

Allergen Databases: Current Status and Perspectives

Adriano Mari, MD, Chiara Rasi, BD, Paola Palazzo, BD, and Enrico Scala, MD

Corresponding author

Adriano Mari, MD
Center for Clinical and Experimental Allergology, IDI-IRCCS,
Via dei Monti di Creta 104, 00167, Rome, Italy.
E-mail: adriano.mari@allergome.org

Current Allergy and Asthma Reports 2009, 9:376–383
Current Medicine Group LLC ISSN 1529-7322
Copyright © 2009 by Current Medicine Group LLC

An increasing number of studies on allergenic molecules have been published during the past 20 years, and the number of proteins reported as allergens is close to 1500 (<http://www.allergome.org>). Collecting, organizing, and displaying data reported in the scientific literature is becoming the major commitment of Web-based databases that organize this knowledge in heterogeneous ways. This heterogeneity prevents the databases from being connected to each other, something that has been done in several other biomedical fields. This review reports on the current status of allergen databases and available tools to study the allergenicity of new compounds. An analysis of what has been done by applying bioinformatics in other medical fields is presented. Suggestions on how to create a common platform in which experimental, clinical, and epidemiologic data could be merged are offered. The model of the Allergome platform and its modules and tools (eg, InterAll, ReTiME, RefArray, and AllergomeBlaster) are used to exemplify interconnectivity and data integration.

Introduction

The human immune system is a complex system that aims to trigger responses to dangerous foreign substances. The word *allergy* came into use in the early 1900s to describe shortness of breath [1]. Allergy is a condition in which IgE antibody-mediated immunologic responses occur in reaction to harmless antigens and are tailored at structures generally considered to be “safe.” Allergy affects more than 30% of the population in developed countries and has major socioeconomic consequences [2•]. Allergenic proteins are responsible for the induction of specific IgE antibody production and cause the release of inflammatory mediators by bridging surface-bound IgE on mast cells or basophils.

Cross-reactive allergenic proteins are very similar in sequence and structure, and the cross-reactivity mechanisms can explain why patients sensitized to pollen allergens may also react to several plant foods [3]. Allergenic molecules can be classified in families [4••], but they do not always share a common molecular structure or biochemical function. Allergenicity may be predicted in many cases by starting from primary sequences. In silico methods based on homology searches with sequenced allergens may suggest potential cross-reacting proteins. Introduction of recombinant proteins into food (genetically modified foods), medicine, and other products could expose consumers to new risks, and allergens could be found in unexpected sources. For this reason, a safety evaluation of those genetically engineered products is mandatory [5••].

When applied to the study of infectious diseases, cancer, allergy, and autoimmunity, high-throughput technologies such as expression arrays and proteomics are making available a vast amount of experimental data. Within bioinformatics, computational immunology (immunoinformatics) has an obvious potential to accelerate the flow of information to and from the repositories, and, according to the requirement identified by researchers and clinicians, could significantly improve the task of searching, retrieving, and integrating such knowledge.

Although many allergen-related databases are currently online, and a battery of bioinformatics tools already have been dedicated to the analysis and modeling of immunologic data, the predictive value of this technology is still hindered by the heterogeneity of data sources and structures. The data integration problem has been addressed by several research groups in an attempt to standardize in silico environments by using knowledge most effectively.

To empower reproducibility, reusability, and interoperability, efforts should focus on adopting standard languages and controlled vocabularies (ontologies). Extensible Markup Language (XML) has become the primary method for the exchange of information [6]. Analysis and tools should implement a new approach to send and accept these semistructured data objects, thus allowing people and machines to communicate more effectively.

As the predictive value of a bioinformatics assessment depends on the completeness of the allergen database [7•], text mining could be a valid option to assist database curators in filtering relevant biomedical literature [8]. The advent of Web 2.0 provides researchers with a completely new environment and collaborative platforms to annotate, explore, and understand allergens and related diseases. This review updates readers on the currently available free-access, noncommercial allergen resources and tools [9]; it also compares them with other well-established resources in other disciplines, such as computational biology and general proteomics.

