



## Allergen databases and allergen semantics

Steven M. Gendel\*

Food and Drug Administration, Center for Food Safety and Applied Nutrition, 5100 Paint Branch Parkway, College Park, MD 20740, United States

### ARTICLE INFO

#### Article history:

Received 20 August 2008

Available online 6 December 2008

#### Keywords:

Food allergy

Food safety

Informatics

Allergen sequences

Database

Semantic web

### ABSTRACT

The efficacy of any specific bioinformatic analysis of the potential allergenicity of new food proteins depends directly on the nature and content of the databases that are used in the analysis. A number of different allergen-related databases have been developed, each designed to meet a different need. These databases differ in content, organization, and accessibility. These differences create barriers for users and prevent data sharing and integration. The development and application of appropriate semantic web technologies, (for example, a food allergen ontology) could help to overcome these barriers and promote the development of more advanced analytic capabilities.

Published by Elsevier Inc.

### 1. Introduction

In the last few years, significant improvements have occurred in our understanding of how to use bioinformatic analysis as part of the allergenicity risk assessment process for bioengineered foods. These improvements include both the development of a more sophisticated understanding of how to analyse amino acid sequences and increasing use of information on protein secondary and tertiary structures. The tools and techniques that have been, and continue to be, developed help not only to assess the potential allergenicity of novel food proteins but may also aid in understanding how allergenic proteins behave during food processing and how to establish regulatory thresholds for food allergens.

The efficacy of any specific bioinformatic analysis of allergenic proteins is dependent on the nature of the data sets that are used in that analysis. As a result, a number of different allergen-related databases have been developed, each designed to meet a different need. For example, some databases include information only on food allergens while others include information on contact, inhalant and injection allergens. Differences also exist in the level of annotation, whether there are linkages to the repository databases that are the sources for sequence and structural data, and in many other characteristics of the databases.

The nature of the existing databases and the differences between them are reviewed here. An approach to enhancing communication between the databases and to leveraging the value of the information contained in them, based on semantic web technologies, is also described.

### 2. Allergen databases

It has long been recognized that food allergens are normal protein components of allergenic foods (Taylor and Lehrer, 1996). As a result, two basic types of food allergen data resources have been developed; those that primarily provide clinical, physiological or epidemiological information on food allergy and those that primarily provide molecular information for allergenic proteins. Several of the molecular databases are linked to, or are components of, sites that provide various types of allergenicity analyses. Examples of the general databases are given in Table 1 and examples of molecular databases are given in Table 2.

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 (Hoffman et al., 1994). This site has recently been redesigned to include more information about each allergen molecule, including amino acid sequences. One critical feature of the allergen naming process implemented through this site is the requirement for clinical information demonstrating allergenic activity.

The AllAllergy site is among the most comprehensive allergy information resources available on the internet. It includes a database of allergenic foods, information on allergen molecules, along with extensive information on (and links to) relevant literature, other organizations, meetings and training programs. The information on this site is supplied by and abridged from the commercial Allergy Advisor database.

\* Fax: +1 301 436 2633.

E-mail address: [steven.gendel@fda.hhs.gov](mailto:steven.gendel@fda.hhs.gov).

**Table 1**  
General databases.

| Database      | URL                                                                                                                                                                                   |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AllAllergy    | <a href="http://allallergy.net/">http://allallergy.net/</a>                                                                                                                           |
| Allergome     | <a href="http://www.allergome.org/">http://www.allergome.org/</a>                                                                                                                     |
| AllergoPharma | <a href="http://www.allergopharma.com/dokumente/en/allergopharma/allergen_db/allergen_db.php">http://www.allergopharma.com/dokumente/en/allergopharma/allergen_db/allergen_db.php</a> |
| InformAll     | <a href="http://www.foodallergens.info/">http://www.foodallergens.info/</a>                                                                                                           |

The Allergome database (Allergy Data Laboratories, Italy) is a transitional form (Mari and Riccioli, 2004). 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 allergen database from AllergoPharma contains extensive information about allergenic species but no molecular data. The information on the site is primarily related to understanding patterns of exposure.

The InformAll database (Institute of Food Research, UK) has replaced and extended the earlier PROTALL database (Gendel and Jenkins, 2006). The current version contains information on some plant and animal food allergens. The site also contains information focused on different audiences. Introductory information intended for a lay audience leads to more technical clinical information and to biochemical data on identified allergenic proteins. This is one of the few sites that uses peer review by a panel of expert referees to ensure data quality.

The Allergen Database for Food Safety (National Institute of Health Science, Japan) site contains a database of allergen and epitope sequences. Users can search these data sets independently as well as accessing allergen sequences by name.

