



ALLERDB database and integrated bioinformatic tools for assessment of allergenicity and allergic cross-reactivity

Zong Hong Zhang ^a, Stephanie C.C. Tan ^b, Judice L.Y. Koh ^a,
Andras Falus ^c, Vladimir Brusic ^{d,*}

^a Institute for Infocomm Research, Singapore 119613, Singapore

^b Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore

^c Department of Genetics, Cell and Immunobiology, Semmelweis University, Budapest, Hungary

^d School of Land and Food Sciences, University of Queensland, Brisbane, Qld 4072, Australia

Received 12 January 2007; accepted 31 January 2007

Abstract

Databases and computational tools are increasingly important in the study of allergies, particularly in the assessment of allergenicity and allergic cross-reactivity. ALLERDB database contains sequences of allergens and information on reported cross-reactivity between allergens. It focuses on analysis of allergenicity and allergic cross-reactivity of clinically relevant protein allergens. The official IUIS allergen data were extracted from the IUIS Allergen Nomenclature Sub-Committee website, and their sequence information from the public databases, and reference publications. The analysis tools assist allergen data analysis and retrieval, and include keyword searching, BLAST, prediction of allergenicity, modification of BLAST that displays cross-reactive allergens, and graphics representation of cross-reactivity data. ALLERDB is new brand of allergen databases with a rich set of tools for sequence comparison, pattern identification, and visualization of results. It is accessible at <http://research.i2r.a-star.edu.sg/Templar/DB/Allergen>.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Allergen nomenclature; Allergenicity; Allergic cross-reactivity; FAO/WHO decision tree

1. Introduction

Bioinformatics is an important tool for managing allergen data, analysis of properties of allergens, assessment of protein allergenicity, and analysis of potential allergic cross-reactivity. Sequences of allergenic proteins are scattered across multiple databases (Table 1), including general-purpose protein databases and specialist allergen databases. Some of these databases contain basic allergen sequences and related information [1–4], some have basic

sequence comparison tools [5–7], while others have additional tools [8,9] for the assessment of allergenicity based on the FAO/WHO expert group recommendations [10]. These databases and their allergen-related content have been described elsewhere [11,12]. Each of these databases has a different focus and they contain different subsets of allergen sequences. The most authoritative data source in terms of content is the IUIS Nomenclature Sub-Committee Allergen Database [4] (IUIS allergens), which defines allergens according to a set of well-defined criteria, but it does not have direct links to allergen sequences. Other databases, such as Allergome and SDAP, contain IUIS allergens, but they may also include putative allergens annotated using sequence similarity to the verified allergens.

Identification of allergic potential is a key issue whenever new proteins are brought into contact with humans,

Abbreviations: BLAST, Basic Local Alignment Search Tool; FAO, Food and Agriculture Organisation; IUIS, International Union of Immunological Societies; Pfam, Protein families database; WHO, World Health Organisation

* Corresponding author.

E-mail address: v.brusic@uq.edu.au (V. Brusic).

Table 1
Online molecular databases of allergens

Database	URL	Reference
*UniProt	www.uniprot.org	[1]
*GenPept	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein	[2]
SwissProt allergens	www.expasy.org/cgi-bin/lists?allergen.txt	[3]
BIFS	www.iit.edu/~sgendel/fa.htm	[4]
Allergome	www.allergome.org	[5]
IUIS allergens	www.allergen.org	[6]
CSL allergen	www.csl.gov.uk/allergen	[7]
FARRP	allergenonline.com	[8]
SDAP	fermi.utmb.edu/SDAP	[9]
ALLERDB	antigen.i2r.a-star.edu.sg/Templar/DB/allergen	This paper

* General-purpose databases.

either through food or other modes of exposure. The current joint recommendation by the FAO and the WHO is a scheme based on a decision tree which involves sequence similarity search against known allergens [10]. Two decision criteria have been proposed for bioinformatic assessment of allergenicity: identity of six or more contiguous amino acids, or minimum 35% identity over an 80 amino-acid segment [13]. While useful in some cases, the six amino-acid identity rule may be not practical because it produces a large number of random hits [8–14]. Several web applications that use the two decision criteria have been reported [8,9,15]. New algorithms based on identification of presence of allergen-specific motifs that show promising results in predicting allergenicity have also been reported [16–18]. The question of what represents a biologically significant sequence similarity match between a query protein and known allergens remains yet unanswered and there is a need for further research in this field. In the meantime, the FAO/WHO recommendations remain a standard for the assessment of allergenicity.

