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.

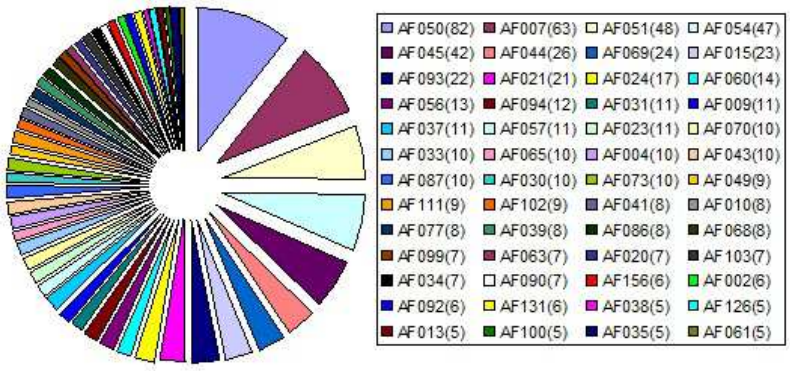


Fig. 1. Pi chart showing the distribution of allergic proteins in different AllFam families. Numbers of constituent allergens have been indicated within brackets.

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)

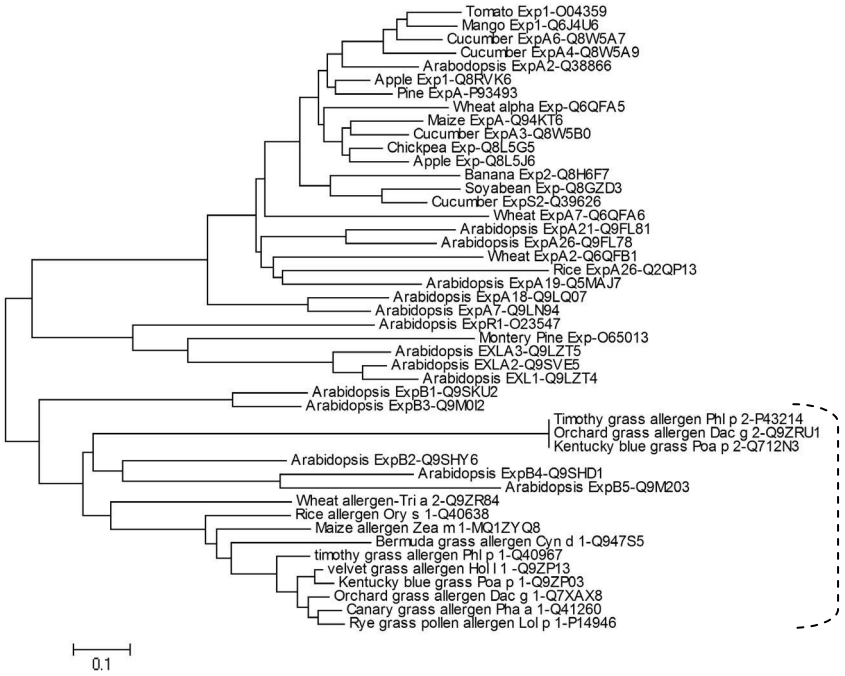
Table 2. Major allergen families (AllFam familes) that contain 10 or more allergens are shown with correspondent pfam domains and examples. Number of allergenic members / allergen family has also been shown.

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.