The Allergen Online database, maintained by the Food Allergy Research and Resource Program of the University of Nebraska, contains a broadly defined set of allergen protein sequences (Hileman et al., 2002; Goodman, 2006). Each entry is identified by source organism, protein name, allergen designation (where available), and is linked through a Gene Identifier [GI] to an accession in the Entrez database at the National Center for Biotechnology Information. Relationships between allergenic molecules are indicated by membership in “allergen groups” which are defined by taxonomic and sequence relationships. This site was among the first to allow users to compare a query sequence to an allergen database online (using FASTA). This site uses peer review to evaluate evidence for inclusion of each molecule in the database.

The oldest of the online molecular databases is the Bioinformatics for Food Safety (BIFS) database at the National Center for Food Safety and Technology, which is no longer being maintained (Gendel, 1998; Gendel and Jenkins, 2006). One of the unique strengths of this database is that it compares the sequences contained in

**Table 2**  
Molecular databases.

| Database                                 | URL                                                                                                                 |
|------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Allergen Database for Food Safety        | <a href="http://allergen.nihs.go.jp/ADFS/">http://allergen.nihs.go.jp/ADFS/</a>                                     |
| Allergen Online (FARRP)                  | <a href="http://allergenonline.com">http://allergenonline.com</a>                                                   |
| AllerDB                                  | <a href="http://sdmc.i2r.a-star.edu.sg/Templar/DB/Allergen/">http://sdmc.i2r.a-star.edu.sg/Templar/DB/Allergen/</a> |
| AllerMatch                               | <a href="http://www.allermatch.org/">http://www.allermatch.org/</a>                                                 |
| Bioinformatics for Food Safety           | <a href="http://www.iit.edu/~sgendel/fa.htm">http://www.iit.edu/~sgendel/fa.htm</a>                                 |
| Central Science Laboratory               | <a href="http://www.csl.gov.uk/allergen/">http://www.csl.gov.uk/allergen/</a>                                       |
| Structural Database of Allergen Proteins | <a href="http://fermi.utmb.edu/SDAP/sdap_ver.html">http://fermi.utmb.edu/SDAP/sdap_ver.html</a>                     |

related accessions among the three repository databases that were used as data sources. This database is also structured in a way that allows identification of complete, non-redundant data sets for food and non-food allergens. There are no analysis tools on this site.

The allergen database of the Central Science Laboratory (UK) (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. Each record in the database contains a field for links to structural information, but this field is not populated in most of the database records. The site has recently added a FASTA search function.

The Structural Database of Allergen Proteins (SDAP) (Univ. of Texas Medical Branch) is currently the most ambitious of the molecular databases (Ivanciuc et al., 2003; Schein et al., 2007). In addition to allergen sequences and structural information, this database has implemented some unique search capabilities. These include a peptide matching function and a peptide similarity search based on a “Property Distance” (PD) value calculated from the physiochemical properties of the amino acids in a peptide.

Several sites have implemented homology searches based on allergenicity assessment criteria suggested by a FAO/WHO expert consultation (FAO/WHO, 2001). This expert consultation proposed that the potential for cross reactivity should be considered when there is a match between a query sequence and a sequence in an allergen database of greater than 35% identity over 80 amino acids or there is an exact match of at least six contiguous amino acids. Searches based on these criteria have been implemented on the AllerPredict, Allergen Database for Food Safety, AllerMatch (Brusic et al., 2003; Fiers et al., 2004), and the SDAP web sites. The Allergen Online database has implemented a version of the 80-mer search, but not the short exact search. Some of these sites permit the user to vary the criteria used to evaluate the matches.

A third set of allergen-related analysis sites is shown in Table 3. These sites provide allergenicity predictions based on sets of sequence motifs or indicator peptide sequences derived by analysis of an allergen sequence database. None of these sites provide direct access to the motif or peptide sequence sets. An allergen motif search is also available on the Allergen Database for Food Safety site along with the searches described above.

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

Table 4 compares several characteristics of the molecular allergen databases. This comparison emphasizes the degree of variation between the different databases. This variation encompasses fundamental characteristics such as the number of sequences contained in the database and the number of different categories of allergen identified in the database. Further, the SDAP database is the only one that permits direct access to allergen sequences independently of a search function.

