

Bioinformatics and Food Allergens

STEVEN M. GENDEL

U.S. Food and Drug Administration, National Center for Food Safety and Technology, 6502 S. Archer Rd, Summit-Argo, IL 60501

Bioinformatics can play an important role in developing improved technology for the detection and characterization of food allergens. However, the full realization of this potential will depend on the development of allergen-specific databases as well as improved methods for data mining within these databases. Examples of existing allergen databases and analysis tools are described, as are the most important issues that need to be addressed in the next stage of database development.

The development of improved technology for the detection, measurement, and control of food allergens will depend on improving our ability to identify and characterize allergenic proteins within foods and food processing environments. In addition, we will need more information on how allergens are affected by processing, cleaning, and sanitation. Bioinformatics will play an important role in organizing, disseminating, and analyzing the data necessary for this technology development.

Bioinformatics can be broadly defined as the application of computers to the collection, characterization, and analysis of biological data (1). Thus, bioinformatics encompasses both the construction of databases and analysis of the data they contain. Because information about allergy and allergens is spread among many databases, one of the most important roles for bioinformatics in food allergen research will be to develop methods for relating and analyzing data from multiple sources. The effective use of data from multiple sources requires an understanding of how each database is designed, how data integrity is maintained, and the semantic relationships among databases (e.g., does the term "allergen" mean the same thing in different databases). Accurate data analysis requires an understanding of the inner workings of the tools that can be used to extract knowledge from these data, particularly when these tools are adapted from other areas of research.

Archival Databases

A great deal of information about food allergens is contained within large archival databases such as GenBank (2), the European Molecular Biology Laboratory (EMBL) database (3), the Protein Information Resource (PIR) database (4), and PubMed (5). On the one hand, this information is difficult to locate. For example, many allergen protein sequences are not identified as such in GenBank (6). On the other hand, searching GenBank using the keyword "allergen" results in a large number of immune system proteins involved in the allergic response (6). Thus, there is no straightforward way to conduct an allergen-specific (much less food allergen-specific) data analysis using GenBank or the other large sequence databases. Similar problems exist with PubMed. For example, a search using the keyword "Ara h2" (a peanut allergen) returned 76 hits, of which more than 65 were unrelated to allergy (most of these involved unrelated genes or chemicals with "ara" as an acronym or abbreviation). In addition, the archival databases are inconsistent in the treatment of some data, such as whether the sequences for alternate alleles are identified as independent sequences. Therefore, bioinformatic analysis of food allergens requires the use of specialized databases. Several examples of allergen-related databases are described below.

Food Allergen Databases

Although a small number of foods account for most cases of food allergy, many foods have been shown to be allergenic. Interestingly, only one published work attempts to identify the less commonly allergenic foods, and this list is not available on-line (7). Therefore, no comprehensive database of allergenic foods exists. However, several databases list known food and non-food allergen proteins. Because all of these databases were designed to address different problems, they include different information and illustrate the difficulty of using multiple data resources in a single analysis. Four examples of food allergen-specific databases are the Biotechnology Information for Food Safety (BIFS) database (6), the Food Allergy Research and Resource Program (FARRP) database, the Allergome database, the Protall database, and the Structural Database of Allergenic Proteins (SDAP; 8). The URLs for each are listed in Table 1.

The BIFS database is the most straightforward of the allergen protein sequence resources (6). It contains a list of

Guest edited as a special report on "Methodological Aspects of Food Allergens" by Kristina Williams.

Corresponding author's e-mail: sgendel@cfsan.fda.gov.

