Allergens are distributed into few protein families and possess a restricted number of biochemical functions

Christian Radauer, PhD,[a] Merima Bublin, PhD,[a] Stefan Wagner, PhD,[a] Adriano Mari, MD,[b,c] and Heimo Breiteneder, PhD[a] *Vienna, Austria, and Latina and Rome, Italy*

Background: Existing allergen databases classify their entries by source and route of exposure, thus lacking an evolutionary, structural, and functional classification of allergens.

Objective: We sought to build AllFam, a database of allergen families, and use it to extract common structural and functional properties of allergens.

Methods: Allergen data from the Allergome database and protein family definitions from the Pfam database were merged into AllFam, a database that is freely accessible on the Internet at http://www.meduniwien.ac.at/allergens/allfam/. A structural classification of allergens was established by matching Pfam families with families from the Structural Classification of Proteins database. Biochemical functions of allergens were extracted from the Gene Ontology Annotation database.

Results: Seven hundred seven allergens were classified by sequence into 134 AllFam families containing 184 Pfam domains (2% of 9318 Pfam families). A random set of 707 sequences with the same taxonomic distribution contained a significantly higher number of different Pfam domains (479 ± 17). Classifying allergens by structure revealed that 5% of 3012 Structural Classification of Proteins families contained allergens. The biochemical functions of allergens most frequently found were limited to hydrolysis of proteins, polysaccharides, and lipids; binding of metal ions and lipids; storage; and cytoskeleton association.

Conclusion: The small number of protein families that contain allergens and the narrow functional distribution of most allergens confirm the existence of yet unknown factors that render proteins allergenic. (J Allergy Clin Immunol 2008;121:847-52.)

Key words: *Allergens, protein families, allergen structures, allergen databases*

Abbreviations used

GO:	Gene ontology
nsLTP:	Nonspecific lipid transfer protein
SCOP:	Structural Classification of Proteins
TIM:	Triosephosphate isomerase
UniProt:	Universal Protein Resource

Since the identification and cloning of the first allergenic proteins in the late 1980s, hundreds of allergens have been identified and their sequences determined. A number of databases that provide molecular, biochemical, and clinical data of allergens were established, such as the Official List of Allergens issued by the International Union of Immunological Societies Allergen Nomenclature Sub-committee (http://www.allergen.org),[1] the Allergome (http://www.allergome.org),[2] the Food Allergy Research and Resource Program Allergen Database (http://www.allergenonline.com), and the InFormAll database (http://foodallergens.ifr.ac.uk/).

The growing number of available allergen sequences together with the advancements of bioinformatics tools and methods enabled scientists to shed light on evolutionary and structural relationships between allergens from different sources.[3] In particular, protein family databases that are linked to protein sequence databases, such as the Pfam database,[4] provided the basis of a novel classification of allergens. Several studies revealed that most allergens can be found in a limited number of protein families.[5-10]

Records of most allergen databases are organized by type of allergen source and route of exposure. Likewise, allergen designations according to the official allergen nomenclature are derived from the scientific name of the allergen source species and a sequential number that in most cases does not reflect evolutionary relationships between allergens. To bring together allergen data stored in allergen databases and evolutionary and structural relationships between allergens established from protein family databases, we constructed AllFam, a database of allergen families. In the present study we used data extracted from AllFam to establish the protein family distribution of allergens and to elucidate common structural and biochemical features of allergens, thus shifting the focus from single allergens or allergen families to a systematic analysis of the complete range of known allergens.

METHODS

Construction of the AllFam database

Data of allergens with known sequences (name, source, routes of exposure, and Universal Protein Resource [UniProt] accession numbers) were downloaded from Allergome,[2] a database based on allergen data

From [a]the Department of Pathophysiology, Center for Physiology, Pathophysiology and Immunology, Medical University of Vienna; [b]the Allergy Data Laboratories, Latina; and [c]the Center for Clinical and Experimental Allergology, IDI-IRCCS, Rome.

Supported by grant SFB-F01802 from the Austrian Science Fund (to H.B.) and an Austrian Programme for Advanced Research and Technology grant from the Austrian Academy of Sciences (to S.W.).

Disclosure of potential conflict of interest: A. Mari is the responsible administrative contact for Allergy Data Laboratories. The rest of the authors have declared that they have no conflict of interest.

Received for publication October 29, 2007; revised January 22, 2008; accepted for publication January 23, 2008.

Reprint requests: Heimo Breiteneder, PhD, Department of Pathophysiology, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria. E-mail: Heimo.Breiteneder@meduniwien.ac.at.

0091-6749/$34.00

doi:10.1016/j.jaci.2008.01.025

TABLE I. Numbers of sequences and protein families of allergens in AllFam

	Sequences	Sequences from known protein families	AllFam families	AllFam families with >1 allergen
All allergens	847	707	134	81
Sources				
Plants	369	338	58	34
Animals	305	268	60	36
Fungi	163	91	37	16
Bacteria	10	10	5	1
Routes of exposure				
Inhalation	479	377	99	59
Ingestion	257	240	48	29
Sting, bite	66	52	14	7
Contact	58	50	35	10
Autoallergen	14	14	14	0
Iatrogenic	11	10	7	2

published in peer-reviewed journals. Data on routes of exposure from Allergome were merged into the following standardized categories: inhalation, ingestion, sting/bite, contact, iatrogenic, and autoallergen.