Allergic cross-reactivity is a significant clinical problem—even distant allergens can cause cross-reactivity [19]. In the latex-fruit syndrome, for example, cross-reactivity between allergens derived from latex and from food plants has been observed [20]. This and similar examples indicate a major need for methods to identify potential cross-reactive allergens, particularly those that have low sequence similarity. The molecular basis for allergic cross-reactivity is in structural similarity among proteins from diverse sources and the clinical relevance appears to be influenced by several factors including the IgE response against the allergen, route of exposure, and the sequence and structure of the allergen [21]. Better understanding of the basis of allergic cross-reactivity is essential for diagnostics and control of severe allergies [22] and the design of immunotherapies [23].

Hereby we report ALLERDB, a database that contains sequences of allergens approved by IUIS Allergen Nomenclature Committee and information on reported cross-reactivity between allergens. ALLERDB focuses on analysis of

allergenicity and allergic cross-reactivity of clinically relevant protein allergens. It contains integrated bioinformatic tools for performing sequence similarity search, assessment of potential allergenicity, and analysis of allergic cross-reactivity. The main purpose of ALLERDB is the support of molecular studies of allergens, the assessment of allergic responses and of allergic cross-reactivity. ALLERDB is accessible at <http://research.i2r.a-star.edu.sg/Templar/DB/Allergen>.

2. Database description

2.1. Data

The allergen and isoallergen lists were extracted from the IUIS allergens website (<http://www.allergen.org>) along with the data on scientific and common names of the source organisms, allergen citation, and allergen public database accession (from GenBank, SwissProt, or PIR). All IUIS allergens that have protein sequences available in the public sequence databases were entered into ALLERDB database. In addition, links to well-defined protein domains in Pfam [24] are included in the records, if available. For IUIS-listed allergens that contain no protein sequence entries in public database but allergen sequence was available in publication reference is available, protein sequence records were entered into the ALLERDB. The same approach was applied to allergens available in the IUIS isoallergen list. IUIS-listed allergens whose sequence information is not available, were not included in the ALLERDB. IUIS allergens that have multiple reference sequences, have these sequences as separate record in the ALLERDB.

ALLERDB records contain fields that have information on allergen sequences as well as the record management information. ‘DBACC’ is a unique number defines identifier of each record in the database. The field ‘Date’ identifies the time when the entry been created. The ‘Name’ field contains the IUIS nomenclature name of the allergen. ‘SOURCE’ field contains the organism source of the allergen. ‘SEQLINK’ field provides hyperlinks to the external database from which the sequence information of the allergen was extracted. The actual sequence is recorded in ‘SEQUENCE’ field. ‘COMMENT’ field shows relevant comments or observations, if any. The ‘Feature’ field has several sub fields such as, ‘Pfam domains’—the protein family, conserved domains, and hyperlinks to their PFAM entries. The ‘XRallergen’ field contains links to reported cross-reactive allergens within ALLERDB. Journal article reference that contains the information is listed in ‘XRreference’. ‘Isoallergen’ field contains isoallergens of individual allergen and provides hyperlinks to their full records.

From the list IUIS Allergen list released on 12 September 2005, we extracted 779 allergens and isoallergens. Among them, we found 671 had sequence information available and were included in the ALLERDB database. It contains entries of 398 allergens, 273 isoallergens, 358

links to Pfam domains, and 128 instances of reported cross-reactivity collected from the literature and verified using a text-mining tool ABK [25].

2.2. Analysis tools

To assist allergen data analysis and retrieval, ALLERDB has five integrated tools, BLAST, Search, XRblast, AllerPredict, and XRgraph.

