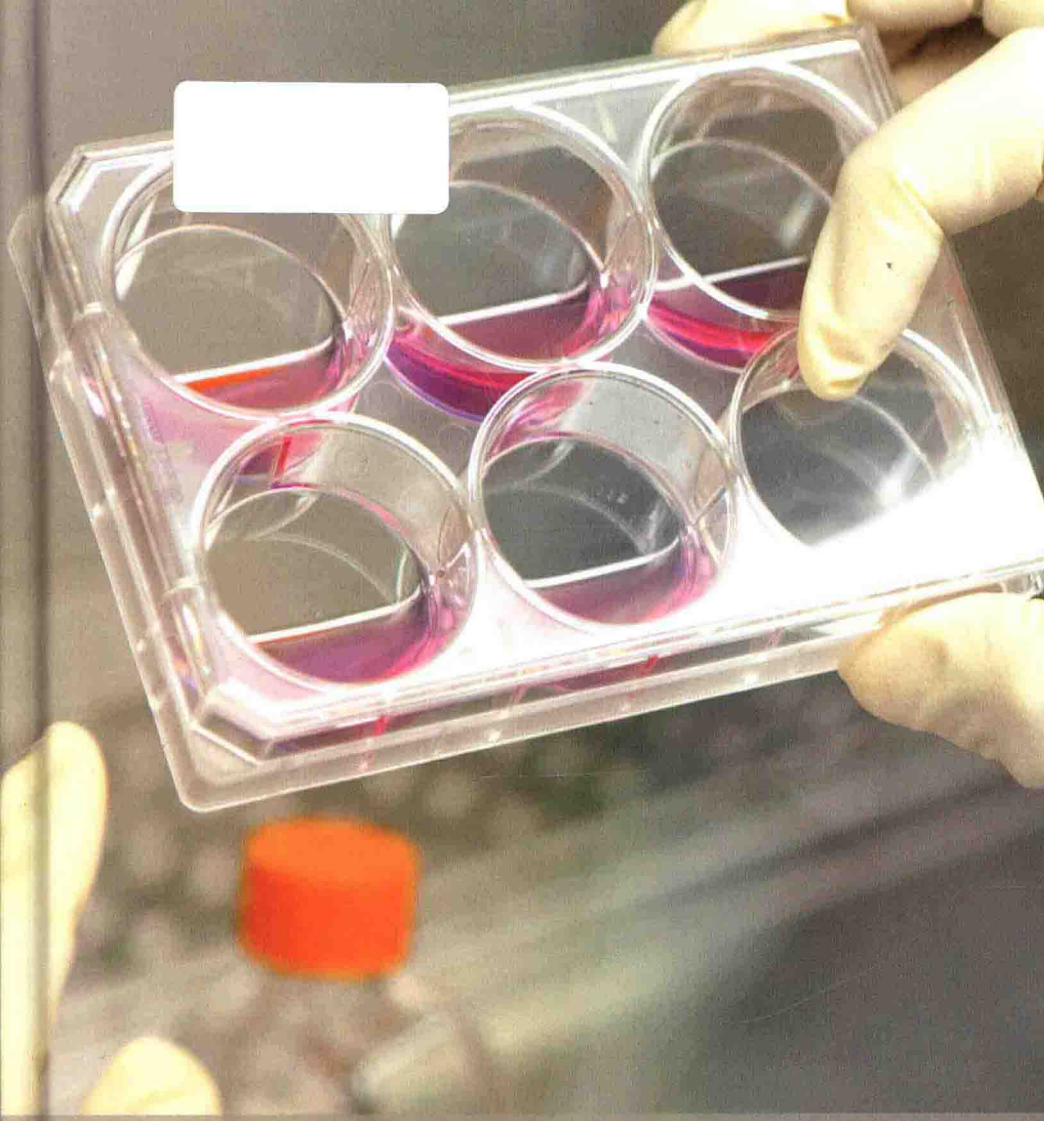


Bioinformatics
Advanced Topics

Gretchen Kenney

Bioinformatics: Advanced Topics

Edited by **Gretchen Kenney**

常州大学图书馆
藏书章

New York

Published by Callisto Reference,
106 Park Avenue, Suite 200,
New York, NY 10016, USA
www.callistoreference.com

Bioinformatics: Advanced Topics
Edited by Gretchen Kenney

© 2015 Callisto Reference

International Standard Book Number: 978-1-63239-097-4 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Copyright for all individual chapters remain with the respective authors as indicated. A wide variety of references are listed. Permission and sources are indicated; for detailed attributions, please refer to the permissions page. Reasonable efforts have been made to publish reliable data and information, but the authors, editors and publisher cannot assume any responsibility for the validity of all materials or the consequences of their use.