UniProt accession numbers were compared with SwissPfam, a database of precomputed protein domain architectures generated by comparing all entries of the UniProt protein sequence database with the Pfam database (version 22.0; July 2007; http://pfam.sanger.ac.uk).[4] For entries that yielded no results, sequences were downloaded and compared with Pfam protein family definitions by using the hmmpfam program from the HMMER 2.3 package (http://hmmer.wustl.edu). This hmmpfam program compares a query sequence with all Pfam protein family definitions, which are stored as hidden Markov models, probabilistic descriptions that are generated from multiple sequence alignments and yield the probabilities of occurrence of all amino acids, as well as of insertions and deletions for each alignment position.

Domain architectures of allergens were translated into AllFam allergen families by using the following criteria. For single-domain proteins, each Pfam family corresponded to an AllFam family. To avoid an artificially high number of allergen families because of counting domains of multidomain proteins as separate families, Pfam domains constituting multidomain proteins were merged into single AllFam families if the constituting domains exclusively occurred in members of a single protein family. Otherwise, each domain was treated as a separate AllFam family.

The AllFam database is freely accessible at http://www.meduniwien.ac.at/allergens/allfam/. It can be queried for lists of allergen families filtered by source and route of exposure. In addition, for each family, the database contains a list of allergens and an allergen family fact sheet with information on biochemical properties and the allergologic significance of its allergenic members. AllFam is cross-linked with the Allergome database and regularly updated.

Protein family distribution of a random set of sequences

Random entries were downloaded from the UniProt database (http://www.expasy.org/uniprot/) and parsed for taxonomic group and Pfam domains. The procedure was repeated until the number of sequences that contained Pfam annotations from plants, animals, fungi, and bacteria reached the numbers of allergens with known protein family memberships in these kingdoms. The number of different Pfam domains found in these sequences was counted. Twenty independent runs of the program were performed. Significance of the difference between the number of protein families found among allergens and among random sequences was tested by using the 1-sample *t* test.

Structural and functional classification of allergens

Structures of allergens and allergen homologues were classified by using the Structural Classification of Proteins (SCOP) database (release 1.71, October 2006; http://scop.mrc-lmb.cam.ac.uk/scop/).[11] AllFam families and SCOP families were matched by using the links to SCOP embedded in the Pfam database.

For a functional classification of allergens using standardized descriptions of biologic functions, all UniProt accession numbers of allergen sequences in AllFam were compared with the Gene Ontology (GO) Annotation Database (http://www.ebi.ac.uk/GOA/).

Sequence conservation within families of allergens

Sequences of representative allergens from the 4 most important families of allergens (prolamins, profilins, tropomyosins, and the EF-hand family) were aligned by using ClustalX 1.83.[12] Sequence identity matrices and neighbor-joining phylogenetic trees were generated from these alignments with ClustalX and visualized with TreeView 1.6.6.[13]

RESULTS

Protein family distribution of allergens

The AllFam database (version of July 18, 2007) contained 847 allergens with known partial or total sequences (Table I). Of these, 707 allergens were classified into 134 AllFam families that contained 184 different Pfam domains. Thus allergens were found in only 2% of all 9318 families in the Pfam database. The list of AllFam families and associated Pfam families can be found in Table E1 in the Online Repository (available at www.jacionline.org). The distribution of allergens was highly biased toward a few protein families. Although the protein family with the highest number of allergens, the prolamin superfamily, contained 59 allergens (8% of all allergens with known protein family) and the 10 most abundant families contained 300 allergens (42%; Fig 1, *A* and *B*), there were 53 families that contained only a single allergen.

Thirty-eight allergen families were grouped by structural similarity or common sequence motifs into 12 superfamilies (termed *clans* in the Pfam database; see Table E2 in the Online Repository at www.jacionline.org). The most important clan, which comprised 7 allergen families, was the triosephosphate isomerase (TIM) barrel glycosyl hydrolase superfamily that contained main allergen families from mites (chitinases from the glycoside hydrolase family 18), plants and fungi (α-amylases, β1, 3-glucanases), and insect venoms (hyaluronidases).

Distribution of protein families among allergens of different sources and routes of exposure

Fig 1, *A* and B, shows the 15 most important families of allergens itemized by source and route of exposure. Most allergen families were confined to a single source kingdom, such as prolamins, profilins, and cupins from plants and tropomyosins, lipocalins, and caseins from animals. A minority of protein families, such as the EF-hand family and the pathogenesis-related proteins (PR-1), contained allergens from multiple kingdoms. A grouping of allergens by route of exposure yielded a different picture. Most protein families contained allergens that sensitize human subjects through different routes. Among these are allergens responsible for cross-reactivity between inhalative allergen sources and foods, such as profilins, Bet v 1–related allergens, and tropomyosins.

Protein family distribution of randomly selected sequences

A comparison of the protein family distribution of allergens with the distribution of random UniProt entries confirmed that the number of protein families among allergens was much smaller