5

Plant protein families as a basis for predicting the allergenicity of food proteins

Peter R. Shewry[#], *John Jenkins*[##], *Frederic Beaudoin*[#] *and E.N. Clare Mills*[##]

Abstract

Plant proteins can be classified into families and superfamilies based on their sequence relationships. Comparison of these families with accepted lists of plant food-protein allergens demonstrates that only a small number of families contain characterized allergens with most of those that sensitize via the gastrointestinal tract falling into only two large superfamilies. These observations are discussed in relation to the use of plant protein family membership to predict the potential allergenicity of novel and GM foods.
Keywords: plant proteins; food allergens; protein families; prediction of allergenicity

Introduction

Plant tissues comprise vast numbers of proteins, although the total number present at any precise time can only be estimated. A theoretical maximum number can be based on the numbers of genes predicted from the genome sequences determined for two plant species, *Arabidopsis thaliana* and rice (*Oryza sativa*). *Arabidopsis* is a 'model' species which was selected for genome sequencing due to its unusually small genome size (approx 115 x 10^6 base pairs of DNA/cell compared with 820 x 10^6 bp in oilseed rape, 2.5 x 10^9 bp in maize and 16 x 10^{12} bp in wheat). However, the genome sequence is of direct relevance to major crops as *Arabidopsis* is a brassica closely related to cultivated species such as cabbage (*Brassica oleracea*), radish (*Raphanus sativus*), mustards (*Sinapis alba, Brassica juncea*) and oilseed rape (also called canola) (*Brassica napus)*. Rice has a larger genome size than *Arabidopsis* (430 x 10^6 bp), although this is still smaller than those of most major crops (as listed above). It was selected because it is a major crop, with particular relevance to developing countries, but the sequence is still not fully confirmed or annotated. It must also be borne in mind that the statistical algorithms used to predict genes that encode proteins (called open reading frames or OFRs), are far from foolproof, being of little value for small proteins (below about 100 amino acids) and failing to discriminate between genes that are expressed and those that are silent (often called pseudogenes). Nevertheless, the analyses of these two genome sequences indicate the presence of about 25,500 and 50,000 genes in *Arabidopsis* and rice, respectively.

[#] Rothamsted Research, Harpenden, Hertfordshire, UK
[##] Institute of Food Research, Norwich, UK

It is clear that many of the proteins encoded by these genes are involved in processes that require only transient expression (e.g. in signalling pathways) while others are present at levels that are too low to be identified using conventional proteomic approaches. Even with these caveats, it is surprising that so few plant proteins appear to be capable of eliciting an allergic response, whether by ingestion, inhalation or contact. Furthermore, the majority of the proteins that are allergenic can be classified on the basis of their amino-acid sequences into a small number of clearly defined groups or families.

These observations raise two important questions: what is it about the sequences, structures or biological properties of some families of proteins that predisposes them to become allergens, and can this information be used to predict the allergenicity of novel proteins?

Classification of plant proteins

Protein classifications have been proposed for almost as long as proteins have been studied, with the criteria used reflecting the level of knowledge that was available at the time. The first systematic attempt to impose a classification on a wide range of plant proteins was that of TB Osborne (see Osborne 1924) who developed a system based on the sequential extraction of proteins in water, dilute saline, alcohol–water mixtures and dilute acids or alkalis. Although this classification is today only widely applied to seed proteins, where the groups still have a high level of biological and functional significance (i.e. in food processing), two of the names that he used are still widely used for other proteins: albumins (soluble in water) and globulins (soluble in dilute saline).

Our increasing knowledge of protein structure and properties has allowed more systematic and scientifically valid classifications to be made (e.g. the EC nomenclature for enzymes) culminating in the recent availability of extensive amino-acid sequences coupled with the more limited availability of 3D structures. This has allowed the classification of proteins into families based on their amino-acid sequences with the presence of conserved sequence motifs and 3D structures allowing more divergent families to be grouped together into superfamilies. The most extensive classification of this type to be applied to plant proteins is the Pfam database compiled at the Wellcome Trust Sanger Institute, Cambridge, UK (http://www.sanger.ac.uk/software/pfam/) (Bateman et al. 2002).

Families of plant protein allergens

The databases Farrp (http://www.allergenonline.com/) and Protall (http://www.ifr.bbsrc.ac.uk/protall/) currently contain over 130 unique protein sequences which are defined as plant food allergens. Of these, over 60% can be assigned to just four of the Pfam families (which currently total almost 4,000). Two of these families comprise proteins that sensitize via inhalation and are outside the scope of this article. The other two families sensitize via the gastrointenstinal tract and are the ‘prolamin superfamily’ and the ‘cupins’.