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy. Furthermore, the publisher ensures that the text paper and cover boards used have met acceptable environmental accreditation standards.

Trademark Notice: Registered trademark of products or corporate names are used only for explanation and identification without intent to infringe.

Printed in China.

Bioinformatics: Advanced Topics

Preface

This book was inspired by the evolution of our times; to answer the curiosity of inquisitive minds. Many developments have occurred across the globe in the recent past which has transformed the progress in the field.

The science of compiling and studying complex biological data, such as genetic codes, is referred to as bioinformatics. This book extensively covers topics related to bioinformatics. It includes a wide variety of topics such as unraveling genetic determinants of complex disorders, functional characterization of inherently unfolded proteins/regions and database resources for protein allergens. It also elaborates topics like protein interaction networks, flexible protein-protein docking and classification and prediction of regulatory motifs. There has also been emphasis on computational means for determining best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments. Computational algorithms have been explained in an easy-to-grasp manner for graduate and undergraduate students apart from researchers studying molecular biology and genetics. The book aims to assist biostatisticians, computational scientists and mathematicians in their academic and research activities in the field of bioinformatics.

This book was developed from a mere concept to drafts to chapters and finally compiled together as a complete text to benefit the readers across all nations. To ensure the quality of the content we instilled two significant steps in our procedure. The first was to appoint an editorial team that would verify the data and statistics provided in the book and also select the most appropriate and valuable contributions from the plentiful contributions we received from authors worldwide. The next step was to appoint an expert of the topic as the Editor-in-Chief, who would head the project and finally make the necessary amendments and modifications to make the text reader-friendly. I was then commissioned to examine all the material to present the topics in the most comprehensible and productive format.

I would like to take this opportunity to thank all the contributing authors who were supportive enough to contribute their time and knowledge to this project. I also wish to convey my regards to my family who have been extremely supportive during the entire project.

Editor

试读结束：需要全本请在线购买： www.ertongbook.com

Contents

	Preface	**VII**
Chapter 1	**Allergen Bioinformatics: Recent Trends and Developments** Debajyoti Ghosh and Swati Gupta-Bhattacharya	**1**
Chapter 2	**Family Based Studies in Complex Disorders: The Use of Bioinformatics Software for Data Analysis in Studies on Osteoporosis** Christopher Vidal and Angela Xuereb Anastasi	**17**
Chapter 3	**Understanding LiP Promoters from *Phanerochaete chrysosporium*: A Bioinformatic Analysis** Sergio Lobos, Rubén Polanco, Mario Tello, Dan Cullen, Daniela Seelenfreund and Rafael Vicuña	**41**
Chapter 4	**Guide to Genome-Wide Bacterial Transcription Factor Binding Site Prediction Using OmpR as Model** Phu Vuong and Rajeev Misra	**65**
Chapter 5	**Relaxed Linear Separability (RLS) Approach to Feature (Gene) Subset Selection** Leon Bobrowski and Tomasz Łukaszuk	**81**
Chapter 6	**Disease Gene Prioritization** Carlos Roberto Arias, Hsiang-Yuan Yeh and Von-Wun Soo	**97**
Chapter 7	**Prediction and Experimental Detection of Structural and Functional Motifs in Intrinsically Unfolded Proteins** Cesira de Chiara and Annalisa Pastore	**117**
Chapter 8	**Flexible Protein-Protein Docking** Sebastian Schneider and Martin Zacharias	**139**

Chapter 9 **Exploiting Protein Interaction Networks to Unravel Complex Biological Questions** 155
Bernd Sokolowski and Sandra Orchard

Permissions

List of Contributors

1

Allergen Bioinformatics: Recent Trends and Developments

Debajyoti Ghosh[1] and Swati Gupta-Bhattacharya[2]
[1]*Division of Allergy, Immunology and Rheumatology, Department of Internal Medicine University of Cincinnati College of Medicine, Ohio*
[2]*Division of Plant Biology Bose Institute Kolkata*
[1]*United States of America*
[2]*India*

1. Introduction

Allergy is a major cause of morbidity worldwide. Allergic reactions result from maladaptive immune responses in predisposed subjects, to otherwise harmless molecules. These allergenic molecules, usually proteins/glycoproteins, can not only elicit specific immunoglobulin E (IgE) in susceptible subjects, but also crosslink effector cell-bound IgE molecules Leading to the release of mediators (e.g. Histamine) and causation of symptoms. From clinical and molecular biological data available in several publicly accessible databases, it is now evident that among hundreds and thousands of proteins that exist in nature, only a few can cause allergy. For example, in more than 500,000 entries (71345 documented at the protein level; Nov, 2010) in swissprot/uniprot database (http://www.uniprot.org), only 686 proteins have been listed in the IUIS allergen nomenclature database (www.allergen.org) as documented allergens. Although about 1500 allergens (including iso-allergens) have been listed in the Allergome database (www.allergome.org), it has been shown that they are distributed into a very limited number of protein families. However, critical feature(s) that makes proteins allergenic is not fully understood. In the present article, we'll discuss recent applications of bioinformatic tools that shaped our current understanding about allergenicity of proteins.

