

Information Management for the Study of Allergies

Vladimir Brusic*

Australian Centre for Plant Functional Genomics, School of Land and Food Sciences, and Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

Abstract: Microarrays and other large-scale screening technologies produce quantities of increasingly complex allergy data. These data link molecular and clinical measurements and observations and provide fertile ground for improving our understanding of the processes involved in allergic reactions. Information technology is employed in gathering, storage, retrieval and analysis of these data. The increasing proportion of allergy data are generated from genomics and proteomics approaches. The major activity focuses on characterization of allergens including IgE reactivity, structural properties, and mapping of IgE and T-cell epitopes. Because of the complexity of allergy data, their utilization requires bioinformatics approaches. Allergen data are stored in the general and specialist databases. At least a dozen of important allergen databases and data repositories have been developed to date. These data are analysed using general and specialist bioinformatics tools. The major applications of bioinformatics include support for allergen characterization, assessment of allergenicity, and identification of allergic cross-reactivity. These applications in turn support the development of vaccines and therapies for allergic disease. In this article we review allergen databases and tools for the analysis of allergens, and discuss the new directions in the field supported by large scale screening involving genomics, proteomics, and bioinformatics support.

Keywords: Allergen databases, allergenicity prediction, allergic cross-reactivity.

INTRODUCTION

Allergies represent a major health problem in the developed world and affect, in some form, up to one third of the general population [1]. The majority of environmental antigens are well-tolerated, but in some individuals these antigens can induce aberrant activation of immune pathways resulting in allergic responses. Allergy is a type of hypersensitive reaction of an organism to antigens which are normally harmless in other people who experience similar quantity and exposure. Although our understanding of the immune regulation is rapidly increasing, the fine details of the processes leading to allergic reactions are still unknown. The most important factors affecting allergic disease are genetic predisposition and environmental effects such as exposure to allergens, air pollution, infections, and diet. Improved understanding of immune pathways, molecular components, and cellular effects of allergic responses is essential for the development of new allergy drugs and treatments. These developments will increasingly depend on the access to accurate and comprehensive data on molecular processes involved in allergic reaction and related clinical observations.

Allergy-related data [2] and the overall body of allergy knowledge are increasing rapidly. The advances in infocomm technologies enable instant access to these data as well as sophisticated analyses. Technological advances in genomics [3, 4], and proteomics [5] enable large-scale screening of allergy related genes and their products. These new data, large in numbers but not well explained, keep stimulating the developments in both information management (bioinformatics) and traditional allergy research aimed at more detailed characterization of the allergic

of allergy-related data and improving the efficiency of allergy research. They provide essential support for generation and interpretation of data generated by large scale screening from genomics and proteomics projects [6], allergen characterization [7], and T-cell epitope screening [8], among others. Allergy-related data are growing in both quantity and complexity.

Allergens originate from the natural environment or from chemical contaminations. The most frequent natural allergen sources are from pollen, mould, fungi, bacteria, house dust mites, epidermis of house pets, food products, and insect venom. The allergens originating from the natural environment are usually proteins of molecular weight greater than 10 kDa. Many allergens are very stable and are resistant to pH changes and high temperature, although there are exceptions. Protein allergens belong to structural families that share sequence similarity, which forms a basis for the study of allergenicity and allergic cross-reactivity. Modern allergy research largely depends on the rapid access to data sources and advanced tools for the analysis of data. Allergen-related data are scattered across various molecular databases, but their annotations typically lack allergy-specific annotations. A number of specialized databases have appeared in the allergy field over the last few years [2]. Bioinformatic tools are used to probe databases for analysis and comparison of sequences, structure, and functional properties [9]. The sequence analysis and comparison tools are used for classification and analysis of confirmed and potential allergens. Predictive tools are used to support structural and functional characterisation of genes and proteins, targeting those with allergenic potential.

The informatics revolution is gradually changing the science and practice of allergology. Allergy informatics provides a basis for integration of the research and clinical arms of the field and also provides support for risk assessment, diagnosis, and therapy for allergic disease. The

*Address correspondence to this author at the Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia; Tel: +61 7 3346 2611; Fax: +61 7 3346 2101; E-mail: v.brusic@uq.edu.au

accumulation of data and information contains clues for understanding the roles of genes and proteins in allergies. This information is complex and includes simple and network interactions, and pathways involved in the development and progression of allergic disease. Three major aspects of allergy data management are developing fast. They include data generation, storage and retrieval (databases), and computational analysis tools. This article contains a review of the bioinformatics framework for allergy research and a discussion of future developments in the field.

ALLERGY-RELATED DATA

Allergy-related data are generated from a variety of sources. Bioinformatics are particularly important for data-intensive technologies. This review focuses on four fields of allergy research that are dependent on bioinformatics for data generation: genomics, proteomics, allergen characterization, and T-cell epitope screening.

Genomics

Genomics refers to the simultaneous study of the structure and function of large numbers of genes – their structure, function, regulation, and interactions. Microarrays (cDNA or oligonucleotide) are used for detection of gene expression from samples that represent specific tissues and specific conditions. Cytokines, chemokines, cytokine and chemokine receptors, major histocompatibility complex molecules, and various transcription factors play roles in allergic reactions, and thus represent molecular targets for treatments [6]. The familial aggregation of allergic disease indicates the importance of genetic factors in allergic disease [4]. Genomics helps our understanding of the molecular mechanisms of allergic disease, by helping decipher the complex interplay of genetic and environmental factors through careful selection of disease and control samples. Distinct gene expression profiles have been identified in T helper cells of types 1 and 2 [10] and differential expression of genes was observed in allergen-specific Th2 cells [11]. cDNA arrays have been used for identification of altered, as compared to healthy subjects, gene expression in peripheral blood mononuclear cells from asthma and allergy patients [12]. Similar studies helped identify susceptibility genes for allergic disease in experimentally induced allergies [13]. More than twenty susceptibility loci for allergic disease have been identified [see 6] providing a fertile ground for the analysis of related genes and network-type interactions. Oligonucleotide microarrays used in a study of the complete transcript of human mast cells helped identify selective expression of nearly 100 transcripts that have potential role in allergic inflammation [14]. The comparison of specific cDNA with genomic sequences was used for the study of gene families which helped characterization of a major peanut allergen Ara h 2 [15].