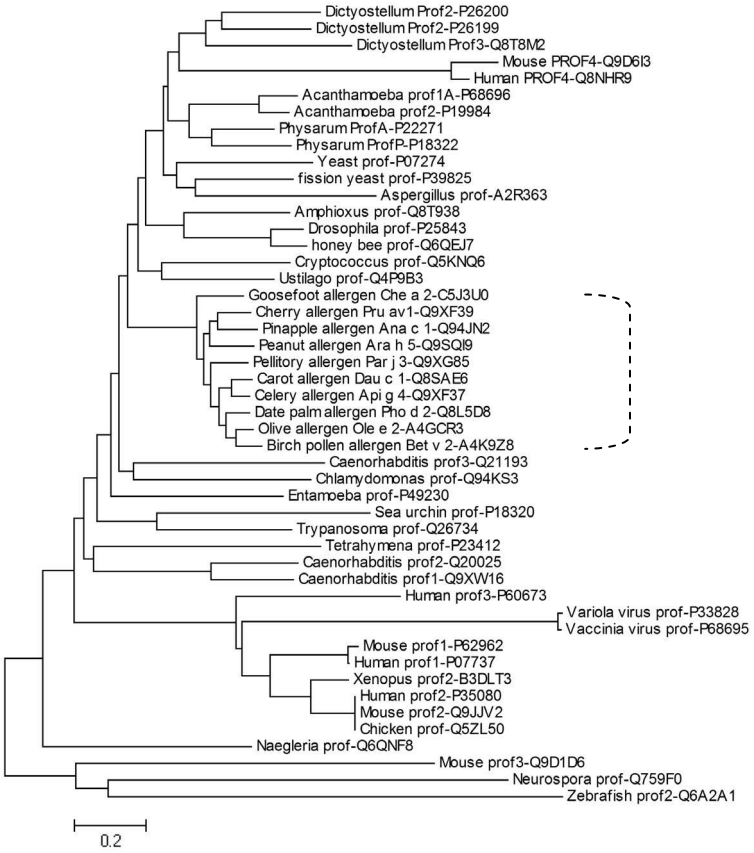


Fig. 2. Phylogenetic trees showing the relationships of two major pollen allergen families: (a) Expansins, (b) Profilins and their respective non-allergenic homologues. Pollen-related plant food allergens such as Ara h 5, Dau c 1 etc are also included. Uniprot accession numbers are shown. Positions of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

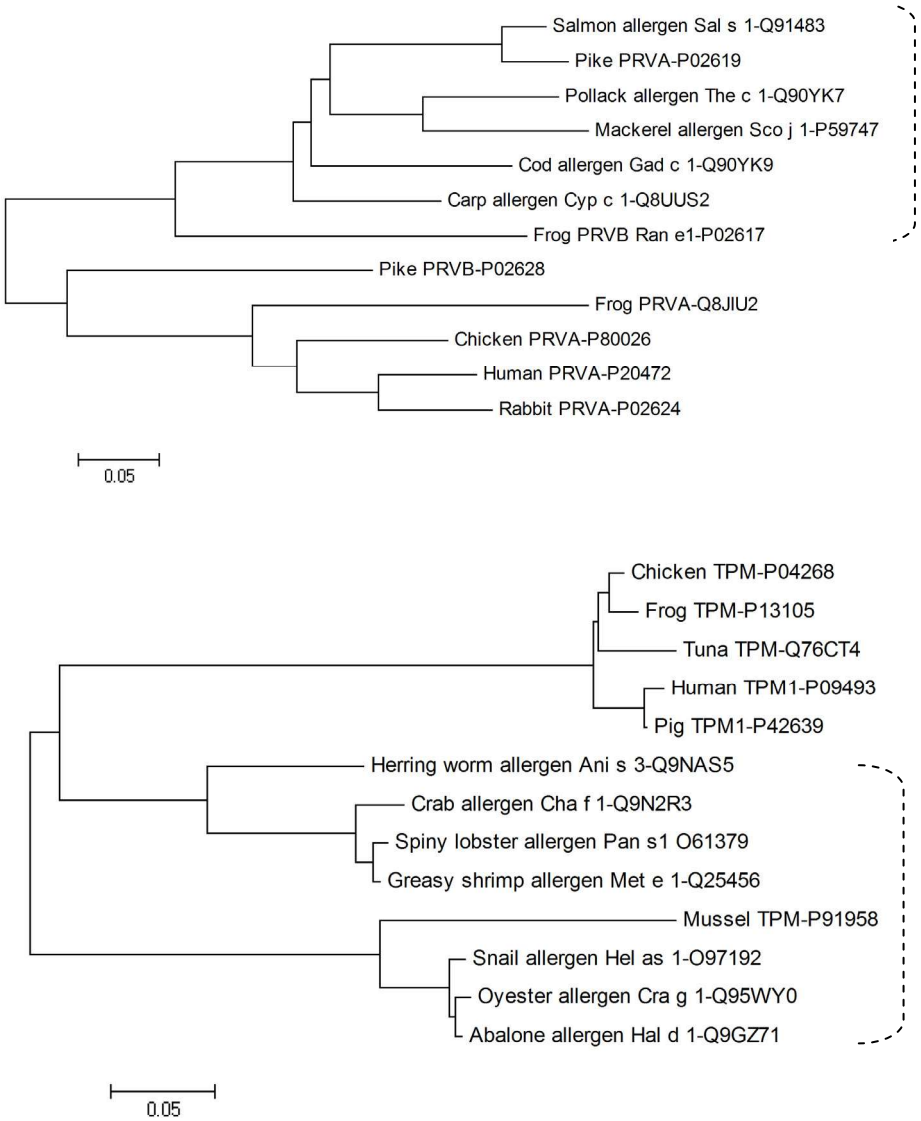


Fig. 3. Dendrogram showing evolutionary relationship among 12 different parvalbumins (a) and 13 different tropomyosins (b) from animals and human. Allergenic proteins and their non-allergenic homologues as well as the closest human homologues are chosen. The Uniprot accession numbers and positions of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