2. Allergen bioinformatics - a need of the hour

Experiments on genetic engineering during the last few decades have led to the production of numerous genetically modified (GM) organisms. So, proteins introduced into GM organisms through genetic engineering must be evaluated for their potential to cause allergic diseases. As a classical example, transgenic soy, that has been genetically engineered to express ground-nut 2S albumin, was found to elicit hypersensitivity reactions in ground-nut allergic people (Nordlee et al., 1996). In 2001, the FAO/WHO suggested a procedure for performing FASTA or BLAST (Basic Local Alignment Search Tool) searches, and a threshold of greater than 35% identity in 80 or greater amino acids to identify potential allergenic cross-reactivity of transgene encoded proteins in genetically enhanced crops (Silvanovich et

al., 2009). Given that this will not exclude all probabilities of a protein to be allergenic (and cross-reactive to known allergens), the codex guidance recognized that the assessment will evolve based on new scientific knowledge (Goodman, 2008).

Bioinformatic tools are key components of the 2009 Codex Alimentarius for an overall assessment of the allergenic potential of novel proteins. Bioinformatic search comparisons between novel protein sequences, as well as potential novel fusion sequences derived from the genome and transgene or from any known allergen(s) are required by all regulatory agencies that assess the safety of genetically modified (GM) products(Ladics et al., 2011).

Allergens were usually seen as an array of proteins with no apparent similarity in structure and function. They come from diverse sources: Plants, animals or fungi and may take different modes of exposure: inhalation, ingestion, sting or contact. They are, like their non-allergenic counterparts, structurally heterogeneous. For example, the major cat allergen Fel d 1, is an alpha-helical tropomyosin, while a major dust mite allergen Der p 2 consists predominantly of beta sheets and the major birch pollen allergen Bet v 1 contains both of these structural elements. Allergen sequences are extensively studied to find out any possible structural element or function associated with allergenicity. However, no such allergen-specific structural / functional element could be identified. High sequence identity between homologous protein allergens may result in common surface patches that may confer cross-reactivity among them. Aalberse pointed out that proteins sharing less than 50% sequence identity are rarely cross-reactive(Aalberse, 2000). In contrast, proteins that share at least 70% identity often show cross-reactivity. Many IgE-binding epitopes have been identified as sequential epitopes, although for many this does not represent the full epitope. Linear epitopes are usually part(s) of conformational epitope(s) responsible for a significant portion of IgE binding. While IgE-binding peptides can consist only of five amino acids (Banerjee et al., 1999), the majority of characterized IgE-linear epitopes are eight amino acids or longer (Chatchatee et al., 2001; Shin et al., 1998). Astwood et al. recommended sequence comparisons to a database of known IgE-binding epitopes. Finally, Ivanciuc and colleagues have recently utilized mixed sequence and structure-based methods to predict IgE-binding sites. This is based on comparison of local sequence and structure to identify common features associated to allergens (Ivanciuc et al., 2009b).

3. Allergen databases

Exponential growth of molecular and clinical data on allergens has created a huge demand for efficient storage, retrieval and analyses of available information. There are numerous allergen databases available on Internet. They are targeted to different aims ranging from easy accessibility of data to novel allergen prediction. A few examples have been provided in table-1.

The IUIS (International Union of Immunological Societies) allergen nomenclature subcommittee has created a unique, unambiguous nomenclature system for allergenic proteins. It maintains an allergen database (www.allergen.org) containing an expandable list of WHO/IUIS -recognized allergen molecules arranged according to Linnean system of classification (Kingdoms: Plantae, Fungi and Animalia and subdivided into lower orders)(Chapman et al., 2007) of the source organism. This database is a precise and convenient source for researchers, since it contains the biochemical name and molecular weight of the allergens and isoallergens (multiple molecular forms of the same allergen showing ≥67% sequence identity). It is searchable by allergen name, source and taxonomic

group. For example, a search using the key word 'Bet v 1' shows about 36 variants (isoallergens) of this allergen, each with genbank, uniprot accession numbers and, if available, with PDB IDs. Each uniprot ID is linked to the original entry in uniprot database. Moreover, once the uniprot IDs are obtained, their sequences can be retrieved in batches using uniprot's 'retrieve' tab.

