BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 7 2009, pages 979–980
doi:10.1093/bioinformatics/btp077

Databases and ontologies

Allergen Atlas: a comprehensive knowledge center and analysis resource for allergen information

Joo Chuan Tong[1], Shen Jean Lim[2], Hon Cheng Muh[3], Fook Tim Chew[3] and Martti T. Tammi[2,3,*]

[1]Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632, [2]Department of Biochemistry, Yong Loo Lin School of Medicine and [3]Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

Received on November 21, 2008; revised and accepted on February 5, 2009

Advance Access publication February 11, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: A variety of specialist databases have been developed to facilitate the study of allergens. However, these databases either contain different subsets of allergen data or are deficient in tools for assessing potential allergenicity of proteins. Here, we describe Allergen Atlas, a comprehensive repository of experimentally validated allergen sequences collected from in-house laboratory, online data submission, literature reports and all existing general-purpose and specialist databases. Each entry was manually verified, classified and hyperlinked to major databases including Swiss-Prot, Protein Data Bank (PDB), Gene Ontology (GO), Pfam and PubMed. The database is integrated with analysis tools that include: (i) keyword search, (ii) BLAST, (iii) position-specific iterative BLAST (PSI-BLAST), (iv) FAO/WHO criteria search, (v) graphical representation of allergen information network and (vi) online data submission. The latest version contains information of 1593 allergen sequences (496 IUIS allergens, 978 experimentally verified allergens and 119 new sequences), 56 IgE epitope sequences, 679 links to PDB structures and 155 links to Pfam domains.

Availability: Allergen Atlas is freely available at http://tiger.dbs.nus.edu.sg/ATLAS/.

Contact: martti@nus.edu.sg.

1 INTRODUCTION

In the last decade, a variety of specialized databases have been developed to facilitate the study of allergens (Brusic *et al*., 2003). Some of these databases contain basic allergen sequences and related information (Bairoch *et al*., 2004; Gendel, 1998), some have basic sequence comparison tools (Wu *et al*., 2006), while others have additional tools (Hileman *et al*., 2002) for the assessment of allergenicity based on the FAO/WHO expert group recommendations (Gendel, 2004). Each of these databases has a different focus with emphasis on different groups of allergens. Databases such as Bioinformatics for Food Safety (Gendel, 1998), Food Allergy Research and Resource Program (Hileman *et al*., 2002) and International Union of Immunological Societies (IUIS) Nomenclature Sub-Committee Allergen Database (King *et al*., 1994) are rich in content but contain different subsets of allergen sequences. The majority of these databases either lack complete information on available allergens or are deficient in bioinformatic tools for analyzing the stored sequences. Other databases, including Allergome (Mari *et al*., 2005) and SDAP (Ivanciuc *et al*., 2003) contain putative allergens annotated using sequence similarity to the verified allergens.

To fill this important gap in existing resources, we report Allergen Atlas, a comprehensive repository of allergen data collected from in-house laboratory, online data submission, literature reports and all existing general-purpose and specialist databases. The database is integrated with a suite of bioinformatic tools to facilitate data analysis, visualization and retrieval, including keyword and sequence similarity searches. The main purpose of Allergen Atlas is to support molecular studies of allergens, assessment of allergic responses and of allergic cross-reactivity.

2 METHODS

2.1 Construction and Implementation

Allergen Atlas is a manually curated PostgreSQL (www.postgresql.org) database hosted on a Linux server. It contains experimentally determined allergen information from in-house laboratory, online submission, literature reports and all existing general-purpose and specialist databases. The most important characteristics of allergens were extracted, manually verified, classified and stored in the database. Each entry is annotated with the following information, where available: (i) allergen name, (ii) scientific and common names of the source organisms, (iii) type, (iv) sequence, (v) bibliographic references, (vi) IgE epitope sequence and (vii) allergen public database accessions that include Swiss-Prot (Boeckmann *et al*., 2003), Protein Data Bank (PDB; Berman *et al*., 2000), Gene Ontology (GO; The Gene Ontology Consortium, 2000), Pfam(Bateman *et al*., 2002) and PubMed. IgE epitopes of known allergens were extracted from Bcipep (Saha *et al*., 2005). Swiss-Prot and PDB provide well-defined sequence and structure information, respectively; GO provides a description of gene and gene product attributes; while Pfam details information about the protein domains and families of existing allergens. This information is included for comprehensiveness in coverage.

3 RESULTS

3.1 Database contents

A total of 1593 experimentally validated allergens with their sequence information were stored in Allergen Atlas. A significant

*To whom correspondence should be addressed.

portion of these sequences (119/1593 or 7.5%) were experimentally determined in our in-house laboratory and remain as unpublished results. In addition, the database contains entries of 496 IUIS allergens, 978 non-IUIS allergens, 56 IgE epitopes, 679 links to PDB structures and 155 links to Pfam domains collected through exhaustive manual searching of primary literature, as well as general and specialist databases.