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

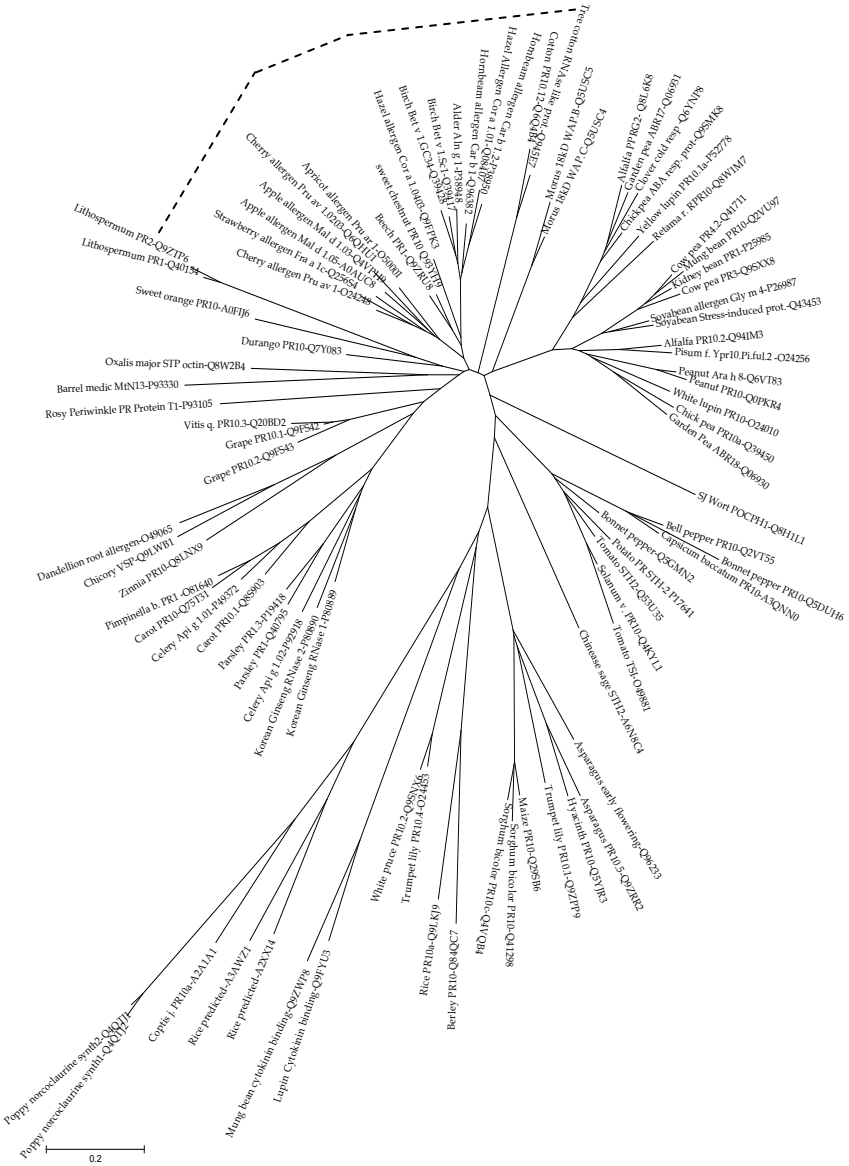


Fig. 4. Un-rooted neighbor-joining tree showing evolutionary relationship among the members of Bet v 1-related plant protein family (containing pfam domain PF00407). Uniprot accession numbers and position of the allergen cluster has been indicated.(Radauer and Breiteneder, 2007). They are (a) the Prolamin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

domain and (d) the Bet v 1 –like proteins containing PF00407 domain. Prolamins are seed storage proteins containing about 82 characterized allergens, with 65 enlisted as ingestants. Figure-4 shows the evolutionary relationship among the Bet v 1-homologous protein family. Twenty-four proteins of this group are known as allergens present in pollen and plant-derived foods responsible for causing allergic sensitization in a large number of people.