Allergome (Mari et al., 2006) is a vast repository of data related to all allergen molecules. It contains data about a larger number of allergens than actually recognized by IUIS/WHO. It also contains links to other databases (eg Uniprot, PDB) and computational resources with additional extensive links to literature. The Allfam database is a useful resource for grouping of allergens into protein families. It utilizes the allergen information from 'Allergome' database and protein family information from pfam database. It can be sorted by source (plants/animals/bacteria/fungi) and route of exposure (inhalation/ingestion/contact/sting etc) or can be searched for specific protein families. Allergen entries are linked to corresponding records in the Allergome database. In addition, each allergen family is linked to a family fact sheet containing descriptions of the biochemical properties and the allergological significance of the family members.

Name (URL)	Purpose
IUIS (http://www.allergen.org/)	Database targated towards systematic nomenclature of allergenic proteins
Allergome (http://www.allergome.org/)	Vast source of information and references about allergen molecules
Allfam (http://www.meduniwien.ac.at/allergens/allfam/)	Database for allergen classification
Allergen Database for Food Safety (ADFS) (http://allergen.nihs.go.jp/ADFS/)	Database with computational allergenicity prediction tool
The Allergen Database (http://allergen.csl.gov.uk//index.htm)	A basic database for allergen structures
Allermatch (http://www.allermatch.org/)	Allergenicity prediction from sequence
AllerTool http://research.i2r.a-star.edu.sg/AllerTool/	Webserver for predicting allergenicity and allergenic cross-reactivity
AlgPred (http://www.imtech.res.in/raghava/algpred/)	*In silico* prediction of allergenicity
WebAllergen (http://weballergen.bii.a-star.edu.sg/)	To predict potential allergenicity of a protein from its sequence
SDAP (http://fermi.utmb.edu/SDAP/)	Database of allergen structure with various resources, links and computational tools

Table 1. A few databases of allergenic proteins and web-servers to predict protential allergenicity from amino-acid sequence.

Although the above-mentioned databases are very useful resources, they do not contain any computational tool to predict allergenicity from amino acid sequences of proteins. However, there are several other databases that can efficiently deal with this aspect. ADFS (Allergen Database for Food Safety) is developed and maintained by Japan's National Institute of Health Sciences. It is a good resource of available information about known allergens (uniprot protein ID, PDB accession number, epitope sequence, presence of carbohydrate, pfam - and interpro domain IDs etc.). Moreover, this website has computational tools to predict allergenicity. Other websites dedicated to allergenicity prediction are Allermatch, AllerTool and Algpred etc. Detailed discussion on these servers is beyond the scope of the present article.

The database which is dedicated to the structural biology of the allergic proteins is SDAP (Structural Database of Allergenic Proteins) hosted by the University of Texas Medical Branch. It integrates a database of allergenic proteins with various computational tools for prediction of allergenicity and epitope sequences on protein allergens.

Analyses of data available in different publicly accessible database have shaped our current understanding about allergens, as discussed in the subsequent sections of this article.

3.1 Allergens seen as proteins without bacterial homolog

Among numerous proteins sequenced till date, only about a thousand has been classified as allergens, although no common structural or biochemical function could be assigned to all allergens. To address this problem, Emanuelson and Spangfort (2007) used 30 randomly selected allergen sequences to search the non-redundant Expasy/SIB and UniProt/TrEMBL databases (subsection Bacteria+Archea) using BLAST (Basic Local Alignment Search Tool) program. For each allergen, an appropriate species-specific non-allergenic control homolog was included. It has been found that 25 out of 30 allergens do not have any bacterial homologues; two other allergens have only a few, while all the non-allergenic controls retrieved numerous bacterial homologues. Moreover, major allergens like Bet v 1, also lack human homolog. The authors, thus, interpreted that the allergens are usually foreign proteins that lack bacterial homologues (Emanuelsson and Spangfort, 2007).

3.2 Allergenic proteins can be organized into families

The first definite interpretation that allergens can be grouped came from arranging allergens into pfam protein families. Pfam classifies proteins into families on the presence of specific domains (pfam domains) identified through multiple sequence alignments and Hidden Markov Models. Pfam 25.0 (latest version; March 2011) contains over 100, 000 protein sequences classified into 12,275 families (Finn et al., 2010). The allergen database that contains pfam domain information is 'AllFam' (http://www.meduniwien.ac.at/allergens/allfam), where allergen sequences are classified into protein families using the Pfam database, and its associated database, SwissPfam. AllFam includes all allergens that can be assigned to at least one Pfam family. But many allergens are multi-domain proteins. The domains of these proteins are merged into a single AllFam family, if the Pfam domains of this allergen occur only in combination with a single other Pfam domain. Figure-1 shows the distribution of allergenic proteins in different Allfam families. The major allergen families (containing 10 or more allergens) with corresponding Pfam domains are shown in Table-2.