Table 1. URLs of on-line allergen resources

URL	
Resource	Archival databases
GenBank (Protein sequences via ENTREZ)	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
EMBL (SwissProt/TrEMBL)	http://us.expasy.org/sprot/sprot-retrieve-list.html
PIR	http://pir.georgetown.edu/
PDB	http://www.rcsb.org/pdb/
PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
Allergen databases	
BIFS	http://www.iit.edu/~sgendel/fa.htm
FARRP	http://www.allergenonline.com/default.asp
Allergome	http://www.allergome.org/index.php
Protall	http://www.ifrn.bbsrc.ac.uk/protall/
SDAP	http://fermi.utmb.edu/SDAP/index.html

Species	Protein	Allergen Name	SW Accession	PIR Accession	ENTREZ Accession	Refs.	Notes
ANIMALS							
Abalone	Tropomyosin		Q9GZ71		AAG08987	1, 20	
Carp	Parvalbumin	Cyc p 1.02			CAC83659	1	
	Parvalbumin	Cyc p 1.01			CAC83658	1	
Cod	Parvalbumin	Gad c1	P02622	A94236		2, 4, 5, 6, 7	
	Parvalbumin		AAK63086		AAK63086	1	
	Parvalbumin		AAK63087		AAK63087	1	
Crab	Tropomyosin	Cha f 1	Q9N2R3		AAF35431	1	
Egg (Chicken)	Ovomucoid	Gal d1	P01005	A92754	AAA48995+	2, 3, 4, 5, 6, 7	Coding?
	Ovalbumin	Gal d2	P01014	A01244	AAA68882	2, 3, 4, 5, 6, 7	
	Ovalbumin	Gal d2		A90455	CAA23716	2, 3, 4, 5, 6, 7	
	Ovalbumin	Gal d2			CAA23682	2, 3, 4, 5, 6, 7	Coding?

Figure 1. An example of part of the BIFS database showing how accessions are cross-correlated within and between the archival databases.

Return to Search Page

Food Allergens Plants Protein List

Total number of records: 137

Species	Common Name	Allergen Name	Comments	Accession No.	Amino Acid Length
Actinidia chinensis	Kiwi	Act c 1	Actinidain (Actinidin)	P00785	380
Anacardium occidentale	Cashew	Ana o 1.0102	Vicilin-like protein	AAM73729	536
Anacardium occidentale	Cashew	Ana o 1.0101	Vicilin-like protein	AAM73730	538
Ananas comosus	Pineapple	Ana c 1	Profilin	AAK54835	131
Apium graveolens	Celery	Api g 4	Profilin	AAD29409	134
Apium graveolens	Celery	Api g 1.0201		CAA99992	159
Apium graveolens	Celery	n/a	Chlorophyll a/b binding protein	P92919	264
Apium graveolens	Celery	Api g 1		S63984	154
Arachis hypogaea	Peanut	Ara h 3	Glycinin	AAC63045	507
Arachis hypogaea	Peanut	Ara h 4	Glycinin	AAD47382	530
Arachis hypogaea	Peanut	Ara h 5	Profilin	AAD55587	131
Arachis hypogaea	Peanut	Ara h 6	Conglutin-like	AAD56337	129
Arachis hypogaea	Peanut	Ara h 7	Conglutin-like	AAD56719	160
Arachis hypogaea	Peanut	Ara h 1	Clone P17	P43237	614
Arachis hypogaea	Peanut	Ara h 1	Clone P41B	P43238	626
Arachis hypogaea	Peanut	Ara h 2		AAK96887	156
Arachis hypogaea	Peanut	n/a		AAM46958	538
Arachis hypogaea	Peanut	n/a	Lectin	S14765	236
Bertholletia excelsa	Brazil nut	Ber e 1	2S sulfur-rich seed storage protein	P04403	146
Brassica juncea	Leaf mustard	Bra j 1	Fragment	ALL1_BRAJU_1	37
Brassica juncea	Leaf mustard	Bra j 1	Fragment	ALL1_BRAJU_2	92
Brassica napus	Rape	n/a	Clone 42	S65144	83
Brassica napus	Rape	n/a	Clone 44	S65145	83

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=protein&term=AAK96887

Figure 2. An example of part of the FARRP database showing the links to GenBank accessions.

ALLERGOME A Database of Allergenic Molecules

HOME SEARCH EXIT HELP LINKS About this site

Allergenic Molecule Monograph: Ara h 2.01

General Information	Native Form	Escherichia coli -
---------------------	-------------	--------------------

LIST NEW ALLERGENS Back to Allergenic Molecules Found

Last update	5/24/2003
Allergome Code	1081
Allergenic Molecule Name	Ara h 2.01
Common Names	Conglutin
Accession Numbers	Ara h (Tax & Im) - 381B, Ara h 2.01 - AY007229
Sources	Arachis hypogaea, Peanuts, Plants

Figure 3. An example of part of the Allergome database showing the links to SwissProt accessions.