3.3 Allergen-associated protein domains

The other allergen database that utilizes the Pfam protein family information is Motifmate (http://born.utmb.edu/motifmate/index.php) (Ivanciuc et al., 2009a). Motifmate assigns pfam domains to the allergens listed in the SDAP (Structural Database of Allergenic Proteins) database developed and maintained by the University of Texas (http://fermi.utmb.edu/SDAP)(Ivanciuc et al., 2003). The recent version of this database contains 679 Proteins (May, 2011). The authors pointed out that all the allergenic protein entries in SDAP could be associated with only 130 pfams (of total 9318 pfams) with only about 31 pfam protein families containing 4 or more number of allergens. [Figure-5]. This outcome supports the previous finding that the allergenic proteins are clustered in few pfam families.

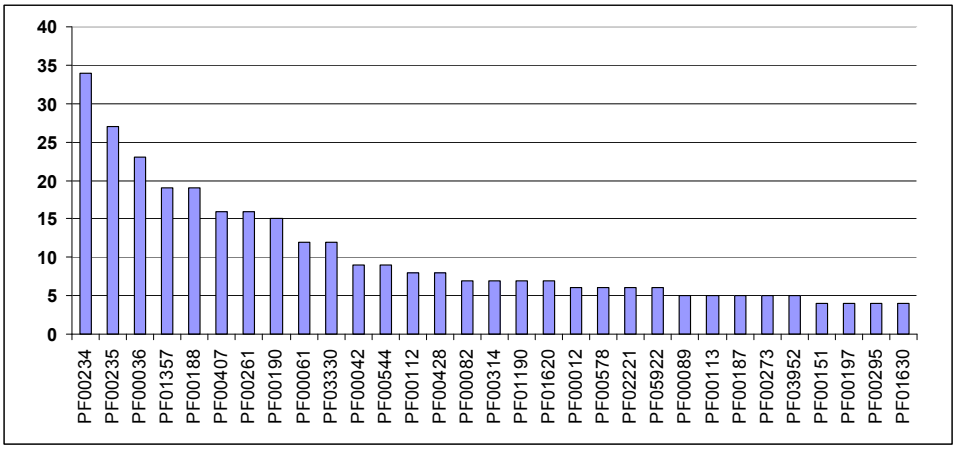


Fig. 5. Distribution of Pfam domains in allergenic proteins according to Motifmate database. lipid transfer proteins (PF0234) highest number of allergns.

4. Insights from structural bioinformatics

After the elucidation of X-ray crystal structure of the birch pollen allergen Bet v 1(Gajhede et al., 1996), structures of several allergens have been solved. Searching the protein databank with the keyword "allergens" returns 321 entries, with occasional presence of multiple entries for one single allergen. Although protein structure gives us valuable insight into their function, structures of several allergens are still not known. More importantly, some allergen families have members with known structures, while others may have very few / no member whose structure(s) have been deduced. Allergen structures are particularly useful to elucidate molecular features related to allergenicity, cross-reactivity and for

designing hypoallergenic derivatives. For example, there are about five structures in protein data bank that correspond to Bet v 1, the major birch pollen allergen : the x-ray structure (1BV1.pdb), the NMR structure (1BTV.pdb), mutants (1B6F.pdb and 1QMR.pdb), complexed with IgG Fab (1FSK.pdb) and the hypoallergenic isoform Bet v 1d (3K78.pdb). On the contrary, several groups, such as the cupin family of seed storage protein allergens are under-represented.

Knowledge about allergen structures is important because it is the over-all structure, not the sequence, which determines the biochemical/immunological properties. Molecular modeling can help us in case the experimentally determined structure of the allergen is not available. Homology modeling, also known as comparative molecular modeling, can predict the 3D model of a given protein from its amino acid sequence using experimentally derived structure(s) (X-ray/NMR) of one or more related homologous protein(s) (called template). This technique is becoming increasingly popular because, if required template selection and alignment criteria are met, it is believed to be the most reliable modeling technique to date (Marti-Renom et al., 2000). It is becoming increasingly useful because although there are millions of proteins in nature, the number of structural folds they can assume is limited (Zhang, 1997) and the number of X-ray/NMR structure of proteins is exponentially increasing providing an increased chance of getting a suitable 'template'. Several authors have successfully utilized this technique of molecular modeling to predict allergen structures and to elucidate the structural basis of cross-reactivity between allergens.