The prolamin superfamily

The prolamin superfamily was first identified by Kreis and co-authors (Kreis et al. 1985) based on the presence of a conserved cysteine skeleton which is characteristic of the major grain storage proteins (prolamins) of wheat and related cereals. This

cysteine skeleton (C-Xn-C-Xn-CC-Xn-CXC-Xn-C-Xn-C) has since been found in a range of small (below *M*r 15000) sulphur-rich proteins, most of which are restricted to seeds and some of which are allergens (Figure 1). The allergenic components include

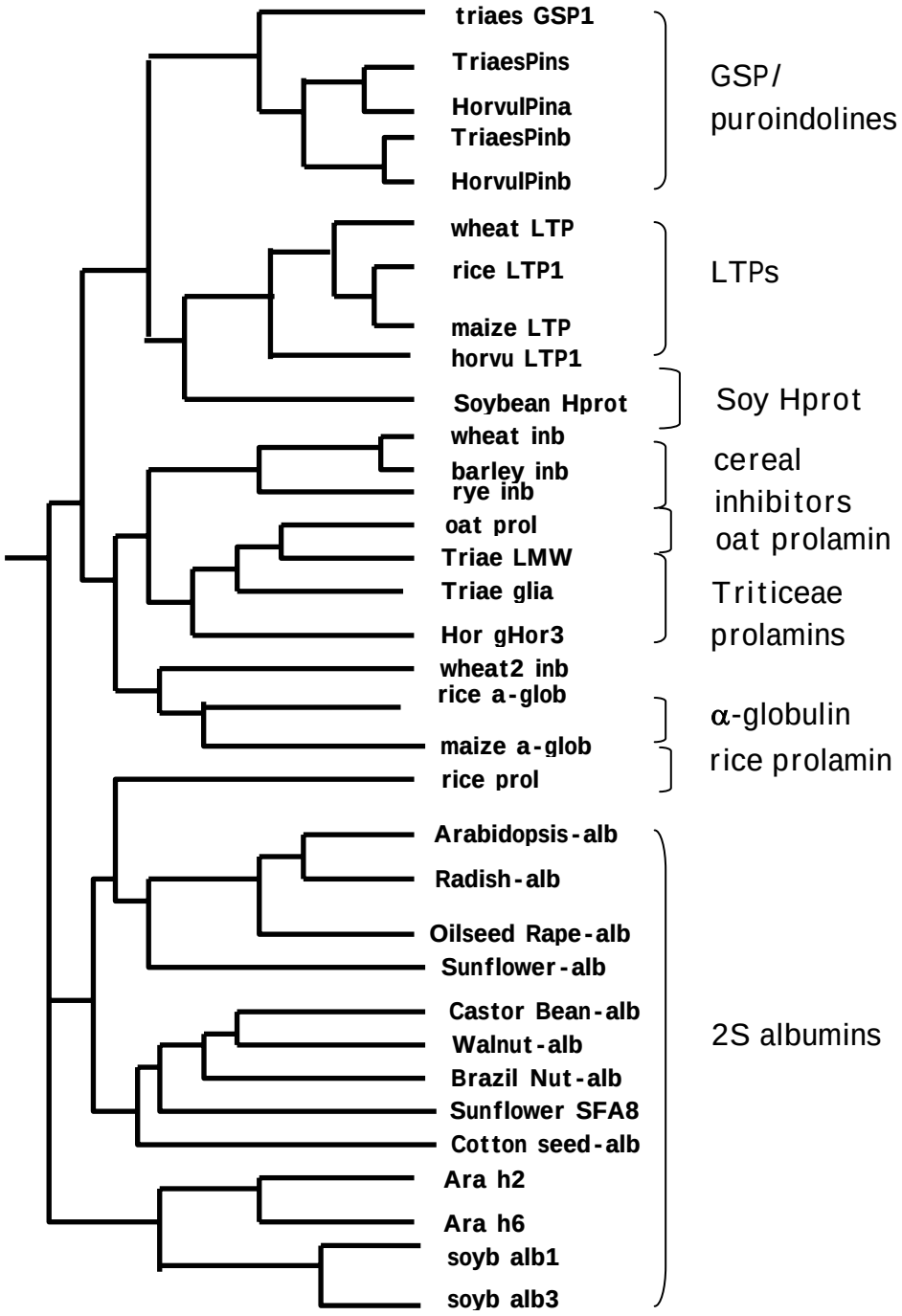

Figure 1. Dendrogram showing the relationships between the amino-acid sequences of selected members of the prolamin superfamily

members of the non-specific lipid-transfer protein family (in fruit of peach, grape, plum, pear, apricot, apple, cherry and seed of chestnut and maize), the 2S albumin storage proteins (in seeds of Brazil nut, castor bean, cotton, peanut, walnut), the cereal α-amylase/trypsin inhibitors (inhalant allergens in wheat, barley and rye flours, dietary allergens in rice seed) and the soybean hydrophobic protein (inhalant allergen

from seed hulls). These individual proteins have little sequence homology apart from the cysteine skeleton but do have highly similar structures. Thus all comprise four or five α-helices which are arranged in a right-handed superhelix (Figure 2), the whole structure being stabilized by four or five intra-chain disulphide bonds. This structure