Search is a retrieval tool which enables user extract record from the database if there is a match to the supplied keyword anywhere in the text of the record. The output of this tool is a table containing accession numbers hyper-linked to their records, source organism, and allergen names (nomenclature). Individual records from the table of matching records can be accessed through hyperlinks attached to the accession numbers. A sample output of Search is given in Fig. 1a and b.

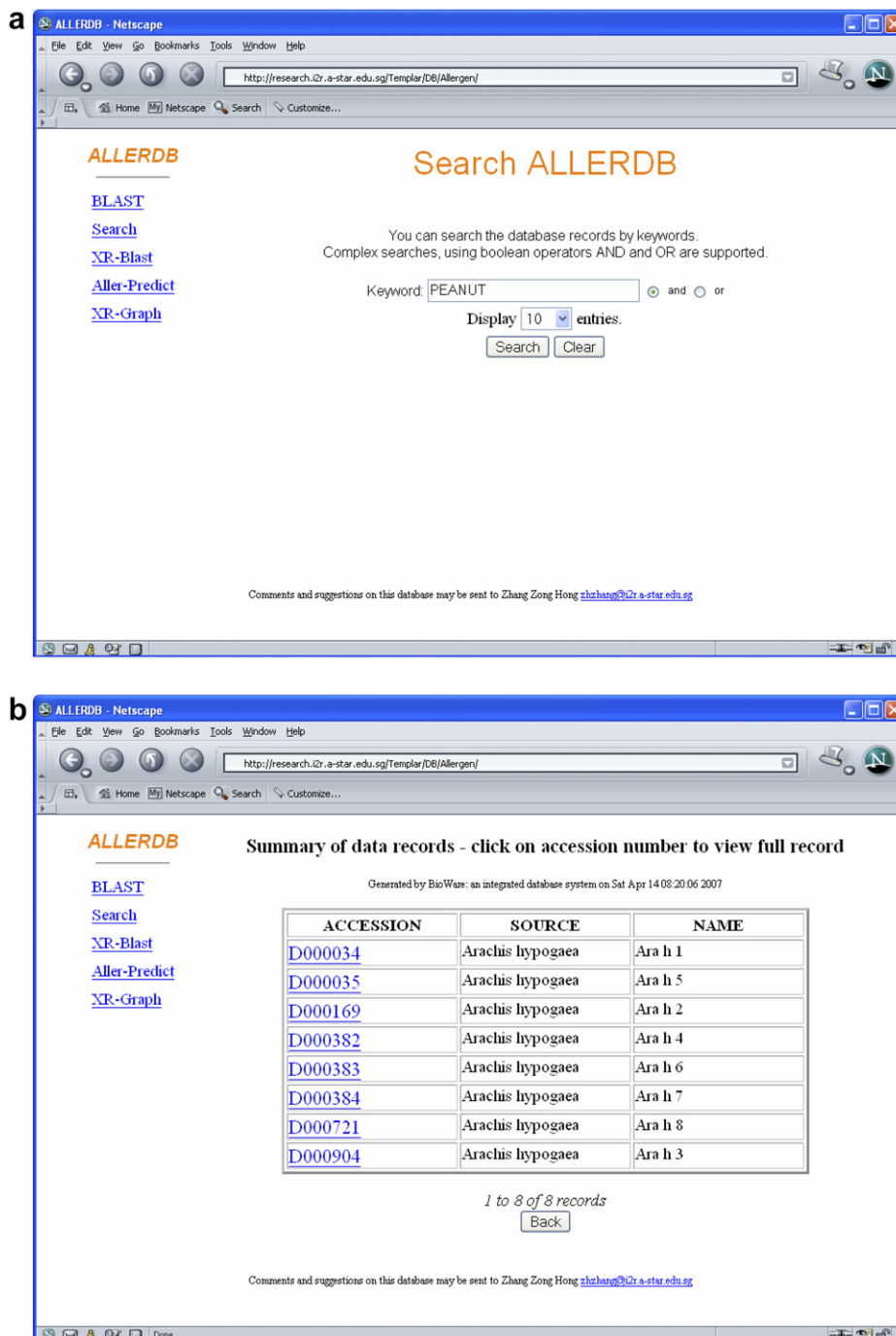


Fig. 1. (a) A Search query “PEANUT” has been entered into the keyword field. (b) Output after query entered in Fig. 1a.

