Review

Allergen sequence databases

Steven M. Gendel[1] and John A. Jenkins[2]

[1]Food and Drug Administration, National Center for Food Safety and Technology, Summit-Argo, Illinois, USA
[2]Institute for Food Research, Norwich Research Park, Norwich, UK

A number of specialized databases have been developed to facilitate studies of human allergens. These include molecular databases focused on protein sequences and structures, informational databases focused on clinical, biochemical and epidemiological data related to protein allergens, a database on allergen nomenclature, and other knowledge bases or informational websites that are peripherally-related to research on allergens. Examples of each type of databases are listed and described briefly in this review. Database construction and maintenance and their impact on database quality and usefulness are also discussed.

Keywords: Allergic / Bioinformatics / Food allergy / IgE

Received: June 9, 2005; revised: February 20, 2006; accepted: April 8, 2006

1 Introduction

Allergen sequence databases are essential tools both for the bioinformatic analyses used during safety assessments for foods produced using modern biotechnology and for analyzing the structural and physiochemical properties of food allergen proteins. It has long been recognized that a complete set of allergen protein sequences can not be easily extracted from large sequence repositories such as Swiss-Prot or GenePept and that no single repository contains a complete set of relevant sequences [1, 2]. Therefore, a number of specialized focused databases have been created that can be used for allergen sequence analysis, several of which are available on-line. Each of these allergen sequence databases has a different structure and includes different sequences, depending on the intended use of the data and the criteria used to select sequences for inclusion. In some cases, the database focuses on molecular information such as protein sequences and structures, while in other cases the molecular data are adjuncts to other biomedical information. We will describe examples of on-line databases in each class and discuss the impacts of the different approaches used to construct these databases.

The one database that does not fit into these categories is the Allergen Nomenclature database of the International Union of Immunological Societies (IUIS) Allergen Nomenclature Sub-Committee (http://www.allergen.org). This database is intended to provide a central resource for insuring that allergen designations are uniform and consistent [3]. Although literature citations and database accession numbers are listed, there are no direct links to the sequence resources.

2 Online allergen databases

2.1 Molecular databases

The allergen databases that focus on molecular information all include basic amino acid sequence information, and may include additional information on protein structure and the location(s) of allergenic epitopes (when known). The seven major molecular databases are listed in Table 1.

The oldest of the databases is the Bioinformatics for Food Safety (BIFS) database at the National Center for Food Safety and Technology [1]. This database takes a broad approach to sequence inclusion because it was initially constructed to support allergenicity assessments for foods produced using modern biotechnology foods and as a tool for

Correspondence: Steven M. Gendel, Food and Drug Administration, National Center for Food Safety and Technology, 6502 S. Archer, Summit-Argo, Illinois 60501, USA
E-mail: sgendel@cfsan.fda.gov
Fax: +1-708-728-4177

Abbreviations: BIFS, Bioinformatics for Food Safety; **CSL,** Central Science Laboratory; **FARRP,** Food Allergy Research and Resource Program; **IUIS,** International Union of Immunological Societies; **SDAP,** Structural Database of Allergen Proteins

Table 1. Online molecular databases

Database	URL
Bioinformatics for Food Safety	http://iit.edu/∼sgendel/fa.htm
Allergen Online (FARRP)	http://allergenonline.com
Central Science Laboratory	http://www.csl.gov.uk/allergen/
Structural Database of Allergen Proteins	http://fermi.utmb.edu/SDAP/sdap_ver.html
AllerPredict	http://sdmc.i2r.a-star.edu.sg/Templar/DB/Allergen/
AllerMatch	http://www.allermatch.org/
International Immunogenetics Information System	http://imgt.cines.fr/

testing and validating query methods. One of the unique strengths of this database is the comparison of accessions between the three repository databases that were used as sources. This database is also structured in a way that allows identification of complete, non-redundant data sets for food and non-food allergens. The online database has been updated several times in the last few years. The current update includes links to the Pfam protein structural database.

