

# Molecular mechanisms in allergy and clinical immunology

(Supported by a grant from Merck & Co, Inc, West Point, Pa)

Series editor: Lanny J. Rosenwasser, MD

## Structural biology of allergens

Rob C. Aalberse, PhD *Amsterdam, The Netherlands*

One of the major challenges of molecular allergy is to predict the allergenic potential of a protein, particularly in novel foods. Two aspects have to be distinguished: immunogenicity and cross-reactivity. Immunogenicity reflects the potential of a protein to induce IgE antibodies, whereas cross-reactivity is the reactivity of (usually preexisting) IgE antibodies with the target protein. In addition to these two issues, the relation between IgE-binding potential and clinical symptoms is of interest. This is influenced by physical properties (eg, stability and size) and immunologic properties (affinity and epitope valence). Discussions on immunogenicity and cross-reactivity of allergens rely on the establishment of structural similarities and differences among allergens and between allergens and nonallergens. For comparisons between the 3-dimensional protein folds, the representation as 2-dimensional proximity plots provides a convenient visual aid. Analysis of approximately 40 allergenic proteins (or parts of these proteins), of which the protein folds are either known or can be predicted on the basis of homology, indicates that most of these can be classified into 4 structural families: (1) antiparallel  $\beta$ -strands: the immunoglobulin-fold family (grass group 2, mite group 2), serine proteases (mite group 3, 6, and 9), and soybean-type trypsin inhibitor (Ole e 1, grass group 11); (2) antiparallel  $\beta$ -sheets intimately associated with one or more  $\alpha$ -helices: tree group 1, lipocalin, profilin, aspartate protease (cockroach group 2); (3) ( $\alpha$ + $\beta$ ) structures, in which the  $\alpha$ - and  $\beta$ -structural elements are not intimately associated: mite group 1, lysozyme/lactalbumin, vespid group 5; and (4)  $\alpha$ -helical: nonspecific lipid transfer protein, seed 2S protein, insect hemoglobin, fish parvalbumin, pollen calmodulin, mellitin from bee venom, Fel d 1 chain 1, serum albumin. Allergens with parallel  $\beta$ -strands (in combination with an  $\alpha$ -helix linking the two strands, a motif commonly found in, for example, nucleotide-binding proteins) seem to be underrepresented. The conclusion is that allergens have no characteristic structural features other than that they need to be able to reach (and stimulate) immune cells and mast cells. Within this constraint, any antigen may be allergenic, particularly if it avoids activation of  $T_H2$ -suppressive mechanisms (CD8 cells and  $T_H1$  cells). (*J Allergy Clin Immunol* 2000;106:228-38.)

**Key words:** Allergen structure, allergenicity, cross-reactivity, epitope, protein folding, food allergen, novel foods, recombinant allergen

### Abbreviations used

CA:	$\alpha$ -Carbon
ER:	Endoplasmic reticulum
MnSOD:	Manganese-dependent superoxide dismutase
PDB:	Protein Data Bank
SCOP:	Structural Classification of Proteins

What makes an antigen an allergen? This question cannot be answered yet, but we are getting clues from different directions. Why is this a relevant question? First, there is the theoretic aspect: if we want to understand allergy, we need to understand allergens. There is a practical aspect as well: an answer to the question would help us decide whether the introduction of a new protein into our environment (particularly into our diet) increases the risk of subsequent allergic symptoms.<sup>1-4</sup> Moreover, modification or replacement of an allergenic protein may be of help for allergic patients, provided that the alternative protein (ie, the modified protein or the substitute) is less allergenic. Finally, if we understand allergens, we may find better ways of producing appropriately modified allergens for treatment.

Protein nomenclature has been confusing the understanding of protein structure for a long time. Proteins with the label *albumin* prove to have very different structures. Egg albumin, the primordial albumin, is structurally very different from serum albumin, milk albumin, fish parvalbumin, or seed 2S albumin. The term *pathogenesis-related protein*<sup>5</sup> is used for a series of plant proteins involved in stress reactions (comparable with the use of *acute-phase proteins* in mammals). Despite their common name, pathogenesis-related proteins have almost no structural relationship at all. Even in recent databases, a nomenclature issue has resulted in mislabeling a serum protein as an allergen. The rat serum prealbumin transthyretin was referred to as the major allergen Rat n 1 because it was labeled as prealbumin. However, Rat n 1 is a urinary prealbumin unrelated to serum prealbumin.

This review deals exclusively with protein allergens. IgE antibodies to nonpeptidic epitopes are known, for example, to classical haptens, such as the penicilloyl group,<sup>6</sup> and to glycosidic side chains of nonmammalian glycoproteins.<sup>7</sup> Protein structure can be described at different levels: primary structure (ie, the amino acid sequence), protein fold, domain structure, and surface structure. The surface structure is the most relevant for antibody binding, particularly the epitope, which is that part of the surface that on an atomic level interacts with the antibody.

From the Department of Allergy, CLB and the Laboratory for Experimental and Clinical Immunology, Academic Medical Center, University of Amsterdam, Amsterdam.

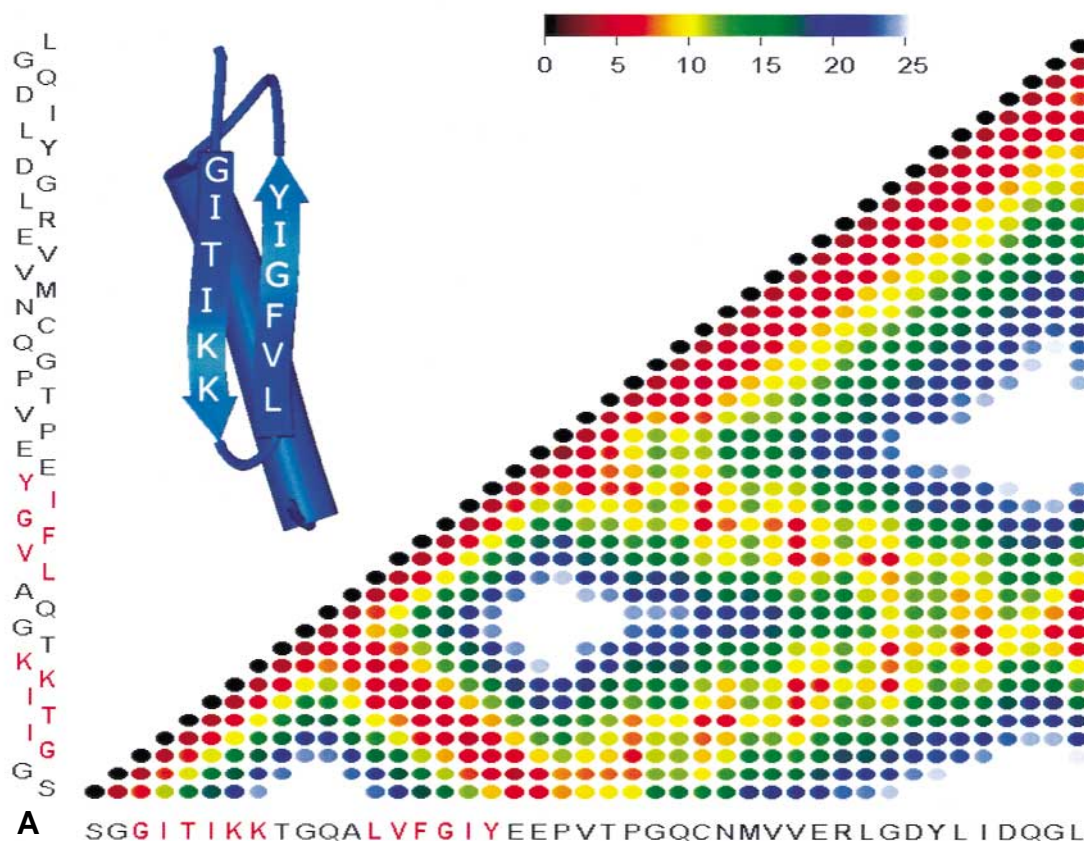
Received for publication Apr 3, 2000; revised Apr 28, 2000; accepted for publication Apr 28, 2000.