**Table 3**  
Motif analysis sites.

| Site        | URL                                                                                                     |
|-------------|---------------------------------------------------------------------------------------------------------|
| Evaller     | <a href="http://bioinformatics.bmc.uu.se/evaller.html">http://bioinformatics.bmc.uu.se/evaller.html</a> |
| WebAllergen | <a href="http://weballergen.bii.a-star.edu.sg/">http://weballergen.bii.a-star.edu.sg/</a>               |
| AllerTool   | <a href="http://research.i2r.a-star.edu.sg/AllerTool">http://research.i2r.a-star.edu.sg/AllerTool</a>   |
| AlgPred     | <a href="http://www.imtech.res.in/raghava/algpred">http://www.imtech.res.in/raghava/algpred</a>         |

**Table 4**

Characteristics of the molecular databases.

| Database                                 | Last update | Inclusion criteria specified? | No. of sequences | Allergen categories |
|------------------------------------------|-------------|-------------------------------|------------------|---------------------|
| Allergen Database for Food Safety        | Jan 2008    | Y                             | 2108             | 8                   |
| Allergen Online                          | Jan 2008    | Y                             | 1313             | 13                  |
| AllerDB                                  | ?           | N                             | ?                | ?                   |
| AllerMatch                               | Jan 2007    | Y                             | 863              | 1                   |
| Bioinformatics for Food Safety           | N/A         | ?                             | ?                | 3                   |
| Central Science Laboratory               | ?           | N                             | ?                | 4                   |
| Structural Database of Allergen Proteins | Jul 2007    | Y                             | 829              | 9                   |

### 3. Database utilization

Diversity between the allergen databases presents several problems for users. First, it can be difficult to determine whether or how a particular data set has changed since it was first established

or described in a publication. Second, it can be difficult to evaluate or quantify the accuracy or applicability of results obtained using a particular dataset. Third, it is impossible to develop “third party” data analysis tools that integrate across or between data sets. Fourth, researchers interested in developing new analytic tools or approaches must devote a significant amount of time and effort to constructing and formatting new task-specific databases. One approach to overcoming these problems, and to increasing the value of the information contained in the databases, is to adapt the concepts and tools of the semantic web.

The semantic web is an enhancement of the existing world wide web that allows data to be shared and reused across application, enterprise and community boundaries (Berners-Lee et al., 2001). These enhancements are achieved by applying enabling technologies including the use of ontologies, descriptive metadata and collaborative working groups.

In informatics, an ontology is a representation of concepts and the relationships between those concepts within a knowledge domain (Gruber, 1993; Rubin et al., 2006). Ontologies provide formal descriptions of concepts, terms and relationships and can be used, for example, to describe types of molecules, molecular functions, biochemical interactions, cellular localizations and any other relevant property. Ontologies can also be used to relate common

#### A

##### Class/Type Details

| General         |                 | Attributes    |                                                       |
|-----------------|-----------------|---------------|-------------------------------------------------------|
| Class/Type Name | <b>Allergen</b> | Definition    | <b>A substance that elicits an allergic reaction.</b> |
| Id              | <b>C62651</b>   | PT            | <b>Allergen</b>                                       |
|                 |                 | primitive     | <b>true</b>                                           |
|                 |                 | CONCEPT_NAME  | <b>Allergen</b>                                       |
|                 |                 | Semantic_Type | <b>Chemical Viewed Functionally</b>                   |
|                 |                 | SYNONYM       | <b>Allergen</b>                                       |

#### B

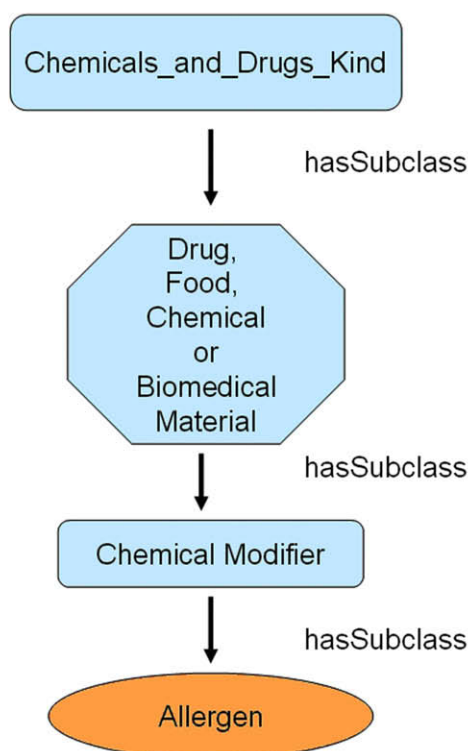


**Fig. 1.** An example of the application of a biomedical ontology related to food allergens. (A) The formal definition of the term “allergen” in the NCI ontology and (B) a diagram of the relationship of this term to other concepts in the ontology.

names or terms to their technical equivalents, identify synonyms and differentiate context-specific multiple meanings attached to a single term. As an example of part of an ontology related to food allergy, Fig. 1 shows the definition and associated relationships for the term “allergen” in the National Cancer Institute (NCI) ontology. This example emphasizes one of the many types of relationships (Subclass) that can be implemented in an ontology. An allergen ontology could be used to provide unambiguous labels for different types of allergenic proteins (food, inhalant, etc.) as well as for concepts such as “putative allergen” or “allergen homolog”.