The Allergen Online database, maintained by the Food Allergy Research and Resource Program of the University of Nebraska also contains a broadly defined set of allergen sequences [4]. As with the BIFS database, each entry is identified by source organism, protein name, allergen designation (if available) and is linked (in this case through a Gene Identifier [GI] number) to an accession in a single repository database (Entrez at the National Center for Biotechnology Information). This site was also the first to allow users to compare a sequence to an allergen database on-line (using FastA)

The allergen database of the Central Science Laboratory (CSL) also contains a set of allergen sequences, although it is not clear what criteria were used to populate the database. This database adds information on epitope sequences where available. There is also a field for links to structural information, but this field is not populated in most of the database records.

The Structural Database of Allergen Proteins (SDAP) is currently the most ambitious of the molecular databases [5]. In addition to allergen sequences and structural links, this database has implemented some unique search capabilities. The most interesting of these is a homology search based on the "rules" suggested by the FAO/WHO expert consultation [6](FAO/WHO, 2001). That is, the SDAP will compare a user's input sequence to the local allergen database to identify matches that are either greater than 35% identical over 80 amino acids using a sliding window or exact contiguous matches of a length specified by the user. Similar search functions based on the FAO/WHO criteria have been implemented as AllerPredict at the Allergen Database site of the Institute for Infocomm Research and at the Allermatch web site [2, 7]. Although the AllerPredict analysis uses a larger allergen sequence database, the AllerMatch provides choice of search sets.

The International Immunogenetics Information System website is focused primarily on information about immune system components, and contains extensive sequence and structural information on immunoglobin and MHC molecules, as well as related genomic information for humans and mouse [8]. This site includes a page devoted to immunoinformatics and a database of food allergen sequences similar to the BIFS and Allergen Online databases.

Finally, although not itself a database, it should be noted that the Swiss-Prot Knowledge Base contains a document with a complete list of accessions for allergen sequences in the Swiss-Prot database (http://www.expasy.org/cgi-bin/lists?allergen.txt). Although based on the IUIS nomenclature list, the Swiss-Prot list contains sequences that are designated as allergens in the accession annotation.

2.2 Databases supplying general information on allergens and allergy

The three databases that link allergen protein information to clinical, biochemical, and epidemiological information are listed in Table 2.

The Allergome database is a transitional form [9]. Although built around a listing of allergen molecules, this database also contains information on the biological functions of the allergen molecules, routes of exposure, epidemiology (prevalence), diagnostic and immunological literature citations, and information on diagnostic reagents. Significantly, this is the only database that lists allergenic sources for which no allergenic molecule(s) have yet been identified.

The AllAllergy site is among the most comprehensive allergy information resources available on the internet. It includes a database of allergenic foods, with some information on allergen molecules, along with extensive information on (and links to) relevant literature, other organizations, meetings and training programs. The major drawback

Table 2. Online general databases

Database	URL
Allergome	http://www.allergome.org/
AllAllergy	http://allallergy.net/
InformAll	http://www.foodallergens.info/

of this site is that only allergen information abridged from the commercial Allergy Advisor database is available.

The InformAll database, established in January 2005, recently replaced and extended the PROTALL database from which it was developed. The first version of InformAll is restricted to plant food allergens with the intension of expanding to encompass all food allergens. The information is divided into a series of layers, focused on different audiences. The introductory layer for a lay audience leads to a layer of clinical information and with biochemical data on identified molecular allergens. The criterion for inclusion is based on articles in the literature rather than on IUIS names. The information consists of structured text and links to sequences, structures, taxonomies and articles. An unusual feature is the use of a panel of expert referees to check the entries in order to try to bring the text towards the level of articles in the refereed literature.

3 How good is the information in the databases?

The many difficulties encountered in attempting to construct and maintain an allergen sequence database have been described previously [1, 2]. Many of these difficulties are common to the assembly of any large complex data set: typographical errors, structural modifications in the source data sets, inconsistencies in annotation among sources, *etc.* However, for allergen databases there are several sets of problems and inconsistencies that reflect underlying scientific problems and differences in philosophy. Some of these can be described with a few specific examples.

Table 3 lists the number of sequences included in several databases for one buckwheat, two peanut, and three soybean allergens. In each case there are differences in the number of allergen sequences listed. Some of these differences result from differences in the number of source databases used, some from different inclusion criteria, and some from differences in the way that duplicate sequences are treated. For example, some buckwheat allergens have not yet been sequenced so they are excluded from the sequence-only databases. Databases based on the IUIS nomenclature list would have included Ara h 8 more than a year before information on this allergen appeared in the literature [10], but would not include soybean glycinin (for which allergenic epitopes have been identified [11–14], but which is not in the IUIS list.