Reprint requests: Rob C. Aalberse, PhD, the Department of Allergy, CLB and the Laboratory for Experimental and Clinical Immunology, Academic Medical Center, University of Amsterdam, PO Box 9190, 1006 AD, Amsterdam, The Netherlands.

Copyright © 2000 by Mosby, Inc.

0091-6749/2000 \$12.00 + 0 1/108434

doi:10.1067/mai.2000.108434



**FIG 1.** Two illustrations on the relationship between 3-dimensional structure and the corresponding proximity plots. The distances are color coded: red indicates a small distance, blue is a longer distance, and white is a distance of more than 2.5 nm. **A**, The C-terminal part of birch profilin (PDB code 1CQA), which consists of two antiparallel  $\beta$ -strands and an  $\alpha$ -helix. The amino acids are indicated in their single-letter abbreviations from left to right and from bottom to top. Red-colored dots indicate that the two amino acids are at a short distance in the 3-dimensional structure; that is, the first amino acid of the first strand (GITIKK) is close to the last amino acid of the second strand (LVFGIY). As a start, it helps to focus on the distances of less than 1.0 nm (red or orange dots). The strands can be distinguished from the helix by counting the number of red-orange dots next to the diagonal: two for the strands and four for the helix. The antiparallel interactions between the two  $\beta$ -strands are visible as the series of adjacent red dots perpendicular to the diagonal. The interaction between the helix and each of the two  $\beta$ -strands produces a characteristic >-shaped checkered pattern; this checkered pattern results from the cyclic structure of the helix. **B**, The N-terminal part of schistosomal glutathione S-transferase (PDB code 1GTA). This fragment also contains two  $\beta$ -strands and one  $\alpha$ -helix, but in this case the helix is between the two strands, and the strands are therefore parallel. The first amino acids in the first strand (ILGYW) are close to the first amino acids of the second strand (EEHLY). The interaction between the  $\beta$ -strands is visible in the lower-right corner as the series of adjacent red squares parallel to the diagonal (rather than perpendicular, as in **A**). This is a common fold, but it is relatively rare in allergens.

The current discussion will be focused on two aspects of allergen structure: a description (in general terms) of what is known about protein folds in allergens and about the characteristics of epitopes.

## IMMUNOLOGIC VERSUS CLINICAL ALLERGENICITY

Semantically, the concept of allergenicity is ill defined. To a clinical allergist, allergenicity reflects the capacity of an antigen to induce symptoms or a skin reaction, whereas to an immunologist, it reflects either a peculiar type of immunogenicity (ie, the capacity of a

protein to induce IgE antibodies) or simply the capacity to bind IgE antibodies.

Similarly, the term *allergen* is used to describe two or three distinct molecular properties: the property to sensitize (ie, induce the immune system to produce high-affinity antibodies, particularly of the IgE class) and the property to elicit an allergic reaction (ie, to trigger allergic symptoms in a sensitized subject). Moreover, it is also used to indicate the property to bind IgE antibodies. Complete allergens have all these properties. Some proteins, however, are known to elicit allergic symptoms but do not usually sensitize.

A well-known example of such a nonsensitizing elicitor (or incomplete allergen) is Mal d 1. This protein is the

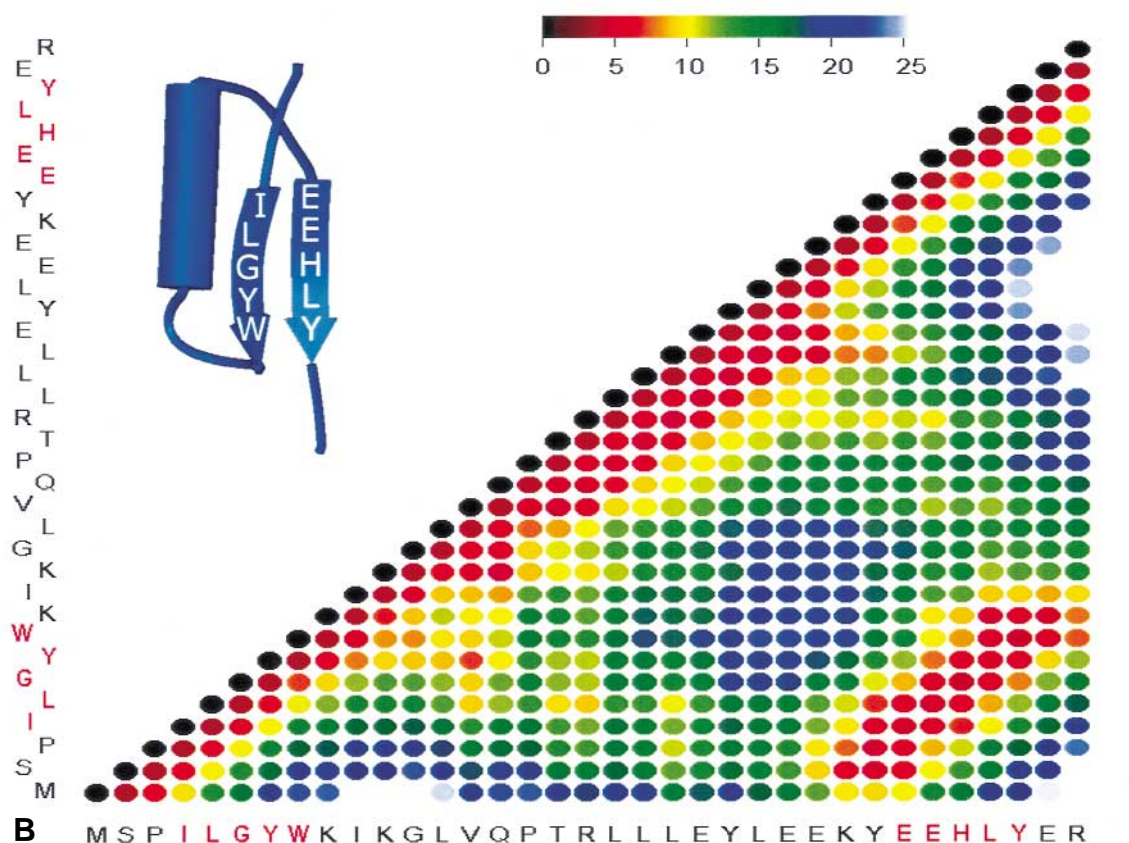


FIG 1. Continued.

apple homologue of the group 1 allergen from birch, Bet v 1. Inhalation of birch pollen induces IgE antibodies to Bet v 1, some of which cross-react with Mal d 1. Ingestion of apple does not induce IgE antibodies but may trigger activation of mast cells that are loaded with preexisting (birch-induced) cross-reactive IgE antibodies. By using RAST inhibition, the complete and incomplete allergens can be distinguished. In the example birch pollen will completely inhibit IgE binding to apple, whereas apple will give only partial inhibition of IgE binding to the birch allergen.

