

Review article

Allergen databases

Allergies represent a significant medical and industrial problem. Molecular and clinical data on allergens are growing exponentially and in this article we have reviewed nine specialized allergen databases and identified data sources related to protein allergens contained in general purpose molecular databases. An analysis of allergens contained in public databases indicates a high level of redundancy of entries and a relatively low coverage of allergens by individual databases. From this analysis we identify current database needs for allergy research and, in particular, highlight the need for a centralized reference allergen database.

**V. Brusic^{1,2}, M. Millot³, N. Petrovsky²,
S. M. Gendel⁴, O. Gigonzac⁵,
S. J. Stelman⁶**

¹Institute for Infocomm Research, Singapore; ²Medical Informatics Centre, University of Canberra, Canberra, ACT, Australia; ³IHCP, Joint Research Centre, Ispra, Italy; ⁴Biotechnology Studies Branch, Food and Drug Administration, National Center for Food Safety and Technology, Summit-Argo, IL, USA; ⁵Bayer CropScience, Evry, France; ⁶Dow AgroSciences, San Diego, CA, USA

Key words: allergen data; bioinformatics; data management; molecular databases.

Associate Professor Vladimir Brusic
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119631

Accepted for publication 26 March 2003

Allergies are amongst the most common causes of chronic ill health in industrialized countries (1), affecting up to one-third of the general population (2). The treatment of allergies is currently predominantly symptomatic rather than targeted at the cause with symptomatic treatments including allergen avoidance, antihistamines, corticosteroids or injectable epinephrine (3–5). Existing targets for allergy therapies include specific immunoglobulin E (IgE), Th2 cells and cytokines (6). Immunotherapies that offer promise in the treatment of allergies include desensitization using anti-IgE antibody therapy or immunomodulatory treatments using engineered proteins, peptides, allergen DNA vaccines, or immunostimulatory DNA (4, 7, 8).

The number of characterized protein allergens is increasing steadily. The Allergen Nomenclature Sub-Committee of the International Union of Immunological Societies (IUIS) maintains a list of 'certified' protein allergens (9). Some 420 protein allergens and 190 isoforms had been classified, as of December 2002. This may represent only a small fraction of the total as individual allergens may have many isoforms (10, 11). Furthermore, while this certified list contains many characterized allergens, it is not exhaustive as unlisted allergens can be found in the literature. Examples of allergens not included in the list are carrot cyclophilin

(12), crab tropomyosin (13), and pollen-specific maize protein (14).

Contemporary research requires rapid access to data sources and the advanced tools for analysis of data. A comprehensive list of databases is available at the Molecular Biology Database Collection (15). Allergen-related data are scattered across various publicly accessible molecular databases. The commonly used protein databases, such as PIR (16) or SWISS-PROT/TrEMBL (17), contain the amino acid sequences of many known allergens, but are not up-to-date (18). In addition, these databases generally do not contain detailed structural and functional information for a particular allergen. As a consequence a number of more specialized allergy-related databases and data repositories have appeared in recent years.

In this paper, we provide an overview of existing allergen databases and data repositories and identify issues in future database development. Rapid accumulation of data and increasing knowledge in the field of allergy requires more sophisticated bioinformatics tools and resources. A systematic approach to allergen data management will create benefits for basic allergy research, food industry applications, and clinical allergology. Given the importance of food allergies, a systematic bioinformatic approach would also be of use to the food

industry. Allergen databases represent a major resource for allergy research and will grow in importance as data accumulates. Various bioinformatics tools used for analysis of allergen data have been reviewed elsewhere (19). Since most, although not all, allergens are proteins we have largely focussed on protein allergens in this review.

Allergen databases and data sources

General-purpose biological databases contain annotated entries of biological sequences. These entries typically contain the sequence, a short description, the source organism, a list of structural or functional features, and literature references. The major public databases are GenBank/GenPept (20), EMBL/TrEMBL (17, 21), DDBJ/DAD (22), PIR (16), SWISS-PROT (17), and PDB (23).