Ara h 1 (vicilin) and Ara h 2 (2S albumin) are seed storage proteins of peanut (*Arachis hypogea*). They are recognized by serum IgE of >90% of peanut-allergic people, thus showing their importance as major peanut allergens (Shin et al., 1998; Stanley et al., 1997). Ara h 1 shows IgE-mediated cross-reactivity with other vicilin allergens such as Len c 1 (from lentil) and Pis s 1 (from sweet pea). Following sequence alignment using ClustalX, structural models of Ara h 1, Len c 1 and Pis s 1 were generated from experimentally derived structure of beta-conglycinin (RCSB protein data bank code: 1IPJ) using programs InsightII, Homology and Discover3 (Accelrys, USA). Electrostatic surfaces of these proteins were also generated using program GRASP (Nicholls et al., 1991). Mapping of linear epitope sequences revealed that nine out of 23 linear B-cell epitopes are located in the N-terminal region. They are unique to Ara h 1. But the remaining B-cell epitopes, situated in the C-terminal part, are well-exposed to the surface, share a high degree of homology and 3D conformation to Len c 1 and Pis s 1. They might be responsible for cross-reactivity among Ara h 1, Len c 1 and Pis s 1 food proteins. Similarly, Ara h 2 and other dietary allergenic 2S albumins Jug r 1 (walnut), Car i 1 (pecan nut), Ber e 1 (Brazil nut) were modeled using the atomic coordinates of homologous Ric c 1(castor bean 2S albumin). Mapping of known epitope sequences on the template and modeled structures revealed no structural homology between allergenic 2S albumins of peanut, walnut, pecan and brazil nut. This indicates that cross-reactivity between Ara h 1 and other 2S albumins, which is less likely, might not be mediated by protein epitopes, but CCDs (cross-reactive carbohydrate determinants). However, the c-terminal epitope region of Jug r 1 showed a clear structural homology with Car i 1 indicating the possibility of their cross-reactivity.

Another important insight was obtained from the homology modeling of allergenic cyclophilins(Roy et al., 2003). Groups of highly homologous cross-reactive allergens such as cyclophilins, profilins, MnSOD are known as pan-allergens. They often cross-react with their

respective human homologues(Crameri et al., 1996) and such cross-reactivity might be responsible for severity and perpetuation of symptoms in the absence of exogenous allergen exposure(Fluckiger et al., 2002). Allergenic cyclophilins (peptidyl-prolyl cis-trans isomerase; PF00160) have been identified from several organisms such as: Periwinkle (pollen allergen Cat r 1), birch (pollen allergen Bet v 7), *Aspergillus fumigatus* (Asp f 11, Asp f 27), *Psilocybe cubensis*, *Malassazia furfur* (Mala s 6, formerly known as Mal f 6) and carrot. IgE-mediated cross-reactivity between Mala s 6, Asp f 11, yeast cyclophilin and human cyclophins has been demonstrated (Fluckiger et al., 2002). The structure of human cyclophilin, which shows high sequence identities to Asp f 11, Mala s 6 and yeast cyclophilin, was known from crystallography (PDB code: 2RMB). Thus, taking this as the template, the molecular models of three other cyclophilins were generated and compared with the human homologue to understand the structural basis of their cross-reactivity.

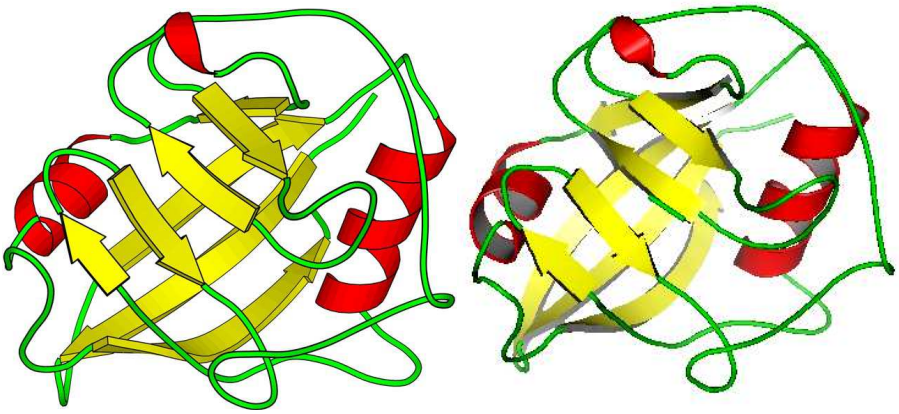


Fig. 6. Structure diagram of the allergenic cyclophilin Mala s 6 as predicted in 2003 from homology modeling (left figure) and as determined by x-ray crystallography (right figure) in 2006 (2CFE.pdb). Alpha helices have been shown in red, while beta sheets are in yellow and loops (smoothed) are in green.