The huge quantity of data produced by structural, functional, and comparative genomics require bioinformatic analyses for their interpretation. Bioinformatics tools are essential for identification of statistically significant changes in gene expression, noise filtering, and co-variation. Several methods for microarray analysis have been described in [16].

Combined with other methods, microarrays provide a platform for linking genetic variation, phenotypic factors, environmental conditions, and various time points in the study of allergic disease.

Proteomics

Proteomics is the study of proteins focusing on a large-scale identification of their structures and functions [see 17]. Proteomics is much more complicated than genomics because protein expression is constantly changing and it will vary depending on the localization, time, and physiological conditions. Proteomics approaches including separation techniques, mass fingerprinting and bioinformatics analysis were used for rapid identification of IgE-binding proteins from complex mixtures of Japanese cedar pollen proteins [18]. This and similar studies have clearly shown that a) allergen variation is far greater than currently captured in literature and databases, b) a significant proportion of protein allergens from mixed allergen sources has yet to be identified, and c) individuals show distinctive response patterns to allergen mixtures. New allergens can be identified using sera of allergic individuals followed by protein identification [e.g. 19]. Proteomic techniques are still at early stages, but the potential for generation precise data on allergens, allergen variation, and individual allergic responses is enormous [20]. The combination of genomics and proteomics data and sophisticated bioinformatics and statistical tools will enable concurrent analysis of genetic, phenotypic, and environmental factors. This results will help translate molecular level results to clinical applications in both diagnostic, monitoring, and therapeutic applications.

Allergen Characterization

Proteomic approaches enable the initial identification of allergens, their variants, and individual response patterns. Additional investigations are required for further assessment of allergenic potential of proteins. The studies of protein allergenic potential focus on three main aspects, the inherent protein allergenicity, allergic cross-reactivity, and clinical symptoms [21]. The inherent allergenicity is the likelihood that a particular allergen would induce an IgE antibody or a T-cell epitope. Mapping of both B- and T-cell epitopes [22, 23, 24, 25] focus on defining recognition sites on allergens and can be done on a large scale. Allergic cross-reactivity is the ability of an IgE clonotype or a T-cell clone, which were induced by one allergen, to react with a different allergen. Studies of insect venom allergens indicate that cross-reactivity would typically occur between allergens that share more than 70% sequence identity [26]. Allergic cross-reactivity has been explored for definition of immunotherapies in food allergy, but those studies are complicated and results are often controversial [27]. These studies require careful design including double-blind, placebo-controlled trials [28] and the results should be used with caution. Studies of inherent allergenicity and cross-reactivity have applications in the development of allergy immunotherapies and vaccines [29, 30, 31, 32, 33]. Extensive clinical and functional data for common allergens have been reported and can be found in literature and public databases.

Allergen characterization for a single individual can be done directly, but comprehensive screening and characterization of allergens, their isoforms, and cross-reactive allergens is expensive and limited by the availability of sera. Miniaturisation of microarrays and recombinant technologies [23] will help improve the efficiency of screening, but the number of isoforms and cross-reactive targets will also keep increasing complicating the analysis. Although allergenicity is a property of a protein only when it comes in contact with an allergic individual, ie involves complex interactions between the allergen and the immune system, some allergens have high prevalence of responses in allergic individuals. According to the International Union of Immunological Societies (IUIS) Allergen Nomenclature Committee recommendations (www.allergen.org), allergens with a prevalence of IgE reactivity above 5% are included in the nomenclature. The generally accepted respective terms "major" and "minor" allergens describe those allergens that induce response in >50% or <50% of tested patients with the allergen-specific IgE in the given test-system [34]. Determination of IgE prevalence is subject to some uncertainty since it includes different test-system and criteria for selection of patients, geographic region, environmental habits and conditions, among others and the recommendation is the inclusion of a substantial number of patients in the analysis [34]. Some proteins are inherently more allergenic than other proteins and the challenge is the identification of structural properties of allergens that confer allergenicity and allergic cross-reactivity. Identification of allergenic determinants is subject to vigorous research. Allergens must contain at minimum B-cell epitopes, the binding sites of IgE can bind, and T-cell epitopes that induce T helper type 2 responses. Additional properties include stability, certain post-translational modifications, and possible enzymatic activity [35].

Because of the critical involvement of B-cell epitopes, protein allergenicity is determined by its three-dimensional (3D) structure. Understanding of allergen structure thus is important for deciphering the molecular basis of allergenicity. A study of known protein folds of 40 allergens suggested that a general structural feature responsible for allergenicity did not exist [21]. However some common structural features have been determined in allergens. For example, a common structural motif comprising a groove located inside an alpha-beta motif has been identified within diverse allergens (dust mite allergen Der p 1, cysteine protease papain, lipocalin Mus m 1, and ragweed allergen Amb a 5) [36]. In addition four major structural families that potentially conferred allergenicity have been identified [21]: a) antiparallel beta-strands: the immunoglobulin-fold family (grass group 2, mite group 2), serine proteases (mite group 3, 6, and 9), and soybean-type trypsin inhibitor (Ole e 1, grass group 11); b) antiparallel beta-sheets intimately associated with one or more alpha-helices: tree group 1, lipocalin, profilin, aspartate protease (cockroach group 2); c) (alpha+beta) structures, in which the alpha- and beta-structural elements are not intimately associated: mite group 1, lysozyme/lactalbumin, vespid group 5; and d) alpha-helical: nonspecific lipid transfer protein, seed 2S protein, insect hemoglobin, fish parvalbumin, pollen calmodulin, mellitin from bee venom, Fel d 1 chain 1, and serum

albumin. Larger number of allergen structures is currently been assembled in the SDAP database [37].