The nucleotide databases, GenBank, EMBL, and DDBJ, focus on collecting, annotating, and providing access to the entries of DNA sequences and the related information. Data are collected, upgraded and exchanged between these three centers on a daily basis. GenPept, TrEMBL, and DAD are protein databases derived from the translations of all coding sequences of the three main nucleotides databases. SWISS-PROT and PIR are protein databases that are manually annotated. Their annotation is in higher detail than GenPept, TrEMBL, and DAD, but they contain fewer entries. PDB is a database of known 3D-structures of proteins.

These databases contain a significant proportion of known allergen and allergen-related sequences. For example, as of December 2002, PIR contained more than 980 protein entries related to known or putative allergens. The list of major databases is shown in Table 1.

Allergen-related data sources

Bioinformatic analysis of allergenicity requires access to credible sources of current data on known allergen sequences and structures. We have reviewed several allergen-specific data sources that have been compiled and made available to the public (Table 2). Most of these

Table 1. General database sources of allergen sequences. This table contains access addresses for six major data sources. Access information to various biological databases is available at the *Nucleic Acids Research* database collection (15) (see Database listing link)

Database	URL (http://)
GenBank/GenPept	www.ncbi.nlm.nih.gov
EMBL	www.ebi.ac.uk/embl
DDBJ	www.ddbj.nig.ac.jp
SWISS-PROT	www.expasy.ch
PIR	www-nbrf.georgetown.edu
PDB	www.rcsb.org/pdb
Database listing	www3.oup.co.uk/nar/database/

Table 2. Major allergen-specific data sources and versions evaluated in this review

Database	URL (http://)	Version or release date
IUIS	www.allergen.org	October 2001
SWISS-PROT	www.expasy.ch/cgi-bin/	November 2002
AllergenIndex	lists?allergen.txt	
BIFS	www.iit.edu/~sgendel/fa.htm	Release 3, 2002
CSL	www.csl.gov.uk/allergen	September 2001
FARRP	www.allergenonline.com	Ver.1.02 September 2001,
PROTALL	www.ifr.bbsrc.ac.uk/Protall	June 2001
ALLALLERGY	www.allallergy.net	2001
Asthma & Allergy	cooke.gsf.de/asthmagen/main.cfm	December 2002

allergen data sources 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. A short description of allergen data sources is given in this section.

IUIS

The list of allergens and isoallergens of clinical relevance developed and maintained by the Allergen Nomenclature sub-committee of the IUIS. These include allergens whose IgE reactivity has prevalence > 5%, or those that showed IgE reactivity in not less than five patients. This list classifies allergens following a proposed nomenclature (24). It is updated two to three times a year and grows at a relatively slow rate (1–20 new entries per update). The submissions are accepted and annotated by the committee members. The current (December 2002) update comprises some 420 proteins allergens from more than 150 species. The list provides scientific and common species names, protein names, nomenclature, the molecular weight, a flag indicating partial or complete sequence data, literature references, and when available, the GenBank or SWISS-PROT accession numbers. IUIS also provides an allergen variant list containing data on some 190 isoallergens from more than 20 species.

SWISS-PROT index of allergen sequences

This index (SPindex) 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.

IFBC/ILSI

The International Food Biotechnology Council (IFBC) and the ILSI Allergy and Immunology Institute sponsored a study (25) on a decision-tree approach for assessment of potential allergenicity of transgenic proteins in crop plants. This study includes a list of protein allergen sequences submitted to public databases. The

allergen list is available in print only, and has not been updated since publication.

BIFS tables

The Biotechnology Information for Food Safety (BIFS) web site (26) contains a nonredundant allergen protein list aimed at facilitating sequence comparisons in assessing potential allergenicity of proteins used in foods. The BIFS tables provide three flat files of food allergens, nonfood allergens, and wheat-gluten proteins. These tables contain lists of sequences published in the three major protein databanks providing for each entry (if available) species of origin, protein name, allergen nomenclature, literature references, and notes. The BIFS tables also contain GenPept, SWISS-PROT, and PIR accession numbers. The food allergen table contains approximately 140 entries corresponding to more than 30 distinct proteins from five animals and 13 plants. The nonfood allergen table contains more than 340 entries corresponding to more than 20 distinct proteins from more than 80 species. The wheat gluten table contains 75 entries corresponding to four distinct proteins.