XRblast is a modification of the BLAST sequence comparison tool. BLAST [26] is a local sequence alignments tool. It enables comparison of query sequence with sequences stored in the ALLERDB database and displays the alignments. The results are reported in a form of a ranked list followed by a series of individual sequence alignments. Two scores, one that indicates similarity score of the match (in bits), and another that indicates the Expectation Value of the match (proxy for statistical significance)

are included in the output. The BLAST version implemented in the current version of ALLERDB is BLAST2.2.3. The XRblast performs BLAST search as well as post processes the output of BLAST and produces enhanced output—standard output of the BLAST search with sequences that have high similarity to the query protein (matches) plus information on allergens that have reported cross-reactivity with the individual matches. Both matches and cross-reactive allergens are hyperlinked to

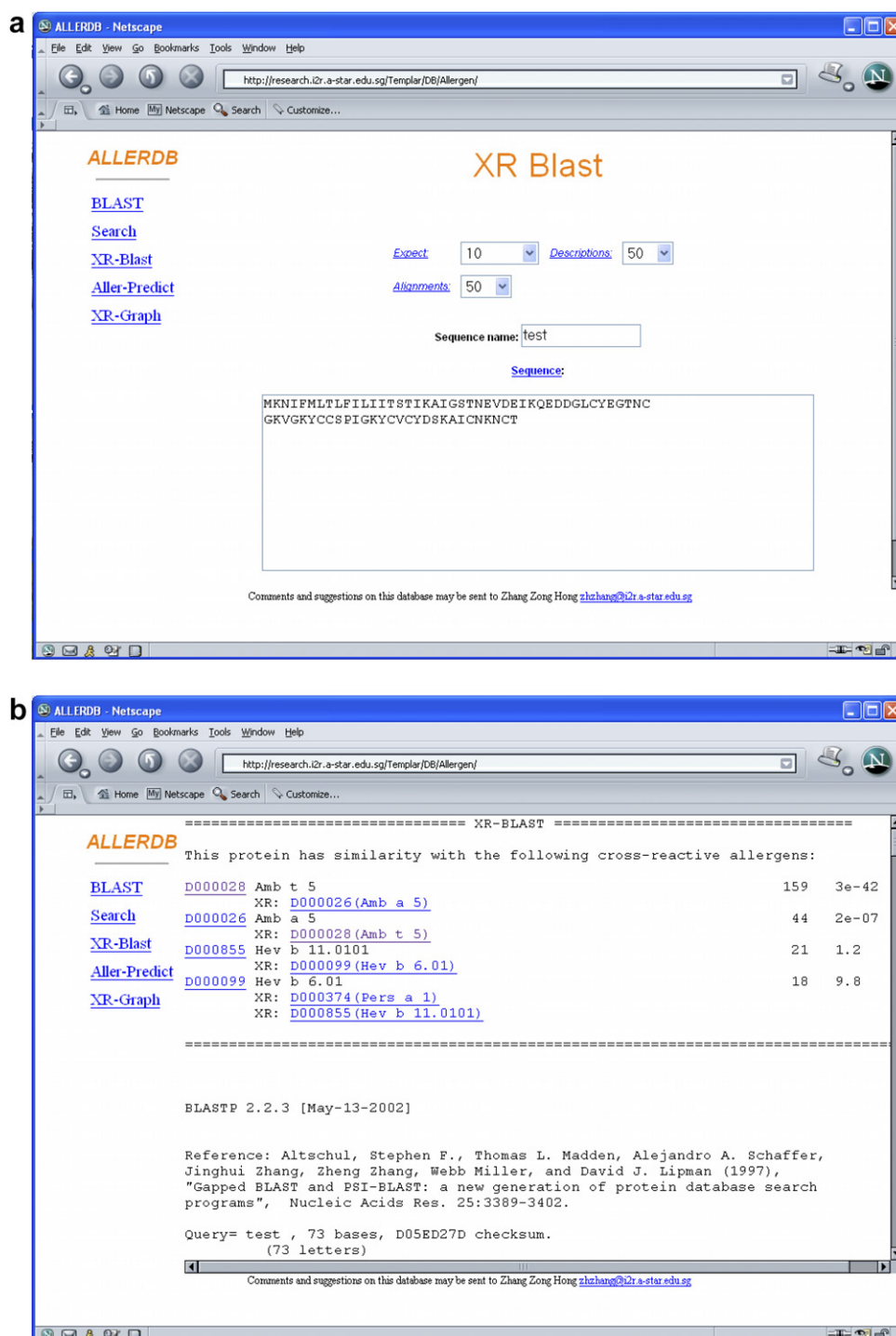


Fig. 2. (a) Input example sequence for XRblast. (b) Output of query in (a).

their records from the XRblast output. A sample output of XRblast is given in Fig. 2a and b.

AllerPredict is a sequence similarity search tool that reports sequence similarity in accordance to FAO/WHO recommendation for the assessment of allergenicity. Both the six contiguous amino acids identity rule, and >35% identity over a stretch of 80 amino acids are implemented. Users

can submit the protein of interest to AllerPredict, which will return the list of matches that satisfy either of the rules. Sample query and output is shown in Fig. 3a and b.

XRgraph is a visual tool which enables users to display graphical representation of the cross-reactivity information on allergens. The graphs are created dynamically from the contents of the database and displays allergens (boxes) that