Primary Databases of Biomedical Interest

The growing need for storing and analyzing large datasets led to the establishment of many databases that collect allergenic structures (Table 1). Depending on the inserted data, inclusion criteria, and different user requirements, it is now possible to find databases focused on protein sequence and structures, allergenic sources, routes of exposure, clinical and epidemiologic data, or general information on allergy. All these records could be inter-linked, making it possible to proceed from proteins to structures, organisms, and literature, with the aim of satisfying different needs on various levels.

The massive amount of scientific data generated during the past decades led to an exponential growth of biological databases. According to the Nucleic Acid Research online Molecular Biology Database Collection [10••], 1170 of these resources existed at the beginning of 2009. Among the primary biological databases, the UniProtKB/Swiss-Prot knowledgebase is a central resource for protein sequences and functional information [11]. Other data repositories (eg, GenBank/EMBL/DDBJ) provide the fundamental starting point for describing and understanding the genetic diversity of organisms [12]. The RCSB Protein Data Bank encompasses all the experimentally determined structures of proteins, DNA, and other complex assemblies [13]. The current allergen nomenclature was first released in 1986 and was updated in 1994 by the Allergen Nomenclature Sub-Committee of the World Health Organization and International Union of Immunological Societies (WHO/IUIS) and has been gradually adopted as a standard by primary and specialized databases [14]. Allergens are named according to the abbreviated Linnean genus (a capitalized three-letter word) followed by the first lowercase letter of the species. The name and number indicate the chronological order of the allergen purification (eg, Bet v 1, the first known allergen from the birch pollen *Betula verrucosa*). However, this nomenclature is not always consistent, and there are allowed variations, such as Asp f1 13 and Cand a 1, for *Aspergillus flavus* and *Candida albicans*, respectively. This unambiguous official allergen nomenclature is contained in a database that is available online (<http://www.allergen.org>).

Allergen Databases

The Biotechnology Information for Food Safety (BIFS, <http://www.iit.edu/~sgendel/fa.htm>) database, released in 1998 by the National Center for Food Safety and Technology in Chicago, was the first attempt to provide a complete, nonredundant list of allergens from food and nonfood sources, with the initial aim of assessing the potential allergenicity of genetically modified foods [15]. Allergen source organisms are linked to the National Center for Biotechnology Information (NCBI) taxonomy database [16], and sequences are available through external links. Although the BIFS website clearly states that the database is updated, since the first release, the last update has not been reported. No computational tools have been implemented on the BIFS website thus far.

Allergen Online (<http://www.allergenonline.org>) was established in 2002 within the Food Allergy Research and Resource Program at the University of Nebraska [17]. Version 9.0 contains a list of 1386 sequences of known and putative allergenic proteins from 236 species. Entries are collected from IUIS, PubMed, individual publications, and other allergen databases. Allergen Online is “well curated” because it only contains protein sequences proven to be allergens or from genetic source models that are associated with allergens. Entries only associated with results from a single research project or test, hypothetical translation products, and “homologues” to allergens with no effective proof of IgE binding to sera from individuals with allergies are considered spurious and therefore are omitted. However, a list of possible candidate allergens categorized as “insufficient” is maintained until new evidence is published. For each allergen, the organism’s species, common name, IUIS official nomenclature (if available), protein length, and a link to the NCBI protein specificity viewer are reported.

The Structural Database of Allergen Proteins (SDAP, <http://fermi.utmb.edu/SDAP>) was released in 2001 [18]. SDAP is a platform formed by a continuously updated allergen database. It has a rich battery of computational tools for predicting the allergenicity of proteins for studying the cross-reactivity between allergens and other tasks. To date, SDAP encompasses 737 allergens and isoallergens, 829 protein sequences, and 22 IgE and IgG epitopes. The database lists IUIS and non-IUIS molecules and parasite antigens together with information on the allergen name and type, species from which the allergen originates, its sequence, literature references, IgE epitopes (when available), Pfam protein domains that group related families of proteins, and three-dimensional protein structures. SDAP is updated three times per year, but the date of last update is not reported on the home page. Searches can be done through a user-friendly interface.