CSL allergen database

This is a searchable database of allergens compiled and maintained at the Food Safety and Quality Directorate of The Central Science Laboratory at Sand Hutton York, UK. It contains 469 allergen entries, with more than 1200 links to isoform sequences. The entries contain the following fields: nomenclature (systematic name), species (scientific and trivial name), allergen type, and textual description. The database contains more than 120 epitope sequences from more than 20 well-studied allergen sequences. The CSL database provides links to full peptide sequences, PDB structural data, and PubMed (27) reference abstracts, whichever is available. This database contains entries from SWISS-PROT, PIR, GenPept, RefSeq (28), PRF (29), and PDB, and in some instances duplicate entries or multiple revisions of the same entries. The CSL allergen database can be searched by keywords, or can provide a list of allergens by type (food, inhalant, contaminant, or other).

FARRP

The FARRP Allergen Database (G. Holleschak and A. Silvanovich, personal communication) is a searchable database containing more than 650 unique sequences of known and putative allergens extracted from GenBank (ver. 119), EMBL (ver. 66), PIR (ver. 56), and PDB (ver. 56), and SWISS-PROT (ver. 38). This database has been developed and maintained by a group at the Monsanto company. Some protein sequences from published literature that have not been deposited in the public

databases have also been included. The FARRP database enables users to view lists of allergens containing: species of origin, common and nomenclature names, and accession number of the source (linked to the Entrez database). The FARRP has a web form for FASTA (30) comparison of protein sequences against the database entries. This is the only basic bioinformatic search tool provided with any of the specialized allergen databases reviewed in this article.

PROTALL

The PROTALL database, hosted at the Institute of Food Research of the United Kingdom was established to bring together clinicians, food scientists and plant biologists to study the problems of food allergy. It contains data on allergenic potential of plant food proteins, their molecular structure and biological activities. This database contains biochemical and clinical information about plant food allergens involved in classical IgE-induced hypersensitivity reactions on 77 allergens from 48 plant species. A PROTALL entry contains general allergen information: name, designation, accession numbers in major databases and links, structural information; properties of the allergen: molecular weight, structure, epitopes, stability, cross-reactivity, protein function, purification information, and references; and clinical data: oral provocation information, skin prick test, IgE determination assay, clinical history, and clinical references. The criteria for inclusion of an allergen into PROTALL include IgE binding demonstrated by immunoassays or immunoblotting. Users can query PROTALL using keyword searching either as a free text search or by specifying a plant.

ALLALLERGY

This 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 (Steinman HA, personal communication). It also includes chemical allergens which explains the huge number of entries compared with IUIS or PROTALL that contain only proteins. Users can extract the comprehensive description on allergens, adverse reactions, cross-reactions, family and taxonomical relationships, and information useful to assist patient assessment. The information in ALLALLERGY is fully referenced and the data have been regularly updated. An abridged version of the ALLALLERGY database is accessible through the Allergy AdvisorTM software product (Zing Solutions, Cape Town, South Africa) that includes information on allergens, additives, preservatives, and naturally occurring substances in food and plants.

Asthma and allergy gene database

This database (31) was developed at the National Research Centre for Environment and Health, Munich, Germany. It provides access to data on linkage and mutation studies data for asthma and allergy related traits. This database includes more than 19 000 linkage studies entries and more than 480 mutation studies entries. The asthma and allergy database can be queried by the chromosome number, study type, output type, and statistical significance. Each linkage study entry has data on chromosome location, marker with link to the Human Genome Database, *P*-value, the phenotype, and the study characteristics (e.g. bibliographical reference). Each mutation studies entry has data on gene name, mutation type, association type, phenotype and the study characteristics (e.g. studied population).

Other data sources

There are many more Internet-accessible data sources that provide access to various aspects of allergen data. A short list of better known specialized sources includes airborne allergens (32), latex allergens (33), European pollen information web site (34), European aeroallergen network pollen database (35), and allergy glossary (36).