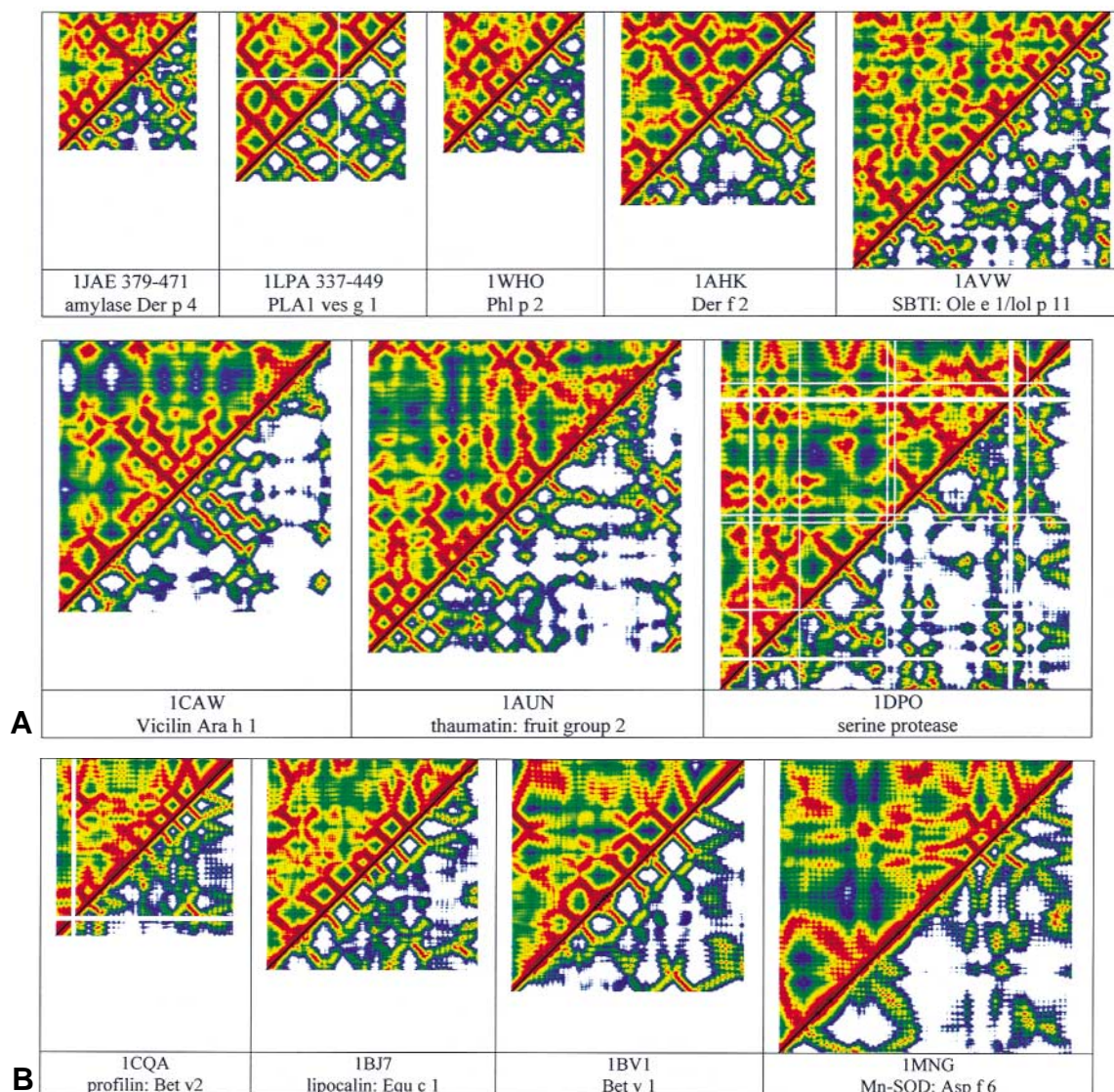
Obviously, oral exposure to Mal d 1 hardly, if ever, triggers the immune system to produce IgE antibodies, whereas inhalation of a very similar protein, Bet v 1, in quantities that are not vastly different often does. It seems likely that Mal d 1 does not have what it takes to make a protein a food allergen. One structural feature that may explain part of the differences between these incomplete food allergens and true food allergens, such as peanut allergens, is the susceptibility to the low pH in the stomach and to proteolytic enzymes, particularly pepsin.<sup>8</sup>

In this review allergenicity will be used in the immunogenicity sense: the induction of IgE antibodies. The most relevant second issue is allergic (in the clinical sense) reactivity of nonallergenic (ie, nonimmunogenic) proteins. This will be discussed under the heading of cross-reactivity.

## WHEN IS A PROTEIN CONSIDERED TO BE A MAJOR ALLERGEN?

The distinction between major and minor allergens is relevant for various reasons but also in relation to the issue of allergenicity. The current definition of major allergen is based on the prevalence of IgE or skin reactivity in subjects that are sensitized (usually very strongly) to the total extract. This definition is unsatisfactory in that it does not reflect the contribution of the allergen to the overall reactivity of the extract. Intuitively, removal of a truly major allergen from an extract is expected to have a noticeable effect on the overall reactivity of that extract: a major allergen should make a difference. Such an interpretation invites a different type of definition. For example, a major allergen is responsible for more than 20% of the allergenic reactivity in more than 20% of the sensitized patients. This requires testing with extracts from which the allergen in question has been selectively removed (eg, with monospecific antibodies). Alternatively, it could be tested serologically by absorbing out all IgE antibodies to the allergen and then testing the residual activity of the absorbed serum. For most allergens, this aspect has not been studied. It is likely, however, that the major allergen claim made for many allergens would need to be reconsidered.

More relevant for the present discussion on allergenicity is to note that people outside the allergen-characteri-



**FIG 2.** Classification of allergen folds. **A**, Group 1: anti-parallel  $\beta$ -strands. Note that the first two structures (1JAE 379-471 and 1LPA 337-449) are C-terminal domains of proteins. The complete proteins are shown in **E**. The antiparallel interaction between  $\beta$ -strands is indicated by the red lines perpendicular to the diagonal. Of vicilin (1CAW) and thaumatin (1AUN), only the N-terminal parts fall in this fold class. **B**, Group 2: antiparallel  $\beta$ -sheets intimately associated with one or more  $\alpha$ -helices. The interaction between the  $\beta$ -sheet and the  $\alpha$ -helix is indicated by the striped patterns caused by the periodicity of the helical structure. The C-terminal helix of Bet v 1 (1BV1) is further away from the  $\beta$ -sheet than the C-terminal helix of profilin (or lipocalin). Of superoxide dismutase (1MNG), only the C-terminal part falls in this class because the N-terminal part consists of two interacting helices. The two domains of trypsin (1DPO) can be seen to have a similar fold. **C**, Group 3:  $\alpha$ + $\beta$  structures, in which the  $\alpha$ - and  $\beta$ -structural elements are not intimately associated. Note the striking similarity between lysozyme (1HEL) and lactalbumin (1HFZ).<sup>40</sup> **D**, Group 4:  $\alpha$ -helical proteins. Interaction between helices can be recognized by the Scottish-kilt patterns caused by the periodicity of the helical structures. **E**, Larger proteins are plotted in a smaller format (25% of their original size). Note the repeating domain structures of serum albumin (3 similar domains) and ovotransferrin (2 similar domains). The white lines in some of the plots indicate gaps in the protein sequence.

zation field may assume that a major allergen is synonymous with a major allergenic risk. This is clearly not the case. For example, patatin has been reported to be a major allergen in potato<sup>9</sup> but is not a major allergenic risk, considering the frequency of exposure (ie, not as a potato protein but possibly as a latex allergen).<sup>10,11</sup> In terms of aller-

genicity, it would be relevant to compare the sensitizing potential of proteins for which the exposure is similar. In this context mite group 1 allergen would be a more major allergen (ie, a higher allergenic risk) than Can f 1, even though the latter is a major allergen. Some allergens are more major than others.