The Allergen Database for Food Safety (<http://allergen.nihs.go.jp/ADFS>), which was launched as a project of the Novel Foods and Immunochemistry of the Japanese National Institute of Health Sciences, released a new version in March 2009. This portal contains a database

Table 1. Allergen databases and their main features

Resource	Official allergen nomenclature of WHO/IUIS	BIFS	Allergen Online
URL	http://www.allergen.org	http://www.iit.edu/~sgendel/fa.htm	http://www.allergenonline.org
Data source	Submission	Not reported	Literature and sequence databases
Allergen protein sequence	Yes and external links	External links	Yes and external links
Allergen protein structure	External links	No	No
Information on protein family and domains	No	No	No
Epitopes	No	No	No
Source organisms	All	Food, nonfood, wheat gluten	Food, inhalant, venom/salivary, contact
References	External links	External links	External links
Internal search engine	Yes	No	Yes
Safety evaluation tools	No	No	Yes
Similarity search tools	No	No	Yes
Experimental data	No	No	No
Web services ontologies	No	No	No
User alert	No	No	No
Last update	May 2009	Not reported	January 2009
ADFS—Allergen Database for Food Safety; BIFS—Biotechnology Information for Food Safety; SDAP—Structural Database of Allergen Proteins; WHO/IUIS—World Health Organization/International Union of Immunological Societies.			

of allergic proteins for food safety and allergenicity prediction tools. It is intended to contain all known allergens and B-cell epitope sequences. It currently is populated by 1340 sequences, 76 epitopes, and 57 molecular three-dimensional structures derived from food, airborne, pathogen-related, and venom/salivary allergens. Protein sequences are provided with GenBank accession, links to sequences sharing the same motifs, external links to RSCB structures, and domain information on the InterPro database at the European Bioinformatics Institute (EMBL-EBI).

The Allergome (<http://www.allergome.org>) platform was released on the Web in 2003 with the aim of classifying allergens, IgE-binding antigens, and non-IgE-binding structures [9]. As of March 2009, it contained 4400 molecules spanning more than 1441 organisms from bacteria to mammals and includes many parasite structures, isoforms, and B epitopes. Real-time statistics for the Allergome database recently were implemented (<http://www.allergome.org/script/statistic.php>). Allergome is updated daily on the basis of published literature. The criteria rely on structural relationships with known allergens and IgE-binding capacity to collect as many molecules as possible. Allergome contains all IUIS allergens, which are marked to distinguish them from non-IUIS structures, thus aiding in archive searches. An Allergenicity Score was recently released for allergenic molecules within the Allergome monographs. The Allergenicity Score is real-time, updated score that reports the current characterization status of a

given allergen. It is easily visualized on top of each monograph page using a three-color, 11-dot sequence. This scoring system is the basis for a sophisticated advanced search tool that allows the finest search and retrieval based on the desired allergen characterization level. Records provide information regarding source names, function, source tissue and organism, taxonomy (external links that redirect to taxonomy pages or to Wikipedia), route of exposure, isoforms, primary sequence, molecular weight, sequence homologues, motifs, post-translational factors, and cross-reactivity. All the abstracts from the literature references are stored in a dedicated section and can be displayed from molecule monographs.

Database Tools for Allergenicity Prediction

Analyzing large datasets and comparing results with established knowledge repositories is often mandatory in routine screening and in validating experiments. However, allergenicity risk assessment for novel proteins expressed in genetically modified organisms includes various steps. One step of utmost importance is to determine whether the structure of the novel protein is similar to structures of allergenic proteins. Because high sequence similarity usually means high structural and often functional similarity, several protocols for risk assessment are based on protein similarity searches to known allergens by means of sequence alignment algorithms. Alignments that cover the whole length of two or more sequences are called

Table 1. Allergen databases and their main features (Continued)

SDAP	ADFS	Allergome
http://fermi.utmb.edu/SDAP	http://allergen.nih.gov/ADFS	http://www.allergome.org
Sequence databases	Not reported	Literature and sequence databases
Yes and external links	External links	Yes and external links
External links	External links	No
Pfam, AutoMotif	Pfam, InterPro	AllFam
Yes	Yes	Yes (in silico)
All	All	All
External links	External links	Abstracts, additional information, external links
Yes	Yes	Yes
Yes	Yes	No
Yes	Yes	Yes
No	No	Yes
No	No	No
Newsletter	No	Newsletter
March 2009	March 2009	June 2009

ADFS—Allergen Database for Food Safety; BIFS—Biotechnology Information for Food Safety; SDAP—Structural Database of Allergen Proteins; WHO/IUIS—World Health Organization/International Union of Immunological Societies.

global alignments, whereas local alignments aim to find shorter regions with higher similarity.

