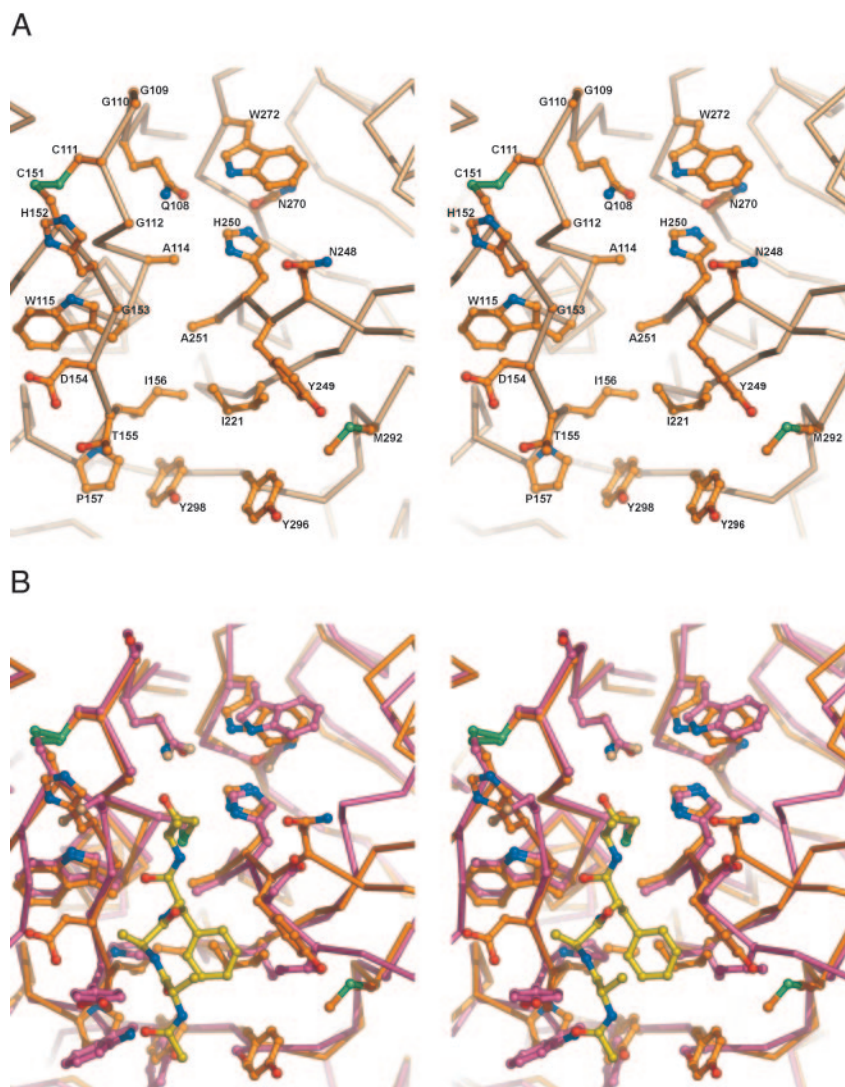
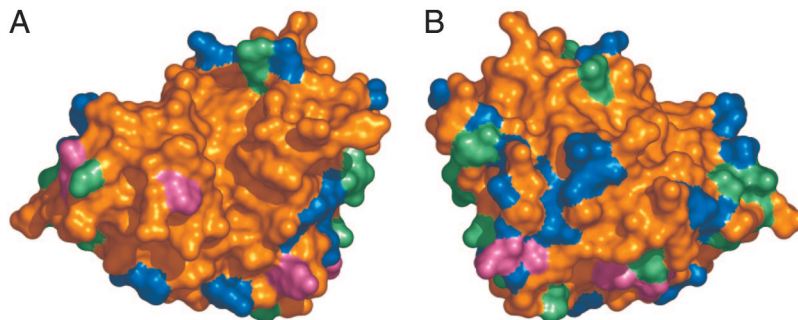


**FIGURE 5.** Sequence differences between Der p 1 and Der f 1 (Swiss-Prot P16311) and/or Eur m 1 (P25780) mapped on the molecular surface of the mature region of the rproDer p 1-CN structure. The orientation of Der p 1 in *A* is identical to that seen in Fig. 2*B*, whereas the molecule has been rotated 180° around a vertical axis in *B* to show the other side. Unchanged residues are colored orange. Residues that differ between the Der p 1 and Der f 1 sequences are green, those that differ between Der p 1 and Eur m 1 are violet, and those that differ between all three proteins are blue.



papain is believed to have seven subsites (S1–S4 and S1'–S3'), each accommodating one side chain. There are many differences in the nature of the side chains of the residues lining the substrate-binding cleft of papain compared with rproDer p 1-CN (Fig. 4). These include S21→G110, N64→H152, G66→D154, Y67→T155, P68→I156, W69→P157, V133→I221, S205→Y296, and F207→Y298 (mature papain and proDer p 1 numbering, respectively). Two additional substitutions, V157→N248 and D158→Y249, are furthermore spatially in a different position because of the different conformation of the helix and loop region containing residues 224–248 in rproDer p 1-CN compared with papain (residues 135–158; Fig. 3*B*).

