Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: An *in silico* analysis

John A. Jenkins, PhD,[a] Sam Griffiths-Jones, PhD,[b] Peter R. Shewry, PhD,[c] Heimo Breiteneder, PhD,[d] and E.N. Clare Mills, PhD[a] *Norwich, Cambridge, and Harpenden, United Kingdom, and Vienna, Austria*

Background: The body of sequence and structural information on allergens and the sequence analysis of whole plant genomes are facilitating the application of bioinformatic approaches to identifying and defining plant allergens.
Objective: An *in silico* approach was used to quantify the distribution of plant food allergen sequences across protein families and to develop and apply a novel means of assessing conserved surface features important for IgE cross-reactivity.
Methods: Plant food allergen sequences were classified into Pfam families on the basis of sequence homology. Contact surface areas of selected proteins were calculated with MOLMOL by using a 1.4-Å probe, corrected by removing contributions from IgE inaccessible main chains and side chains forming the ligand binding sites.
Results: A set of 129 food allergen sequences were classified into only 20 of 3849 possible Pfam families, with 4 families accounting for more than 65% of food allergens. Structural bioinformatic analysis of conserved exterior main chains and amino acid side chains in cross-reactive homologues of Bet v 1 and nonspecific lipid transfer proteins showed higher levels of similarity than shown by simple sequence comparisons. Thus, 75% of the Mal d 1 surface is likely to bind anti–Bet v 1 antibodies, compared with a sequence identity of ~56%.
Conclusion: Most plant food allergens belong to only 4 structural families, indicating that conserved structures and biological activities may play a role in determining or promoting allergenic properties. Structural bioinformatic analysis shows that conservation of 3-dimensional structure should be included in any assessment of potential IgE cross-reactivity in, for example, novel proteins. (J Allergy Clin Immunol 2005;115:163-70.)

Key words: *Allergen bioinformatics, distribution of plant food allergens across protein families, plant food allergen structures, Bet v 1 family cross-reactivity, nsLTP family cross-reactivity*

One of the major challenges of molecular allergology is to pinpoint and evaluate characteristics of a protein that confer its allergenic properties when it comes into contact with an atopic immune system. It is clear that in some cases, the biochemical properties of the protein itself are important, such as the stability and proteolytic activity of the inhalant allergen Der p 1 from house dust mite.[1] However, our current knowledge of food allergens is much less complete. For example, we do not know why peanuts pose so much more of an allergenic risk than peas, despite their close botanical relationship. These gaps in understanding prevent us from giving rational advice on diverse issues, including the allergenic potential of novel foods and processes such as genetically modified organisms.

Both the ability of a substance to sensitize a naive individual and to elicit an allergic response in individuals who are already sensitized to a given protein are encompassed in its allergenicity. Although the molecular basis for these effects is still not completely understood, it is becoming evident that the level of exposure and the properties of the allergen itself play important roles in determining the allergenic potential.[2]

The last 10 years have seen an explosion in the identification and sequencing of food allergens. The sequences of more than 100 different food allergens have now been published, corresponding to 70% to 80% of the identified food allergens. When isoallergens are included, the number of reported sequences rises to more than 200 (www.ifr.ac.uk/protall, www.allergen.org, www.allergome.org, www.allergenonline.com). Bioinformatic tools are available to classify proteins into families on the basis of their shared amino acid sequences and conserved 3-dimensional structures. These form the basis of several protein family databases, including Pfam.[3] Previous analyses of plant food allergens have indicated that most belong to only a small number of protein superfamilies,[4] contributing to the development of a more molecular approach to the classification of plant food allergens.[5] These studies have facilitated the application of *in silico* approaches to predict the allergenic potential of proteins directly from their amino acid sequence. However, Hileman et al[6] reported that searches using a 6–amino acid window wrongly predicted that 41 of 50 corn proteins

From [a]the Institute of Food Research, Norwich; the [b]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge; [c]Rothamsted Research, Harpenden; and [d]Department of Pathophysiology, Medical University of Vienna.

Partly supported by EU through FAIR CT-98-4356 (PROTALL), QLK1-CT-2000-01394 (SAFE), and BBSRC through the competitive strategic grant to the Institute of Food Research and Rothamsted Research. Dr Griffiths-Jones is supported by the Wellcome Trust. Dr Breiteneder is supported by Austrian Science Fund grant SFB F01802.

Received for publication August 31, 2004; revised October 6, 2004; accepted for publication October 7, 2004.

Reprint requests: Heimo Breiteneder, PhD, Department of Pathophysiology, Medical University of Vienna, AKH-EBO-3Q, Waehringer Guertel 18-20, 1090 Vienna, Austria. E-mail: Heimo.Breiteneder@meduniwien.ac.at.

0091-6749/$30.00

doi:10.1016/j.jaci.2004.10.026

Abbreviations used
nsLTP: Nonspecific lipid transfer protein
PLAT: Polycystin-1-lipoxygenase-α-toxin

were potential allergens. In contrast, the FASTA algorithm identified only 9 corn proteins which shared >35% identity over 80 amino acids with known allergens, with only 6 them similar over their entire length. Soeria-Atmadja et al[7] and Stadler and Stadler[8] have used local alignment and MEME, a motif discovery tool, in automated learning approaches to achieve good predictions and relatively low rates of false-positive identifications of allergens.