Molecular modeling was done using program Modeller (Sali and Blundell, 1993). The structures were energy-minimized using program Discover with Consistent Valence Force Fields(Hagler et al., 1979) and their steriochemical qualities were checked using Procheck(Laskowski, 1993). Several empirical/semi-empirical programs were used to predict the antibody binding sites (B-cell epitopes) on these proteins and residue-wise solvent accessibility values of these predicted epitopes were calculated using program NACCESS(Hubber, 1992). The cyclosporine-binding site of these proteins were also identified by aligning the sequences (using ClustalW) and structures. This study revealed large conserved solvent-exposed patches on the surfaces of these proteins strongly suggesting their cross-reactivity. The x-ray crystal structure of Mala s 6 (PDB code: 2CFE), published three years later (Glaser et al., 2006), very much resembles its predicted model [Figure-6]. The domain-swapped structure of Asp f 11 dimer (PDB code: 2C3B), also published at the same time, showed similar structural fold. Asp f 11 dimer (resulting from increased protein concentration) seems to be enzymatically inactive, since the active sites of both its subunits are blocked due to dimerization. However the constituent monomers

retained the basic cyclophilin structure. More recently, a comprehensive 3D structural modeling of allergens, with no known structure, has been conducted followed by surface accessibility calculation and mapping of known IgE-binding epitope sequences. It has been found that Ala, Asn, Gly and Lysine have a high propensity to occur in the IgE-binding sites on the surface of allergenic proteins (Oezguen et al., 2008).

Finally, techniques of structural bioinformatics have also been applied to assess features critically required for allergenicity and cross-reactivity. This has been done by analyzing the predicted structure of protein T1, the naturally occurring non-allergenic member of the Bet v 1 allergen family (Ghosh and Gupta-Bhattacharya, 2008). Protein T1 shows considerable sequence similarity with the proteins of Bet v 1 allergen family, but it is neither allergenic, nor cross-reactive to the Bet v 1 group (Laffer et al., 2003). Comparative molecular modeling, solvent accessibility calculations and mapping of surface electrostatic potential showed substantial difference in antigenic surface that can be responsible for the loss of cross-reactivity. Solvent-accessible surface area and electrostatics calculations were done using program DSSP (dictionary of secondary structure of proteins) and program APBS (Adoptive Poissopn-Blotzman solver) respectively (Baker et al., 2001; Kabsch and Sander, 1983). Although, as suggested by ligand docking, it should be able to perform its biological function as a brassinosteroid carrier.

5. Conclusion

Allergy is a world-wide problem. Allergic symptoms are elicited following exposure to a structurally diverse group of proteins known as allergens. Understanding allergenicity at the molecular level has wide application in food safety and in treating allergic diseases. What makes a protein allergic is not yet understood. However, advanced tools of bioinformatics have been applied to address this problem. It has been found that allergens are usually foreign proteins with few/no bacterial homologue. They are clustered into few protein families (associated with a limited number of protein domains) opposing the idea that any protein can be allergenic. Methods have been developed to predict probable allergenicity from protein sequence, although more works need to be done for better and more precise prediction. The structural difference between IgG-binding and IgE-binding epitopes is still not very clear, but homology modelling in combination with residue-wise solvent-accessibility of monomers and biological assemblies of allergens certainly gives valuable information about antigenic determinants on protein allergens.

6. References

Aalberse R. C. (2000) Structural biology of allergens. *J Allergy Clin Immunol* 106, 228-38.

Baker N. A., Sept D., Joseph S., Holst M. J. and McCammon J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98, 10037-41.

Banerjee B., Greenberger P. A., Fink J. N. and Kurup V. P. (1999) Conformational and linear B-cell epitopes of Asp f 2, a major allergen of Aspergillus fumigatus, bind differently to immunoglobulin E antibody in the sera of allergic bronchopulmonary aspergillosis patients. *Infect Immun* 67, 2284-91.