The unknown 3D structure of allergens can be modeled from homology to proteins of known structure is known. For example, the Der p 1 dust mite allergen, modelled upon the known structure of cysteine proteinases [38], cat cystatin was modelled upon human cystatin structure [39], and the Blo t 13 allergen of the *Blomia tropicalis* mite was modelled upon cytosolic lipid transport proteins [40].

Screening of Allergenic Epitopes

Both B-cell and T-cell epitopes are essential determinants of allergenic responses and are important for understanding allergenic potential of a protein. A number of studies reported systematic identification of both linear and conformational B-cell IgE epitopes on allergen surfaces, including Asp f 4 (fungi) [41], Hev b 6 (rubber) [42], Jug l r (walnut) [43], Chi t 1 (midge) [44], Cry 1 and Cry 2 (Japanese cedar) [18], among others. Mutagenesis studies of IgE epitopes are used for finding low-allergenic variants and minimization of allergic cross-reactivity [45]. IgE epitopes were shown to differ between patients and are highly specific for individuals [18, 22, 46], making engineering of hypoallergens a difficult task. The study of overlapping peptides from peanut allergens Ara h 1, Ara h 2, and Ara h 3 with sera of 77 patients and 15 controls showed that 97% of patients had IgE specific for at least one of these allergens, and 87% showed detectable responses to linear epitopes [22]. Microarray technology using recombinant allergens is expected to replace tests with crude natural extracts and allow comprehensive analysis of individual IgE reactivity profiles [4]. IgE epitope prediction is mainly based on structure modeling [45] or sequence similarity [37]. Knowledge of the structural requirements for effector cell activation may also be used for conversion of allergens into hypoallergenic immunotherapeutics by reorientation of IgE epitopes [47]. Computational predictions of IgE epitopes are still at very early stages, and microarray-based identification is currently the method of choice for their mapping.

CD4+ T helper 1 (Th1) cells produce cytokines that activate macrophages and are involved in delayed-type hypersensitivity, while Th2 cells mediate antibody responses [48] and are important for IgE-mediated allergic inflammatory responses. Allergen-specific T-cell clones have been studied and T-cell epitopes have been identified for a variety of allergens [49]. However, this information is often of limited use because the human leukocyte antigen (HLA - human MHC) specificity of these T-cell epitopes has not been identified. HLA molecules show huge diversity with than 500 variants of HLA-DR, -DP, and -DQ molecules characterized to date [50]. Human individuals have different combinations of HLA molecules and T-cell responses of individuals will often target different T-cell epitopes, even if they share the same HLA alleles. The influence of HLA to allergic responses is even more complicated because certain HLA alleles have been reported to confer susceptibility or protection against specific allergies [51]. Experimental T-cell epitope mapping involves binding or functional assays of multiple overlapping peptides. Systematic T-cell epitope mapping of an allergen requires an extensive experimental

effort even for a study of a single HLA molecule. The systematic experimental characterisation of T-cell epitopes of a particular allergen for multiple HLA molecules is impractical and these studies are generally performed through combination of computational predictions and experimental validation [8]. However, highly promiscuous T-cell epitopes have been identified in allergens [52] and some of them are involved in allergic cross-reactivity [25]. Such promiscuous T-cell epitopes are promising targets in the design of immunotherapies for allergy. Computational prediction methods significantly reduce the number of experiments required for mapping of T-cell epitopes. Binding of peptide to a HLA molecule is a prerequisite for T-cell recognition and thus can be used for prediction of candidate T-cell epitopes [53].

ALLERGEN DATABASES

The diversity and quantity of allergy-related data is increasing, mainly from the microarray and other large-scale screening studies. Clinical and functional data for common allergens have been reported extensively in the literature while public sequence databases contain entries of allergen sequences [54, 55], structures [56], or T-cell epitopes [57, 58]. However, these data sources are neither complete nor up-to-date with respect to allergens. The number of characterized protein allergens is increasing steadily. The Allergen Nomenclature Sub-Committee of the International Union of Immunological Societies (IUIS) certifies allergens and maintains a list of official protein allergens (www.allergen.org). Some 450 protein allergens and 220 of their isoforms had been classified, as of June 2004. This may represent only a small fraction of the total as individual allergens may have many isoforms [59, 60]. While this certified list contains many characterized allergens it is not exhaustive – an increasing number of unlisted allergens can be found in the literature. Major public databases that contain records of allergen sequences are listed in Table 1. These databases generally do not contain detailed structural and functional information for a particular allergen. A number of specialized allergy-related databases and data

Table 1. The List of General Database Sources that Contain Data of Allergen Sequences. This Table Contains Access Addresses for Seven Major Databases. Access Information to Various Biological Databases is Available at the Nucleic Acids Research Database Collection (www3.oup.co.uk/nar/database/)

Database	URL
NCBI GenBank/GenPept	www.ncbi.nlm.nih.gov
EMBL	www.ebi.ac.uk/embl
DDBJ	www.ddbj.nig.ac.jp
SWISS-PROT	www.expasy.ch
UNIPROT	www.uniprot.org
PIR	www-nbrf.georgetown.edu
PDB	www.rcsb.org/pdb

repositories have appeared recently. Earlier allergen data repositories and databases (IUIS, SWISS-PROT AllergenIndex, BIFS, CSL, FARRP, Protall, Allallergy, Asthma & Allergy), see Table 2, are in a tabular form providing lists of allergen names, and, to various degrees, descriptions of allergen structural or functional properties, accession numbers and links to entries in major public databases, and literature references, and with an exception of FARRP [61], do not include analysis tools.