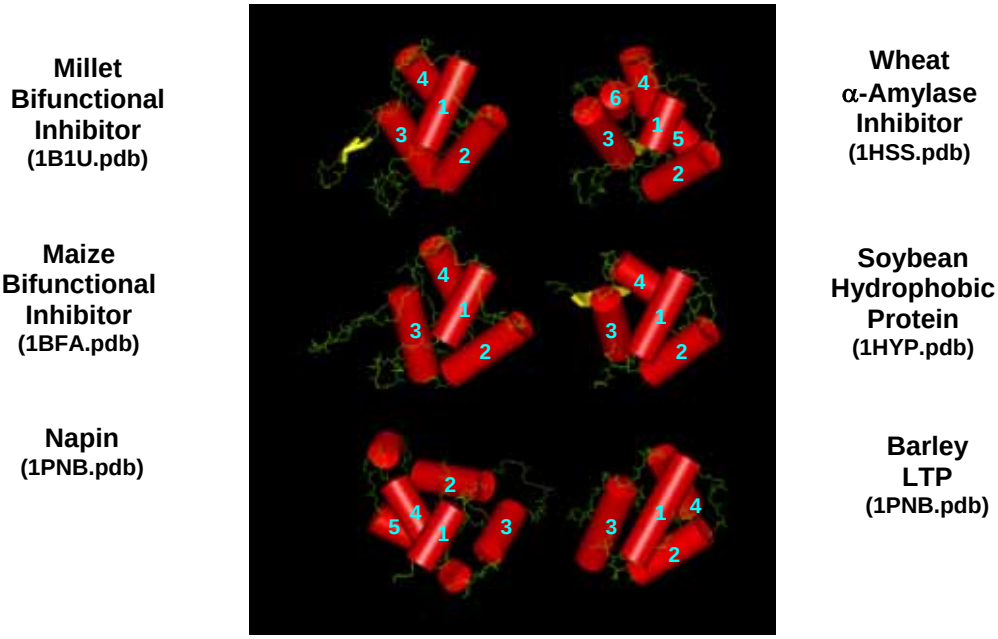

Figure 2. Schematic three-dimensional structures of small S-rich proteins of the prolamin superfamily, with the α-helices shown as cylinders

is extremely stable to both thermal denaturation and enzymic cleavage, characteristics which may contribute to high stability during digestion in the stomach. Less is understood about the identities of the epitopes that react with IgE, but in some cases these appear to be present in variable loop regions between the α-helices.

The cupins

The second major superfamily of plant food allergens, the cupins, differs from the prolamins in several respects. Firstly, they are widely distributed in *Archaea*, *Eubacteria* and *Eukaryota*, the latter including plants and animals. They vary widely in sequence but are all characterized by two short consensus sequence motifs and a core structural feature, a barrel-like double-stranded β-helix or jellyroll. Hence the name cupins, which is based on the Latin *cupa* which means small barrel (Dunwell 1998).

Despite their wide distribution allergenic cupins appear to be restricted to two groups of seed storage proteins called 7S and 11S globulins. These proteins are trimers (7S) or hexamers (11S) of subunits with masses averaging about 50,000 to 70,000, meaning that the protein masses range from about 150,000 to 450,000. The subunits of the 11S globulins are post-translationally processed to give acidic and basic chains associated by a single disulphide bond, but apart from this the protein structures are stabilized solely by non-covalent forces. Allergenic proteins are present in soybean, peanut, cashew (7S and 11S in all species), lentil and walnut (both 7S), although other allergenic forms may also occur. Multiple epitopes appear to be present on the 7S and 11S allergens of peanut (Ara h 1 and Ara h 3, respectively).

Furthermore, several of the Ara h 3 epitopes also occur in the allergenic glycinin G 1 (11S globulin) of soybean (Beardslee et al. 2000).