Accession Number web link(s)	http://ca.expasy.org/cgi-bin/niceprot.pl?P43237 http://ca.expasy.org/cgi-bin/niceprot.pl?P43238
Allergen 3D structure accession number	Not determined
Molecular weight	65,000
Allergen epitopes	Immunodominant epitopes: Ara h 1 25-34 (KSSPYQKK) Ara h 1 65-74 (EYDPRLVY) Ara h 1 89-98 (ERTRGRQP) Ara h 1 498-507 (RRYTARLKEG) Other IgE epitopes: Ara h 1 48-57 (QEPDDLKQKA) Ara h 1 89-98 (ERTRGRQP) Ara h 1 97-105 (GDYDDRRR) Ara h 1 107-116 (RREEGGRW) Ara h 1 123-132 (EREEDWRQ) Ara h 1 134-143 (EDWRRPSHQQ) Ara h 1 143-152 (PRKIRPEG) Ara h 1 294-303 (PGQFEDFF) Ara h 1 311-320 (YLQEFSSRN) Ara h 1 325-334 (FNAEFNEIRR) Ara h 1 344-353 (QEERGQRR) Ara h 1 393-402 (DITNPINLRE) Ara h 1 409-418 (NNFGKLFVK) Ara h 1 461-470 (GNLELY) Ara h 1 525-534 (ELHLLGGIN) Ara h 1 539-548 (HRIFLAGDKD) Ara h 1 551-560 (IDQIEKQAKD) Ara h 1 559-568 (KDIAFPGSGE) Ara h 1 578-587 (KESHFVSARP) Ara h 1 597-606 (EKESPEKED)

Figure 4. An example of part of the Protall database showing sequences of the epitopes for Ara h1.

accessions for, and links to, food and non-food allergen protein sequences in the GenBank, SwissProt/TrEMBL, and PIR databases. To insure that the BIFS database contains as complete a set of allergen sequences as possible, all sequences containing the keyword "allergen" in the annotation in any of the source databases and all allergen sequences identified through literature searches are included. The major strengths of this database lie in the fact that it includes all accessions for each allergen protein, including all variants and isoforms, in each of the source databases and that related sequences are cross-correlated both within each source database and between databases. Figure 1 shows an example of part of the BIFS database. Accessions appearing on a single line are identical, while accessions for the same protein that appear on different lines have different sequences. This is the only database that provides complete correlation among multiple primary sources.

The FARRP database is also a straightforward allergen sequence database, providing a single link to a GenBank accession for each food and non-food allergen or isoform (Figure 2). The criteria for inclusion in this database are similar to those in the BIFS database. The greatest strength of the FARRP database lies in its ability to perform a FASTA search of a query sequence against the entire database, including the ability to control some of the FASTA input

parameters. It is not possible to restrict the search to a subset of sequences (such as only the food allergens).

The Allergome database contains a more limited set of sequences but adds biochemical and immunological information. Only molecules that have been shown to cause an IgE-mediated reaction are included. This database provides links to EMBL accessions for each molecule (Figure 3). In addition, the Allergome database provides links to PubMed literature resources for each allergen. These include papers related to both the biochemistry of each allergen molecule and medically oriented publications. Interestingly, the database contains information on allergenic sources without identified allergen molecules, but there is no way to access this information directly. For example, a keyword search for "maple" returns links to information on exposure and the notation "No Allergenic Molecules Identified." However, there is no way to recover a complete list of such entries. The major strengths of this database are the ability to perform Boolean keyword searches and links to sources of immunochemical reagents.

The databases described above provide links to complete sequences for each allergen protein. The Protall database (Figure 4) extends these data by providing epitope sequences, when these are known. Protall, like the Allergome database, contains additional information about each allergen, including links to accessions in both the GenBank and EMBL databases,

as well as biochemical and clinical information. Although literature citations are listed, the citations are not linked to a source database. It is not clear whether this database is currently being updated.