Papain has been crystallized with the inhibitor acetyl-Ala-Ala-Phe-methylenyl-Ala bound covalently to the active site Cys residue (PDB ID 1PAD) (45). The inhibitor binds in the same direction as substrates. The Phe side chain of the inhibitor molecule marks the location of the important S2 subsite (Fig. 4*B*). The structural differences in the substrate-binding cleft of rproDer p 1-CN compared with papain have a large impact on this subsite. The major differences are the substitution of V133 in papain with I221 in Der p 1 and the altogether different structure of residues 157–158 and 248–249 in papain and rproDer p 1-CN, respectively. The P1 Ala residue, in contrast, points out of the cleft, resulting in the absence of a S1 subsite. The S1' subsite as defined by analogy to procaricain and procathepsin L (see above) is occupied by the L62 side chain in rproDer p 1-CN. This subsite seems suitable for binding a large hydrophobic side chain because it is lined by residues I224, L227, F230, N248, H250, and W272. The side chain amide group of N248 is rotated away from L62, exposing C<sub>β</sub> to the S1' subsite.

#### *The effect of species variation on the possible Ab binding sites*

The crystal structure of rproDer p 1-CN enables mapping of the sequence differences between this allergen and the homologous Der f 1/Eur m 1 allergens from the taxonomically related species *D. farinae*/Euroglyphus maynei on the structure of the mature region. If the C114 and N132 mutations and the one residue insertion between residues 87 and 88 in Der p 1 present in both Der f 1 and Eur m 1 is disregarded, then these proteins contain 40 and 34 substitutions relative to Der p 1-CN, respectively. The substitutions are not evenly distributed throughout the structure but concentrated on the surface of the protein with the important exception of the substrate-binding cleft (Fig. 5). The only buried substitutions to nonsimilar amino acids are A285Q in Der f 1 and A90L in Eur m 1. Residues I238, Q240, V263, I288, and L290 all have at least one side chain atom within 5 Å of the A285 C<sub>β</sub> atom. Of these, Q240 and I288 are solvent exposed on the surface of the Der p 1-CN structure. V263 and I288 are substituted by Asp and Asn, respectively, in Der f 1, making the area considerably more hydrophilic. Furthermore, N287 located between A285 and I288 is substituted with a Gly adding flexibility to the region. The other buried substitution, A90, is located in the stretch of amino acids

that adopt an extended conformation in the rproDer p 1-CN structure between the mature N terminus and the first secondary structure element (starting at residue 94). This stretch contains a total of five substitutions in both Der f 1 and Eur m 1 and is generally solvent exposed, so a relatively high degree of flexibility must be expected. In conclusion, it is likely that the two buried substitutions can be accommodated without structural changes outside the immediate vicinity. The many surface substitutions are likely to have a pronounced effect on Abs binding to areas containing the substitutions. Other surface areas, on the other hand, are conserved between species or the substitutions may be sufficiently conservative to allow binding of cross-reactive Abs. This explains the observation that some Abs cross-react between the group 1 allergens from different species, whereas others do not (6, 7).

#### *In situ CLIE shows differential Ab binding between pro and mature rDer p 1*

The crystal structure of rproDer p 1-CN shows that the pro-peptide covers 1356 Å<sup>2</sup> of the solvent-accessible surface area of mature Der p 1-CN, which corresponds to 12.9% of the total solvent-accessible surface area of 10481 Å<sup>2</sup>. In comparison, the larger pro-peptides in procaricain and procathepsin L cover 20.7 and 18.2%, respectively.

CLIE was used to visualize differences in polyclonal anti-nDer p 1 Ab binding between pro and mature forms of rDer p 1 to determine the extent to which the pro-peptide covers Ab-binding epitopes. When rDer p 1-N was present in the intermediate gel, only a line fused with an arc was observed, indicating that all the anti-nDer p 1 Abs recognized this mature protein (Fig. 6*A*). However, if rproDer p 1-CN was present in the intermediate gel (Fig. 6*B*), an arc of nDer p 1 was observed below the precipitate line representing rproDer p 1-CN. The arc arose from Abs that could not be precipitated in the rproDer p 1-CN line and hence did not bind to this protein. These Abs were precipitated by nDer p 1