试读结束：需要全本请在线购买： www.ertongbook.com

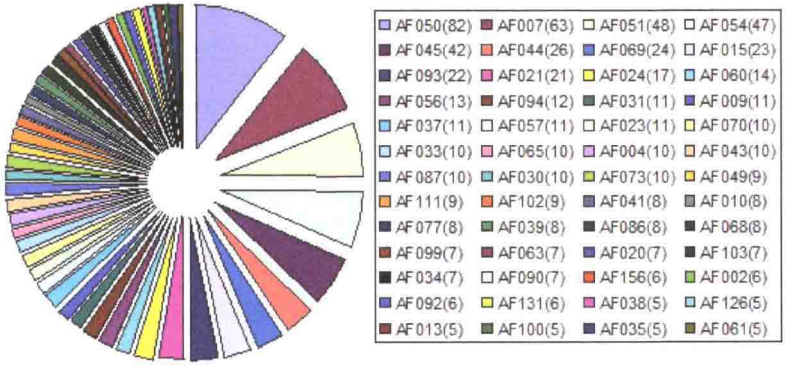


Fig. 1. Pi chart showing the distribution of allergic proteins in different AllFam families. Numbers of constituent allergens have been indicated within brackets.

Allfam ID	Name	Pfam ID	Allergens	Examples
AF050	Prolamin superfamily	PF00234	82	Amb a 6 (Ragweed) Ana o 3 (cashewnut)
AF007	EF hand domain	PF00036 PF01023	63	Aln g 4 (Alder) Bet v 3 (birch)
AF051	Profilin	PF00235	48	Bet v 2 (birch) Ana c 1 (pinapple)
AF054	Tropomyosin	PF01357	47	Der p 10 I(mite) Hom a 1 (lobster)
AF045	Cupin superfamily	PF00190 PF04702	42	Ara h 1 (peanut) Gly m 5 (soyabean)
AF044	CRISP/PR-1/venom group 5 allergen family	PF00188	26	Pol d 5 (wasp venom), Art v 2 (mungwort)
AF069	Bet v 1-related protein	PF00407	24	Bet v 1 (birch) Api g 1 (celery)
AF015	Lipocalin	PF00061 PF08212	23	Can f 1 (dog) Bos d 2 (domestic cattle)
AF093	Expansin, C-terminal domain	PF01357	22	Phl p 1 (timothy grass) Tri a 1 (wheat)
AF021	Subtilisin-like serine protease	PF00082 PF02225 PF05922	21	Asp f 13 (fungal) Pen c 1 (fungal)
AF024	Trypsin-like serine protease	PF00089 PF02983 PF09396	17	Der f 3 (mite) Blo t 3 (mite)
AF060	Thaumatin-like protein	PF00314	14	Mal d 2 (apple) Pru av 2 (Cherry)
AF056	Serum albumin	PF00273	13	Can f 3 (dog) Fel d 2 (cat)

AF094	Expansin, N-terminal domain	PF03330	12	Phl p 1 (timothy grass) Ory s 1 (rice)
AF031	Enolase	PF00113 PF3952	11	Alt a 6 (fungal) Cha h 6 (fungal)
AF009	Globin	PF00042	11	Chi t 1 (midge) Chi t 2 (midge)
AF043	Hevein-like domain	PF00187	11	Hev b 6 (rubber latex) Mus a 2 (banana)
AF073	Pectate lyase	PF00544	11	Amb a 1 (ragweed) Cry j 1 (Japanese cedar)
AF057	Polygalacturonase	PF00295	11	Cry j 2 (Japanese Cedar) Jun a 2 (mountain Cedar)
AF037	Lipase	PF00151 PF01477	11	Pol a 1 (wasp) Sol i 1 (ant)
AF070	60S acidic ribosomal protein	PF00428	10	Alt a 5 (fungal) Cla h 6 (fungal)
AF033	Alpha amylase	PF00128 PF02806 PF07821 PF09154 PF09260	10	Blo t 4 (mite) Der p 4 (mite)
AF065	Alpha/beta casein	PF00363	10	Bos d 8 alphaS1 (domestic cattle) Ovi a casein alphaS1 (sheep)
AF004	Eukaryotic aspartyl protease	PF00026 PF07966	10	Asp f 10 (fungal) Bla g 2 (cockroach)
AF030	Papain-like serin protease	PF00112 PF08246	10	Der p 1 (mite) Blo t 1 (mite)
AF023	Thioredoxin	PF00085	10	Alt a 4 (fungal) Fus c 2 (fungal)
AF073	Pectate lyase	PF00544	10	Amb a 1 (short ragweed) Jun a 1 (mountain ceder)