Allergen data issues

Individual allergen-related data sources have been constructed for specific purposes and thus have different criteria for inclusion of entries. For example, the IUIS database, a list of proteins with assigned allergen nomenclature, contains only proteins showing IgE reactivity, whereas the BIFS tables, constructed for assessment of food protein allergens, use a broader definition of allergen that includes proteins involved in celiac disease. Some of these data sources (such as the CSL allergen database) include redundant data extracted from multiple protein databases and even multiple revisions of the same data. Other data sources (such as IUIS or BIFS) provide nonredundant data sets. To assess the quantity and the quality of allergen databases we (S. Stelman, September 2001) performed the analysis and classification of database entries of allergen sequences.

Construction of a nonredundant allergen data set

To analyse and compare the content of available allergen databases, the entries from each database were cross-referenced into a single composite nonredundant allergen data set (Fig. 1). Comparison and removal of redundant entries consisted of several steps.

Validation and checking of accession numbers. Protein databases were queried to verify that the accession

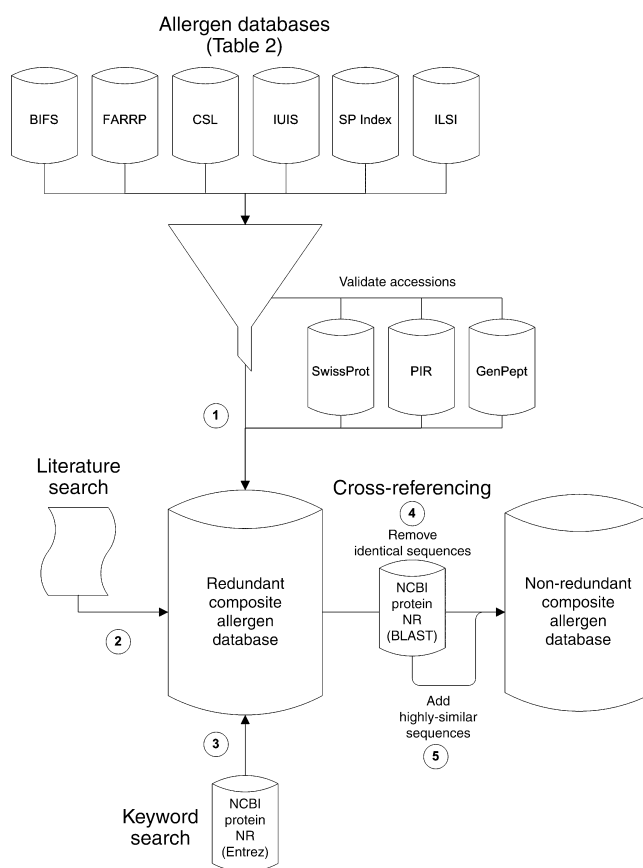


Figure 1. The compilation process for building a nonredundant composite allergen database. The entries from various allergen databases (step 1), keyword searches from major public databases (step 2), and literature (step 3) were compared with entries from major databases and cross-referenced (step 4) and enriched by similar sequences (step 5). Finally data were cleaned of redundancies to form the reference data set for analysis.

number for each allergen was valid. Entries whose accessions had been merged or combined in their source database were combined. Entries that had been completely removed from their source database were noted. Invalid accessions were corrected where typographical errors were obvious.

Addition of sequences not submitted to major protein databases. The 124 published references for entries in the IUIS database for which no accession numbers were recorded were examined for their protein or nucleotide sequences. Sequences found in the literature were searched against the NCBI nonredundant protein database (nr) using BLAST search (37) and accessions of proteins with 100% identity over the entire sequence length of both molecules were entered in the allergen data set. Published sequences lacking 100% identity with a protein sequence in the public databanks were assigned a unique identifier and entered in the allergen data set.

Additional sequences identified by keyword search. The NCBI nr database was queried through the Entrez browser with four keywords: 'allergen', 'allergens', 'isoallergen' and 'isoallergens'. A total of 2423 entries were returned and entered into the reference data set. These included 875 unannotated GenPept entries from 30 issued US patents, and six GenBank entries that do not contain protein translations.