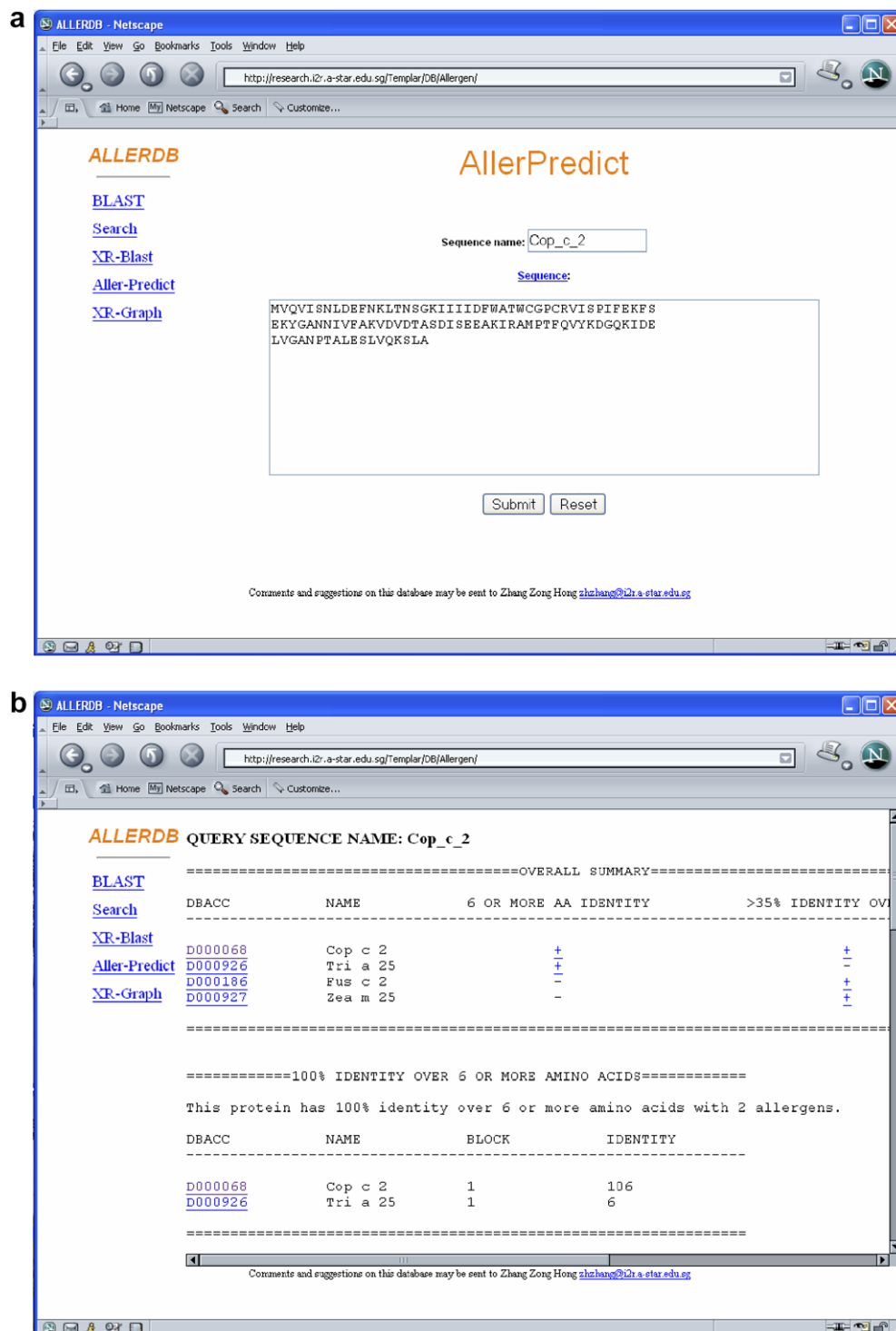


Fig. 3. (a) Example query for AllerPredict. (b) Output of Fig. 3a, shown the sequences which have six contiguous amino acids identity with the query sequence, or >35% identity over a stretch of 80 amino acids comparing to the query sequence.