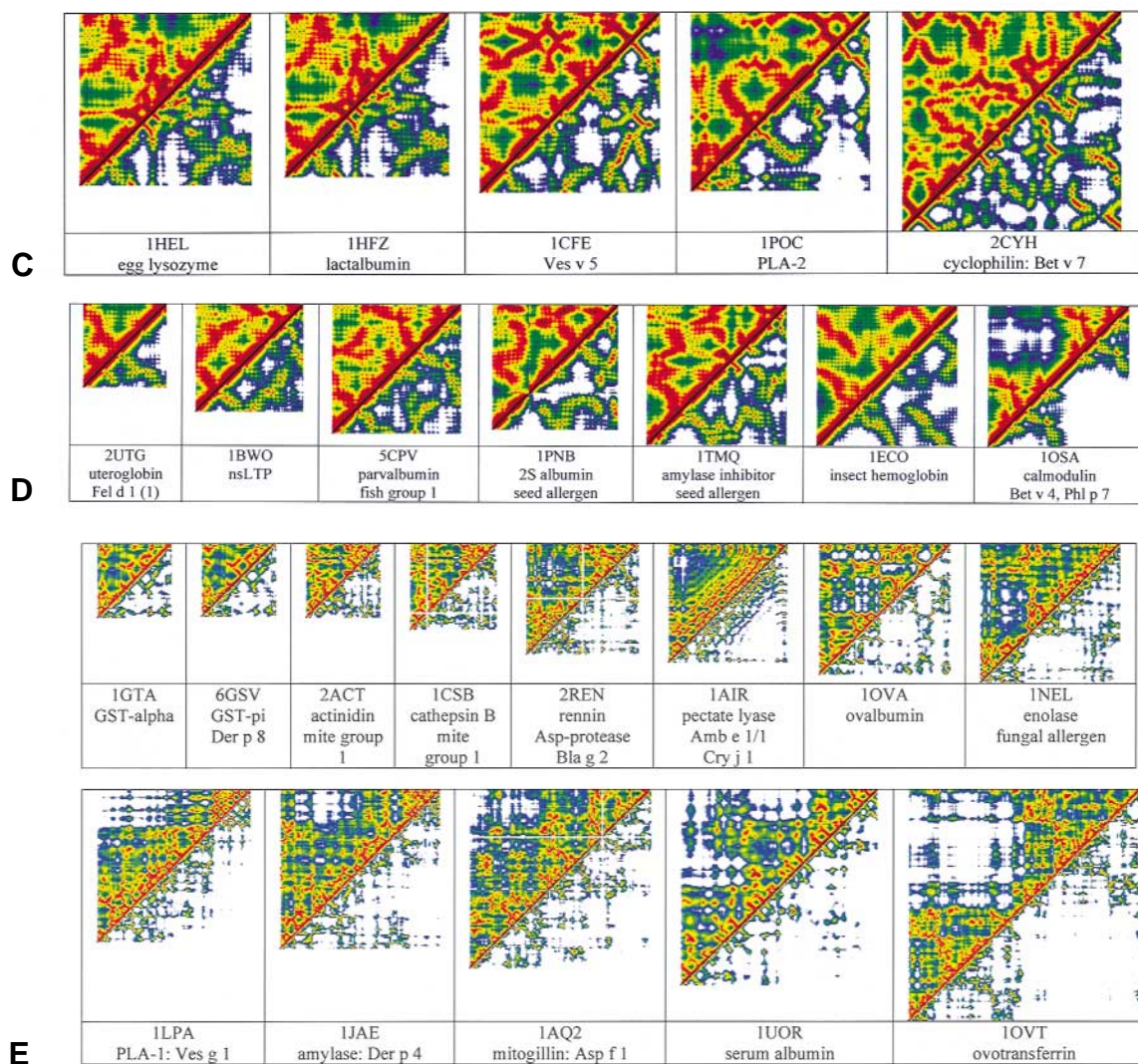


FIG 2. Continued.

## DETERMINANTS OF ALLERGENICITY

Aspects of protein structure likely to be relevant for allergenicity are solubility, stability, size, and the compactness of the overall fold. These aspects reflect dependency of allergenicity on transport over mucosal barriers and susceptibility to proteases. Size and solubility of the intact protein would be relevant factors for airborne allergens more than for food allergens (for which limited proteolysis might enhance mucosal transport and hence allergenicity) or parenteral antigens, such as insect venoms, insect salivary allergens, invasive organisms (helminths and fungi), vaccines, or therapeutic proteins.

Posttranslational modification may affect allergenicity in different ways. It may induce new epitopes and it may affect solubility, stability, size, and susceptibility toward proteases. Moreover, uptake and processing by antigen-presenting cells are also known to be markedly influenced. Although glycosylation affects many of

these processes, it is not a critical factor for allergenicity in general. Many allergens are not glycosylated, whereas some important allergens (eg, Gal d 1 [ovomucoid]) are heavily glycosylated.

It will become clear from this overview that few, if any, structural features are currently known to be common for allergens in general, even though most allergens can be grouped into a small number of structural classes. Some of these classes have been suggested to be intrinsically more allergenic, for example, 2S albumin from seeds<sup>12</sup> and lipocalin.<sup>13</sup> These will be discussed in more detail below.

It is, however, likely that features other than structure are more relevant for allergenicity. In addition, the search for common structural features relevant for allergenicity will become more relevant as we learn about the various pathways of allergic sensitization. It is not realistic to assume that the requirements for sensitization to food