BLAST [19] and FASTA [20] algorithms are two approximation methods of the Smith-Waterman algorithm [21], but they are at least 50 times faster and less demanding in terms of computational load. Although they offer the same potentiality in various comparisons, both heuristic methods have advantages and disadvantages: FASTA provides more accurate full-length alignments between sequences but is generally slower than BLAST [22].

In 2003, the Joint Food and Agriculture Organization of the United Nations (FAO)/WHO Expert Consultation on Foods Derived From Biotechnology adopted a risk assessment procedure that involves a recommended scheme for *in silico* methodology. Two criteria are suggested for assessing allergenicity of proteins, both of which are based on similarity searches. The first criterion states that a protein is likely to be an allergen if it shares a sequence of six contiguous peptides with a known allergen. This procedure attempts to recognize common linear epitopes but has been criticized [23] and is not included in the guidance of Codex Alimentarius. Another guideline was released in 2006 by the European Food Safety Authority that recommends that amino acid residues used in the search setting instead be based on a scientifically justified rationale. This approach is used to minimize the potential for false-negative or false-positive results (European Food Safety Authority guidelines for the assessment of genetically modified plants [May 2006]). The second FAO/WHO criterion focuses on the risks entailed due to the presence of conformational epitopes in the observed protein. As reported by Aalberse [24], because cross-reactivity is observed between proteins that share at least 50% iden-

tity (70% in most cases), the proposed method involves the segmentation of the protein into a set of 80–amino acid overlapping sequences. Each set is then compared with a database of allergenic sequences. If a sequence identity greater than 35% is observed, the protein is likely to be an allergen. This method has been optimized to account for identical and similar amino acid residues via the use of the substitution matrices available on the protein similarity search tools. A sequence similarity search is just the first step in the allergenicity evaluation of novel proteins. Further procedures involve serum screenings, *in vitro* digestibility, and tests in animal models. Several of the aforementioned allergen databases include one or more computational tools for alignment and prediction of allergenic structures, which enables rapid identification of novel structures to populate the database.

Allergen Online includes search routines for comparisons of query sequences with database entries. From the website, it is possible to scan the sequence archive using the FASTA3 algorithm [22] (or the FAO/WHO 2001 recommendation) and, by using the same algorithm, search for an identity greater than 35% over every possible 80–amino acid segment of the query protein. In the latter situation, identity scores are adjusted for fragments containing gaps or numbers or for those that are shorter than 80 amino acids.

On the SDAP website, an implementation of the FAO/WHO recommendation is available for protein searches as well. The query sequence may be compared for homology against the SDAP database, used to perform a full FASTA alignment, or used to search for high-significance scores of alignments with 80-mer sliding windows. The sequence identity cutoff is not strictly 35%, as it can be chosen by

the submitter. Even if the default is six, the number of contiguous amino acids in exact matches search can be customized. Instead of a whole protein, users may choose to assess the potential allergenicity of smaller fragments. Sequences shorter than 30 amino acids in length can be submitted to the SDAP and to the UniProtKB databases to search for exact matches, or SDAP can be scanned for proteins with similar physicochemical properties to the protein of interest. This approach is based on a classification of all SDAP allergen subdomains according to their closest Pfam protein families.

A collaboration between the Institute of Immunology Bern Inselspital in Switzerland and the Allergome team led to the use of an original algorithm in an attempt to define and display the mutual relationship between allergens in terms of their motif recognition [25]. The motif discovery tool is running every month using the most updated Allergome list of molecules. Results derived from this automated process are displayed within the allergen monographs in Allergome and are generating new experimental data [26]. The Allergome platform was recently provided with the AllergomeBlaster tool [27], a BLAST algorithm-based utility that performs similarity searches on a regular basis on the Allergome and UniProtKB sequence databases. UniProtKB BLAST results with an identity of 100% with one or more sequences in the Allergome are collected and integrated into the platform as new putative *in silico* allergens, even if they were not previously reported as allergens in the literature. BLAST results with other Allergome molecules are reported as well. The same algorithm can be invoked from within each sequence entry. By setting AllergomeBlaster filters, users can choose the range of sequence similarity to display the results. Correspondences retrieved from the UniProtKB archive could be stripped of entries already reported in the Allergome to display only “new” molecular alignments. Users also can scan the Allergome and UniProtKB archives for similarity matches by submitting their own sequences through the Allergome tools page.