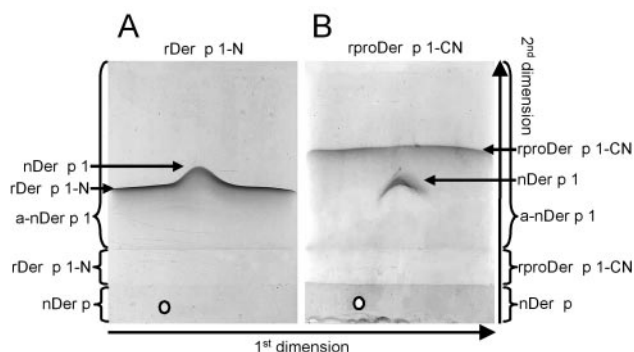
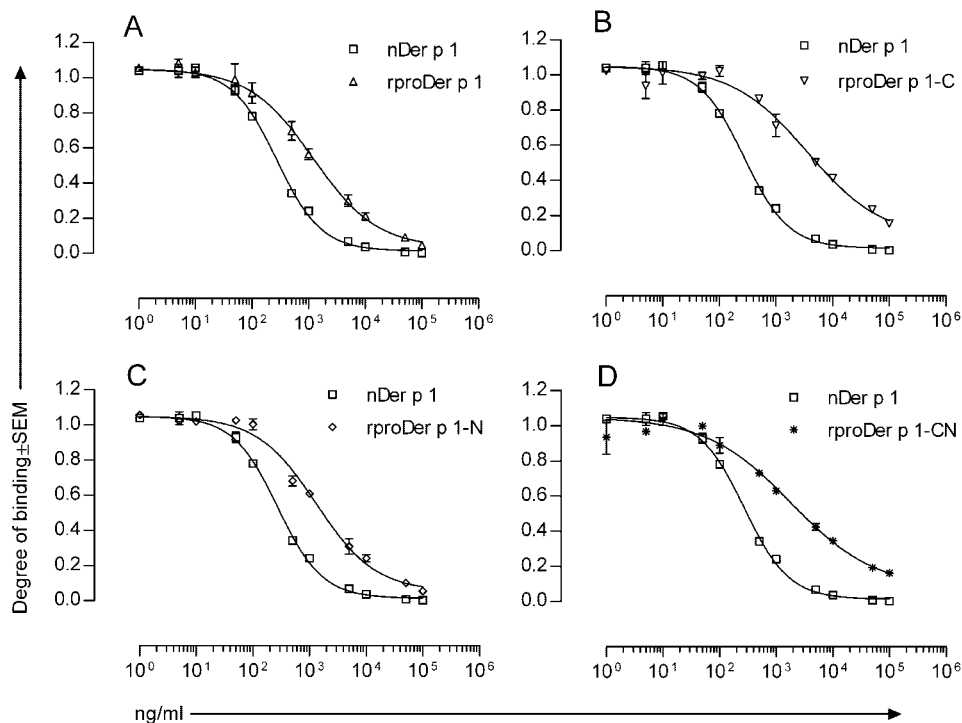


**FIGURE 6.** CLIE of rDer p 1-N (*A*) and rproDer p 1-CN (*B*). The contents in the gels are described on the figure.

**FIGURE 7.** IgE inhibition curves of rproDer p 1 variants compared with nDer p 1. The inhibitor competed with biotinylated nDer p 1 for binding to IgE. Both the fitted curve and the data points are shown for each protein. Each point represents the mean of two or three independent experiments, each constituting a dependent triplicate measurement. The error bars are not visible for many of the points because the bars are so short that they are covered by the symbol. The variant inhibition curves were nonparallel in pairwise comparisons with nDer p 1, with  $p$  values of 0.0031 (A), 0.0014 (B), 0.0029 (C), and 0.0001 (D), respectively.



(without pro-peptide) from the extract, indicating that they recognized epitopes obstructed at least partially by the pro-peptide.

#### IgE inhibition

The various preparations of recombinant pro-forms of Der p 1 (rproDer p 1, rproDer p 1-C, rproDer p 1-N, and rproDer p 1-CN) all inhibited the interaction between human IgE and biotinylated nDer p 1 (Fig. 7). The fit showed that the bottom asymptotic value can be regarded as zero for all four inhibition curves, indicating that each preparation contained all the IgE-binding epitopes on nDer p 1 recognized by the serum pool used. However, the recombinant pro-forms each exhibited statistically significant nonparallel inhibition curves, indicating that the epitope composition of the pro-forms is altered compared with that of nDer p 1.