**TABLE I.** Classification of protein folds in allergens

(1) Antiparallel $\beta$ -strands	
The immunoglobulin-fold family	
Grass group 2 ( <b>1BMW</b> , <b>1WHO</b> , <b>1WHP</b> )	
Grass group 1 (C-terminus)	
Grass group 3	
Mite group 2 ( <b>1A9V</b> , <b>1AHK</b> , <b>1AHM</b> )	
Serine proteases (example: 1DPO, trypsin)	
Mite group 3	
Mite group 6	
Mite group 9	
Soybean Kunitz-type trypsin inhibitor (1AVW)	
Ole e 1	
Grass group 11	
Fruits group 2: thaumatin (1AUN)	
Vicilin: peanut Ara h 1 (1CAW, 1DGR, 1DGW)	
(2) Antiparallel $\beta$ -sheets intimately associated with one or more $\alpha$ -helices	
Tree group 1 ( <b>1BTV</b> , <b>1BV1</b> )	
Lipocalin	
Milk $\beta$ -lactoglobulin ( <b>1BLG</b> )	
Mouse ( <b>1MUP</b> ) and rat urinary protein ( <b>2A2G</b> , <b>2A2U</b> )	
Dog Can f 1	
Dog Can f 2	
Bovine Bos d 1	
Horse Equ c 1 ( <b>1BJ7</b> )	
Cockroach Bla g 4	
Cystatin: cat allergen 4 <sup>30</sup> (1A67, 1CEW)	
Profilin ( <b>1CQA</b> )	
Aspartate protease (2REN)	
Cockroach Bla g 2	
(3) ( $\alpha$ + $\beta$ ) structures, in which the $\alpha$ - and $\beta$ -structural elements are not intimately associated	
Mite group 1 (2ACT, 1CSB)	
Lysozyme ( <b>1HEL</b> )/lactalbumin ( <b>1HFZ</b> )	
Vespid group 5 (1CFE)	
Ovotransferrin = conalbumin ( <b>1OVT</b> )	
Cyclophilin (2CYH)	
Grass group 4	
Tree group 7	
Phospholipase A2 ( <b>1POC</b> )	
(4) $\alpha$ -Helical	
Nonspecific lipid transfer protein (1BWO)	
Seed 2S albumin (1PNB)	
Insect hemoglobin ( <b>1ECO</b> )	
Fish parvalbumin ( <b>1CPD</b> , <b>5CPV</b> )	
Calmodulin (1OSA)	
Bet v 4	
Jun o 2	
Phl p 7	
Mellitin from bee venom ( <b>1MLT</b> )	
Fel d 1 chain 1 (2UTG)	
Serum albumin (1UOR)	
(5) Other structures	
$\beta$ -Helix: pectate lyase (1AIR, 2PEC)	
Amb e 1	
Amb e 2	
Cry j 1	
Serine protease inhibitor (Serpins-family)	
Ovalbumin ( <b>1OVA</b> )	
PLA1 1LPA	
Glutathione S-transferase (1HNB, <b>1GTA</b> )	
Cockroach group 5	
Mite group 8	
Schistosomal glutathione S-transferase	
Mitogillin: Asp f 1 ( <b>1AQ2</b> )	
MnSOD Asp f 6 (1MNG)	
Enolase (1NEL)	
Amylase (1JAE)	
Ovotransferrin (1OVT)	
Coiled coil: tropomyosin (1C1G, 1TMZ, 2TMA)	
Shrimp group 1	
Mite group 10	
Cockroach	
Small proteins	
Ovomucoid (third domain only) 1OMU, 1OVO, 1CT4	
Hevein <b>1HEV</b>	
Amb e 5 <b>1BBG</b> , <b>2BBG</b> , <b>3BBG</b>	

PDB codes printed in bold are coordinate files of the allergen itself; the other PDB files represent homology models.

allergens in early childhood are very similar to those for late-onset sensitization to airborne occupational allergens.

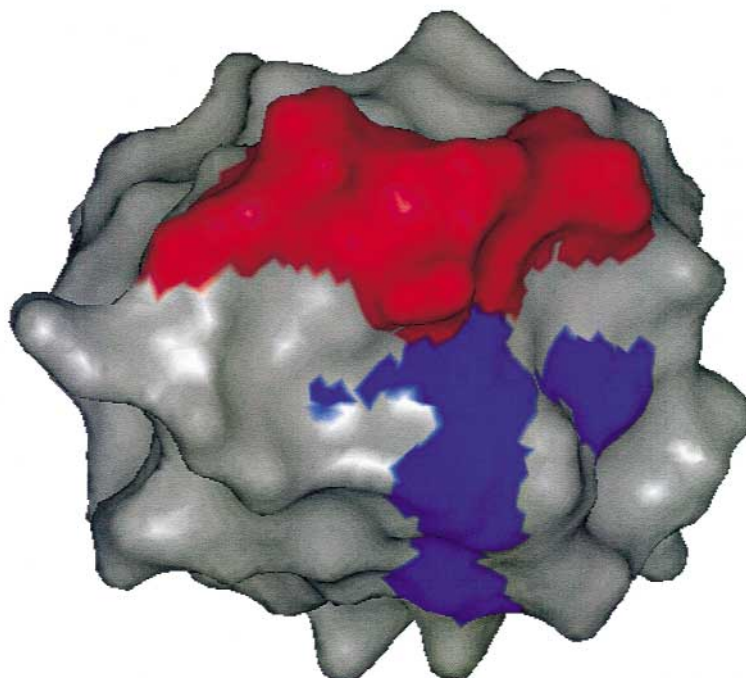
## DETERMINANTS OF CROSS-REACTIVITY

In contrast to allergenicity, cross-reactivity is largely determined by structural aspects: two proteins are cross-reactive only (almost) if they share structural features. What are the few exceptions to this rule?

Antibody affinity is an important consideration: low-affinity antibodies (particularly of the IgM class) have been found to be reactive with antigens with very little structural similarity. However, IgE-producing plasma cells evolved from a strongly T cell-dependent immune reaction in which only high-affinity B cells survive. Moreover, IgE-mediated cellular reactions are triggered by trace amounts of allergens. For these reasons, the discussion will be focused on high-affinity antibodies. It should be kept in mind, however, that information on the

threshold of the affinity required for triggering allergic reactions by using cross-reactive allergens is still lacking. Another apparent exception is the cross-reactivity between glycoprotein allergens with unrelated protein structures.<sup>7</sup> In this situation cross-reactivity is due to shared glycans and thus similarity in structure.

Apart from these exceptions, all IgE cross-reactions described so far have been found to reflect shared features on the level of both primary and tertiary structure of the cross-reactive proteins. Whereas all cross-reactive proteins have a similar fold, the reverse is not true: proteins with a similar fold are not necessarily cross-reactive. This is partially due to immunologic tolerance induced by autologous proteins with a similar fold. More importantly, protein folding is liberal with respect to single amino acid substitutions for many positions in the sequence. Such substitutions may markedly affect the outer protein surface and thus reduce antibody reactivity. Similar protein folds are found with as little as 25%



**FIG 3.** The epitope of egg white lysozyme as recognized by mAb D1.3 (PDB code 1FDL).<sup>31</sup> The amino acids that contribute to the N-terminal part of the epitope (18, 19, 22, 23, 24, and 27) are indicated in blue. The amino acids that contribute to the C-terminal part of the epitope (116-121, 124, and 125) are indicated in red. This figure was produced by using WebLabViewerPro (Molecular Simulations Inc).

amino acid identity, whereas cross-reactivity is rare below 50% identity. In most situations cross-reactivity requires more than 70% identity.

The likelihood of cross-reactivity depends not only on the similarity between the primary sequences of the two antigens. If a human homologue exists and is available to the immune system, this is likely to induce immune tolerance to the common epitopes. In some instances tolerance to the human protein appears to be absent, resulting in autoreactive IgE antibodies. The first description of such an autoreactive IgE antibody presumably induced by contact with a cross-reactive environmental allergen was the autoreactivity described for human profilin in patients with pollinosis.<sup>14</sup> Another striking example is the reactivity of IgE antibodies induced by the fungal allergen Asp f 6 (manganese-dependent superoxide dismutase [MnSOD]).<sup>15,16</sup> The recombinant human protein was found to give strongly positive skin reactions in subjects with reactivity to the *Aspergillus* protein.

