

Bioinformatics for characterisation of allergens, allergenicity and allergic crossreactivity

Vladimir Brusic^{1,2} and Nikolai Petrovsky²

¹Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

²Medical Informatics Centre, University of Canberra, Bruce ACT 2601, Australia

Allergies are a major cause of chronic ill health in industrialised countries with the incidence of reported cases steadily increasing. This Research Focus details how bioinformatics is transforming the field of allergy through providing databases for management of allergen data, algorithms for characterisation of allergic crossreactivity, structural motifs and B- and T-cell epitopes, tools for prediction of allergenicity and techniques for genomic and proteomic analysis of allergens.

Allergies affect approximately a third of the general population and are a major cause of chronic ill health in developed countries [1]. The list of protein allergens certified by the Allergen Nomenclature Sub-Committee of the International Union of Immunological Societies (IUIS) contains >400 allergens and nearly 200 isoforms (as of January 2003). Given the rapid growth in allergy data, bioinformatics methods are becoming indispensable for the management of allergen data and detailed analysis of their sequences, structures, functional properties and allergenicity.

Major sources of allergen data include the specialised allergen databases (listed in Table 1). The Food Allergy Research and Resource Program (Farrp) [2] and the Structural Database of Allergenic Proteins (SDAP) [3] are searchable databases that contain allergen sequences and enable sequence comparison analyses. The SDAP database enables access to and cross-referencing of sequences, structures and immunoglobulin E (IgE) epitopes of allergenic proteins from the IUIS allergen list. SDAP provides computational tools for assessment of potential allergenicity and allergic crossreactivity, based on the analysis of short peptides and similarity to known IgE epitopes. In addition, SDAP tools enable structural analysis of the surface of a potential allergen. Although potentially useful for the identification of IgE epitopes, SDAP does not conform to the established standards for assessment of allergenicity, which are described later. A recent comparative analysis of allergen data sources [4] showed that the individual allergen data sources focus on unique aspects of allergen annotation with the consequence that the available sources are incomplete and not up-to-date. From this analysis it was concluded that there is a great need for a complete, well annotated and unified allergen database. Such a unified database should be comprehensive, contain only non-redundant entries, comply with the IUIS allergen

classification and enable analyses of allergen sequence, structure, allergenicity, crossreactivity and B- and T-cell epitopes of allergens [4].

Allergen characterisation

Analysis of a novel allergen typically involves cloning, expression and molecular characterisation, followed by evaluation of allergenicity. Molecular characterisation involves sequence comparison and three-dimensional (3D) structural analysis. For example, bioinformatic analysis of a recently characterised dust mite allergen Blo t 1 [5] from *Blomia tropicalis* revealed 35% identity to mite cysteine proteases. The 3D structure modelling of Blo t 1 showed the conserved cysteine protease secondary and tertiary structure. This simple bioinformatics analysis, therefore, was useful for determining the function and structure of Blo t 1.

Protein allergenicity is determined by its 3D structure and structural knowledge can, therefore, help provide an insight into the molecular basis of allergenicity and elucidation of biological function of allergens. Nearly 50 allergen structures are included in the Protein Data Bank (PDB) database (as of January 2002) [6]. The deposited structures include those of ovalbumin, milk allergens, lysozymes, several pollen allergens and various animal and fungal allergens. For example, major birch allergen Bet v 1 belongs to a family of pathogenesis-related proteins (PR-10) and until recently, the biological function of Bet v 1 was unknown. Structural analysis of a hypoallergenic isoform Bet v 1 revealed similarity to a steroid-binding domain in the human protein MLN64 and showed its biological function to be a plant-steroid carrier [7]. Multiple sequence alignment identified variable amino-acids position among 15 PR-10 proteins. Because nine Bet v 1 isoforms are known to differ in their binding affinity to IgE and in the induction of allergic symptoms, it was possible to design a six-point mutant protein that has reduced allergenicity and that might be useful in allergic tolerance immunotherapy [7]. A similar point mutation study of major rye grass pollen allergens enabled identification of modified proteins that have reduced allergenicity, and which are suitable targets for allergen-specific immunotherapy [8]. Secondary structure analysis of a set of cross-reactive plant allergens was used to identify structural elements and the common P-loop domain of pollen and related food allergens [9]. This group of sequences displayed strong allergenic crossreactivity and the presence