These allergenic globulins are generally less stable to thermal denaturation and enzymic digestion than the small, highly disulphide-bonded proteins of the prolamin superfamily, but it is possible that stable intermediate forms persist in the small intestine. Interactions with other food components may also increase their stability to proteolysis and absorbance by the small intestine (as discussed by Mills, Jenkins and Shewry 2003).

Although a wide range of legumes are consumed allergenic globulin storage proteins are restricted to a few species with other species which are widely consumed (e.g. peas and beans) rarely if ever leading to allergy. Similarly, globulin storage proteins are widely distributed and highly abundant in seeds of many other plants and are consumed in vast quantities in raw and processed forms. Nevertheless reports of allergenic globulins in species other than legumes are rare although the same species may contain allergenic 2S albumins (e.g. in Brazil nut, castor bean, mustards). The globulins are clearly not an intrinsically allergenic family to the same extent as the 2S albumins and LTPs.

Other families of plant food allergens

When double counting is eliminated (i.e. only one related sequence per species is included) only the prolamin superfamily and cupins contain more than five characterized allergens, with some Pfam families containing only single characterized allergenic proteins (for example, the storage proteins of maize seeds (zein) and potato tubers (patatin)). However, one family is of interest as it contains closely related proteins from a diverse range of foods: there are the cysteine proteinases from kiwi fruit (actinidin Act c 1), pineapple (bromelain), papaya (papain), fig (ficin) and soybean (Gly m Bd 30k). These proteins have similar three-dimensional structures and are all synthesized as precursors with long prodomains which block the active site until removed proteolytically (see Shewry et al. 2003).

Conclusions: use of protein families for prediction of allergenicity

There is currently a high level of interest in predicting the allergenicity of novel foods and food proteins, which is partly related to the regulation of modified foods produced using genetic engineering. Current recommendations include comparison of the amino-acid sequence of the novel protein with those of known allergens, with a homology level of 35% being quoted by FAO-WHO (2001; 2002) as significant.

Analysis of the prolamin superfamily demonstrates that allergenicity is exhibited by proteins that have little sequence homology beyond the conserved cysteine skeleton although their three-dimensional structures are highly conserved. Hence it is probably valid to consider all small, sulphur-rich members of this family (i.e. not the major prolamin storage proteins) as potential allergens and therefore to avoid using them in transgenic plants. In contrast, in the cupins allergenicity appears to be essentially restricted to the 7S and 11S storage globulins, and to only some components within these groups. The reason for this is not known but it could relate to the extent of modification during processing as well as to the volume of consumption and the properties of the proteins themselves. In view of the vast amounts of legumes that are consumed throughout the world it is clearly not realistic to treat them all as potentially allergenic and a case-by-case approach is more logical.

Beyond these two superfamilies only the cysteine proteinases are consistently allergenic in a range of species and hence should again be avoided.

References

Bateman, A., Birney, E., Cerruti, L., et al., 2002. The Pfam protein families database. *Nucleic Acids Research,* 30 (1), 276-280.

Beardslee, T.A., Zeece, M.G., Sarath, G., et al., 2000. Soybean glycinin G1 acidic chain shares IgE epitopes with peanut allergen Ara h 3. *International Archives of Allergy and Immunology,* 123 (4), 299-307.

Dunwell, J.M., 1998. Cupins: a new superfamily of functionally diverse proteins that include germins and plant storage proteins. *Biotechnology & Genetic Engineering Reviews,* 15, 1-32.

FAO and WHO, 2001. *Evaluation of allergenicity of genetically modified foods: report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, 22–25 January 2001*. FAO, Rome. [http://www.who.int/foodsafety/publications/biotech/en/ec_jan2001.pdf]

FAO and WHO, 2002. *Report of the third session of the Codex Ad Hoc Intergovernmental Task Force on Foods derived from Biotechnology, Yokohama, Japan, 4-8 March 2002*. FAO, Rome.

Kreis, M., Forde, B.G., Rahman, S., et al., 1985. Molecular evolution of the seed storage proteins of barley, rye and wheat. *Journal of Molecular Biology,* 183 (3), 499-502.

Mills, E.N.C., Jenkins, J.A. and Shewry, P.R., 2003. The role of common properties in determining plant protein allergenicity. *In:* Mills, E.N.C. and Shewry, P.R. eds. *Plant food allergens*. Blackwell, Oxford, 158-170.

Osborne, T.B., 1924. *The vegetable proteins*. 2nd edn. Longmans Green, London. Monographs on Biochemistry.

Shewry, P.R., Jenkins, J., Beaudoin, F., et al., 2003. The classification, functions and evolutionary relationships of plant proteins in relation to food allergies. *In:* Mills, E.N.C. and Shewry, P.R. eds. *Plant food allergens*. Blackwell, Oxford, 24-41.