The lists of allergens and isoallergens of clinical relevance have been developed and maintained by the Allergen Nomenclature Sub-committee of the IUIS. SWISS-PROT index of allergen sequences lists more than 280 SWISS-PROT entries of 235 IUIS-designated protein allergen sequences. The list contains allergen names, accession numbers, and links to SWISS-PROT entries. The International Food Biotechnology Council (IFBC) and the ILSI Allergy and Immunology Institute sponsored a study on a decision-tree approach for assessment of potential allergenicity of transgenic proteins in crop plants [77]. This

Table 2. Major Allergen-Specific Data Sources

Database	URL	Version or Release Date
IUIS	www.allergen.org	March 2004
SWISS-PROT AllergenIndex	www.expasy.ch/cgi-bin/lists?allergen.txt	May 2004
BIFS	www.iit.edu/~sgendel/fa.htm	Oct. 2003
CSL	www.csl.gov.uk/allergen	Sep. 2001
FARRP	www.allergenonline.com	Ver. 4, Apr. 2003
Protall	www.ifr.bbsrc.ac.uk/Protall	Jun. 2001
Allallergy	www.allallergy.net	2002
Asthma&Allergy	Cooke.gsf.de/asthmagen/main.cfm	Dec. 2003
Latex allergens	dmd.nih.gov/jp/latex/allergen-c.html	-
SDAP	http://fermi.utmb.edu/SDAP	May 2004
Allergome	www.allergome.org	May 2003
Allergen	research.i2r.a-star.edu.sg/Templar/DB/Allergen	Sep. 2003

Table 3. Protein Sequence Content for Four Data Sources (as of October 2003). These Sources are Representative of the Number of Allergens Used for Studies of Allergenicity

	IUIS	SDAP	SWISS-PROT	ALLERGEN
No protein sequence	234	200	0	0
With protein sequence	330	536	283	524
Total	564	736	283	524

study included a list of protein allergen sequences which is available in print only, and has not been updated since publication. The Biotechnology Information for Food Safety (BIFS) web site [62] contains a non-redundant allergen protein list aimed at facilitating sequence comparisons in assessing potential allergenicity of proteins used in foods. CSL allergen database contains food allergens compiled and maintained at the Food Safety and Quality Directorate of The Central Science Laboratory at Sand Hutton York, UK. The FARRP Allergen Database [61] is a searchable database containing more than 650 unique sequences of known and putative allergens. The PROTALL database contains data on allergenic potential of plant food proteins, their molecular structure and biological activities. The ALLALLERGY database comprises more than 4500 entries. It can be queried using the name or the allergen category (e.g. cereals, fish, fruit, pollens) and returns the background information, the function of the substance and the adverse reactions. Asthma and allergy gene database [63] provides access to data on linkage and mutation studies data for asthma and allergy related traits. A more detailed description of these databases is available in [2].

SDAP Database

SDAP (fermi.utmb.edu/SDAP) [64] provides access to the sequences, structures, and IgE epitopes of protein allergens. It allows access to these data, extraction, and analysis of allergen sequences. A SDAP entry contains allergen name, source organism, protein sequence, IgE epitopes, and literature references and easy links to the major protein. The data in SDAP originate from SWISS-PROT/TrEMBL, PIR-ALN, NCBI Taxonomy Browser, PDB SDAP, and literature. These sources of data, if available, are accessible through hyperlinks in the SDAP interface. SDAP has a rich set of tools for viewing and analysis of allergen data. They include sequence analysis tools FASTA [65], a peptide matching tool, peptide similarity by properties, and entries listing tool. In addition, SDAP has a number of protein structure analysis tools – for calculation of accessible surface area, low-energy conformations, homology modeling, and structure-based sequence alignment. SDAP can be used for prediction of allergenicity, but does not conform to the FAO/WHO guidelines for the assessment of allergenicity [66] and allergic cross-reactivity [67].

Allergome

Allergome - A Database of Allergenic Molecules (www.allergome.org) contains information on allergens that cause IgE-mediated disease (e.g. anaphylaxis, asthma, atopic dermatitis, conjunctivitis, rhinitis, and urticaria). The data

have been selected from more than 4000 scientific articles including name, links to sequence data, sources and tissues of allergens, routes of exposure, prevalence data, and references, tests and reagents, and basic biochemical data. ALLERGOME allows keyword searching and has four divisions: allergenic sources without identified allergens, IUIS official nomenclature allergens, allergens not yet sanctioned by IUIS, and isoforms. Molecules causing non-IgE allergic diseases (e.g. contact dermatitis, drug allergy, celiac disease), and other substances causing pseudo-allergic reactions or intolerance (e.g. food intolerance) are excluded from the Allergome database.

ALLERGEN DATABASE

The ALLERGEN database (research.i2r.a-star.edu.sg/Templar/DB/Allergen) facilitates retrieval of information on official allergens and data analysis in the assessment of protein allergenicity. ALLERGEN records were compiled from major public databases and literature using the IUIS Allergen List as reference reference. This database focuses on prediction of allergenicity and allergic cross-reactivity. It aims at collecting all allergen sequences from the official IUIS list and currently contains approximately 90% of the total number. ALLERGEN database is quality controlled: it is non-redundant has high content of protein sequences (Table 3). Allergen database has information on allergen proteins, source organisms, and allergic cross-reactivity. More than one hundred ALLERGEN database records have information on observed allergic cross-reactivity, extracted from the literature. Search tools in the Allergen database include keyword search, and sequence alignment tool BLAST [68]. In addition, two new analysis tools have been developed for use with the allergen database: XR BLAST, which combines standard BLAST output with allergen cross-reactivity data, and ALLERPREDICT that performs bioinformatics assessment of potential allergenicity and allergic cross-reactivity of a protein, using FAO/WHO guidelines [66, 67].