Metadata are data about data. That is, metadata include information that can be used to describe either an individual datum or a data set. For example, the common practice of describing a literature citation by author, title, year, journal, etc. is a form of structured metadata about a published article. The annotation that accompanies sequence data in the GenBank or EMBL databases can also be considered metadata related to the sequence information. In this context, descriptive information defined in an allergen ontology could be easily attached to the contents of existing databases as metadata through the use of extensible markup language (XML) coding. This would facilitate the exchange and integration of data from different sources and promote the development of integrative cross-platform analyses.

At this time, the most significant technical obstacle to the use of metadata and ontologies for protein allergens is that existing ontologies are not well developed in this area. For example, the Gene Ontology does not include the term “allergen” although an associated Human Disease ontology includes terms such as “allergy” and “food allergy” (The Gene Ontology Consortium, 2000). On the other hand, the NCI ontology does include the term allergen, but does not differentiate between allergen types. “Allergen” is also a Medical Subject Heading (MeSH) term used by the National Library of Medicine (<http://www.nlm.nih.gov/mesh/>).

The third component needed for the successful application of semantic web technology is widespread collaboration within the field. For example, in addition to collaborating on the development of a useful ontology, database developers would need to provide greater data accessibility. None of the allergen databases, except for the limited data in BIFS, allow users direct access to the database contents. In some cases, users can “drill down” to information on individual allergens but can see only information about a single molecule at any one time. This makes it extremely difficult to write software to automatically query a data set or to carry out integrative or comparative analyses. One easy step toward this goal is for developers to separate the databases from the analysis tools so that external users can access either independently.

#### 4. Conclusions

The design, construction, and maintenance of any bioinformatic database are difficult tasks. This is particularly true for specialized databases intended to support analytical tools rather than to act as sequence repositories. The variety of 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 an 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 the development and use of an allergen ontology and metadata tags could increase the value of these resources. For example, it would be valuable to develop common descriptive terms related to the evidence used to identify a molecule as an allergen or a particular sequence as an allergenic epitope. Realization of the full value of this approach will require that developers “unbundle” databases and analysis tools and provide a path to the data for external applications.

#### Conflict of interest statement

The author declares that there are no conflicts of interest.

#### References

- Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic web. *Scientific American* 284, 34–43.
- Brusic, V., Millor, M., Petrovsky, N., Gendel, S., Gigonzac, O., Stelman, S., 2003. Allergen databases. *Allergy* 58, 1093–1100.
- FAO/WHO, 2001. Evaluation of allergenicity of genetically modified foods. Report of a joint FAO/WHO consultation on food derived from biotechnology, Rome, Italy, 22–25 January 2001.
- Fiers, M., Kleter, G., Nijland, H., Peigneburn, A., Nap, N., Van Ham, R., 2004. Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO codex alimentarius guidelines. *BMC Bioinformatics* 5, 133.
- Gendel, S., 1998. Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Advances in Food and Nutrition Research* 42, 63–92.
- Gendel, S., Jenkins, J., 2006. Allergen sequence databases. *Molecular Nutrition and Food Research* 50, 633–637.
- Goodman, R., 2006. Practical and predictive bioinformatics methods for the identification of potentially cross-reactive protein matches. *Molecular Nutrition and Food Research* 50, 655–660.
- Gruber, T., 1993. A translational approach to portable ontology specifications. *Knowledge Acquisition* 5, 199–220.
- Hileman, R., Silvanovich, A., Goodman, R., Rice, E., Holleschak, G., Astwood, J., Hefle, S., 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *International Archives of Allergy and Immunology* 128, 280–291.
- Hoffman, D., Lowenstein, H., Marsh, D., Platts-Mills, T., Thomas, W., 1994. Allergen nomenclature. *Bulletin of the World Health Organization* 72, 796–806.
- Ivanciu, O., Schein, C., Braun, W., 2003. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Research* 31, 359–362.
- Mari, A., Riccioli, D., 2004. The allergome web site – a database of allergenic molecules. Aim, structure and data of a web-based resource. *Journal of Allergy and Clinical Immunology* 113, S301.
- Rubin, D., Lewis, S., Mungall, C., Misra, S., Westerfield, M., Ashburner, M., Sim, I., Chute, C., Solbrig, H., Story, M., Smith, B., Day-Richter, J., Noy, N., Musen, M., 2006. National center for biomedical ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS: A Journal of Integrative Biology* 10, 185–198.
- Schein, C., Ivanciu, O., Braun, W., 2007. Bioinformatics approaches to classifying allergens and predicting cross-reactivity. *Immunology and Allergy Clinics of North America* 27, 1–27.
- Taylor, S., Lehrer, S., 1996. Principles and characteristics of food allergens. *Food Science and Nutrition* 36, S91–S118.
- The Gene Ontology Consortium, 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25, 25–29.