#### Discussion

The crystal structure of proDer p 1 was solved yielding a high-resolution structure of this important allergen. No interpretable electron density was observed for residues 1–3, 77–81, and 307–312. It is not surprising that the three N-terminal residues and the C-terminal His-tag were disordered because of flexibility. The disorder of residues 77–81 can also be explained by flexibility, because proDer p 1 must be cleaved between residues 80 and 81 to mature into the active enzyme. The same region is disordered in the crystal structure of procaricain, in which residues 90p–100p (PDB ID 1PCI numbering) yield weak electron density (42). The glycerol molecule placed in the active site cleft between the pro-peptide and the mature region of rproDer p 1-CN only forms two hydrogen bonds, and the binding site does not, in general, seem optimal to accommodate this molecule. It is thus likely that the very high glycerol concentration in the crystallization solution (20%) caused binding to a low-affinity site. Furthermore, the mutation of the catalytic C114 to the smaller Ala side chain influenced the electrostatic characteristics of the active site. Similarly, the conformation of the residues surrounding the active site may have been influenced by the mutation of the catalytic H159 to Gly in procaricain (PDB ID 1PCI) and the oxidation of the catalytic C25 to cysteinesul-

fonic acid in procathepsin L (1CS8). It is thus possible that the architecture of the pro-peptide-bound active site is slightly different in the naturally occurring molecule in all three cases.

The longer helix  $\alpha 2$  in procaricain and procathepsin L allows the formation of strand  $\beta 1$ , generating a helix-turn-strand hairpin motif, which contributes to the larger surface area shielded by these pro-peptides compared with rproDer p 1-CN. The C-terminal part of the pro region forms a fourth  $\alpha$ -helix in rproDer p 1-CN, whereas this region adopts an extended conformation in procathepsin L and procaricain. This is of importance because it influences the position of the residues that cover the active site Cys residue. Active site accessibility possibly affects intramolecular processing, if any, which has been suggested to be of importance for activation at low pH in papain (13). The only residue from the ERFNIN motif that is conserved in rproDer p 1-CN is F46p, whereas the functionality of N49p is preserved by a Ser residue. The ERFNIN motif is thus not conserved in rproDer p 1-CN, excluding it from this subfamily. However, rproDer p 1-CN cannot be assigned to the cathepsin B-like enzymes, because cathepsin B has a different pro-peptide conformation and completely lacks helix  $\alpha 1$ . Judging from the pro-peptides, Der p 1 can therefore not be assigned to either group, and we propose a new third subfamily of the C1 family characterized by an 80-aa pro-peptide with four  $\alpha$ -helices and no  $\beta$ -strands.

The similarity of the fold of the mature region of rproDer p 1-CN when compared with papain crystallized without the pro-peptide is evident from Fig. 3B. Furthermore, the crystal structures of caricain and cathepsin K have been determined for both the pro and mature forms of these proteins. The structure of the mature region within the pro form is virtually identical to that of the mature form in both cases (42, 46). The mature region of the proDer p 1 structure described here is thus, in all probability, in the same (active) conformation as the natural mature molecule. This is supported by the CLIE results. When the pro-peptide was present, as in rproDer p 1-CN, some of the Abs did not bind; however, removing the pro-peptide, as in rDer p 1-N, restored the ability to bind all the specific Abs in a polyclonal nDer p

1 antiserum. This indicates that the structure of rproDer p 1-CN represents a correctly folded protein, with the pro-peptide covering at least two B cell epitopes. Furthermore, this demonstrates that rDer p 1-N can be processed during fermentation and remain correctly folded. Finally, it suggests that a single recombinant isoallergen harbors most of the B cell epitopes present in all the isoallergens available in the nDer p 1 extract used here.