Comments on Allergy Databases

Each allergen-related data source was constructed for a specific purpose and each has different criteria for the inclusion of entries. The IUIS list of proteins with assigned allergen nomenclature, contains only proteins showing IgE reactivity, whereas the BIFS tables contains also proteins involved in celiac disease. Some of these data sources (such as the CSL allergen database) have a large degree of redundancy, while others (such as ALLERGEN) are non-redundant. It has been estimated that there are more than 2500 reported allergenic proteins [2], but only some 600

have been added to the official list of allergens. The quality control of allergen data in public databases, both general and specialised is lacking. Some allergen sequences have been published in the literature, but are not available in sequence databases.

A reference allergen database was proposed in [2]. The reference allergen database should contain a comprehensive, accurate and expert-annotated data and a set of specialist tools for extraction of complex information and analysis of allergens. The reference allergen database should:

- Comprise comprehensive non-redundant entries of allergens, with detailed description of allergen properties,
- Comply with the official classification of allergens (IUIS, nomenclature),
- Have a set of tools for classification, characterization, and analysis of new and known allergens,
- Enable detailed analysis of existing allergens and identification of common sequence and structural properties and characterisation of those properties that are likely to be related to allergenicity,
- Enable identification of B-cell and T-cell epitopes in allergens, and
- Support identification of cross-reactivity among known and novel allergens.

A comprehensive reference database will support research on molecular basis of allergies, design of immunotherapies, risk assessment in patients, and assessment of the safety of novel foods. Such uses will require the development and implementation of sophisticated bioinformatics tools.

BIOINFORMATIC ANALYSIS OF ALLERGENS

Bioinformatic tools are available for study of allergens [9]. The main applications include prediction of allergenicity [69], allergen cross-reactivity [69], and T-cell epitopes [8]. Characterisation of allergens is increasingly performed through genomics and proteomics approaches supported by bioinformatics analyses. The main approaches include sequence comparison [70], structural approaches employing secondary [71], and tertiary structure analysis [21, 36, 72], and various specialist tools for analysis of allergens and their properties. Sequence and structure analysis are sometimes used in combination for detailed characterization of allergens and their properties [73]. The tools for T-cell epitope prediction have been reviewed in [74] and their applications to the study of allergens in [9]. Tools for sequence comparison and structural analysis and their use for study of allergens have been described elsewhere [9] and in this section we focus on the latest specialist allergen analysis tools.

A number of allergen analysis tools have appeared in the last two years and we expect increased activity in this area in the coming years. Introduction of novel foods is surrounded by controversy and among the main objections are safety issues [75], including potential introduction of new allergens into the food chain [76]. The world bodies are working on defining the framework and standards [66, 67]. Criteria and

rules for identification of potential allergenicity based on sequence identity or similarity have been proposed. The earliest criterion for significant similarity was the identity of eight contiguous amino acids between a query protein and any of the known allergens [77]. The minimal length of IgE epitopes in the Ara h 2 peanut allergen [78] were shown to be six amino acids. Such considerations were behind the second criterion for establishment of potential allergenicity which requires identity of six contiguous amino acids [see 66]. The *in vitro* mutation studies showed that amino acids of IgE epitopes of allergens are not necessarily contiguous, and that replacement of amino acids of IgE epitopes may increase, decrease, or not affect the affinity of IgE specific for that site [79]. Additional criterion requiring 35% identity over eighty amino acids to a known allergen was introduced as a threshold for potential allergenicity based on high probability of allergic cross-reactivity [67].

Analyses of allergen data have indicated that these criteria do not provide particularly accurate assessment of allergenicity. The analysis of selected non-allergenic food proteins and randomly selected maize proteins [61] demonstrated that six contiguous amino acid identity rule produces a large number of irrelevant matches. It was shown [80] that two thirds of SwissProt proteins would be ranked as potential allergens by using the six amino acids identity rule. Data on diversity of IgE responses to mutated B-cell epitopes [79] and the diversity of individual responses to allergens [18] indicate that sequence comparison, being only an approximate approach, is not sufficient for capturing the determinants of allergenicity. A number of novel methods have been reported. The ALLALLERGY site enables FASTA searches against a list of some 800 allergens. Prediction of allergenicity based on six or more contiguous amino acid identity, as well as minimum 35% identity over eighty amino acids segment against some 600 official allergens can be performed at the ALLERGEN database site. A different sequence similarity search method, based on five descriptors (dimensional and of amino acid properties), has been implemented in the SDAP database. This method is useful for identification of cross-reactive IgE epitopes, as demonstrated by rapid identification of IgE sites on known shellfish and insect allergens, similar to an IgE epitope from shrimp tropomyosin. A learning system [81] that uses FASTA3 algorithm [82] and the k-Nearest-Neighbour classification was applied to learning allergenic determinants useful for prediction of allergenicity. This system was further enhanced by introduction of additional learning algorithms [83]. The two extracted features are the alignment score and the alignment length which are used to map a sequence to the prevalent class in the nearest set of known allergens. For improvement of allergenicity prediction, relative to the FAO/WHO guidelines, a method based on motif identification has been developed by Stadler and Stadler [80]. They used the MEME motif discovery tool [84] for extraction of allergen-specific motifs, and produced 52 statistically relevant motifs that matched 644 of 779 allergens in their database. Finally, a method using clustering of allergens, combined with the wavelet analysis was used for identification of conserved motifs, followed by building Hidden Markov Profiles for prediction of potential allergens [85]. Although these latest methods use more sophisticated algorithms (than the simple rules that are officially

recommended), their utility is limited by the small number of allergen sets that they use. Future developments will largely depend on the availability of large and well characterized sets of allergens.