3.2 Capabilities

The Allergen Atlas web interface allows for keyword search as well as sequence similarity searches of stored allergens.

A text-based keyword search function permits general survey of specific allergens stored in the database. Users can query the database based on allergen name, organism name, pathway or bibliographical references. Cross-reference searches can be performed using GO accession, Pfam accession or Swiss-Prot accession.

BLAST (Altschul *et al*., 1990) is a local sequence comparison tool that outputs information on allergens containing similar regions with the query sequence. BLAST searches allow users to identify matching or similar sequences and display the results in the form of a table. A variant of BLAST, the position-specific iterative BLAST (PSI-BLAST; Altschul *et al*., 1997), is also included in Allergen Atlas to facilitate the identification of weak relationships of the query sequence to annotated entries in the database which may not detected by a BLAST search. Recent studies have shown that this approach can predict allergens with up to 95.02% accuracy (Lim *et al*., in press).

The FAO/WHO criteria search is a sequence similarity search tool for assessing potential allergenicity of proteins in accordance to the current FAO/WHO Codex alimentarius guidelines which comprises of two rules—rule 1: a sequence identity of six consecutive amino acids between the sequences of the query protein and an experimentally verified allergen; or rule 2: a sequence identity of >35% over a stretch of 80 amino acids (FAO/WHO, 2003). This approach has been adopted by numerous research groups including Fiers *et al*. (2004) and Gendel(1998). However, concerns have been raised that the precision was reportedly low for methods solely relying on the six amino acid rule (Silvanovich *et al*., 2006). Allergen Atlas allows for more stringent searches of the FAO/WHO protocol by allowing users to define input parameters such as the number of contiguous amino acids for screening (rule 1) as well as the sequence identity threshold (rule 2).

To facilitate data interpretation, users are provided capabilities for displaying the relationships of allergen data using a graphical visualization module. Given a list of selected entries returned from a search query, users can select a list of display options including (i) allergen name, (ii) organism name, (iii) Swiss-Prot accession, (iv) PDB accession, (v) GO accession, (vi) Pfam accession, (vii) IgE epitope and (viii) IgG epitope. The graphical visualization module allows for the display of an allergen information network based on the selected entries and annotations.

4 CONCLUSION

Allergen Atlas has been developed to facilitate research in allergology. To enhance the usefulness of this resource, newly validated allergen sequences from our in-house laboratory, and primary literature will be constantly added to the database. We look forward to the day when researchers worldwide will voluntarily share their experimental data and upload their findings to an online repository, such as ours, much as today where we upload our own experimental data to share with the research community. An online submission website was specifically developed for such a purpose. In addition to high quality data, essential analytical tool resources that are widely accepted in the scientific community are also provided in Allergen Atlas. The list of bioinformatic tools will be periodically updated based on input from the scientific community. With advances in clinical allergology, genomics and proteomics, we envision a future in which large amounts of data will be available for the study of allergens, which will be included in Allergen Atlas and provided to the research community.

Funding: National University of Singapore (R154000265112).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A. *et al*. (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.*, **5**, 39–55.

Bateman,A. *et al*. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.

Brusic,V. *et al*. (2003) Allergen databases. *Allergy*, **58**, 1093–1100.

FAO/WHO. (2003) *Codex Principles and Guidelines on Foods derived from Biotechnology*. Joint FAO/WHO Food Standards Programme, Rome, Italy.

Fiers,M.W. *et al*. (2004) Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics*, **5**, 133.

Gendel,S.M. (2004) Bioinformatics and food allergens. *J. AOAC Int.*, **87**, 1417–1422.

Gendel,S.M. (1998) Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food. Nutr. Res*., **42**, 63–92.

Hileman,R.E. *et al*. (2002) Bioinformatic methods for allergenicity assessment using a comprehensive ALLERGEN database. *Int. Arch. Allergy Immunol.*, **128**, 280–291.

Ivanciuc,O. *et al*. (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.*, **31**, 359–362.

King,T.P. *et al*. (1994) Allergen nomenclature. WHO/IUIS allergen nomenclature subcommittee. *Int. Arch. Allergy Immunol*., **105**, 224–233.

Lim,S.J. *et al*. The value of position-specific scoring matrices for assessment of protein allergenicity. *BMC Bioinformatics*, **9**, 12:S21.

Mari,A. *et al*. (2005) Allergome – a database of allergenic molecules: structure and data implementations of a web-based resource. *J. Allergy Clin. Immunol.*, **115**, S87.

Saha,S. *et al*. (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics*, **6**, 79.

Silvanovich,A. *et al*. (2006) The value of short amino acid sequence matches for prediction of protein allergenicity. *Toxicol. Sci*., **90**, 252–258.

The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Wu,C.H. *et al*. (2006) The universal protein resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.