Chapman M. D., Pomes A., Breiteneder H. and Ferreira F. (2007) Nomenclature and structural biology of allergens. *J Allergy Clin Immunol* 119, 414-20.

Chatchatee P., Jarvinen K. M., Bardina L., Beyer K. and Sampson H. A. (2001) Identification of IgE- and IgG-binding epitopes on alpha(s1)-casein: differences in patients with persistent and transient cow's milk allergy. *J Allergy Clin Immunol* 107, 379-83.

Crameri R., Faith A., Hemmann S., Jaussi R., Ismail C., Menz G. and Blaser K. (1996) Humoral and cell-mediated autoimmunity in allergy to Aspergillus fumigatus. *J Exp Med* 184, 265-70.

Emanuelsson C. and Spangfort M. D. (2007) Allergens as eukaryotic proteins lacking bacterial homologues. *Mol Immunol* 44, 3256-60.

Finn R. D., Mistry J., Tate J., Coggill P., Heger A., Pollington J. E., Gavin O. L., Gunasekaran P., Ceric G., Forslund K., Holm L., Sonnhammer E. L., Eddy S. R. and Bateman A. (2010) The Pfam protein families database. *Nucleic Acids Res* 38, D211-22.

Fluckiger S., Fijten H., Whitley P., Blaser K. and Crameri R. (2002) Cyclophilins, a new family of cross-reactive allergens. *Eur J Immunol* 32, 10-7.

Gajhede M., Osmark P., Poulsen F. M., Ipsen H., Larsen J. N., Joost van Neerven R. J., Schou C., Lowenstein H. and Spangfort M. D. (1996) X-ray and NMR structure of Bet v 1, the origin of birch pollen allergy. *Nat Struct Biol* 3, 1040-5.

Ghosh D. and Gupta-Bhattacharya S. (2008) Structural insight into protein T1, the non-allergenic member of the Bet v 1 allergen family-An in silico analysis. *Mol Immunol* 45, 456-62.

Glaser A. G., Limacher A., Fluckiger S., Scheynius A., Scapozza L. and Crameri R. (2006) Analysis of the cross-reactivity and of the 1.5 A crystal structure of the Malassezia sympodialis Mala s 6 allergen, a member of the cyclophilin pan-allergen family. *Biochem J* 396, 41-9.

Goodman R. E. (2008) Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. *Food Chem Toxicol* 46 Suppl 10, S24-34.

Hagler A. T., Lifson S. and Dauber P. (1979) Consistant force field studies of intermolecular forces in hydrogen-bonded crystals.2. A benchmark for the objective comparison of alternative force fields. *J. Amer. Chem. Soc.* 101, 5122-5130.

Hubber S. (1992) ACCESS: A program for calculating Accessibilities. *Dept. of Biochemistry and Molecular Biology, Univ Col of London*.

Ivanciuc O., Garcia T., Torres M., Schein C. H. and Braun W. (2009a) Characteristic motifs for families of allergenic proteins. *Mol Immunol* 46, 559-68.

Ivanciuc O., Schein C. H. and Braun W. (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* 31, 359-62.

Ivanciuc O., Schein C. H., Garcia T., Oezguen N., Negi S. S. and Braun W. (2009b) Structural analysis of linear and conformational epitopes of allergens. *Regul Toxicol Pharmacol* 54, S11-9.

Jenkins J. A., Breiteneder H. and Mills E. N. (2007) Evolutionary distance from human homologs reflects allergenicity of animal food proteins. *J Allergy Clin Immunol* 120, 1399-405.

Kabsch W. and Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-637.

Ladics G. S., Cressman R. F., Herouet-Guicheney C., Herman R. A., Privalle L., Song P., Ward J. M. and McClain S. (2011) Bioinformatics and the allergy assessment of agricultural biotechnology products: Industry practices and recommendations. *Regul Toxicol Pharmacol*. 60, 46-53.

Laffer S., Hamdi S., Lupinek C., Sperr W. R., Valent P., Verdino P., Keller W., Grote M., Hoffmann-Sommergruber K., Scheiner O., Kraft D., Rideau M. and Valenta R.

(2003) Molecular characterization of recombinant T1, a non-allergenic periwinkle (Catharanthus roseus) protein, with sequence similarity to the Bet v 1 plant allergen family. *Biochem J* 373, 261-9.

Laskowski R. A. (1993) PROCHECK: A program to check the stereochemistry of protein structure. *J. Appl. Cryst.* 26, 283-291.