CONCLUSION AND DISCUSSION

Allergen data are rapidly increasing in quantity, complexity, and accessibility. Increasing amounts of allergen data and accumulating knowledge in the allergy field provide a rich resource in support of basic and applied allergy research. Significant preliminary work has already been done in collection and classification of allergens. New computational algorithms are heralding an era where comprehensive computational analysis will complement large-scale experimental research. The necessary infrastructure in support of biotechnology-driven allergy research include the development of a comprehensive and expansive database of allergens, and definition of high quality test sets which must include allergens and carefully selected non-allergens. Non-allergens must include both proteins that do not look as allergens, but also hypoallergenic sequences related to known allergens. We have previously [2, 3] identified the need for a central reference allergen database integrated with specialist search and analysis tools that could provide a core resource for the study of allergens. This, in turn, would enhance the growth and development of other allergy-related databases that specialize in various aspects of allergen-related studies. The computational infrastructure and high-throughput technologies for experimental measurement of allergens and allergic responses will mark the developments in the allergy research. Practical applications of these developments will enable new allergy treatments including allergen modification and engineering [29, 86], vaccines [30], and immunotherapies [87, 88], offering a great hope for the important clinical field where currently the treatments are predominantly symptomatic (allergen avoidance, antihistamines, corticosteroids, or injectable epinephrine) [89, 90], rather than targeted at the cause.

REFERENCES

- Ono, S.J. *Annu. Rev. Immunol.*, **2000**, *18*, 347-366.
- Brusic, V.; Millot, M.; Petrovsky, N.; Gendel, S.M.; Gigonza, O.; Stelman, S.J. *Allergy*, **2003**, *58*, 1093-1100.
- Cookson, W. *Immunol. Rev.*, **2002**, *190*, 195-206.
- Deinhofer, K.; Sevcik, H.; Balic, N.; Harwanegg, C.; Hiller, R.; Rumpold, H.; Mueller, M.W.; Spitzauer, S. *Methods*, **2004**, *32*, 249-254.
- Pineiro, C.; Barros-Velazquez, J.; Vazquez, J.; Figueras, A.; Gallardo, J.M. *Proteome Res.*, **2003**, *2*, 127-135.
- Toda, M.; Ono, S.J. *Immunology*, **2002**, *106*, 1-10.
- Puc, M. *Ann. Agric. Environ. Med.*, **2003**, *10*, 143-149.
- de Lalla, C.; Starniolo, T.; Abbruzzese, L.; Hammer, J.; Sidoli, A.; Sinigaglia, F.; Panina-Bordignon, P. *J. Immunol.*, **1999**, *163*, 1725-1729.
- Brusic, V.; Petrovsky, N.; Gendel, S.M.; Millot, M.; Gigonza, O.; Stelman, S.J. *Allergy*, **2003**, *58*, 1083-1092.
- Hamalainen, H.; Zhou, H.; Chou, W.; Hashizume, H.; Heller, R.; Lahesmaa, R. *Genome Biol.*, **2001**, *167*, 3057-3063.
- Li, X.D.; Essayan, D.M.; Liu, M.C.; Beatty, T.H.; Huang, S.K. *Genes Immun.*, **2001**, *2*, 88-98.
- Brutsche, M.H.; Brutsche, I.C.; Wood, P.; Mogulkoc, N.; Custovic, A.; Egan, J.; Woodcock, A. *Clin. Exp. Immunol.*, **2001**, *123*, 181-187.
- Karp, C.L.; Grape, A.; Schadt, E.; Ewart, S.L.; Keane-Moore, M.; Cuomo, P.J.; Kohl, J.; Wahl, L.; Kuperman, D.; Germer, S.; Aud, D.; Peltz, G.; Wills-Karp, M. *Nat. Immunol.*, **2000**, *1*, 221-226.
- Saito, H.; Nakajima, T.; Matsumoto, K. *Int. Arch. Allergy Immunol.*, **2001**, *125*, 1-8.
- Viquez, O.M.; Summer, C.G.; Dodo, H.W. *J. Allergy Clin. Immunol.*, **2001**, *107*, 713-717.
- Pawliczak, R.; Shelhamer, J.H. *Allergy*, **2003**, *58*, 973-980.