Table 1. Major allergen databases and allergen lists

Database	URL location
IUIS Allergen Nomenclature (IUIS)	www.allergen.org
Structural Database of Allergenic Proteins (SDAP)	fermi.utmb.edu/SDAP
SWISS-PROT Allergen Index	www.expasy.ch/cgi-bin/lists?allergen.txt
Biotechnology Information for Food Safety (BIFS)	www.iit.edu/~sgendel/fa.htm
Food Allergy Research and Resource program (Farrp)	www.allergenonline.com
Food Allergens of Plant Origin (Protall)	www.ifr.bbsrc.ac.uk/Protall
Central Science Laboratory Allergen Database (ADB)	www.csl.gov.uk/allergen
Asthma and Allergy Gene Database	cooke.gsf.de/asthmagen/main.cfm

of common and specific epitopes. 3D analysis helped to identify a common structural motif, comprising a groove inside an $\alpha\beta$ motif, within diverse allergens (dust mite allergen, cysteine protease papain, lipocalin and ragweed allergen) [10]. The analysis of forty allergen structures revealed four structural families: antiparallel β -strands, antiparallel β -sheets associated with α -helices, $\alpha\beta$ structures, and α -helical structure [11].

These examples demonstrate the use of basic bioinformatics tools in allergy research. Common sequence comparison methods for database searching are the FASTA [12] and Basic Local Alignment Tool (BLAST) [13] algorithms. Multiple sequence alignment programs focus on global alignment, such as CLUSTAL [14], or local alignment algorithms [15]. Protein sequence analysis focuses on finding conserved motifs, prediction of protein structure and prediction of protein function [16].

Prediction of allergenicity and allergic crossreactivity

Identification of allergic potential is a crucial issue whenever new proteins are brought into contact with humans, either through food, cosmetics or other modes of contact. Various schemes have been introduced for the assessment of protein allergenicity. The current recommendation sanctioned by the World Health Organisation (WHO) and the Food and Agriculture Organisation (FAO) is a scheme based on a decision tree (Fig. 1). A crucial step in the assessment of allergenicity of novel proteins is sequence similarity search against known allergens. The FAO–WHO decision tree has been used for the assessment of allergenicity of an ice structuring protein (ISP) type III from arctic fish [17]. The method involved BLAST searches against five major protein sequence databases and the PDB. Local alignments with database entries were identified using BLAST search parameters optimised for identification of short alignments. Also, all possible 6-, 7- and 8-mer sequences from ISP were generated and searched for identity with the Protein Information Resource (PIR) database (www-nbrf.georgetown.edu). The authors reported finding a seven amino-acid identity over a nine amino-acid stretch but no overall homology with a known allergen Asp f 6. BLAST searches produced 355 alignments, none of which contained more than five contiguous amino-acid identities. The authors also reported 1674, 1771 and 2442 exact database entry matches of 6-, 7- and 8-mer peptides respectively but none with known allergens. A subsequent study [18] using the FAO–WHO decision tree and twenty individuals allergic to fish established that the ISP produced positive results in both skin prick tests and *in vitro* allergenicity tests but did not show any allergic activity in

functional biological assays from the same cohort. The results provide a 99% level of confidence that the ISP III is unlikely to sensitise susceptible individuals. These experimental findings are consistent with lack of sequence homology between ISP type III and any known allergens.

The question of what represents a significant sequence similarity match between a query protein and known allergens remains unanswered. Two decision criteria have been proposed: identity of six or more contiguous amino acids, or minimum 35% identity over an 80 amino-acid segment [11]. The six amino-acid identity rule is not practical because it produces a large number of hits, as reported in a recent study [2]. The eight amino-acid identity between two proteins rarely occurs by chance [19] and such an occurrence might have biological significance. However, it is not clear what level of sequence similarity is appropriate to account for similar conformational epitopes and there is a need for further research in this field.