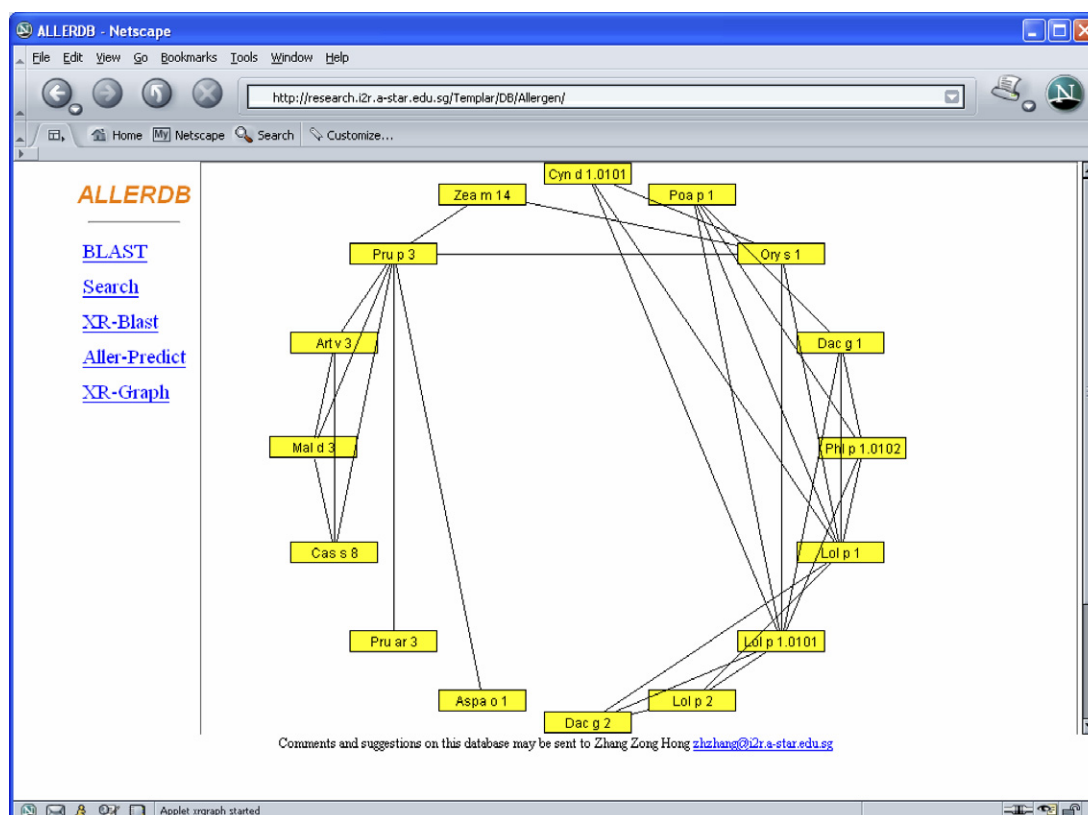


Fig. 4. An example of XR Graph output for Lol p 1.

are related by reported cross-reactivity (links). This visual tool enables user to establish possible allergen cross reactivity relationships not reported before. This tool might be use-

ful in the development of novel allergy diagnostics approaches. A sample output of XRgraph is shown in Fig. 4.

2.3. Database construction

The process of building ALLERDB database is shown in Fig. 5. It includes data acquisition and inclusion into a database, additional annotations, and integration of analysis tools. The entire ALLERDB database has been created and maintained using BioWare [27] data warehousing system.

3. Conclusion

ALLERDB is a database of allergenic proteins with integrated bioinformatic tools for the assessment of allergenicity and allergic cross-reactivity. It facilitates the assessment of allergenic