The SDAP database is similar to Protall in that it provides links to GenBank and EMBL databases as well as epitope sequences when known (8). Although there is no structural information at the site, there are links to several structural analysis programs and to the Protein Data Bank (PDB) and the

National Center for Biotechnology Information (NCBI) Molecular Modeling Data Base (9).

In the future, structural analysis will become as important (or more important) as sequence analysis for understanding the physiology and biochemistry of food allergens. Allergen proteins exist as 3-dimensional structures in foods, they interact with the immune system as 3-dimensional structures, and food processing and cleaning treatments may affect this structure. A search of the PDB in May 2003 using the keyword "allergen" resulted in 55 entries, including 14 food allergens, of which 10 were hen egg white lysozyme and 4 were bovine *beta*-lactoglobulin. Similarly, a search using the keyword "peanut" found 12 entries, of which 6 were for peanut lectin. Clearly, there is a great deal of room for growth in the field of allergen structural databases and analysis.

Data Analysis Resources

Most of the common bioinformatic data analysis tools can be adapted for use with food allergens. However, it is important to understand the limitations of each tool and the effects of any user-controlled parameters.

Protein sequence analysis is used to determine whether a protein is evolutionarily related to a known allergen, whether evolutionarily unrelated proteins share short sequence similarities that might be cross-reactive epitopes, and to examine sequence conservation within allergen families. The first 2 forms of analysis are generally carried out by searching a database with a single query sequence using either FASTA or Basic Local Alignment Search Tool (BLAST; 1). The NCBI, the EMBL, and the PIR all provide BLAST searches of large archival protein databases. However, there is no way to restrict these searches to allergen sequences or to easily parse the results to identify allergen sequences in the output. The FARRP database is the only on-line resource with the ability to search an allergen sequence database.

In most cases, BLAST and FASTA are used to find regions of sequence homology in evolutionarily related proteins. This information is potentially useful in understanding patterns of allergenic cross-reactivity and in predicting the effects of processing and other treatments on allergens. Because these programs were originally designed for this type of homology analysis, the default parameters are usually appropriate. However, the on-line versions of these programs are not suitable for finding very short regions of sequence similarity that might suggest the existence of cross-reactive epitopes in evolutionarily unrelated proteins, which is essential for assessing the potential allergenicity of novel proteins in bioengineered foods. This type of analysis can only be effectively carried out using allergen databases and analysis programs that give full control of parameters such as the scoring matrix and expectation values.

The third form of sequence analysis, the study of sequence conservation, can be approached by using a multiple sequence alignment tool such as ClustalW, a version of which is available at the PIR (1). In this case, the sequences of a number of related sequences are aligned to each other, and the degree of sequence conservation at each site in the sequence is

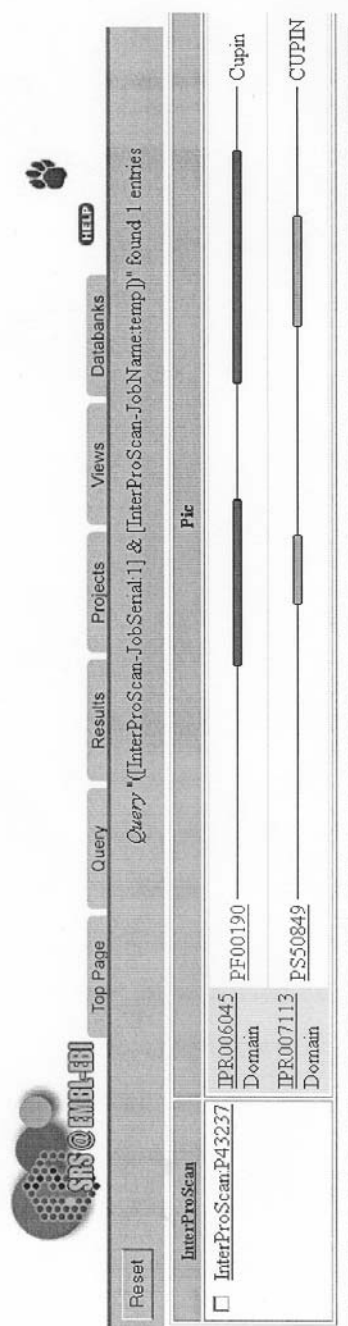


Figure 5. An example of an InterPro query for Ara h1 showing the locations of cupin domains within the protein.

determined. It may be possible to correlate sequence conservation with allergenic cross-reactivity to identify immunologically important regions of the proteins.