In some cases, the intended use of the database governs the philosophy used to determine the criteria for inclusion. For example, the celery chlorophyll a-b binding protein was found to bind IgE in sera from a single patient and was given the designation “Api g 3” when the sequence was deposited in Swiss-Prot (P92919) (K. Hoffmann-Sommergruber, pers. comm.). As a result, the BIFS, FARRP and SDAP databases all include this “allergen”. However, this protein clearly does not fulfill the criteria for an IUIS name and thus “Api g 3” is not included in databases that contain only allergens recognized by IUIS. Similarly, celery NADP-dependent malate dehydrogenase and phosphoglyceromutase are included in some allergen databases because of their GenBank annotations despite the fact that no clinical characterizations of either have been published (H. Breiteneder, pers. comm.). In general, databases such as BIFS, FARRP, and SDAP that have a regulatory focus include all possible allergens to avoid any chance that a relationship between a query sequence and an allergen might be missed. Other databases, particularly those with a more clinical focus, use more restrictive criteria for inclusion. An extreme example of this is the inclusion of a Bet v 1 homologue from *Catharanthus roseus*, which is known from an article that explicitly noted the absence of cross-reactivity with Bet v 1, in at least one database [15].

Table 3. Number of sequences for specific allergens in various databases

Database	Allergen					
	Peanut			Soybean		Buckwheat
	Ara h 1	Ara h 2	Gly m 1	Gly m Bd30K	Glycinin	Fag e 1
BIFS	2	3	2	2	18	3
FARRP	2	2	1	3	11	3
SDAP	2	2	2	3	13	3
CSL	6	5	2	0	0	0
Allergome	2	3	2	1	2	4
Allerpredict	1	1	2	0	2	0
InformAll	2	7	0	3	5	3

Nomenclature can also become a source of confusion. For example, the nomenclature of isoallergens can be inconsistent. The allergens Api g 1 and Api g 2 have been designated as isoallergens of Api g 1 by the IUIS. Conversely, the IUIS has accepted Ara h 3 and Ara h 4 as separate allergens despite the fact that they have a higher degree of identity than do the Api g 1 isoallergens. Further, one entry in Entrez has been annotated as Ara h 3/4 because it is as closely related to both Ara h 3 and Ara h 4 as they are to each other. The designation Fag e 1 has not been approved by IUIS and is used inconsistently by the allergen databases, making it difficult to identify related entries in the different databases.

The databases containing epitope sequences also face problems. For example, different epitope sequences have been reported for soybean glycinin by different groups [11, 12, 14, 16]. These differences reflect both differences in the methods used and in the size and composition of the serum pools available. It is also not clear how to treat epitopes identified in regions of a sequence that is cleaved during maturation such as for soybean Gly m Bd30K [17, 18] or Ara h 1 [19, 20].

Unusually complex sequence entries can present problems for both database administrators and users. A good example is the celery Api g 5 sequence. Although listed as a single accession and presented as a single sequence in Swiss-Prot (P81943), this accession actually consists of several noncontiguous sub-sequences, the N-terminal sequence of the intact protein and the N-termini of three peptides produced by cyanogen bromide cleavage. Ironically, in this case there is clear evidence that post-transcriptional glycosylation (rather than the primary sequence) is critical for allergenicity [21].

4 Conclusions

The design, construction, and maintenance of any bioinformatic database is a difficult process. This is particularly true for specialized databases intended for use as analytical tools rather than as sequence repositories. The variety in the databases described here shows how different approaches and philosophies result in significantly different databases, and emphasizes the value of multiple options. The existence of diverse data resources provides the opportunity to use diverse approaches to allergen analysis as well as to carry out comparative studies of search and analysis procedures.

This variety also suggests that it will be difficult to develop simple data interchange procedures. The use of stable sequence identifiers such as GenBank or Swiss-Prot accessions (which can be used to track the history of a sequence entry in each of these databases), a standard sequence format (such as the FASTA format), and recognized species names (rather than common names) will facilitate data interchange and comparison. In addition, there are a few principles that will facilitate widespread use of these datasets. These include clear statements of the criteria for inclusion of proteins in the database, the repository database(s) used, whether sequences are stored locally, the timing of the last update, a description of update criteria (*e.g.*, do updates consist of additions only or are previously added sequences re-verified), a description of any quality control procedures used, and information on whether earlier versions are available for comparative analysis.