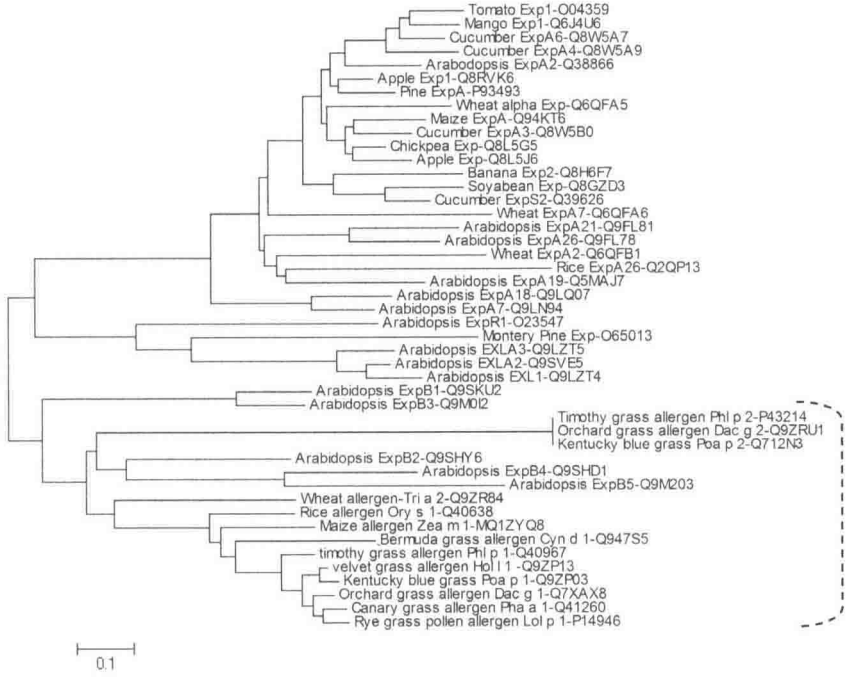
Table 2. Major allergen families (AllFam familes) that contain 10 or more allergens are shown with correspondent pfam domains and examples. Number of allergenic members / allergen family has also been shown.

AllFam takes the allergen information from "Allergome", the comprehensive allergen database. In the latest version of Allfam (May, 2011), 950 allergens have been arranged into 150 allergen families (AllFam families). It has been found that the allergens are distributed in a really skewed manner with about 30% members belonging to only 5 families (Prolamin, Profilins, EF hands, tropomyosin and cupins) and showing few restricted biological functions such as hydrolysis, storage or binding to cytoskeleton [6]. Moreover, allergens contain about 245 pfam domains in total, which is only about 2.0% of all domains identified to date.

AllFam gave us an opportunity to retrieve and sort allergen data according to source (plant/animal/fungi/bacteria), route of exposure (inhalation/ingestion/contact etc) and

Pfam/AllFam family identities. This analysis combined with the study of evolutionary relationship among the proteins has led to the following valuable insights:

i. Pollen allergens (Inhalant plant allergens) are restricted into few protein families(Radauer and Breiteneder, 2006). They populate only 29 out of more than 7000 protein families, with (a) Expansins (b) Profilins and (c) calcium-binding proteins (with EF-hand domains) consisting most of the pollen allergens followed by Bet v 1 related /pathogenesis-related proteins (PR10 family). Figure-2 shows the evolutionary relationship between several allergenic and non-allergenic members of (a) expansins and (b) profiling families. The evolutionary history was inferred using the Neighbor-Joining method(Saitou and Nei, 1987). The evolutionary distances were computed using the Poisson correction method(Zuckerkandl and Pauling, 1965) and the phylogenetic analyses were conducted in MEGA4(Tamura et al., 2007). Similar method has been followed in the subsequent sections of the present article. Allergens of the expansin family are clustered as highly identical proteins as shown in the figure. Allergenic plant profilins also constitute a conserved homologous group with high sequence identities (70-85%) among themselves, while showing low identities (30-40%) with non-allergenic profilins from other eukaryotes including human(Radauer and Breiteneder, 2006). About 10 of the 29 pollen allergen families are also present in plant-derived foods.