potential of anonymous proteins that might be introduced in food chain or potential close contact with allergic individuals. This database is along with SDAP, a representative of new brand of allergen databases which include a rich set of tools, including those for sequence comparison, pattern identification, and visualization of results. With advances in clinical allergology, genomics and proteomics, large amounts of data will be available for the study of allergens and bioinformatics tools will be essential in handling this data. We anticipate the emergence of more advanced tools for prediction of aller-

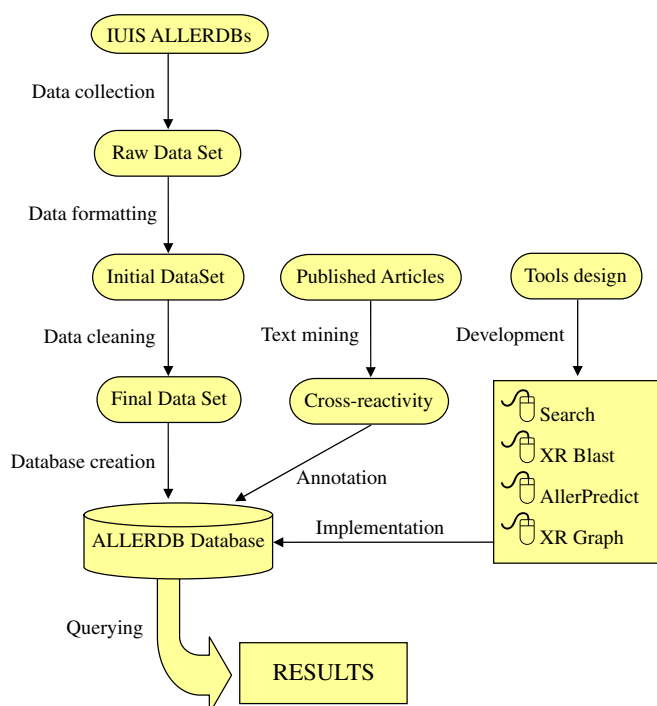


Fig. 5. The construction of ALLERDB database.

genicity, which will be included in ALLERDB and provided to the research community.

Acknowledgment

V.B. acknowledges the support of the ImmunoGrid project, under EC contract FP6-2004-IST-4, No. 028069.