Cross-referencing between databases and removal of redundant entries. Using the BIFS tables as a benchmark nonredundant database, each sequence that was also listed in another allergen database was compared against the NCBI protein nr database using BLAST search. Identical sequences from the same species were removed and their accession numbers were merged into a single entry, except in cases where multiple identical sequences are found in one database (e.g. same sequence submitted by different authors, multiple sequences submitted by same author). In those instances, each sequence was maintained as a unique entry unless reference and annotation information provided unambiguous support for removal from the reference list. Entries in the BIFS database with multiple accessions found not to be identical in the current NCBI protein nr database were divided into multiple entries.

Addition of similar sequences. Sequences showing BLAST hits with significant similarity to query sequences that were from the same organism and had similar protein names and functional annotations were also entered in the reference data set.

Analysis of the reference allergen data set

The total composite database (Tables 3 and 4), which included entries from all available allergen databases, Entrez keyword searches, and a sample BLAST analysis, contained 3051 valid database accessions, 105 sequences that were not present in any of the protein databases, and as many as 96 referenced allergens (data not shown) for which no sequence information is known. The total nonredundant set contained 2643 protein sequences, a

Table 3. Cross-checking of the number of sequences from various data sources (September 2001)

	IUIS	SP index	ILSI	BIFS	CSL	FARRP	ENTREZ	BLAST	Total
Original entries	514	220	197	557	377	654	n/a	n/a	n/a
Nonredundant entries	589	266	190	559	805	646	1981	118	2643
Unsubmitted sequences	100	0	0	0	0	18	n/a	n/a	105
Unavailable sequences	84	2	1	2	45	7	n/a	n/a	96

Original entries refers to the number of records reported by each allergen database. Nonredundant entries include the total number of unique sequences. Unsubmitted sequences include entries found in literature but not in protein databases. Unavailable sequences indicates that a peptide sequence could not be obtained (includes entries with incorrect or lacking sequence or reference information). The total column includes the number of unique entries in the composite database.

Table 4. The number of valid accession numbers from major public databases corresponding to entries in allergen data sources (September 2001)

	IUIS	SPindex	ILSI	BIFS	CSL	FARRP	ENTREZ	BLAST	Total
SWISS-PROT	42	266	7	285	222	149	358	3	491
PIR	4		75	344	344	133	290	21	453
GenPept	285		114	375	375	352	1700	110	2006
GenBank only	8						6		10
PRF					19		28		28
RefSeq					3		3		4
PDB					60	2	38		59
Total valid accessions	339	266	196	1004	1023	636	2423	134	3051

number of entries significantly higher than any one allergen set available in individual databases.

The ILSI and FARRP databases, both of which include only one accession per molecule, showed a decrease in the number of entries after the removal of redundant sequences. In other cases, nonredundant entries exceed total entries, because of sequence variants originally reported as single entries. Most notable is the CSL database, which reports only 377 entries, yet contains over 1000 accessions of which 805 were unique protein sequences.

Curation errors in the allergen databases were numerous. Obsolete entries and typographical errors were common. Each allergen data source contained at least one incorrect or obsolete accession. IUIS, an authoritative source in the field, lacked retrievable reference information for more than 10% of the entries to which official nomenclature had been assigned. The BIFS tables, which are a useful model for cross-indexing redundant data present in different protein databases, are especially prone to obsolescence if sequence updates from major databases are not regularly monitored.

Using public sequence databases

All the reviewed allergen data sources point to sequence entries in public protein databases and provide listings of accession numbers as text or hyperlinks. The advantage of using established public databases is the significant infrastructure required for collecting and maintaining data, and the requirement for interfaces for data extraction and analysis. Public databases also provide, to an extent, the curation and quality assurance of data. They are centrally accessible for submission, updating, and retrieval of included data entries. Standardized formats and fields facilitate automation of extraction and application of bioinformatics tools.