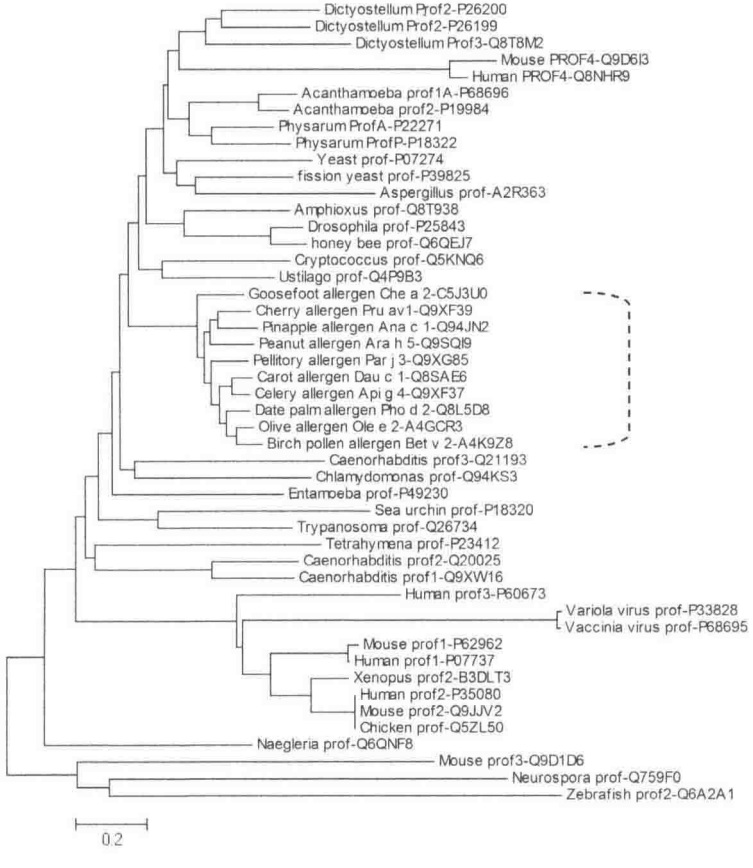


Fig. 2. Phylogenetic trees showing the relationships of two major pollen allergen families: (a) Expansins, (b) Profilins and their respective non-allergenic homologues. Pollen-related plant food allergens such as Ara h 5, Dau c 1 etc are also included. Uniprot accession numbers are shown. Positions of allergens are indicated by dotted lines.

ii. In case of major animal food protein families evolutionary distance from human homologue reflects their allergenicity (Jenkins et al., 2007). This has been demonstrated in major food allergen families like (a) parvalbumins, (b) casins and (c) tropomyosins.