References

- [1] S.M. Gendel, *Adv. Food Nutr. Res.* 42 (1998) 63–92.
- [2] A. Mari, V. Mari, A. Ronconi, *J. Allergy Clin. Immunol.* 115 (2005) S87.
- [3] The Central Science Laboratory, www.csl.gov.uk/allergen.
- [4] D. Hoffman, H. Lowenstein, D. Marsh, T. Platts-Mills, W. Thomas, *Bull. World Health Organ.* 72 (1994) 796–806, www.allergen.org.
- [5] A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, *Brief. Bioinform.* 5 (2004) 39–55.
- [6] C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, et al., *Nucleic Acids Res.* 34 (2006) D187–D191.
- [7] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, et al., *Nucleic Acids Res.* 34 (2006) D173–D180.
- [8] R.E. Hileman, A. Silvanovich, R.E. Goodman, E.A. Rice, G. Holleschak, J.D. Astwood, S.L. Hefle, *Int. Arch. Allergy Immunol.* 128 (2002) 191–280.
- [9] O. Ivanciuc, C.H. Schein, W. Braun, *Nucleic Acids Res.* 31 (2003) 359–362.
- [10] FAO/WHO. Evaluation of allergenicity of genetically modified foods. Report of a joint FAO/WHO consultation on food derived from biotechnology, Rome, Italy, January 2001, pp. 22–25.
- [11] V. Brusic, M. Millot, N. Petrovsky, S.M. Gendel, O. Gigonzac, S.J. Stelman, *Allergy* 58 (2003) 1093–1100.
- [12] S.M. Gendel, *J. AOAC Int.* 87 (2004) 1417–1422.
- [13] R.C. Aalberse, *J. Allergy Clin. Immunol.* 106 (2000) 228–238.
- [14] A. Silvanovich, M.A. Nemeth, P. Song, R. Herman, L. Tagliani, G.A. Bannon, *Toxicol. Sci.* 90 (2006) 252–258.
- [15] M.W. Fiers, G.A. Kleter, H. Nijland, A.A. Peijnenburg, J.P. Nap, R.C. van Ham, *BMC Bioinformatics* 5 (2004) 133.
- [16] M.B. Stadler, B.M. Stadler, *FASEB J.* 17 (2003) 1141–1143.
- [17] A.K. Bjorklund, D. Soeria-Atmadja, A. Zorzet, U. Hammerling, M.G. Gustafsson, *Bioinformatics* 21 (2005) 39–50.
- [18] T. Riaz, H.L. Hor, A. Krishnan, F. Tang, K.B. Li, *Bioinformatics* 21 (2005) 2570–2571.
- [19] V. Brusic, N. Petrovsky, *Trends Immunol.* 24 (2003) 225–228.
- [20] C. Blanco, *Curr. Allergy Asthma Rep.* 3 (2003) 47–53.
- [21] F. Ferreira, T. Hawranek, P. Gruber, N. Wopfner, A. Mari, *Allergy* 59 (2004) 243–267.
- [22] S.T. Bolhaar, R. van Ree, Y. Ma, C.A. Bruijnzeel-Koomen, S. Vieths, K. Hoffmann-Sommergruber, A.C. Knulst, L. Zuidmeer, *Int. Arch. Allergy Immunol.* 136 (2005) 45–52.
- [23] M. Jutel, L. Jaeger, R. Suck, H. Meyer, H. Fiebig, O. Cromwell, *J. Allergy Clin. Immunol.* 116 (2005) 608–613.
- [24] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmanns, S. Moxon, M. Marshall, A. Khanna, R. Durbin, et al., *Nucleic Acids Res.* 34 (2006) D247–D251.
- [25] O. Miotto, T.W. Tan, V. Brusic, *Genome Inform.* 16 (2005) 32–44.
- [26] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, *Nucleic Acids Res.* 25 (1997) 3389–3402.
- [27] J.L.Y. Koh, S.P.T. Krishnan, S.H. Seah, P.T. Tan, A. Khan, M.L. Lee, V. Brusic, BioWare: a framework for bioinformatics data retrieval, annotation and publishing. ACM SIGIR Workshop on Search and Discovery in Bioinformatics (SIGIRBIO), Sheffield, UK, July 2004.