To predict potential cross-reactivity of a new allergen (eg, an engineered food protein), it is informative to compare protein folds. In the absence of similarity in folding with allergens, protein cross-reactivity is virtually excluded (apart from the above-mentioned exceptions). If similarity in folding is observed, cross-reactivity needs to be investigated. The knowledge of the folding family allows a search restricted to subjects with sensitization to allergens with that particular protein fold. However, antibodies are highly idiosyncratic. Therefore absence of cross-

reactivity of one antibody (or even many antibodies) does not exclude cross-reactivity of another antibody, even with the same epitope specificity. As with all risk assessments, statistical evaluation on the basis of a sufficiently large number of cases is important.

## PROTEIN STRUCTURE

Many details on protein structure are given by Branden and Tooze.<sup>17</sup> The basis is the primary structure, the amino acid sequence. This information is becoming more easily obtainable through sequencing of complementary DNA. The amino acid sequence deduced from the nucleotide sequence is usually not the complete and final primary structure because cotranslational and posttranslational modifications are common. Most allergens are extracellular proteins. Secretion of a protein by a cell requires the presence of an N-terminal leader peptide for transport of the protein through the membrane of the endoplasmic reticulum (ER). This peptide is cleaved by a protease in the ER. In the ER other modifications occur, particularly glycosylation of asparagines, serines, and/or threonines. This glycosylation is actually a multistep process involving many glycosyl transferases and trimming glycosidases. These reactions often result in heterogeneity among the secreted glycoproteins. Particularly for plant glycoproteins, an important modification is the hydroxylation of prolines and lysines. These hydroxylated amino acids are also targets for glycosylation. The list of posttranslational modifications is

much longer and still growing, but this aspect of protein structure is beyond the scope of this review.

After *in vitro* unfolding, most proteins will fold at least partially back into their native structure. Therefore the primary sequence contains all the clues for the 3-dimensional structure, and we should be able to predict the 3-dimensional structure of the protein. In practice this is still a distant dream. However, progress in this field is rapid because of the rapidly increasing number of 3-dimensional structures that have been and are being solved and the increased computational power. Homology modeling proves to be reliable for sequences that can be aligned well. With sequence homology as low as 25%, overall backbone folds can be predicted often with high accuracy, with the exception of some of the more complicated loops. Configurations of solvent-exposed side chains are far more difficult to predict, and therefore even if the 3-dimensional structure of two proteins with similar folds is known, prediction of cross-reactivity is not yet possible. Fold recognition is, however, an important help in the search for potential cross-reactivity.

Fold recognition on the basis of the primary sequence is dependent on the classification of folds, which can be automated to a certain extent,<sup>18-24</sup> but remains to some degree a subjective process. For this review, I have used the Structural Classification of Proteins (SCOP).<sup>25</sup> To visually compare the folds of allergens, these folds are represented in the format of a proximity matrix.<sup>26</sup> The rationale for this procedure is described below.

## PROXIMITY PLOT: COMPARING PROTEIN FOLDS AND DOMAIN INTERACTIONS

The structural basis of a folded protein is its peptide backbone. This consists of linear repeats of the 3 atoms (the amino-nitrogen, the  $\alpha$ -carbon, and the carbonyl carbon) that form the basic peptide triplet  $(-N-CA-CO-)_n$ . Because the goniometric angles between the atoms are fixed, the distance between two neighboring  $\alpha$ -carbons (CAs) is virtually constant (0.38 nm). The degree of contraction of the protein backbone largely depends on torsion angles (ie, the angles along the axis between two atoms). The two most relevant torsion angles are torsion along the N-CA axis ( $\Phi$ ) and torsion along the CA-CO-axis ( $\Psi$ ). In its most extended form (the  $\beta$ -strand), the length of the backbone increases by approximately 0.33 nm for each subsequent amino acid. In its contracted form (ie, the well-known  $\alpha$ -helix), the length increases by approximately half this distance. Intermediates between these two extremes (the extended  $\beta$ -strand and the contracted  $\alpha$ -helix) are referred to as turns or coils (depending, among others, on the number of amino acids involved). These secondary structure elements are stabilized by hydrogen bonds between  $\beta$ -strands (resulting in  $\beta$ -sheets) or within  $\alpha$ -helices, turns, or coils.

Interactions between these elements of secondary structure results in the tertiary structure: domains of usually 80 to 200 amino acids often stabilized by intrachain disulfide bonds. In many enzymes two domains interact to form a substrate-binding cleft.

These 3-dimensional structures are usually presented in the form of simplified cartoons. This works particularly well on a computer screen with programs such as RASMOL (see Appendix). However, a visual comparison between more than two such 3-dimensional structures is unrewarding, partially because of the limitations of the visual system, partially also because of the dependency on the rotation of the protein structures that are to be compared.

The information present in the x, y, and z coordinates (which are the basis of the structure in space) can be presented as a rotation-independent 2-dimensional plot by using distances between the CAs of the amino acids rather than their actual position. Because secondary and tertiary structures depend on contacts between amino acids, the usual 2-dimensional representation of these distances emphasizes proximity rather than distance: the shorter the distance between 2 amino acids, the more emphasis. Hence the use of the term *proximity plot*.

For a protein of 100 amino acids, the 10,000 distances between the CAs are calculated and presented as a symmetrical square ( $100 \times 100$ ) matrix. For the diagonal, the distances are obviously zero. All the values next to the diagonal are also unrevealing because these are all 0.38 nm (see above). The values next off the diagonal are already more interesting because values of more than 0.65 nm indicate a  $\beta$ -strand, whereas values smaller than 0.55 nm indicate an  $\alpha$ -helix (or a turn). Cells in the matrix that are close to the diagonal thus provide information on the secondary structure. Low values (ie, small distances) in cells that are away from the diagonal indicate contact between elements of secondary structure; that is, they provide information on the tertiary structure (or in multichain structures on interactions between protein modules). Such a distance matrix contains all the information necessary to rebuild the 3-dimensional structure (apart from chirality because the mirror image of the protein will give the same distance matrix). For easy viewing, the distances relative to some cut-off values are color coded and plotted. In the examples presented here, cut-off values of 5.0 and 2.5 nm have been used.

One of the convenient aspects of these proximity plots is that they enable a quick, objective, and rotation-independent comparison between protein folds. In Fig 1 two examples are shown in more detail to illustrate the basic principles. Both consist of protein fragments containing two  $\beta$ -strands and one  $\alpha$ -helix connected in a simple up-and-down fashion. In the first example (the C terminus of birch profilin<sup>27</sup>) the sequence is  $\beta$ - $\beta$ - $\alpha$ , which implies that the  $\beta$ -strands run antiparallel. In the second example (the N-terminus of glutathione S-transferase<sup>28</sup>) the sequence is  $\beta$ - $\alpha$ - $\beta$ , and therefore the  $\beta$ -strands run in parallel.

Even if the structural interpretation of these proximity plots (Fig 2) requires some practice, one message can easily be derived from these plots by using them as fingerprints: the more different the plots, the more different the folds. Comparisons on the basis of the plots are liberal, and therefore proteins with similar plots may have significant differences in folding. However, if two proteins have different plots, they definitely have different folds.