The applicability of these databases will also improve as the underlying problems, such as the nomenclature difficulties described above, are solved. For example, the process of assigning new allergen names could be modified to allow assignments based on the literature even when the authors have not contacted the IUIS. Further, procedures can be developed to resolve sequence conflicts between related accessions in different repository databases.

Overall, the value of the bioinformatics approach to allergen analysis depends on the use of fully-described, publicly available data sets and well characterized search procedures and analysis algorithms.

John Jenkins was supported by EU through the concerted action QLRT-2001-02284 (InformAll) and the BBSRC through the competitive strategic grant to the IFR. Steven Gendel was partially supported by Cooperative Agreement FD000431 between the FDA and the National Center for Food Safety and Technology.

5 References

[1] Gendel, S., *Adv. Food Nutr. Res.* 1998, *42*, 63–92.

[2] Brusic, V., Millor, M., Petrovsky, N., Gendel, S. *et al.*, *Allergy* 2003, *58*, 1093–1100.

[3] Hoffman, D., Lowenstein, H., Marsh, D., Platts-Mills, T., Thomas, W., *Bull. of the World Health Organ.* 1994, *72*, 796–806.

[4] Hileman, R., Silvanovich, A., Goodman, R., Rice, E., *et al.*, *Int. Arch. Allergy Immunol.* 2002, *128*, 280–291.

[5] Ivanciuc, O., Schein, C., Braun, W., *Nuc. Acids Res.* 2003, *31*, 359–362.

[6] FAO/WHO, *Evaluation of Allergenicty of genetically modified foods. Report of a joint FAO/WHO consultation on food derived from biotechnology*, Rome, Italy, 2001.

[7] Fiers, M., Kleter, G., Nijland, H., Peignenburn, A., *et al.*, *BMC Bioinformatics* 2004, *5*, 133.

[8] LeFranc, M.–P., *Nuc. Acids Res.* 2003, *31*, 307–310.

[9] Mari, A., Riccioli, D., *J. Allergy Clin. Immunol.* 2004, *113*, S301.

[10] Mittag, D., Akkerdaas, J., Ballmer-Weber, B. K., Vogel, L., *et al.*, *J. Allergy Clin. Immunol.* 2004, *114*, 1410–1417.

[11] Beardslee, T., Zeece, M., Sarath, G., Markwell, J., *Int. Arch. Allergy Immunol.* 2000, *123*, 299–307.

[12] Xiang, P., Beardslee, T., Zeece, M., Markwell, J., Sarath, G., *Arch. Biochem. Biophys.* 2000, *408*, 51–57.

[13] Helm, R., Cockrell, G., Connaughton, C., Sampson, H., *et al.*, *Int. Arch. Allergy Immunol.* 2000, *123*, 205–212.

[14] Helm, R., Cockrell, G., Connaughton, C., Sampson, H., *et al.*, *Int. Arch. Allergy Immunol.* 2000, *123*, 213–219.

[15] Carpin, S., Laffer, S., Schoentgen, F., Valenta, R., *et al.*, *Plant Mol. Biol.* 1998, *36*, 791–798.

[16] Helm, R., Cockrell, G., Connaughton, C., West, C., *et al.*, *J. Allergy Clin. Immunol.* 2000, *105*, 378–384.

[17] Helm, R., Cockrell, G., Herman, E., Burks, A., *et al.*, *Int. Arch. Allergy Immunol.* 1998, *117*, 29–37.

[18] Kalinski, A., Melroy, L., Dwivedi, R., Herman, E., *J. Biol. Chem.* 1992, *267*, 12068–12076.

[19] Burks A., Shin, D., Cockrell, G., Stanley, J., *et al.*, *Eur. J. Biochem.* 1997, *245*, 334–339.

[20] Wichers, H., De Beijer, T., Savelkoul, H., Van Amerongen, A., *J. Agric. Food Chem.* 2004, *52*, 4903–4907.

[21] Bublin, M., Radauer, C., Wilson, I., Kraft, D., *et al.*, *FASEB J.* 2003, *17*, 1697–1699.