A major concern was the quality control of data. The data entered by hundreds of different individuals will be of variable quality and accuracy, and leave room for propagation of errors that cannot be easily rectified by public database annotators. Problems arising from the reliance on public databases include:

- Some allergen sequences have been published but not submitted to sequence databases (Table 3).
- Sequences in process – submitted and assigned an accession number, but not released (unclear whether an accession has been correctly reported).
- Variant sequence data may be listed only in comments, but not as separate entries (common in SWISS-PROT).
- Obsolete accession numbers may no longer be retrievable (cross-referencing problem; common with PIR entries when accessions are combined).
- Obsolete accessions may still be retrievable from public databases (redundant entries; occurs when NCBI's GenInfo Identifier (GI number) is referenced rather than the accession number).
- Sequence corrections may not necessarily be submitted to public databases.
- Lack of synchronization between databases.
- Incorrect or ambiguous annotations and keywords (curation problem; particularly with allergen-binding proteins and patent-related sequences).

The listed problems are common for data in public databases. One way to ensure data quality and consistency is to use the data warehouse paradigm (38). In data warehouses a major effort is directed towards ensuring the relevance and quality of data.

Current needs

The growth of allergy-related data is entering an exponential phase because of advances in genomic, proteomic, and molecular biology techniques. Bioinformatics provides tools for managing this data and is complementary to wet-lab research. The purpose of allergy-related databases is to facilitate the collection, access to, and use of such data. General sequence databases provide basic molecular, structural, taxonomy and reference information. Specialist databases focus on more specific issues and contain more detailed information. An ideal allergy-related molecular database – *a reference database* – should have a comprehensive, accurate and expert-annotated content. It should also contain a set of specialist tools that provide for extraction of complex information. Such databases and corresponding search tools must be accurate and robust. Given the current level of sophistication of bioinformatics, such a database should contain or provide for the following:

- Comprise comprehensive nonredundant entries of allergens, with their biochemical, structural, functional, and clinical features including, but not limited to, their sequences, structures, sources, references, and relevant links.
- Comply with the official classification of allergens (9).

- Provide the basis for further classification of allergens according to their structural and functional properties, as well as clinical relevance.
- Allow for detailed analysis of existing allergens and identification of common sequence and structural properties and characterization of those properties that are likely to be related to allergenicity.
- Have tools for bioinformatics-driven identification and detailed characterization of novel allergens, and assessment of potential allergenicity of proteins.
- Enable rapid identification of both promiscuous, and patient-specific targets of allergic responses (B-cell epitopes) and potential regulatory elements (T-cell epitopes).
- Support accurate identification of cross-reactivity among known and novel allergens.
- Include references to historical and obsolete published data, with rationale for retractions.

A database that enables these analyses must encompass diverse data formats, including strings, numerical, categorical or high-level data such as complete structural descriptions of the molecules. Potential uses of this database include the support of research towards better understanding of the molecular basis of allergies, the design of immunotherapy for allergies, patient risk assessment, and assessment of the safety of novel foods (Fig. 2). The uses of this database require not only simple extraction of data, but also the extraction of high-level information (data mining, knowledge discovery), data visualization, and generation of reports. In addition to the content requirements, a modern molecular database should be robust, simple to use, easy to build and update and relatively inexpensive. It should allow for easy data transfer and integration from various sources, and provide for efficient checking and expert annotation of these data. A further desirable feature is the ease of use and integration of diverse and complex analysis tools into

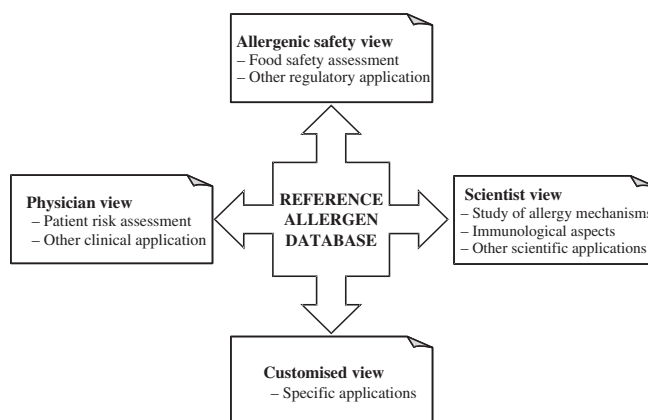


Figure 2. Multiple uses for the proposed Reference Allergen Database. Each view will have a different set of tools for data extraction and analysis.