Glycosylations on natural and recombinant (*P. pastoris*) Der p 1 and Der f 1, resulting in a large molecular mass smear for the recombinant proteins, have been reported by a number of groups (16, 17, 47–49). By mutating the potential glycosylation site at N132 and N133 in the mature region of proDer p 1 and proDer f 1, respectively, this smear can be shifted into distinct bands with a lower molecular mass in SDS-PAGE (17, 49). There are four possible glycosylation sites in proDer p 1: N16, N82, N132, and N195. The glycosylation pattern of rproDer p 1-CN and rproDer p 1-N was investigated by analyzing the proteins before and after treatment with *N*-glycanase using SDS-PAGE and silver stain or glyco-binding dye. Although both proteins had the correct N terminus, two major glycosylated bands still remained after removal of the N132 glycosylation site (disregarding the 36 kDa rproDer p 1-N band). In Der f 1 N16 cannot be glycosylated due to an S18N substitution, and only a single band is observed in preparations of rproDer f 1-N133Q (17, 48, 49). Furthermore, *N*-glycanase treatment resulted in a single albeit still glycosylated major band for both rproDer p 1-N and rproDer p 1-CN. This indicates that N16 was glycosylated in a fraction of the rproDer p 1-N and rproDer p 1-CN molecules giving rise to the dual bands observed in SDS-PAGE representing molecules with (40-kDa band) and without (38-kDa band) a glycan on N16. The sharpness of the bands suggests that N16 carried a small but significant and homogenous glycan that was susceptible to *N*-glycanase cleavage. The lower intensity of the 40-kDa bands in silver-stained SDS-PAGE indicates that the majority of the molecules had no glycosylation on N16 (Fig. 1A). Hence, the lack of electron density for the N16 glycan in the rproDer p 1-CN crystal structure can be explained by low occupancy combined with flexibility. The persistence of a glycosylated band in rproDer p 1-N, rproDer p 1-CN, and nDer p 1 after *N*-glycanase treatment and in rDer p 1-N indicates a third glycosylated site. The crystal structure shows that this was at N195 and furthermore reveals no evidence for any glycosylation on N82. However, this carbohydrate seems to be resistant to treatment with *N*-glycanase. Interestingly, the 40- and 38-kDa bands in the rproDer p 1-N sample could be separated by Con A affinity chromatography as the 40-kDa band was retained on the column (data not shown). The N195 carbohydrate thus furthermore did not bind this lectin. In the glyco-binding dye-stained gel, the 40-kDa rproDer p 1-N and rproDer p 1-CN bands stain more intensely than the 38-kDa bands although present in smaller amounts according to the silver-stained gel. This is in line with the above argumentation, because the 40-kDa proteins contain two glycosylations and consequently bind more dye than the 38-kDa proteins. The nature of the 36-kDa rproDer p 1-N band remains to be determined, but it may be an N-terminally truncated variant as observed previously for rproDer f 1 (49) and rproDer p 1 (17).

The substrate specificity of Der p 1 is currently being established. The most important subsite in determining specificity in papain as well as most proteases of family C1 is S2, which displays a preference for bulky hydrophobic side chains like Phe. In contrast, subsite S1 is less selective, presumably because of the lack of a clearly defined binding pocket. Only little is known about the specificities of the S' subsites, but it seems to be relatively broad. Generally, papain has a rather broad substrate specificity (18). Der p 1 has been shown to exhibit preference for the small aliphatic residues

Ala or Val in subsite S2 (50–52). This observation can be explained from the crystal structure. The I221 side chain is larger, and the entire residue is placed more into the active site cleft compared with V133 in papain, which renders a smaller albeit still hydrophobic S2 subsite. The fact that the P1 residue points out of the active site in the papain inhibitor complex IPAD suggests that only a long side chain would be capable of interacting directly with the protein. Mutating the P1 residue to an Arg in the superposition shown in Fig. 4B and choosing one of the five most observed side chain rotamers in the graphics program O places the N<sub>η2</sub> atom 2.5 Å from N248 O<sub>δ1</sub> in the rproDer p 1-CN structure (data not shown). N248 is not hydrogen bonded to any residue in the mature region of rproDer p 1-CN, and it is therefore possible that it is flexible and able to interact with other types of hydrophilic residues at the P1 site. This could explain the reported preference for charged residues such as Arg, Lys, and Glu at this site (51, 52). The proposed preference for small aliphatic residues and charged residues in subsites S2 and S1, respectively, is of particular interest because proDer p 1 is cleaved after the A79-E80 sequence during maturation. It is therefore plausible that mature Der p 1 is capable of activating other proDer p 1 molecules, which experiments using nDer p 1 to process rproDer p 1 have shown (47). The theoretical isoelectric point of mature wt Der p 1 (Swiss-Prot P08176) is 5.6 in contrast to the classical papaya cysteine endopeptidases, which have a basic theoretical pI (e.g., 9.6 for papain, P00784). The physicochemical properties in and around the active site cleft of Der p 1 and papain are thus very different, suggesting altogether different substrate specificities.

It has been suggested that Der p 1 should contain serine protease activity in addition to the cysteine protease activity (53). However, none of the 12 Ser residues present in the mature region of the rproDer p 1-CN structure are located in the environment including the other two members of the catalytic triad, a His and an Asp residue, that normally constitute the active site in serine proteases. We have recently shown that the observed serine protease activity can be ascribed to another mite allergen, Der p 3, which copurifies with nDer p 1 (J. Sauer, unpublished information).

It is evident from Fig. 5 that the residues forming the substrate-binding cleft are conserved between the three mite species *D. pteronyssinus*, *D. farinae*, and *E. maynei*. This suggests that the substrate specificity is conserved among the different isoallergens from the same species as well as homologous allergens from different mite species. It furthermore suggests that one inhibitor (e.g., E-64) is sufficient to inhibit all mite group 1-mediated proteolytic activity of a natural or recombinant preparation of mite allergens from the species mentioned above.