## CLASSIFICATION OF PROTEIN FOLDS IN ALLERGENS

Technical details can be found in the Appendix. The table contains a selection of allergens (fully listed on the World Wide Web at [www.allergen.org/List.htm](http://www.allergen.org/List.htm)) for which the structure has been determined directly or for which the structure can be predicted with a highly variable degree of accuracy from the primary sequence by homology with proteins with a known structure. The literature references for these structures can be found in the Protein Data Bank (PDB) files.<sup>29</sup>

Whereas the basis for the classification in Table I<sup>30</sup> was provided by the SCOP database,<sup>19</sup> which is also an excellent source for finding PDB codes, the classification used was slightly adapted to combine allergens with similar features into a single class. The main issue was the group of proteins with one or two prominent antiparallel  $\beta$ -sheets in conjunction with one or two  $\alpha$ -helices. Depending on the size of the helix, SCOP classifies some allergens as  $\beta$  (lipocalin and aspartate protease) and others as  $\alpha+\beta$  (Bet v 1, profilin, and cystatin).

Some structures occur infrequently in the allergens investigated. Some strikingly unique allergens are tropomyosin (a long, filamentous, coiled, coil protein very different from the usual globular structures found for most allergens) and the allergens belonging to the pectate lyase family (eg, Amb e 1 and Cry j 1). The structure for this group of proteins is a long series of  $\beta$ -strands wound into a helix in which these  $\beta$ -strands form parallel sheets.

Many allergens are small, single-domain structures; some are so small that classification is problematic (Amb e 5 and hevein). However, complex multifold allergens do exist, particularly among the allergens that do not have to pass the airway mucosal barrier. Examples are the fungal proteins mitogillin, enolase, and MnSOD and venom phospholipase A.

## EPITOPES: LINEAR OR CONFORMATIONAL?

The part of the surface of the allergen that interacts with the Fab fragment of the antibody (ie, the epitope) is of particular interest.

The atomic details of the interaction between antibody and antigen are known for more than 30 antibody-antigen combinations. The interaction between two proteins is usually described in terms of buried surface (ie, the surface area that becomes inaccessible to water molecules as a consequence of the formation of the immune complex). This value is calculated from crystallographic data by using a spherical probe the size of a water molecule. The review by Padlan<sup>31</sup> gives a range of 54.0 to 89.0 nm<sup>2</sup> for the immune complexes with intact protein antigens (as opposed to peptides or haptens) that were available at that time. For a typical globular 20-kd allergen, the radius is approximately 2.0 nm, with a spherical surface area of 500.0 nm<sup>2</sup>. The water-accessible surface area varies depending on the compactness of the folded struc-

ture but will usually be approximately twice this value or 1000.0 nm<sup>2</sup> in this example. Thus the interaction with a Fab fragment of an antibody will bury approximately 5% to 10% of the surface of such a 20-kd allergen. The surface area that is accessible to a water molecule is obviously larger than the surface area accessible to a Fab fragment (and even more so for a complete antibody). It is therefore likely that a single allergen of this size cannot accommodate more than 5 to 10 antibodies at the same time.

Before crystallographic information was available, information on the structure of epitopes was derived, among others, from the reactivity of the antibody with peptides derived from the linear sequence of the allergen or with the unfolded allergen on an immunoblot. On the basis of this type of information, epitopes were classified as either linear or conformational. An antibody was assumed to react with a linear epitope if it was reactive with a small peptide (usually 8-15 amino acids) derived from the linear sequence or if it was reactive (in an immunoblot) after the allergen had been unfolded by boiling in SDS in the presence of a reducing agent. However, from the crystallographic studies, a general picture emerges that does not support the dichotomy of linear versus conformational epitopes: all protein epitopes are conformational. In all examples studied this buried surface area involves amino acids from different parts of the linear sequence; that is, all these epitopes are conformational to some extent. Fig 3 shows the epitope of lysozyme for mAb D1.3.<sup>32</sup>

How can this observation be reconciled with the observation that many antibodies react with small peptides derived from the linear sequence of the antigen? The main factor is the huge difference in affinity. On a molar basis, the peptides are very inefficient compared with the intact antigen, inhibiting the interaction between antigen and antibody. The peptide represents a fraction of the epitope, and the strength of the interaction with the antibody is a fraction of the strength with the complete antigen. The strength of interaction with the peptide is usually decreased even more because of the higher flexibility of the free peptide compared with the peptide in the complete antigen, resulting in a larger loss of entropy on binding. For the interpretation of the immunoblot data, it is relevant to appreciate that many blotted proteins refold extensively on removal of the SDS, particularly if reoxidation of any disulfides that might be present is allowed.

## CONCLUSIONS

The information on the atomic details of allergen structures indicates that allergens are heterogeneous, also from a structural point of view. Even if some folds are less prevalent among the currently known allergens, none of the protein folds seem to be incompatible with allergenicity. The current review is, however, biased toward allergens with a well-defined, stable structure.

What then determines allergenicity? Why are some allergens more major than others? The level and route of

exposure are obviously relevant. Similarity to human homologous proteins needs to be considered, but this relation is likely to be complex (depending on the accessibility of the human protein to the tolerizing mechanisms of the immune system). For food allergens, digestibility is clearly a factor to consider. For airborne allergens, size and solubility are important.

The high allergenicity of castor bean dust is presumably due to a negative adjuvant function of one of its components, the antiproteolytic activities of ricin.<sup>33</sup> The IgE-inducing potential of helminths is likely to reflect similar mechanisms. Inhaled proteases, particularly mite group 1, may act in the same way<sup>34-36</sup> or might enhance the local permeability of the airway mucosa.<sup>37-39</sup> Because most allergens do not possess protease activity, it would be important to investigate whether inhaled mite group 1 allergens have an IgE-promoting adjuvant effect. Because the prevalence of allergy is not strikingly lower in areas with low mite numbers, the case for an obligatory connection between enzymatic activity and allergenicity is as yet unconvincing.

We clearly need more information. However, a shift in focus may be needed. The main lack of information is not on major allergens but on their counterparts, the nonallergenic antigens. We need to identify and characterize antigens that are able to reach (and stimulate) immune cells and yet are not allergenic. The current data suggest the following working hypothesis: within the constraint mentioned, allergenicity depends on a single specific property to avoid activation of T<sub>H</sub>2-suppressive mechanisms.

I thank Ronald van Ree and Kaare Meno for helpful discussions. Ed Nieuwenhuys, Walter Langhorst, and Piet Schrama have been helping out whenever the computers did not do what I had in mind.