the database. It is difficult to fulfil all these requirements because of the complexity of biological data, diversity of data formats, problems with incompleteness and high level of data errors, lack of standards, and the diversity of tools for analysis of biological data. Because relevant allergy-related data are scattered across various sources and are incomplete for comprehensive analysis, there is a need for tight integration of data into a single, trusted database. The bioinformatics field is developing rapidly and it is difficult to predict all future analysis needs. Therefore, it is convenient to have analysis tools in a modular form, so that the tools can be upgraded, replaced or combined without affecting the integrity of the system.

The latest developments in bioinformatics including databases and data warehousing approaches (38, 39) provide convenient tools for the design and maintenance of a specialist molecular, allergy-related database with multiple data views and flexible data-analysis tools. Data warehousing provides a means for integration of diverse data using a single global schema, a single interface and access to a single data source. It also provides for multiple views to data and for complex searches using sophisticated bioinformatics tools.

Conclusion

Rapidly increasing amounts of data and accumulating knowledge in the allergy field provides a rich source of

data to support basic and applied allergy research. Significant preliminary work has already been performed to collect and classify allergy-related data. The next step is the integration of allergy resources and the creation of bioinformatics infrastructure to enable a systematic approach to the management of allergen data. We have 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.

Two new allergen databases have been made available since this article was accepted for publication. Allergome database (<http://www.allergome.org>) contains mainly information relevant for allergenicity of proteins. The Structural Database of Allergenic Proteins (<http://fermi.utmb.edu/SDAP>) contains structural and epitope-related information on allergens.

Acknowledgments

Dr Gendel's work was partially supported by Cooperative Agreement FD-000431 between the US Food and Drug Administration and the National Center for Food Safety and Technology.

References

1. LARCHE M. Specific immunotherapy. *Br Med Bull* 2000;**56**:1019–1036.
2. ONO SJ. Molecular genetics of allergic diseases. *Annu Rev Immunol* 2000;**18**:347–366.
3. WOODCOCK A, CUSTOVIC A. Allergen avoidance: does it work? *Br Med Bull* 2000;**56**:1071–1086.
4. SAMPSON HA. Food anaphylaxis. *Br Med Bull* 2000;**56**:925–935.
5. GREINER DK. Risk management in allergen immunotherapy. *J Allergy Clin Immunol* 1996;**98**(6 Pt 3):S330–S334.
6. BURKS W, BANNON G, LEHRER SB. Classic specific immunotherapy and new perspectives in specific immunotherapy for food allergy. *Allergy* 2001;**56** Suppl 67:121–124.
7. VALENTA R, VRTALA S, FOCKE-TEJKE M, BUGAJSKA-SCHREITER, BALL T, TWARDOZ A et al. Genetically engineered and synthetic allergen derivatives: candidates for vaccination against type I allergy. *Biol Chem* 1999;**380**:815–824.
8. BROIDE DH. Molecular and cellular mechanisms of allergic disease. *J Allergy Clin Immunol* 2001;**108** Suppl 2: S65–S71.
9. Allergen Nomenclature Sub-Committee, International Union of Immunological Societies. Allergen Nomenclature. (www.allergen.org/List.htm).
10. SWOBODA I, JILEK A, FERREIRA F, ENGEL E, HOFFMAN-SOMMERGRUBER K, SCHEINER O et al. Isoforms of Bet v 1, the major birch pollen allergen, analyzed by liquid chromatography, mass spectrometry, and cDNA cloning. *J Biol Chem* 1995;**270**:2607–2613.
11. KETTNER A, HUGHES GJ, FRUTIGER S, ASTORI M, ROGGERO M, SPERTINI F, CORRADIN G. Api m 6: a new bee venom allergen. *J Allergy Clin Immunol* 2001;**107**:914–920.
12. FUJITA C, MORIYAMA T, OGAWA T. Identification of cyclophilin as an IgE-binding protein from carrots. *Int Arch Allergy Immunol* 2001;**125**:44–50.
13. LEUNG PS, CHEN YC, GERSHWIN ME, WONG SH, KWAN HS, CHU KH. Identification and molecular characterization of *Charybdis feriatus* tropomyosin, the major crab allergen. *Allergy Clin Immunol* 1998;**102**:847–852.
14. HEISS S, FLICKER S, HAMILTON DA, KRAFT D, MASCARENHAS JP, VALENTA R. Expression of Zm13, a pollen specific maize protein, in *Escherichia coli* reveals IgE-binding capacity and allergenic potential. *FEBS Lett* 1996;**381**:217–221.
15. BAXEVANIS AD. The molecular biology database collection: 2002 update. *Nucleic Acids Res* 2002;**30**:1–12.
16. WU CH, HUANG H, ARMINSKI L, CASTRO-ALVEAR J, CHEN Y, HU ZZ et al. The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res* 2002;**30**:35–37.