An alternative approach to allergen classification may be developed from the observation that most plant food allergens belong to only a small number of protein superfamilies.[4,5] We have therefore exploited profile-based sequence homology methods[3] to classify plant food allergens into families and then used a structural bioinformatic approach to analyzing conserved surface features that are thought to be important determinants of IgE cross-reactivity at the molecular level.

METHODS

Plant food allergen databases

Lists of known plant food allergens were obtained from the PROTALL (www.ifr.ac.uk/protall) and FARRP (www.allergenonline.com) databases. The version of FARRP used defined proteins from plant foods as food allergens even when they had been found to act only as aeroallergens, such as Gly m 1. The PROTALL set was limited to proteins shown to bind serum IgE from at least 3 patients with clinical allergy to the food from which the allergen originated. Removing duplications between the 2 databases left a protein set composed of 136 entries mapped to SWISS-PROT and TrEMBL entries.[9] A separate set of known pollen allergens was obtained from FARRP and mapped to 152 nonredundant SWISS-PROT and TrEMBL entries.

Protein family assignment in plant genomes

Arabidopsis thaliana. The *Arabidopsis* protein set was downloaded from the Proteome Web site (www.ebi.ac.uk/proteome).[10] The 2 allergen protein sets and the *Arabidopsis* protein set were searched against the Pfam 7.3 collection of hidden Markov model profiles characteristic of a family[3] (www.sanger.ac.uk/Software/Pfam/) by using the HMMER software package, version 2.2g (hmmer.wustl.edu/). The set of Pfam thresholds chosen using the lowest score for sequences known to belong to the family and the highest score for sequences not belonging to the family were used to give protein family classifications. The Pfam database classifies plant protein sequences into families on the basis of sequence homology, which is related to conserved 3-dimensional structures and possible function. Pfam 7.3 contains 3849 families, which include more than 70% of all plant protein sequences reported in the SWISS-PROT and TrEMBL databases. Where multiple domains are found within a single protein, these are counted only once, but nonidentical domains contribute to the scores of both families.

Oryza sativa. Currently there is no public domain protein set available for rice analogous to that available for *Arabidopsis*. Consequently, the assembled DNA sequence of *O sativa*[11] was obtained from Genbank (accession AAAA00000000) and searched against Pfam 7.3 by using the WISE2 software package version 2.1.22c (www.sanger.ac.uk/Software/Wise2/). The GENEWISEDB program was used to search for high-scoring (≥30 bits) full-length matches (global search mode) by using the alternate splice model (GT/AG), as suitable for plant genomes. Families were ranked only for *A thaliana* because the rice genome could be searched only with existing families that we had, rather than with the whole protein family universe. Each set is available together with the results of our analysis as supplementary information in the Journal's Online Repository (and also available online at www.sanger.ac.uk/Users/sgj/JACI_suppl).

Analysis of allergen surfaces

The 3-dimensional structures and contact surface areas of proteins were calculated with MOLMOL[12] by using a 1.4-Å radius probe. Surface areas were assigned to the nearest atom, and main chains and side chains were manually checked both for sequence conservation and to exclude main chains and side chains that contributed only to the surface of the IgE inaccessible lipid binding cavities (residues with no exposure to a probe of 3.5-Å radius included all IgE inaccessible residues in nonspecific lipid transfer proteins (nsLTPs) and Bet v 1a but also several in shallow depressions on the exterior surface). Sequences were aligned with T-Coffee[13] for Bet v 1 homologues, and the insertion in corn nsLTP was positioned by comparing the wheat and barley structures.

RESULTS

Bioinformatic analysis of allergen sequences

The Pfam database was used to assign 136 plant food allergen sequences compiled by using the FARRP and PROTALL databases. Most of these sequences are of clinically proven food allergens, but because inclusion criteria differ slightly between the databases, some sequences of proteins that are not proven food allergens were included in the data set. All but 7 sequences could be unambiguously assigned to Pfam families, with the 7 unmatched sequences mainly corresponding to short sequenced fragments of proteins. It is striking that the 129 matched sequences fall into only 20 of a possible 3849 Pfam families, with only 4 of these families accounting for more than 65% of the sequences (Fig 1, *A, black bars*). These 4 families are the cereal prolamin superfamily (composed of cereal storage proteins [*eg,* the ω-5 gliadin allergen of wheat], nonspecific lipid transfer proteins [such as Pru p3 from peach and Zea m 14 from corn], 2S storage albumins [*eg,* Ber e 1 from Brazil nut] and inhibitors of trypsin and α-amylase); the cupins (including the 7S and 11S globulin storage proteins of seeds, to which many allergens belong, notably Ara h 1, 3, and 4 from peanut); homologues of the major birch pollen allergen, Bet v 1 (such as Mal d 1 in apple and Api g 1 in celery); and profilins (such as Api g 4 from celery).

Allergens from the same species with greater than 67% sequence identity have been defined by the Allergen Nomenclature Committee as isoallergens.[14] To ensure that the distribution of sequences was not biased by the inclusion of isoallergens (as is the case, for example, for