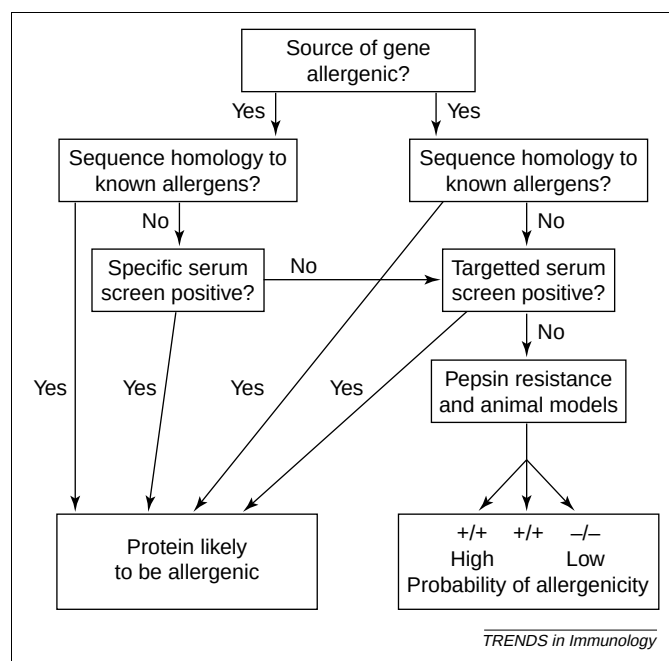


Fig. 1. The Food and Agriculture Organisation (FAO)–WHO (World Health Organisation) decision tree (Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO–WHO Expert Consultation on allergenicity of Foods Derived from Biotechnology (2001) WHO, Rome). The assessment of sequence homology is a key step in the decision process. When similarity of the query protein to the established 'target' allergens is determined, the sera from patients allergic to the target allergen are tested to find whether the query protein will induce reactions in allergic patients (left side of the tree). If negative, sera from patients allergic to foods related to target allergens are tested to find if the query protein will sensitise susceptible patients (right side of the tree). The resistance to proteolytic degradation and allergenicity in animal models (++) indicates high probability, one of the two tests positive and another negative (+/+) indicate moderate probability and both tests negative (--) indicate low probability of allergenicity.

Factors to be taken into account for assessment of allergenicity by sequence similarity include the design and content of the sequence database, the analysis strategy and the criteria for evaluating the results of sequence similarity [19]. Allergic crossreactivity is a significant problem and even distant allergens can cause crossreactivity. For example, in latex-fruit syndrome, crossreactivity between latex and plant-derived food allergens has been observed [20]. This indicates a major need for a method for identifying potential crossreactive allergens of low homology. A method based on sequence motifs in form of generalised profiles [21] might assist in these situations (M.B. Stadler, unpublished). This technique involves multiple sequence alignment of known allergens, assignment of weights to each sequence in the alignment, creation and calibration of a generalised profile and prediction of allergenicity based on matches to this profile.

Computational methods for identification of B- or T-cell epitopes assist better understanding of complex allergic reactions. Methods based on conserved properties of amino-acid side-chains have been developed for the identification of portions of query proteins similar to the regions of known allergens and reported B-cell epitopes [3]. Early bioinformatic methods for identification of potential B-cell epitopes used sequence and amino-acid properties (such as hydrophobicity) and were of low accuracy. More recent methods use 3D structure modeling in addition to the earlier strategies. Structural requirements for effector-cell activation were used for conversion of allergens into hypoallergenic immunotherapeutics by reorientation of IgE epitopes [22].

Highly promiscuous human leukocyte antigen (HLA) allele-specific T-cell epitopes with potential applications in immunotherapeutics have been identified in allergens through bioinformatics approaches [23,24]. Computational prediction methods dramatically reduce the number of experiments required to identify relevant T-cell epitopes. Sophisticated artificial intelligence techniques (such as artificial neural networks or hidden Markov models) can be used for large-scale screening of HLA binding peptides [25]. Similar methodologies with only a slight variation can be applied to the prediction of both HLA class I and class II T-cell epitopes. Examples of the current state of the art in T-cell epitope prediction include quantitative approaches to computational vaccinology [26] and prediction of promiscuous T-cell epitopes [27,28]. Molecular modeling or molecular dynamic simulations [29] have also been used for predicting candidate HLA class II epitopes. The relatively small number of well characterised 3D structures and the computationally intensive nature of this methodology currently prevent the use of structure-based methods for systematic prediction of both B- and T-cell epitopes.

Conclusion

With the explosive growth of allergy-related data through functional genomics [30], gene microarrays [31] and proteomics [32], bioinformatics has become an essential tool for allergy data management, interpretation of results and experimental planning. Significant work has gone into developing specialised methods for data management

and analysis of allergens. An important next step is the integration of existing allergy data resources into a unified well annotated allergen database complete with appropriate bioinformatics analysis tools. Sequence comparison tools are already essential in the determination of allergenicity and allergic crossreactivity [2,17–19]. Sequence motif analysis [21] is emerging as an important methodology for determination of allergic crossreactivity involving low-homology allergens. Structural bioinformatics has an important role in deciphering the structural basis of allergenicity of individual allergens, identification of IgE recognition sites and engineering of hypoallergenic proteins [7,8,10,11]. Collectively, bioinformatics tools are already crucial to allergy research and will accelerate the design of vaccines and therapies for allergic diseases [8,22,24].