Mari A., Scala E., Palazzo P., Ridolfi S., Zennaro D. and Carabella G. (2006) Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. *Cell Immunol* 244, 97-100.

Marti-Renom M. A., Stuart A. C., Fiser A., Sanchez R., Melo F. and Sali A. (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29, 291-325.

Nicholls A., Sharp K. A. and Honig B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11, 281-96.

Nordlee J. A., Taylor S. L., Townsend J. A., Thomas L. A. and Bush R. K. (1996) Identification of a Brazil-nut allergen in transgenic soybeans. *N Engl J Med* 334, 688-92.

Oezguen N., Zhou B., Negi S. S., Ivanciuc O., Schein C. H., Labesse G. and Braun W. (2008) Comprehensive 3D-modeling of allergenic proteins and amino acid composition of potential conformational IgE epitopes. *Mol Immunol* 45, 3740-7.

Radauer C. and Breiteneder H. (2006) Pollen allergens are restricted to few protein families and show distinct patterns of species distribution. *J Allergy Clin Immunol* 117, 141-7.

Radauer C. and Breiteneder H. (2007) Evolutionary biology of plant food allergens. *J Allergy Clin Immunol* 120, 518-25.

Roy D., Ghosh D. and Gupta-Bhattacharya S. (2003) Homology modeling of allergenic cyclophilins: IgE-binding site and structural basis of cross-reactivity. *Biochem Biophys Res Commun* 307, 422-9.

Saitou N. and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4, 406-25.

Sali A. and Blundell T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234, 779-815.

Shin D. S., Compadre C. M., Maleki S. J., Kopper R. A., Sampson H., Huang S. K., Burks A. W. and Bannon G. A. (1998) Biochemical and structural analysis of the IgE binding sites on ara h1, an abundant and highly allergenic peanut protein. *J Biol Chem* 273, 13753-9.

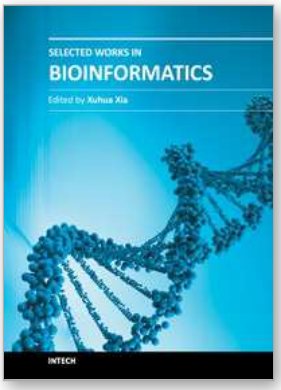
Silvanovich A., Bannon G. and McClain S. (2009) The use of E-scores to determine the quality of protein alignments. *Regul Toxicol Pharmacol* 54, S26-31.

Stanley J. S., King N., Burks A. W., Huang S. K., Sampson H., Cockrell G., Helm R. M., West C. M. and Bannon G. A. (1997) Identification and mutational analysis of the immunodominant IgE binding epitopes of the major peanut allergen Ara h 2. *Arch Biochem Biophys* 342, 244-53.

Tamura K., Dudley J., Nei M. and Kumar S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24, 1596-9.

Zhang C. T. (1997) Relations of the numbers of protein sequences, families and folds. *Protein Eng* 10, 757-61.

Zuckerkandl E. and Pauling L. (1965) Evolutionary divergence and convergence in proteins. *in Evolving Genes and Proteins, edited by V. Bryson and H.J. Vogel. Academic Press, New York.*, 97-166



Selected Works in Bioinformatics

Edited by Dr. Xuhua Xia

ISBN 978-953-307-281-4
Hard cover, 176 pages
Publisher InTech
Published online 19, October, 2011
Published in print edition October, 2011

This book consists of nine chapters covering a variety of bioinformatics subjects, ranging from database resources for protein allergens, unravelling genetic determinants of complex disorders, characterization and prediction of regulatory motifs, computational methods for identifying the best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments, functional characterization of inherently unfolded proteins/regions, protein interaction networks and flexible protein-protein docking. The computational algorithms are in general presented in a way that is accessible to advanced undergraduate students, graduate students and researchers in molecular biology and genetics. The book should also serve as stepping stones for mathematicians, biostatisticians, and computational scientists to cross their academic boundaries into the dynamic and ever-expanding field of bioinformatics.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Debajyoti Ghosh and Swati Gupta-Bhattacharya (2011). Allergen Bioinformatics: Recent Trends and Developments, Selected Works in Bioinformatics, Dr. Xuhua Xia (Ed.), ISBN: 978-953-307-281-4, InTech, Available from: http://www.intechopen.com/books/selected-works-in-bioinformatics/allergen-bioinformatics-recent-trends-and-developments

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821