The IgE inhibition assays show that the presence of a glycosylation on N132 had no significant effect on Ab binding. The pro-peptide, on the other hand, clearly changed the epitope pattern of the molecule, presumably by shielding a subset of the epitopes. However, all four variants were capable of inhibiting all IgE binding to nDer p 1. This suggests that either the IgE Abs were able to displace the pro-peptide or a small amount of the mature rDer p 1 variants were present in the samples. The CLIE results support the latter explanation. Fig. 7 shows that the two variants with a mutation at the active site Cys residue have an even more altered epitope pattern. A single amino acid substitution is unlikely to have this effect, but this remains to be investigated further.

The crystal structure reported here enables the design of Der p 1 variants with modified Ab-binding properties and additional studies of the substrate specificity of Der p 1. Furthermore, an elucidation of the interactions between the pro and mature regions of proDer p 1 provides insight into the activation process and specific inhibition of this important allergen with a view to large-scale production of correctly folded recombinant mature protein, and to the

design of small molecule inhibitors that will inhibit the protease activity by interacting directly with the mature protein or its pro form.

## Acknowledgments

We thank Annette Giselsson, Gitte N. Hansen, Michael G. Jensen, and Gitte K. Koed for excellent technical assistance and Dr. Jette Sandholm Kastrup for helpful discussions during structure refinement.

## Disclosures

K. Meno, P. B. Thorsted, H. Ipsen, J. N. Larsen, M. D. Spangfort, and K. Lund are employees at ALK-Abello A/S and H. Ibsen is also a stockholder.