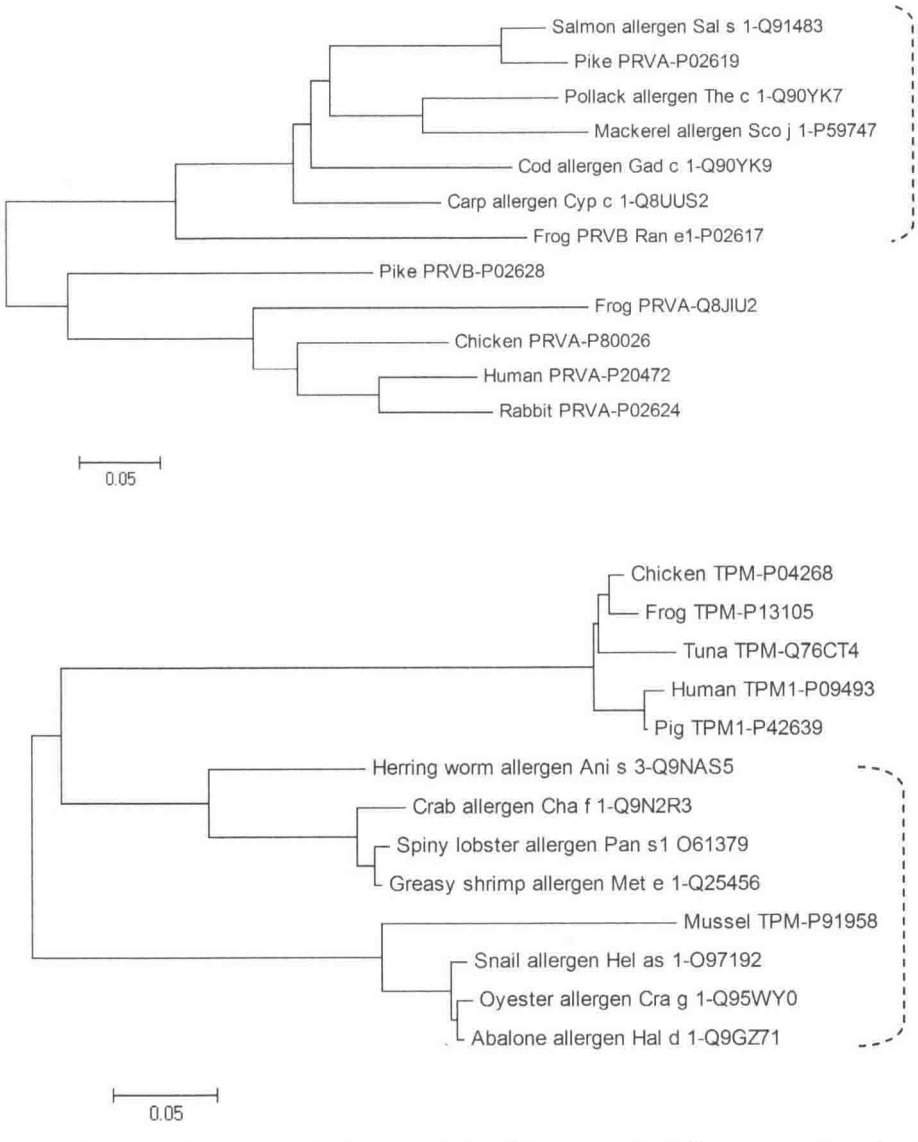


Fig. 3. Dendrogram showing evolutionary relationship among 12 different parvalbumins (a) and 13 different tropomyosins (b) from animals and human. Allergenic proteins and their non-allergenic homologues as well as the closest human homologues are chosen. The Uniprot accession numbers and positions of allergen clusters are indicated.

iii. Plant food allergens are clustered into only four major protein families

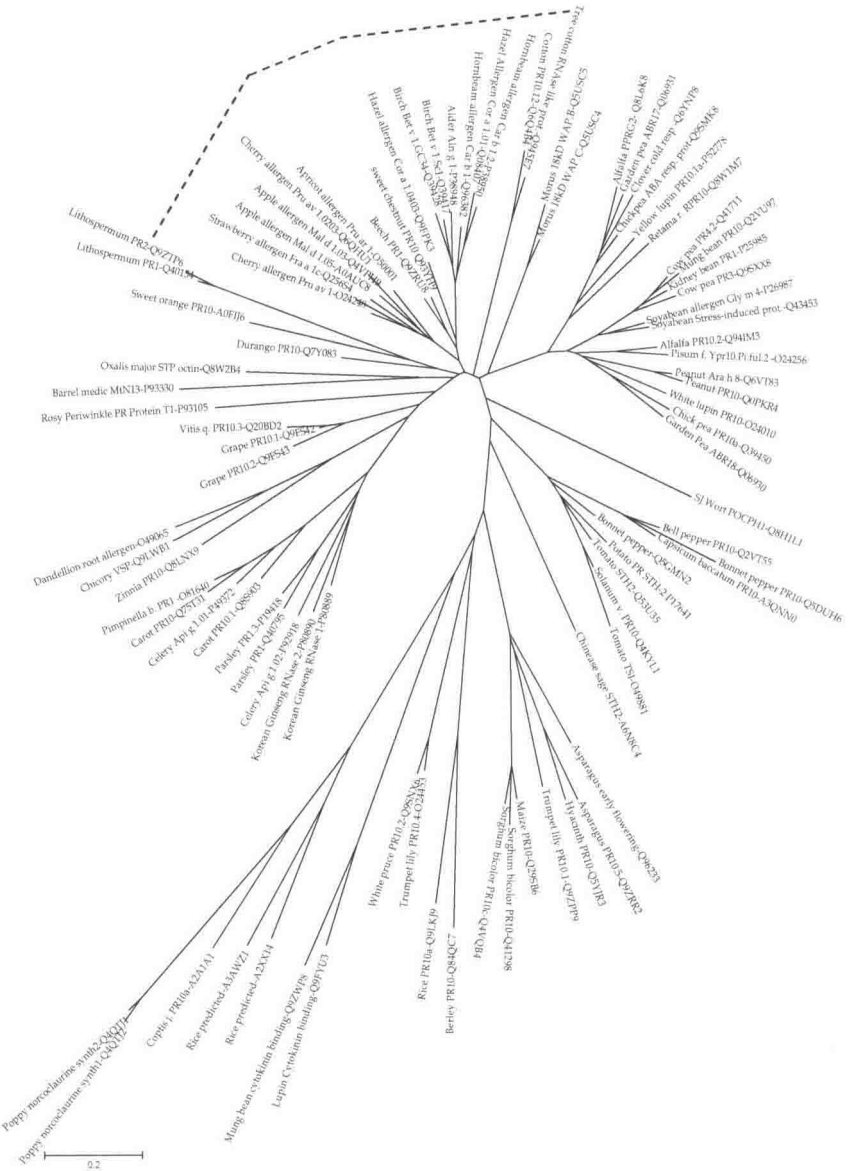


Fig. 4. Un-rooted neighbor-joining tree showing evolutionary relationship among the members of Bet v 1-related plant protein family (containing pfam domain PF00407). Uniprot accession numbers and position of the allergen cluster has been indicated.(Radauer and Breiteneder, 2007). They are (a) the Prolamin superfamily with PF00234 domain (b) the cupin superfamily with PF00190 and PF04702 domains (c) the Profilins with PR00235

试读结束：需要全本请在线购买： www.ertongbook.com