Repositories of Experimental Data: The Importance of Sharing and Integrating Knowledge

One example of sharing data among allergen Web-based resources is the AllFam database (<http://www.meduni-wien.ac.at/allergens/allfam>), which was created at the Vienna Medical University in Austria and is based on collaboration with the Allergome database. The aim of AllFam is to classify allergens into protein families, as has been already done for proteins in the Pfam database. Data are directly downloaded from the Allergome archives and processed, and every new set of allergen family is then displayed again on the Allergome pages. Analysis of the generated data is yielding knowledge on how allergens are distributed differently than protein families [4••].

Because allergen-specific antibodies play a key role in eliciting allergic reactions (type allergy), singleplex diag-

nostics were adapted for detection and quantification of IgE against allergenic molecules in allergic patients' sera. The ability to produce microarrays was the driving force for the creation of more sophisticated diagnostic tools aimed at recognition of the molecular nature of the allergen. Allergen biochips consist of allergenic molecules spotted onto a surface in an array fashion whereby a multitude of allergens can be simultaneously screened for IgE binding after incubation with a small quantity of patient serum [28]. This approach offers the possibility of defining an allergic profile depending on the number of molecules spotted on the chip. The high-throughput capacity of allergen biochips represents a promising technology. As new allergenic molecules are discovered, allergen chips are enriched with new allergen probes. Many studies on allergen microarrays have been published in recent years, and allergen chips are producing vast amounts of experimental data.

To make maximum use of data generated via similar high-throughput methodologies, several repositories of experimental data have been set up in other research fields. Gene Expression Omnibus [29], developed by the NCBI, and the ArrayExpress database of microarray [30], developed by EBI, are examples of how to provide infrastructures and tools for sharing and analyzing microarray data. They also offer facilities to submit and retrieve experimental data and data processing protocols.

Although a great effort has been made to populate allergen databases with records derived from the literature, there is still a gap between these data and the amount of results originating from the routine tests conducted in hospitals and research centers. Allergen databases could be further extended with clinical and epidemiologic data to provide better insight into the variety of phenomena associated with allergy. Allergome-InterAll is a Web-based software created by Allergy Data Laboratories (Latina, Italy) and experimentally implemented at the Center of Clinical and Experimental Allergology, IDI-IRCSS (Rome, Italy) in an attempt to integrate the aforementioned data. It was developed for full, real-time connectivity between medical patient records, laboratory instruments, (namely the Immuno Solid-phase Allergen Chip [ISAC] IgE microarray system [VBC-Genomics, Vienna, Austria]), and the Allergome platform. InterAll is used and shared by allergists and laboratory and administrative personnel. Routine and experimental data are saved in the highest structured form in order to use them in subsequent analyses without interfering with the daily work flow. InterAll is interfaced with the Allergome database, receives real-time updates from the website (eg, information on new entered allergens), and transfers data to the website that directly feeds the Allergome database. Routine tests are performed daily on patients' sera using the ISAC system to detect the IgE specificities against a panel of 103 allergens. Once screening is over, ISAC test results populate the InterAll database in real time, thus enriching the patients' history. Each patient's

data (once entered into the system) are represented by an exhaustive and dedicated monograph that can be edited and expanded instantly by the clinical allergist. The patient's record spans basic demographic data, including geographic coordinates of birth and living places and the patient's familial relationships to other patients. As long as more information is available and new tests are performed, the monographs' information grows, thus assisting in the allergy diagnosis. InterAll is not only a model platform for the allergy-related data management, but it can be customized to include test templates created by users who are not allergists, and thus becomes a powerful instrument for hospitals and research centers in general. InterAll data are connected with the Allergome database and hyperlinks that are available to redirect users at any time to the master database information. User-friendly tools have been implemented that allow visualizing of microarray IgE results by using logical groups or selected organism sources.