## References

- Bousquet, J., P. van Cauwenberge, and N. Khaltaev. 2001. Allergic rhinitis and its impact on asthma. *J. Allergy Clin. Immunol.* 108: S147–S334.
- Janson, C., J. Anto, P. Burney, S. Chinn, R. de Marco, J. Heinrich, D. Jarvis, N. Kuenzli, B. Leynaert, C. Luczynska, et al. 2001. The European Community Respiratory Health Survey: what are the main results so far? *European Community Respiratory Health Survey II. Eur. Respir. J.* 18: 598–611.
- Holm, J., M. Gajhede, M. Ferreras, A. Henriksen, H. Ipsen, J. N. Larsen, L. Lund, H. Jacobi, A. Millner, P. A. Wurtzen, et al. 2004. Allergy vaccine engineering: epitope modulation of recombinant Bet v 1 reduces IgE binding but retains protein folding pattern for induction of protective blocking-antibody responses. *J. Immunol.* 173: 5258–5267.
- Fernandez-Caldas, E., L. Puerta, L. Caraballo, and R. F. Lockey. 2004. Mite allergens. *Clin. Allergy Immunol.* 18: 251–270.
- Thomas, W. R., W. A. Smith, B. J. Hales, K. L. Mills, and R. M. O'Brien. 2002. Characterization and immunobiology of house dust mite allergens. *Int. Arch. Allergy Immunol.* 129: 1–18.
- Horn, N., and P. Lind. 1987. Selection and characterization of monoclonal antibodies against a major allergen in *Dermatophagoides pteronyssinus*: species-specific and common epitopes in three dermatophagoides species. *Int. Arch. Allergy Appl. Immunol.* 83: 404–409.
- Lind, P., O. C. Hansen, and N. Horn. 1988. The binding of mouse hybridoma and human IgE antibodies to the major fecal allergen, Der p I, of *Dermatophagoides pteronyssinus*: relative binding site location and species specificity studied by solid-phase inhibition assays with radiolabeled antigen. *J. Immunol.* 140: 4256–4264.
- King, C., R. J. Simpson, R. L. Moritz, G. E. Reed, P. J. Thompson, and G. A. Stewart. 1996. The isolation and characterization of a novel collagenolytic serine protease allergen (Der p 9) from the dust mite *Dermatophagoides pteronyssinus*. *J. Allergy Clin. Immunol.* 98: 739–747.
- Lind, P. 1985. Purification and partial characterization of two major allergens from the house dust mite *Dermatophagoides pteronyssinus*. *J. Allergy Clin. Immunol.* 76: 753–761.
- Schulz, O., H. F. Sewell, and F. Shakib. 1999. The interaction between the dust mite antigen Der p 1 and cell-signalling molecules in amplifying allergic disease. *Clin. Exp. Allergy* 29: 439–444.
- Chua, K. Y., G. A. Stewart, W. R. Thomas, R. J. Simpson, R. J. Dilworth, T. M. Plozza, and K. J. Turner. 1988. Sequence analysis of cDNA coding for a major house dust mite allergen, Der p 1: homology with cysteine proteases. *J. Exp. Med.* 167: 175–182.
- Chua, K. Y., P. K. Kehal, and W. R. Thomas. 1993. Sequence polymorphisms of cDNA clones encoding the mite allergen Der p I. *Int. Arch. Allergy Immunol.* 101: 364–368.
- Vernet, T., H. E. Khouri, P. Laflamme, D. C. Tessier, R. Musil, B. J. Gour-Salin, A. C. Storer, and D. Y. Thomas. 1991. Processing of the papain precursor: purification of the zymogen and characterization of its mechanism of processing. *J. Biol. Chem.* 266: 21451–21457.
- Ikemura, H., H. Takagi, and M. Inouye. 1987. Requirement of pro-sequence for the production of active subtilisin E in *Escherichia coli*. *J. Biol. Chem.* 262: 7859–7864.
- Greene, W. K., J. G. Cyster, K. Y. Chua, R. M. O'Brien, and W. R. Thomas. 1991. IgE and IgG binding of peptides expressed from fragments of cDNA encoding the major house dust mite allergen Der p I. *J. Immunol.* 147: 3768–3773.
- Jacquet, A., M. Magi, H. Petry, and A. Bollen. 2002. High-level expression of recombinant house dust mite allergen Der p 1 in *Pichia pastoris*. *Clin. Exp. Allergy* 32: 1048–1053.
- Takai, T., R. Mineki, T. Nakazawa, M. Takaoka, H. Yasueda, K. Murayama, K. Okumura, and H. Ogawa. 2002. Maturation of the activities of recombinant mite allergens Der p 1 and Der f 1, and its implication in the blockade of proteolytic activity. *FEBS Lett.* 531: 265–272.
- Barrett, A. J., N. D. Rawlings, and J. F. Woessner, eds. 1998. Cysteine peptidases. In *Handbook of Proteolytic Enzymes*. Academic Press, London, p. 543–798.
- Menard, R., J. Carriere, P. Laflamme, C. Plouffe, H. E. Khouri, T. Vernet, D. C. Tessier, D. Y. Thomas, and A. C. Storer. 1991. Contribution of the glutamine 19 side chain to transition-state stabilization in the oxyanion hole of papain. *Biochemistry* 30: 8924–8928.
- Storer, A. C., and R. Menard. 1994. Catalytic mechanism in papain family of cysteine peptidases. *Methods Enzymol.* 244: 486–500.
- Ho, S. N., H. D. Hunt, R. M. Horton, J. K. Pullen, and L. R. Pease. 1989. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* 77: 51–59.
- Otwinski, Z., and W. Minor. 1997. Processing of X-ray diffraction data collected in oscillation mode. In *Methods in Enzymology: Macromolecular Crystallography*, Vol. 276, Pt. A. C. W. Carter Jr. and R. M. Sweet, eds. Academic Press, New York, p. 307–326.
- Collaborative Computational Project, Number 4. 1994. The CCP4 suite: programs for protein crystallography. *Acta Cryst. D* 50: 760–763.
- Vagin, A., and A. Teplyakov. 1997. MOLREP: an automated program for molecular replacement. *J. Appl. Cryst.* 30: 1022–1025.
- Maes, D., J. Bouckaert, F. Poortmans, L. Wyns, and Y. Looze. 1996. Structure of chymopapain at 1.7 Å resolution. *Biochemistry* 35: 16292–16298.