References

- 1 Ono, S.J. (2000) Molecular genetics of allergic diseases. *Annu. Rev. Immunol.* 18, 347–366
- 2 Hileman, R.E. *et al.* (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.* 128, 280–291
- 3 Ivanciuc, O. *et al.* (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.* 31, 359–362
- 4 Brusic, V. *et al.* Allergen databases. *Allergy* (in press)
- 5 Mora, C. *et al.* (2003) Cloning and expression of Blo t 1, a novel allergen from the dust mite *Blomia tropicalis*, homologous to cysteine proteases. *Clin. Exp. Allergy* 33, 28–34
- 6 Westbrook, J. *et al.* (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 31, 489–491
- 7 Markovic-Housley, Z. *et al.* (2003) Crystal structure of a hypoallergenic isoform of a major birch pollen allergen Bet v 1 and its likely biological function as a plant steroid carrier. *J. Mol. Biol.* 325, 123–133
- 8 Singh, M.B. and Bhalla, P.L. (2003) Hypoallergenic derivatives of major grass pollen allergens for allergy vaccination. *Immunol. Cell Biol.* 81, 86–91
- 9 Scheurer, S. *et al.* (1999) Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen. *Mol. Immunol.* 36, 155–167
- 10 Furmonaviciene, R. and Shakib, F. (2001) The molecular basis of allergenicity: comparative analysis of the three dimensional structures of diverse allergens reveals a common structural motif. *Mol. Pathol.* 54, 155–159
- 11 Aalberse, R.C. (2000) Structural biology of allergens. *J. Allergy Clin. Immunol.* 106, 228–238
- 12 Pearson, W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185–219
- 13 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
- 14 Jeanmougin, F. *et al.* (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23, 403–405
- 15 Heringa, J. (2002) Local weighting schemes for protein multiple sequence alignment. *Comput. Chem.* 26, 459–477
- 16 Fasman, K.H. and Salzberg, S.L. (1998) An introduction to biological sequence analysis. In *Computational methods in molecular biology* (Salzberg, S.L. *et al.*, eds), pp. 29–42, Elsevier, Amsterdam
- 17 Baderschneider, B. *et al.* (2002) Sequence analysis and resistance to pepsin hydrolysis as part of an assessment of the potential allergenicity of ice structuring protein type III HPLC 12. *Food Chem. Toxicol.* 40, 965–978
- 18 Bindslev-Jensen, C. *et al.* (2003) Assessment of the potential allergenicity of ice structuring protein type III HPLC 12 using the FAO/WHO 2001 decision tree for novel foods. *Food Chem. Toxicol.* 41, 81–87
- 19 Gendel, S.M. (2002) Sequence analysis for assessing potential allergenicity. *Ann. N.Y. Acad. Sci.* 964, 87–98
- 20 Blanco, C. (2003) Latex-fruit syndrome. *Curr. Allergy Asthma Rep.* 3, 47–53
- 21 Bucher, P. *et al.* (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem.* 20, 3–23

- 22 Valenta, R. and Kraft, D. (2001) Recombinant allergen molecules: tools to study effector cell activation. *Immunol. Rev.* 179, 119–127
- 23 Friedl-Hajek, R. *et al.* (1999) Identification of a highly promiscuous and an HLA allele-specific T-cell epitope in the birch major allergen Bet v 1: HLA restriction, epitope mapping and TCR sequence comparisons. *Clin. Exp. Allergy* 29, 478–487
- 24 de Lalla, C. *et al.* (1999) Cutting edge: identification of novel T-cell epitopes in Lol p5a by computational prediction. *J. Immunol.* 163, 1725–1729
- 25 Yu, K. *et al.* (2002) Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol. Med.* 8, 137–148
- 26 Doytchinova, I.A. and Flower, D.R. (2002) Quantitative approaches to computational vaccinology. *Immunol. Cell Biol.* 80, 270–279
- 27 Sturniolo, T. *et al.* (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* 17, 555–561
- 28 Brusic, V. *et al.* (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol. Cell Biol.* 80, 280–285
- 29 Toh, H. *et al.* (2000) Magnitude of structural changes of the T-cell receptor binding regions determine the strength of T-cell antagonism: molecular dynamics simulations of HLA-DR4 (DRB1*0405) complexed with analogue peptide. *Protein Eng.* 13, 423–429
- 30 Saito, H. *et al.* (2001) Human mast cell transcriptome project. *Int. Arch. Allergy Immunol.* 125, 1–8
- 31 Harwanegg, C. *et al.* (2003) Microarrayed recombinant allergens for diagnosis of allergy. *Clin. Exp. Allergy* 33, 7–13
- 32 Yu, C.J. *et al.* (2003) Proteomics and immunological analysis of a novel shrimp Allergen, Pen m 2. *J. Immunol.* 170, 445–453