- Ticha, M.; Pacakova, V.; Stulik, K. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, **2002**, *771*, 343-353.
- Fujimura, T.; Shigeta, S.; Kawamoto, S.; Aki, T.; Masubuchi, M.; Hayashi, T.; Yoshizato, K.; Ono, K. *Int. Arch. Allergy Immunol.*, **2004**, *133*, 126-135.
- Yu, C.J.; Lin, Y.F.; Chiang, B.L.; Chow, L.P. *J. Immunol.*, **2003**, *170*, 445-453.
- Halapi, E.; Hakonarson, H. *Curr. Opin. Pulm. Med.*, **2004**, *10*, 22-30.
- Aalberse, R.C. *J. Allergy Clin. Immunol.*, **2000**, *106*, 228-238.
- Shreffler, W.G.; Beyer, K.; Chu, T.H.; Burks, A.W.; Sampson, H.A. *J. Allergy Clin. Immunol.*, **2004**, *113*, 776-782.
- Laukkanen, M.L.; Makinen-Kiljunen, S.; Isohironen, K.; Haastela, T.; Soderlund, H.; Takkinen, K. *J. Immunol. Methods*, **2003**, *278*, 271-281.
- de Silva, H.D.; Gardner, L.M.; Drew, A.C.; Beezhold, D.H.; Rolland, J.M.; O'Hehir, R.E. *Clin. Exp. Allergy*, **2004**, *34*, 611-618.
- Bohle, B.; Radakovics, A.; Jahn-Schmidt, B.; Hoffmann-Sommergruber, K.; Fischer, G.F.; Ebner, C. *Eur. J. Immunol.*, **2003**, *33*, 3303-3310.
- King, T.P.; Spangfort, M.D. *Int. Arch. Allergy Immunol.*, **2000**, *123*, 99-106.
- Bolhaar, S.T.; Tiemeessen, M.M.; Zuidmeer, L.; van Leeuwen, A.; Hoffmann-Sommergruber, K.; Braijneel-Koomen, C.A.; Taams, L.S.; Knol, E.F.; van Hoffen, E.; van Ree, R.; Knulst, A.C. *Clin. Exp. Allergy*, **2004**, *34*, 761-769.
- Van Ree, R. *Curr. Opin. Allergy Clin. Immunol.*, **2004**, *4*, 235-240.
- Spangfort, M.D. *Arch. Paul Ehrlich Inst. Bundesamt Sera Impfstoffe Frankfurt A. M.*, **2003**, *94*, 151-154.
- Hartl, A.; Weiss, R.; Hochreiter, R.; Scheibhofer, S.; Bauer, R.; Valenta, R.; Ferreira, F.; Leitner, W.; Thalhammer, J. *Arch. Paul Ehrlich Inst. Bundesamt Sera Impfstoffe Frankfurt A. M.*, **2003**, *94*, 279-298.
- Westrich, K.; Focke, M.; Verdino, P.; Goossens, W.; Keller, W.; Twardosz, A.; Mari, A.; Horak, F.; Wiedermann, U.; Hartl, A.; Thalhammer, J.; Spert, W.R.; Valent, P.; Valenta, R. *J. Immunol.*, **2004**, *172*, 5684-5692.
- Akdis, C.A.; Blaser, K. *Arch. Paul Ehrlich Inst. Bundesamt Sera Impfstoffe Frankfurt A. M.*, **2003**, *94*, 219-227.
- Cromwell, O.; Kahlert, H.; Schramm, G.; Suck, R.; Nandy, A.; Weber, B.; Fiebig, H. *Arch. Paul Ehrlich Inst. Bundesamt Sera Impfstoffe Frankfurt A. M.*, **2003**, *94*, 156-161.
- King, T.P.; Hoffman, D.; Lowenstein, H.; Marsh, D.G.; Platts-Mills, T.A.E.; Thomas, W. *Bull. WHO*, **1994**, *72*, 797-806.
- Huby, R.D.J.; Dearman, R.J.; Kimber, I. *Toxicol. Sci.*, **2000**, *55*, 235-246.
- Furmonaviciene, R.; Shakib, F. *Mol. Pathol.*, **2001**, *54*, 155-159.
- Ivanciu, O.; Schein, C.H.; Braun, W. *Bioinformatics*, **2002**, *18*, 1358-1364.
- Topham, C.M.; Srinivasan, N.; Thorpe, C.J.; Overington, J.P.; Kalsheker, N.A. *Protein Eng.*, **1994**, *7*, 869-894.
- Ichikawa, K.; Vailes, L.D.; Pomes, A.; Chapman, M.D. *Int. Arch. Allergy Immunol.*, **2001**, *124*, 55-56.
- Puerta, L.; Kennedy, M.W.; Jimenez, S.; Caraballo, L. *Int. Arch. Allergy Immunol.*, **1999**, *119*, 181-184.
- Ramachandran, H.; Banerjee, B.; Greenberger, P.A.; Kelly, K.J.; Fink, J.N.; Kurup, V.P. *Clin. Diagn. Lab. Immunol.*, **2004**, *11*, 261-265.
- Karisola, P.; Alenius, H.; Mikkola, J.; Kalkkinen, N.; Helin, J.; Penttinen, O.T.; Repo, S.; Reunala, T.; Turjanmaa, K.; Johnson, M.S.; Palosuo, T.; Kulomaa, M.S. *J. Biol. Chem.*, **2002**, *277*, 22656-22661.
- Robotham, J.M.; Teuber, S.S.; Sathe, S.K.; Roux, K.H. *J. Allergy Clin. Immunol.*, **2002**, *109*, 143-149.
- van Kampen V.; Liebers V.; Sander I.; Chen Z.; Baur X.; Raulf-Heimsoth M.; Falkenberg F.W. *Allergy*, **2001**, *56*, 118-125.