Future Directions

Food allergen databases have the potential to become important resources for understanding the properties and behavior of known allergens and for predicting the allergenic potential of novel food proteins. However, full realization of this potential will require improved allergen-specific databases and better methods for integrated analysis. For example, each of the allergen sequence databases described above uses different criteria for determining which sequences to include. It should be possible to develop a common vocabulary that can be used in all allergen databases to indicate why each entry is present in each database, and the degree of certainty associated with that determination. Similarly, common vocabularies are needed for the various clinical and epidemiological concepts that are included in some allergen databases.

The two most important short-term improvements in allergen databases will be the inclusion of information about epitope sequences and establishing links between the allergen sequence data and structural databases. The description of allergenic epitopes is another example of an area where a common vocabulary is needed to identify the method(s) used to locate each epitope.

An increasing number of protein structural databases are becoming available, and (as described above) linking the allergen-specific databases to these resources will become increasingly important. For example, in addition to sequence homology, proteins can share similar structural domains and therefore be considered members of the same protein family. Because proteins interact with the immune system as 3-dimensional structures, the ability to place a protein within a family is potentially useful for understanding allergenicity. Although there are a number of different protein family classification systems and databases available, integrated access to many of these can be obtained through the interPro database (10). Figure 5 is an example of a domain scan for the peanut allergen Ara h1 from EMBL.

Full realization of the potential for bioinformatic analysis of data on food allergens will depend on encouraging on-going coordination between different databases. Although it is neither necessary nor desirable to construct a single

allergen database, it is desirable that each be designed and constructed using a common vocabulary and that methods be developed for coordinated analysis of these databases, and perhaps for using these databases as focused points of entry into the large archival databases.

Acknowledgments

This work was supported by a Cooperative Agreement between the U.S. Food and Drug Administration and the National Center for Food Safety and Technology.

References

- (1) Brown, S. (2000) *Bioinformatics*, Eaton Publishing Co., Natick, MA
- (2) Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., & Wheeler, D. (2003) *Nucleic Acids Res.* **31**, 23–27
- (3) Stoesser, G., Baker, W., van den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., Nardone, F., Stoehr, P., Tuli, M.A., Tzouvara, K., & Vaughan, R. (2003) *Nucleic Acids Res.* **31**, 17–22
- (4) Wu, C., Yeh, L., Huang, H., Arminski, L., Castro-Alvea, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R., Suzek, B., Vinayaka, C., Zhang, J., & Barker, W. (2003) *Nucleic Acids Res.* **31**, 345–347
- (5) Wheeler, D., Church, D., Federhen, S., Lash, A., Madden, T., Pontius, J., Schuler, G., Schriml, L., Sequeira, E., Tatusova, T., & Wagner, L. (2003) *Nucleic Acids Res.* **31**, 28–33
- (6) Gendel, S. (1998) *Adv. Food Nutr. Res.* **42**, 63–92
- (7) Hefle, S., Nordlee, J., & Taylor, S. (1996) *Crit. Rev. Food Sci. Nutr.* **36S**, S69–S89
- (8) Ivanciuc, O., Schein, C., & Braun, W. (2003) *Nucleic Acids Res.* **31**, 359–362
- (9) Westbrook, J., Feng, Z., Chen, L., Yang, H., & Berman, H. (2003) *Nucleic Acids Res.* **31**, 489–491
- (10) Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S., Pagni, M., Peyruc, D., Ponting, C., Selengut, J., Servant, F., Sigrist, C., Vaughan, R., & Zdobnov, E. (2003) *Nucleic Acids Res.* **31**, 315–318