Even if an increasing number of institutions possess patient electronic records with a vast amount of clinical data, they cannot be provided with an appropriate computing infrastructure for data analysis. A possible solution for these problems could be the organization of research centers and medical institutions into grid networks. The objective would be to improve access to the shared resources and take advantage of a distributed computational environment. Under this voluntary participation scenario, members from selected institutions would share data and computational resources, idle computers, and dedicated machines to process these large amounts of data and become nodes that exchange data structured according to the same standards [31]. The most important immunology-related grid platform to date was developed in 2006 and funded by the European Commission within the European Immunogrid Project (<http://www.immunogrid.org>). Immunogrid is a human immune system simulator that integrates complex processes at molecular, cellular, and organ levels [32••].

Because data repositories, allergen databases included, commonly have different data formats, access mechanisms, and Web interfaces, the adoption of a common semantic language would greatly enhance their accessibility and usability. This would allow better integration and better data exchange and retrieval. The introduction of high-throughput technologies (eg, microarrays) in routine allergy testing also requires adoption of common standards to make tests and experiments reproducible and sharable. The recently released Clinical and Laboratory Standards Institute document I/LA20-A2 introduced the first single-code system for allergens and describes for the first time in an official document the multiplex (microarray) approach to IgE detection [33••]. Appendix A of that document, using the Allergome classification system, is the first attempt to establish a shared identification system for allergenic sources and molecules. This system could be adopted by

authorities, institutions, health care systems, clinicians, and researchers.

Database Curation: The Importance of Text Mining

The absolute volume of published biomedical research is impressive. To date, the MEDLINE database contains more than 19 million records and spans more than 5000 journal titles [16]. This constitutes a major challenge for database curators, who have to deal with this huge amount of biomedical knowledge and identify abstracts and articles of interest. The curators' task is to convert unstructured scientific data derived from the literature into structured database records. Allergy-related database curators are interested in curating and indexing articles dealing with allergic expressions.

The Allergome database encompasses a manually curated additional database that stores an allergy-related bibliography. It aims to offer better insight on allergen data by integrating information available in the literature. Data contained in the text of most of the published articles are manually extracted and inserted into allergenic source- and molecule-related monographs. By means of a free, user-friendly search tool (Allergome RefArray), selected papers can be scanned for relevant information, and the which can be retrieved by means of a cross-sectional view of the Allergome monograph referencing system. Input fields may be author's name, journal name, publication date, or one or more molecules available in the Allergome archives. Search results on more than 14,000 references (Allergome statistics page) are returned and arranged by topics and allergens. Information displayed comprises both molecular/biochemical data and immunologic, experimental, and diagnosis/immunotherapy-related references.

In most cases, the aim of text mining is the extraction of useful information from unstructured data and not the discovery of new knowledge. It involves three major steps: information retrieval, information extraction, and data mining. The latter operates on structured data and is responsible for the associations among the pieces of information that are extracted by means of machine learning and natural language-processing algorithms.

Although recent success has been achieved in developing text and literature mining tools with applications in the genomic field [34] and clinical domain [35], those systems are not yet as accurate as human curators. Only a few specialized databases rely on the use of text mining tools [34] and, excluding isolated reports [36], there are no specific tools available for literature mining to support automatic curation of allergen databases. A machine learning system that attempts to "filter" relevant documents for curators has been suggested and was recently used as an example in the curation task of an allergy database [8]. In this approach, a text-mining, curator-aided system scans the major databases (eg, PubMed)

for allergen cross-reactivity information, provides a document categorization system that selects the relevant literature, and then analyzes documents for keywords, identifiers, patterns, and ontologies. The next step, the curator's annotation, becomes the basis for the classifier training. Machine learning algorithms can significantly reduce the workload of database curators, but their adoption depends on the availability of user-friendly text-mining tools and on full access to documents.