1471-4906/03/\$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S1471-4906(03)00075-9

Letters

β₂-microglobulin as a potential initiator of inflammatory responses

Jin Xie and Qing Yi

Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences, 4301 West Markham Street, Slot 776, Little Rock, Arkansas 72205, USA

Ankylosing spondylitis (AS) is a chronic systemic inflammatory rheumatic disorder of uncertain aetiology that primarily affects the axial skeleton (sacroiliac joints and spine). Although the pathogenesis of the disease is poorly understood, immune-mediated mechanisms involving human leukocyte antigen (HLA)-B27, inflammatory cellular infiltrates, cytokines, such as tumor necrosis factor (TNF)-α and interleukin-10 (IL-10), as well as genetic and environmental factors, are thought to have key roles [1]. Based on the strong association between HLA-B27 and the disease, Uchanska-Ziegler and Ziegler

recently proposed that AS could be a β₂-microglobulin (β₂m)-deposition disease, with parallels to β₂m-amyloidosis observed in long-term dialysis patients [2]. The essence of their hypothesis is that cells bearing AS-associated HLA-B27 subtypes might exhibit a higher rate of β₂m-dissociation from surface-expressed HLA-B27 molecules; β₂m molecules might accumulate and become trapped in synovia where they bind to collagen, form amyloid deposits or interact with synovial fibroblasts in which they induce the synthesis and secretion of proteins involved in tissue destruction. Some of the deposited β₂m

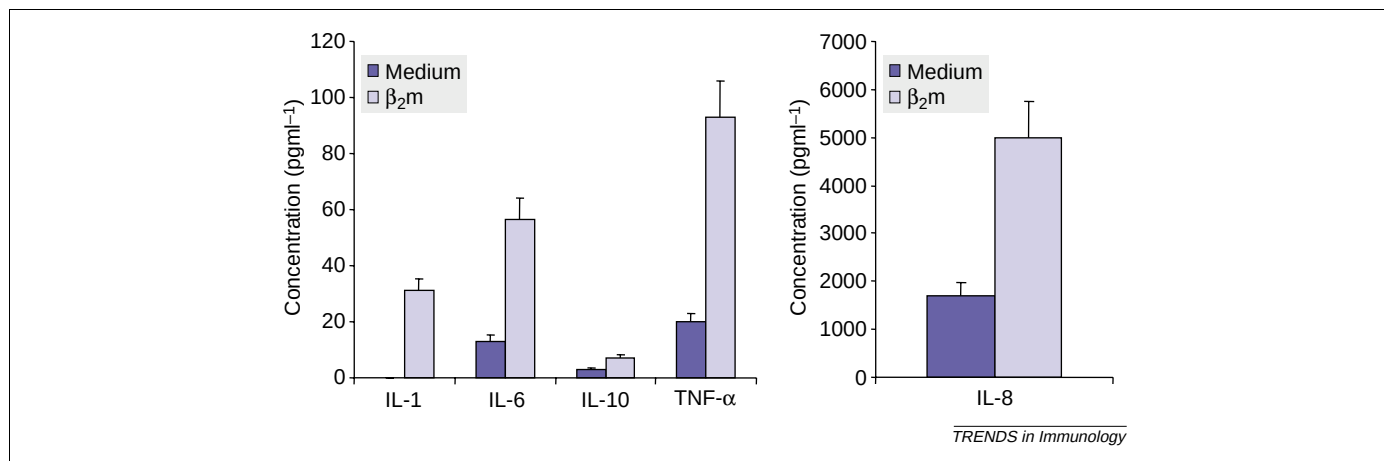


Fig. 1 Cytokine secretion profile of monocytes cultured for 24 hours with or without addition of 20 μg ml⁻¹ of β₂m. A cytometric bead array analysis was used to measure the concentration (pg ml⁻¹) of various cytokines in cell culture medium. Mean ± SEM of 4 experiments are shown. Abbreviations: β₂m, β₂-microglobulin; IL-1, interleukin-1; TNF-α, tumor necrosis factor-α.

Corresponding author: Qing Yi (YiQing@uams.edu).