- Lamzin, V. S., A. Perrakis, and K. S. Wilson. 2001. The ARP/WARP suite for automated construction and refinement of protein models. In *Crystallography of Biological Macromolecules*, Vol. F. M. G. Rossman and E. Arnold, eds. Kluwer Academic Publishers, Dordrecht, The Netherlands, p. 720–722.
- Jones, T. A., J.-Y. Zou, S. W. Cowan, and M. Kjeldgaard. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Cryst. A* 47: 110–119.
- Murshudov, G. N., A. A. Vagin, and E. J. Dodson. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst. D* 53: 240–255.
- Brünger, A. T. 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355: 472–475.
- Perrakis, A., T. K. Sixma, K. S. Wilson, and V. S. Lamzin. 1997. wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Cryst. D* 53: 448–455.
- Krøll, J. 1973. Crossed-line immunoelectrophoresis (73, 76). *Scand. J. Immunol.* 2(Suppl. 1): 79–81.
- Harboe, N., and A. Ingild. 1973. Immunization, isolation of immunoglobulins, estimation of antibody titre. *Scand. J. Immunol.* 2(Suppl. 1): 161–164.
- Weeke, B. 1973. A manual of quantitative immunoelectrophoresis: methods and applications. 1. General remarks on principles, equipment, reagents and procedures. *Scand. J. Immunol.* 2(Suppl. 1): 15–35.
- Petersen, A. B., P. Gudmann, P. Milvang-Gronager, R. Morkeberg, S. Bogestrand, A. Linneberg, and N. Johansen. 2004. Performance evaluation of a specific IgE assay developed for the ADVIA centaur immunoassay system. *Clin. Biochem.* 37: 882–892.
- Seppala, U., P. Hagglund, P. A. Wurtzen, H. Ipsen, P. Thorsted, T. Lenhard, P. Roepstorff, and M. D. Spangfort. 2005. Molecular characterization of major cat allergen Fel d 1: expression of heterodimer by use of a baculovirus expression system. *J. Biol. Chem.* 280: 3208–3216.
- DeLano, W. L. 2002. The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, www.pymol.org.
- Fraczkiewicz, R., and W. Braun. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* 19: 319–333.
- Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55: 379–400.
- Dilworth, R. J., K. Y. Chua, and W. R. Thomas. 1991. Sequence analysis of cDNA coding for a major house dust mite allergen, Der f I. *Clin. Exp. Allergy* 21: 25–32.
- Karrer, K. M., S. L. Peiffer, and M. E. DiTomas. 1993. Two distinct gene subfamilies within the family of cysteine protease genes. *Proc. Natl. Acad. Sci. USA* 90: 3063–3067.
- Krissinel, E., and K. Henrick. 2003. Protein structure comparison in 3D based on secondary structure matching (SSM) followed by Cα alignment, scored by a new structural similarity function. In *Proceedings of the 5th International Conference on Molecular Structural Biology*. A. J. Kungl and P. J. Kungl, eds. Vienna, p. 88.
- Groves, M. R., M. A. Taylor, M. Scott, N. J. Cummings, R. W. Pickersgill, and J. A. Jenkins. 1996. The prosequence of procaricain forms an α-helical domain that prevents access to the substrate-binding cleft. *Structure* 4: 1193–1203.
- Coulombe, R., P. Grochulski, J. Sivaraman, R. Menard, J. S. Mort, and M. Cygler. 1996. Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. *EMBO J.* 15: 5492–5503.
- Groves, M. R., R. Coulombe, J. Jenkins, and M. Cygler. 1998. Structural basis for specificity of papain-like cysteine protease proregions toward their cognate enzymes. *Proteins* 32: 504–514.
- Drenth, J., K. H. Kalk, and H. M. Swen. 1976. Binding of chloromethyl ketone substrate analogues to crystalline papain. *Biochemistry* 15: 3731–3738.
- LaLonde, J. M., B. Zhao, C. A. Janson, K. J. D'Alessio, M. S. McQueney, M. J. Orsini, C. M. Debouck, and W. W. Smith. 1999. The crystal structure of human procathepsin K. *Biochemistry* 38: 862–869.
- van Oort, E., P. G. de Heer, W. A. van Leeuwen, N. I. Derksen, M. Muller, S. Huveners, R. C. Aalberse, and R. van Ree. 2002. Maturation of *Pichia pastoris*-derived recombinant pro-Der p 1 induced by deglycosylation and by the natural cysteine protease Der p 1 from house dust mite. *Eur. J. Biochem.* 269: 671–679.
- Takahashi, K., T. Takai, T. Yasuhara, T. Yokota, and Y. Okumura. 2001. Effects of site-directed mutagenesis in the cysteine residues and the N-glycosylation motif in recombinant Der f 1 on secretion and protease activity. *Int. Arch. Allergy Immunol.* 124: 454–460.
- Yasuhara, T., T. Takai, T. Yuuki, H. Okudaira, and Y. Okumura. 2001. Biologically active recombinant forms of a major house dust mite group 1 allergen Der f 1 with full activities of both cysteine protease and IgE binding. *Clin. Exp. Allergy* 31: 116–124.
- Schulz, O., H. F. Sewell, and F. Shakib. 1998. A sensitive fluorescent assay for measuring the cysteine protease activity of Der p 1, a major allergen from the dust mite *Dermatophagoides pteronyssinus*. *Mol. Pathol.* 51: 222–224.
- Shakib, F., O. Schulz, and H. Sewell. 1998. A mite subversive: cleavage of CD23 and CD25 by Der p 1 enhances allergenicity. *Immunol. Today* 19: 313–316.
- Harris, J., D. E. Mason, J. Li, K. W. Burdick, B. J. Backes, T. Chen, A. Shipway, G. Van Hecke, L. Gough, A. Ghaemmaghami, et al. 2004. Activity profile of dust mite allergen extract using substrate libraries and functional proteomic microarrays. *Chem. Biol.* 11: 1361–1372.
- Hewitt, C. R., H. Horton, R. M. Jones, and D. I. Pritchard. 1997. Heterogeneous proteolytic specificity and activity of the house dust mite proteinase allergen Der p 1. *Clin. Exp. Allergy* 27: 201–207.