Conclusions

High-throughput technologies such as microarrays and proteomics are making available a huge amount of biomedical data. Intuitive access to this information is crucial for scientists and physicians. For an in-depth analysis of data, an integration of resources from several areas is necessary. Research results need to be integrated with patient clinical records, epidemiologic data, and current medical and molecular knowledge. Several allergy-related databases for collecting allergen-related information have been created with different inclusion criteria: protein sequence and structures, allergenic sources, routes of exposure, clinical and epidemiologic data, or general information on allergy. Creating allergy database designs and constructions is difficult, and their maintenance is a time- and money-consuming process. Many allergen databases have been implemented and then deserted once a project was completed or the funding ended. Although efforts have been made to provide some standards for data recognition and exchange, allergy databases still provide users with heterogeneous data formats. The adoption of a semantic language would enhance exchange and retrieval of information from disparate databases. The introduction of a common standard in reporting results of microarrays or other molecular-based allergy screening tests in electronic medical records would guarantee their unambiguous interpretation and reproducibility. In this scenario, reorganization of research centers and medical institutions working in the allergy domain into networks would enable a fruitful collaboration with the aim of sharing data and computational resources. To acquire new knowledge, collective and collaborative platforms for data creation should be investigated, thereby allowing a pluralistic growth of data repositories [37]. Literature-mining tools, based on controlled vocabularies, could be an equally valid option to help feed allergy databases.

Acknowledgments

The Allergome project receives funding from the following institutions: University of Nebraska Food Allergy Research and Resource Program, University of Salzburg Priority Programme BioScience and Health, and University Gabriele D'Annunzio di Chieti e Pescara Allergy and Clinical Immunology Section.

Disclosure

The Allergome project is supported by unrestricted grants from Phadia AB, Indoor Biotechnologies, Siemens Medical Solutions Diagnostics, VBC-Genomics, Biomay, Probelte Pharma, and HAL Allergy.

Dr. Mari is the creator of the Allergome platform. He acts as chief scientific and business officer of Allergy Data Laboratories, a nonprofit company devoted to managing allergy-related data. No other potential conflicts of interest relevant to this article were reported.

References and Recommended Reading

Papers of particular interest, published recently, have been highlighted as:

- Of importance
 - Of major importance
1. Cohen SG, King JR: **Skin tests: a historic trail.** *Immunol Allergy Clin North Am* 2001, **21**:191–249.
 - 2.• Kamble S, Bharmal M: **Incremental direct expenditure of treating asthma in the United States.** *J Asthma* 2009, **46**:73–80.
- This is the most recent study on cost-related data in allergic diseases.
3. Ferreira F, Hawranek T, Gruber P, et al.: **Allergic cross-reactivity: from gene to the clinic.** *Allergy* 2004, **59**:243–267.
 - 4.•• Radauer C, Bublin M, Wagner S, et al.: **Allergens are distributed into few protein families and possess a restricted number of biochemical functions.** *J Allergy Clin Immunol* 2008, **121**:847–852.
- This was the first study on allergen classification into families comparing results with protein distribution within protein families. It is an example of established collaboration on allergens between a database resource and a research center.
- 5.•• Thomas K, Macintosh S, Bannon G, et al.: **Scientific advancement of novel protein allergenicity evaluation: an overview of work from the HESI Protein Allergenicity Technical Committee (2000–2008).** *Food Chem Toxicol* 2009, **47**:1041–1050.
- This was a comprehensive report on the current status of knowledge on allergenicity evaluation of novel proteins.
6. Li X, Zhang Y: **Bioinformatics data distribution and integration via Web Services and XML.** *Genomics Proteomics Bioinformatics* 2003, **1**:299–303.
 - 7.• Madsen CB, Hattersley S, Buck J, et al.: **Approaches to risk assessment in food allergy: report from a workshop “Developing a framework for assessing the risk from allergenic foods.”** *Food Chem Toxicol* 2009, **47**:480–489.
- This was a full update on strategies for risk assessment evaluation of allergenic foods.
8. Miotto O, Tan TW, Brusica V: **Supporting the curation of biological databases with reusable text mining.** *Genome Inform* 2005, **16**:32–44.
 9. Mari A, Scala E, Palazzo P, et al.: **Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model.** *Cell Immunol* 2007, **244**:97–100.
 - 10.•• Galperin MY, Cochrane GR: **Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009.** *Nucleic Acids Res* 2009, **37**:D1–D4.
- This is the latest and most comprehensive list of worldwide biomedical databases and bioinformatics tools.
11. Boutet E, Lieberherr D, Tognolli M, et al.: **UniProtKB/Swiss-Prot.** *Methods Mol Biol* 2007, **406**:89–112.
 12. Kanz C, Aldebert P, Althorpe N, et al.: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Res* 2005, **33**:D29–D33.