## REFERENCES

- Lehrer SB, Horner WE, Reese G. Why are some proteins allergenic? Implications for biotechnology. *Crit Rev Food Sci Nutr* 1996;36:553-64.
- Kimber I, Kerkvliet NI, Taylor SL, Astwood JD, Sarlo K, Dearman RJ. Toxicology of protein allergenicity: prediction and characterization. *Toxicol Sci* 1999;48:157-62.
- Metcalfe DD, Astwood JD, Townsend R, Sampson HA, Taylor SL, Fuchs RL. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit Rev Food Sci Nutr* 1996;36(Suppl):S165-86.
- Astwood JD, Fuchs RL. Allergenicity of foods derived from transgenic plants. *Monogr Allergy* 1996;32:105-20.
- Van Loon LC, Van Strien EA. The families of pathogenesis-related protein, their activities, and comparative analysis of PR-1 types proteins. *Physiol Mol Plant Pathol* 1999;55:85-97.
- Baldo BA. Penicillins and cephalosporins as allergens—structural aspects of recognition and cross-reactions. *Clin Exp Allergy* 1999;29:744-9.
- Aalberse RC, Van Ree R. Crossreactive carbohydrate determinants. *Clin Rev Allergy Immunol* 1997;15:375-87.
- Astwood JD, Leach JN, Fuchs RL. Stability of food allergens to digestion in vitro. *Nat Biotechnol* 1996;14:1269-73.
- Seppala U, Alenius H, Turjanmaa K, Reunala T, Palosuo T, Kalkkinen N. Identification of patatin as a novel allergen for children with positive skin prick test responses to raw potato. *J Allergy Clin Immunol* 1999;103:165-71.
- Kostyal DA, Hickey VL, Noti JD, Sussman GL, Beezhold DH. Cloning and characterization of a latex allergen (Hev b 7): homology to patatin, a plant PLA2. *Clin Exp Immunol* 1998;112:355-62.
- Breiteneder H, Sowka S, Wagner S, Krebitz M, Hafner C, Kinaciyan T, et al. Cloning of the patatin-like latex allergen Hev b 7, its expression in the yeast *Pichia pastoris* and its immunological characterization. *Int Arch Allergy Immunol* 1999;118:309-10.
- Teuber SS, Dandekar AM, Peterson WR, Sellers CL. Cloning and sequencing of a gene encoding a 2S albumin seed storage protein precursor from English walnut (*Juglans regia*), a major food allergen. *J Allergy Clin Immunol* 1998;101:807-14.
- Virtanen T, Zeiler T, Mantylajarvi R. Important animal allergens are lipocalin proteins: why are they allergenic? *Int Arch Allergy Immunol* 1999;120:247-58.
- Valenta R, Duchene M, Pettenburger K, Sillaber C, Valent P, Bettelheim P, et al. Identification of profilin as a novel pollen allergen; IgE autoreactivity in sensitized individuals. *Science* 1991;253:557-60.
- Mayer C, Appenzeller U, Seelbach H, Achatz G, Oberkofler H, Breitenbach M, et al. Humoral and cell-mediated autoimmune reactions to human acidic ribosomal P2 protein in individuals sensitized to *Aspergillus fumigatus* P2 protein. *J Exp Med* 1999;189:1507-12.
- Cramer R, Faith A, Hemmann S, Jaussi R, Ismail C, Menz G, et al. Humoral and cell-mediated autoimmunity in allergy to *Aspergillus fumigatus*. *J Exp Med* 1996;184:265-70.
- Branden C, Tooze J. Introduction to protein structure. 2nd ed. New York: Garland Publishing; 1999.
- Holm L, Sander C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res* 1999;27:244-7.
- Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. *Nucleic Acids Res* 1999;27:254-6.
- Aalberse RC, Van der Zee JS, Vlug A. IgG4 antibodies in atopic allergy. *Lab Management* 1985;23:19-32.
- Pietrokowski S, Henikoff JG, Henikoff S. The Blocks database—a system for protein classification. *Nucleic Acids Res* 1996;24:197-200.
- Corpet F, Gouzy J, Kahn D. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res* 1999;27:263-7.
- Holm L, Sander C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res* 1999;27:244-7.
- Holm L, Sander C. Dictionary of recurrent domains in protein structures. *Proteins* 1998;33:88-96.
- Seiberler S, Natter S, Hufnagl P, Binder BR, Valenta R. Characterization of IgE-reactive autoantigens in atopic dermatitis. 2. A pilot study on IgE versus IgG subclass response and seasonal variation of IgE autoreactivity. *Int Arch Allergy Immunol* 1999;120:117-25.
- Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993;233:123-38.
- Fedorov AA, Ball T, Mahoney NM, Valenta R, Almo SC. The molecular basis for allergen cross-reactivity: crystal structure and IgE-epitope mapping of birch pollen profilin. *Structure* 1997;5:33-45.
- McTigue MA, Bernstein SL, Williams DR, Tainer JA. Purification and crystallization of a schistosomal glutathione S-transferase. *Proteins* 1995;22:55-7.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235-42.
- Chapman MD, Ichikawa K, Vailes LD. Molecular cloning of a new cat allergen, cystatin, a cysteine protease inhibitor [abstract]. *J Allergy Clin Immunol* 1999;103:S122.
- Padlan EA. X-ray crystallography of antibodies. *Adv Protein Chem* 1996;49:57-133.
- Fischmann TO, Bentley GA, Bhat TN, Boulot G, Mariuzza RA, Phillips SE, et al. Crystallographic refinement of the three-dimensional structure of the FabD1.3-lysozyme complex at 2.5-A resolution. *J Biol Chem* 1991;266:12915-20.
- Underwood SL, Kemeny DM, Lee TH, Raeburn D, Karlsson JA. IgE production, antigen-induced airway inflammation and airway hyperreactivity in the brown Norway rat: the effects of ricin. *Immunology* 1995;85:256-61.
- Shakib F, Schulz O, Sewell H. A mite subversive: cleavage of CD23 and CD25 by Der p 1 enhances allergenicity. *Immunol Today* 1998;19:313-6.
- Schulz O, Sewell HF, Shakib F. The interaction between the dust mite antigen Der p 1 and cell-signalling molecules in amplifying allergic disease. *Clin Exp Allergy* 1999;29:439-44.
- Gough L, Schulz O, Sewell HF, Shakib F. The cysteine protease activity of the major dust mite allergen Der p 1 selectively enhances the immunoglobulin E antibody response. *J Exp Med* 1999;190:1897-902.
- Robinson C, Kalsheker NA, Srinivasan N, King CM, Garrod DR, Thompson PJ, et al. On the potential significance of the enzymatic activ-

- ity of mite allergens to immunogenicity. Clues to structure and function revealed by molecular characterization. *Clin Exp Allergy* 1997;27:10-21.
38. Stewart GA. Dust mite allergens. *Clin Rev Allergy Immunol* 1995;13:135-50.
39. Wan H, Winton HL, Soeller C, Tovey ER, Gruenert DC, Thompson PJ, et al. Der p 1 facilitates transepithelial allergen delivery by disruption of tight junctions [see comments]. *J Clin Invest* 1999;104:123-33.
40. Smith SG, Lewis M, Aschaffenburg R, Fenna RE, Wilson IA, Sundaralingam M, et al. Crystallographic analysis of the three-dimensional structure of baboon alpha-lactalbumin at low resolution. Homology with lysozyme. *Biochem J* 1987;242:353-60.

## APPENDIX

### Technical considerations

The 4-character PDB code of the protein structure gives access to the file that contains the coordinates of most of the atoms in the structure. For example, one of the files for the birch allergen Bet v 1 has the PDB code 1BV1. The complete filename is

"PDB1BV1.ENT". The PDB<sup>29</sup> can be freely accessed through the Internet via SearchLite at <http://www.rcsb.org/pdb/searchlite.html>. Coordinate files, such as PDB1BV1.ENT, can be downloaded and opened in either a text viewer by using a text editor, such as WORDPAD, or a word processor. They can also be opened in a molecular viewer, such as RASMOL, which is also freely available through the Internet at <http://www.umass.edu/microbio/rasmol/>. The PDB file contains not only the x, y, and z coordinates for each atom in Ångströms (1 Ångstrom = 0.1 nm), but also background information, including the authors and literature references.

SCOP is accessible through <http://scop.mrc-lmb.cam.ac.uk/scop/>.

The PDB files contain the information required to produce the proximity plots for which an option is available in several molecular modeling packages. The plots in this review have been produced by using PROXIMA.EXE, a small (51 kB) DOS-based program, available for noncommercial use from the author by means of e-mail ([aalberse@clb.nl](mailto:aalberse@clb.nl)).