17. O'DONOVAN C, MARTIN MJ, GATTIKER A, GASTEIGER E, BAIRICH A, APWEILER R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 2002;**3**:275–284.
18. Joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology. Evaluation of allergenicity of genetically modified foods. Rome, Italy: FAO/WHO, 22–25 January 2001.
19. BRUSIC V, PETROVSKY N, GENDEL SM, MILLOT M, GIGONZAC O, STELMAN SJ. Computational tools for the study of allergens. *Allergy* 2003;**58**:1083–1092.
20. BENSON DA, KARSCH-MIZRACHI I, LIPMAN DJ, OSTELL J, RAPP BA, WHEELER DL. GenBank. *Nucleic Acids Res* 2002;**30**:17–20.
21. STOESSER G, BAKER W, VAN DEN BROEK A, CAMON E, GARCIA-PASTOR M, KANZ C et al. The EMBL nucleotide sequence database. *Nucleic Acids Res* 2002;**30**: 21–26.
22. TATENO Y, IMANISHI T, MIYAZAKI S, FUKAMI-KOBAYASHI K, SAITOU N, SUGAWARA H, GOJOBORI T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res* 2002;**30**:27–30.
23. WESTBROOK J, FENG Z, JAIN S, BHAT TN, THANKI N, RAVICHANDRAN V et al. The protein data bank: unifying the archive. *Nucleic Acids Res* 2002;**30**:245–248.
24. KING TP, HOFFMAN D, LOWENSTEIN H, MARSH DG, PLATTS-MILLS TA, THOMAS W. Allergen nomenclature. *Allergy* 1995;**50**:765–774.
25. METCALFE DD, ASTWOOD JD, TOWNSEND R, SAMPSON HA, TAYLOR SL, FUCHS RL. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit Rev Food Sci Nutr* 1996;**36**: S165–S186.
26. GENDEL SM. Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv Food Nutr Res* 1998;**42**:63–92.
27. NCBI PubMed. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>
28. PRUITT KD, MAGLOTT DR. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 2001;**29**:137–140.
29. Protein Research Foundation. <http://www.prf.or.jp/en/>
30. PEARSON WR, LIPMAN DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;**85**:2444–2448.
31. WJST M, IMMERSVOLL T. An internet linkage and mutation database for the complex phenotype asthma. *Bioinformatics* 1998;**14**:827–828.
32. <http://www.niaid.nih.gov/publications/allergens/title.htm>
33. <http://dmd.nihs.go.jp/latex/allergen-e.html>
34. <http://www.cat.at/pollen/index.en.html>
35. <http://www.ean.cc/>
36. <http://www.hon.ch/Library/Theme/Allergy/Glossary/allergy.html>
37. ALTSCHUL SF, GISH W, MILLER W et al. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–410.
38. SCHÖNBACH C, KOWALSKI-SAUNDERS P, BRUSIC V. Data warehousing in molecular biology. *Brief Bioinform* 2000;**1**:190–198.
39. SHOOP E, SILVERSTEIN KA, JOHNSON JE, RETZEL EF. MetaFam: a unified classification of protein families. II. Schema and query capabilities. *Bioinformatics* 2001;**17**:262–271.