13. Berman HM, Westbrook J, Feng Z, et al.: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235–242.
 14. King TP, Hoffman D, Lowenstein H, et al.: **Allergen nomenclature.** *J Allergy Clin Immunol* 1995, **96**:5–14.
 15. Gendel SM: **Assessing the potential allergenicity of new food proteins.** *Food Biotechnol* 1998, **12**:175–185.
 16. Sayers EW, Barrett T, Benson DA, et al.: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, **37**:5–15.
 17. Hileman RE, Silvanovich A, Goodman RE, et al.: **Bioinformatic methods for allergenicity assessment using a comprehensive allergen database.** *Int Arch Allergy Immunol* 2002, **128**:280–291.
 18. Ivanciuc O, Schein CH, Braun W: **SDAP: database and computational tools for allergenic proteins.** *Nucleic Acids Res* 2003, **31**:359–362.
 19. Altschul SF, Gish W, Miller W, et al.: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
 20. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**:2444–2448.
 21. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195–197.
 22. Pearson WR: **Comparison of methods for searching protein sequence databases.** *Protein Sci* 1995, **4**:1145–1160.
 23. Silvanovich A, Nemeth MA, Song P, et al.: **The value of short amino acid sequence matches for prediction of protein allergenicity.** *Toxicol Sci* 2006, **90**:252–258.
 24. Aalberse RC: **Structural biology of allergens.** *J Allergy Clin Immunol* 2000, **106**:228–238.
 25. Stadler MB, Stadler BM: **Allergenicity prediction by protein sequence.** *FASEB J* 2003, **17**:1141–1143.
 26. Marti P, Truffer R, Stadler MB, et al.: **Allergen motifs and the prediction of allergenicity.** *Immunol Lett* 2007, **109**:47–55.
 27. Rasi C, Palazzo P, Carabella G, Mari A: **AllergomeBlaster: Auto-collecting new potential in silico allergens for the Allergome platform.** *Allergy* 2008, **63**:S400.
 28. Hiller R, Laffer S, Harwanegg C, et al.: **Microarrayed allergen molecules: diagnostic gatekeepers for allergy treatment.** *FASEB J* 2002, **16**:414–416.
 29. Barrett T, Suzek TO, Troup DB, et al.: **NCBI GEO: mining millions of expression profiles—database and tools.** *Nucleic Acids Res* 2005, **33**:D562–D566.
 30. Brazma A, Parkinson H, Sarkans U, et al.: **ArrayExpress—a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31**:68–71.
 31. Loewe L: **Global computing for bioinformatics.** *Brief Bioinform* 2002, **3**:377–388.
 32. Pappalardo F, Halling-Brown MD, Rapin N, et al.: **ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization.** *Brief Bioinform* 2009, **10**:330–340.
- This is the first example of collaborative network in the immunology field.
33. Matsson PNJ, Hamilton RG, Esch RE, et al.: **Analytical Performance Characteristics and Clinical Utility of Immunological Assays for Human Immunoglobulin E (IgE) Antibodies and Defined Allergen Specificities; Approved Guideline, edn 2.** Wayne, PA: Clinical and Laboratory Standards Institute; 2009.
- The updated release of approved guidelines on IgE detection officially reports for the first time about coding of allergenic molecules and allergenic sources and on microarray testing in the allergy field.
34. Donaldson I, Martin J, de Bruijn B, et al.: **PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine.** *BMC Bioinformatics* 2003, **4**:11.
 35. Cheng D, Knox C, Young N, et al.: **PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites.** *Nucleic Acids Res* 2008, **36**:W399–W405.
 36. Valarakos AG, Karkaletsis V, Alexopoulou D, et al.: **Building an allergens ontology and maintaining it using machine learning techniques.** *Comput Biol Med* 2006, **36**:1155–1184.
 37. Gendel SM: **Allergen databases and allergen semantics.** *Regul Toxicol Pharmacol* 2008 Dec 6 (Epub ahead of print).