- [45] García-Casado, G.; Pacios, L.F.; Díaz-Perales, A.; Sánchez-Monge, R.; Lombardero, M.; García-Selles, F.J.; Polo, F.; Barber, D.; Salcedo, G. *J. Allergy Clin. Immunol.*, **2003**, *112*, 599-605.
- [46] Cramer, R. *Biochem. J.*, **2003**, *376*(Pt 1), c1-2.
- [47] Valenta, R.; Kraft, D. *Immunol. Rev.*, **2001**, *179*, 119-127.
- [48] Romagnani, S. *Immunol. Today*, **1997**, *18*, 263-266.
- [49] van Noerven, R.J.; Ebner, C.; Yssel, H.; Kapsenberg, M.L.; Lamb, J.R. *Immunol. Today*, **1996**, *17*, 526-532.
- [50] Robinson, J.; Malik, A.; Parham, P.; Bodmer, J.G.; Marsh, S.G. *Tissue Antigens*, **2000**, *55*, 280-287.
- [51] Kim, Y.K.; Oh, H.B.; Oh, S.Y.; Cho, S.H.; Kim, Y.Y.; Min, K.U. *Clin. Exp. Allergy*, **2001**, *31*, 110-115.
- [52] Friedl-Hajek, R.; Spangfort, M.D.; Schou, C.; Breiteneder, H.; Yssel, H.; Joost van Noerven, R.J. *Clin. Exp. Allergy*, **1999**, *29*, 478-487.
- [53] Bian, H.; Reidhaar-Olson, J.F.; Hammer, J. *Methods*, **2003**, *29*, 299-309.
- [54] Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Wheeler, D.L. *Nucl. Acids Res.*, **2004**, *32*, D23-26.
- [55] Hainoch, A.; Boeckmann, B.; Ferro, S.; Gasteiger, E. *Brief. Bioinform.*, **2004**, *5*, 39-55.
- [56] Bourne P.E.; Westbrook J.; Berman H.M. *Brief. Bioinform.*, **2004**, *5*, 23-30.
- [57] Brusic, V.; Rudy, G.; Harrison, L.C. *Nucl. Acids Res.*, **1998**, *26*, 368-371.
- [58] Schönbach, C.; Koh, J.L.; Flower, D.R.; Wong, L.; Brusic, V. *Nucl. Acids Res.*, **2002**, *30*, 226-229.
- [59] Swoboda, I.; Jilek, A.; Ferreira, F.; Engel, E.; Hoffmann-Sommergruber, K.; Scheiner, O.; Kraft, D.; Breiteneder, H.; Pittenauer, E.; Schmid, E.; Vicente, O.; Heberle-Boes, E.; Ahoen, H.; Breitenbach, M. *J. Biol. Chem.*, **1995**, *270*, 2607-2613.
- [60] Kettner, A.; Hughes, G.J.; Frutiger, S.; Astori, M.; Roggero, M.; Spertini, F.; Corradin, G. *J. Allergy Clin. Immunol.*, **2001**, *107*, 919-920.
- [61] Hileman, R.E.; Silvanovich, A.; Goodman, R.E.; Rice, E.A.; Holleschak, G.; Astwood, J.D.; Hefle, S.L. *Int. Arch. Allergy Immunol.*, **2002**, *128*, 280-291.
- [62] Gendel, S.M. *Adv. Food Nutr. Res.*, **1998**, *42*, 63-92.
- [63] Wjst, M.; Immervoll, T. *Bioinformatics*, **1998**, *14*, 827-828.
- [64] Ivancic, O.; Schein, C.H.; Braun, W. *Nucl. Acids Res.*, **2003**, *31*, 359-362.
- [65] Pearson, W.R. *Methods Mol. Biol.*, **1994**, *25*, 365-389.
- [66] Joint FAO/WHO Expert Consultation. *Evaluation of Allergenicity of Genetically Modified Foods*. FAO/WHO, Rome, **2001**. (http://www.who.int/foodsafety/publications/biotech/en/ice_jan2001.pdf).
- [67] Codex Alimentarius Commission. *Codex Principles and Guidelines on Foods Derived from Biotechnology*. (<http://ftp.fao.org/codex/standard/en/CodexTexts/BiotechFoods.pdf>). FAO prepublication, **2003**.
- [68] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. *Nucl. Acids Res.*, **1997**, *25*, 3389-3402.
- [69] Brusic, V.; Petrovsky, N. *Trends Immunol.*, **2003**, *24*, 225-228.
- [70] Luttkopf, D.; Müller, U.; Skov, P.S.; Ballmer-Weber, B.K.; Wuthrich, B.; Skamstrup Hansen, K.; Poulsen, L.K.; Kastner, M.; Hausteiner, D.; Vieths, S. *Mol. Immunol.*, **2002**, *38*, 515-525.
- [71] Scheurer, S.; Son, D.Y.; Boehm, M.; Karamloo, F.; Franke, S.; Hoffmann, A.; Hausteiner, D.; Vieths, S. *Mol. Immunol.*, **1999**, *36*, 155-167.
- [72] Furtado, P.B.; Furmonaviciene, R.; McElveen, J.; Sewell, H.F.; Shakib, F. *Proteins*, **2001**, *43*, 134-144.
- [73] Smith, A.M.; Benjamin, D.C.; Hozie, N.; Derewenda, U.; Smith, W.A.; Thomas, W.R.; Gafvelin, G.; van Hage-Hamsten, M.; Chapman, M.D. *J. Allergy Clin. Immunol.*, **2001**, *107*, 977-984.
- [74] Brusic, V.; Flower, D.R. *Drug Discov. Today: Biosilico*, **2004**, *2*, 18-23.
- [75] Uzogara S.G. *Biotechnol. Adv.*, **2000**, *18*, 179-206.
- [76] Gendel, S.M. *Ann. N.Y. Acad. Sci.*, **2002**, *964*, 87-98.
- [77] Metcalfe, D.D.; Astwood, J.D.; Townsend, R.; Sampson, H.A.; Taylor, S.L.; Fuchs, R.L. *Rev. Food Sci. Nutr.*, **1996**, *36* Suppl, S165-186.
- [78] Stanley, J.S.; King, N.; Burks, A.W.; Huang, S.K.; Sampson, H.; Cockrell, G.; Helm, R.M.; West, C.M.; Bannon, G.A. *Arch. Biochem. Biophys.*, **1997**, *342*, 244-253.
- [79] Shin, D.S.; Compadre, C.M.; Maleki, S.J.; Kopper, R.A.; Sampson, H.; Huang, S.K.; Burks, A.W.; Bannon, G.A. *J. Biol. Chem.*, **1998**, *273*, 13753-13759.
- [80] Stadler, M.B.; Stadler, B.M. *FASEB J.*, **2003**, *17*, 1141-1143.
- [81] Zorzet, A.; Gustafsson, M.; Hammerling, U. *In Silico Biol.*, **2002**, *2*, 525-534.
- [82] Pearson, W.R. *Methods Mol. Biol.*, **2000**, *132*, 185-219.
- [83] Soeria-Atmadja, D.; Zorzet, A.; Gustafsson, M.G.; Hammerling, U. *Int. Arch. Allergy Immunol.*, **2004**, *133*, 101-112.
- [84] Bailey, T.L.; Elkan, C.; *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1995**, *3*, 21-29.
- [85] Li, K.B.; Isaac, P.; Krishnan, A. *Bioinformatics*, **2004**, *20*, 2572-8.
- [86] Larche, M.; Ferreira, F.; Mohapatra, S.S. *Clin. Allergy Immunol.*, **2004**, *18*, 641-661.
- [87] Alexander, C.; Kay, A.B.; Larche, M. *Curr. Drug Targets Inflamm. Allergy*, **2002**, *1*, 353-361.
- [88] Larche, M. *Br. Med. Bull.*, **2000**, *56*, 1019-1036.
- [89] Woodcock, A.; Custovic, A. *Br. Med. Bull.*, **2000**, *56*, 1071-1086.
- [90] Sampson, H.A. *Br. Med. Bull.*, **2000**, *56*, 925-935.