

Review article

Computational tools for the study of allergens

Allergy is a major cause of morbidity worldwide. The number of characterized allergens and related information is increasing rapidly creating demands for advanced information storage, retrieval and analysis. Bioinformatics provides useful tools for analysing allergens and these are complementary to traditional laboratory techniques for the study of allergens. Specific applications include structural analysis of allergens, identification of B- and T-cell epitopes, assessment of allergenicity and cross-reactivity, and genome analysis. In this paper, the most important bioinformatic tools and methods with relevance to the study of allergy have been reviewed.

**V. Brusic^{1,2}, N. Petrovsky²,
S. M. Gendel³, M. Millot⁴,
O. Gigonzac⁵, S. J. Stelman⁶**

¹Institute for Infocomm Research, Singapore;
²Centre for Medical Informatics, University of
Canberra, Canberra, ACT, Australia; ³Biotechnology
Studies Branch, Food and Drug Administration,
National Center for Food Safety and Technology,
Summit-Argo, IL, USA; ⁴IHCP, Joint Research Centre,
Ispra, Italy; ⁵Aventis CropScience, Evry, France;
⁶Dow AgroSciences, San Diego, CA, USA

Key words: allergen; molecular modelling; prediction
of allergenicity; prediction of cross-reactivity; sequence
analysis.

Vladimir Brusic
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119631

Accepted for publication 26 March 2003

Allergic responses result from activation of immune pathways involving interactions between multiple molecular and cellular components. Better understanding of such pathways is crucial for understanding the nature of allergic responses and, in particular, for the development of new allergy treatments. Genetic studies show that multiple genes are involved in allergy susceptibility and the chromosomal locations of some allergy-related genes have been identified (1, 2). Genomic and proteomic technologies that have been applied to allergy research include gene expression arrays (3) and proteomics (4). The number of characterized protein allergens is increasing rapidly. The Allergen Nomenclature Sub-Committee of the International Union of Immunological Societies (IUIS) maintains a list of 'certified' protein allergens (5). By December 2002, some 420 protein allergens had been classified. While this list contains numerous allergens it is not exhaustive as unlisted allergens can be found in the existing literature. Furthermore, although an additional 190 isoforms of protein allergens are included on the official allergen list, this may represent only a small fraction of the total as individual allergens may have tens of isoforms (6, 7).

The assessment of protein allergenic potential focuses on three main aspects: inherent protein allergenicity,

cross-reactivity, and clinical symptoms (8). Inherent allergenicity refers to the likelihood of an immunoglobulin E (IgE) antibody or T-cell response to a particular allergen. Studies of B-cell (9, 10) and T-cell (11, 12) epitopes focus on defining recognition sites on allergens. Cross-reactivity refers to the ability of an IgE clonotype or a T-cell clone, which were previously induced by one allergen, to react with another allergen. Studies of stinging insect venom allergens have shown that cross-reactivity between allergens having <70% sequence identity is uncommon (13). Studies of inherent allergenicity and cross-reactivity have applications in the development of allergy immunotherapies and vaccines (14–16). Clinical and functional data for common allergens have been reported extensively and public databases contain entries of allergen sequences (17, 18), structures (19), or T-cell epitopes (20, 21). However, these data sources are neither complete nor up-to-date.

Modern research requires rapid access to rich data sources and advanced tools for data analysis. Unfortunately, allergen-related data are currently scattered across various publicly accessible molecular databases. A detailed description of allergen databases was given in a recent review (22).

The computational and bioinformatic tools used to probe databases and data repositories belong to various categories: sequence analysis and comparison, prediction and structural analysis. The sequence analysis and comparison tools focus on classification and characterization of genes and proteins of confirmed and putative allergens (23, 24). Prediction tools focus on functional and structural analysis of genes and proteins leading to identification of those with allergenic potential. High accuracy predictions can guide experimental design, facilitating fast and cost-effective experimental validation of potential allergens. Bioinformatic tools are available for prediction of allergenicity (25), allergen cross-reactivity (26), and T-cell epitopes (12). Structural analysis tools help identify structural properties of proteins, such as secondary structure (27), or tertiary structure (8, 28) that may impart allergenicity, e.g. through affecting IgE binding sites.

In this paper, we provide an overview of bioinformatic applications available for allergy research and discuss bioinformatic advances needed to fulfil current and future allergy research needs with a focus on protein allergens. We propose that a systematic approach to allergen data management would create immediate benefits for both basic allergen research and clinical allergology.

Bioinformatic tools

Multiple bioinformatic tools have been used in the study of allergy. The allergic cascade is a series of chemical reactions that start with allergen recognition by T- and B-cells which eventually results in the release of histamines and other inflammatory molecules from the mast cells. The bioinformatic tools and applications that assist the study of specific steps in the allergic cascade have been shown in Fig. 1.

Common bioinformatic tools

The common characteristics of allergens such as structural, functional, or biochemical properties that explain their ability to elicit allergic responses remain unclear. Both general and specialized bioinformatic tools have been used in allergy-related studies. Standard protein sequence analysis is commonly used for characterization of DNA and protein sequences of allergens, and identification of their common features. Standard techniques use sequence comparison, sequence alignment, database searching, motif and pattern searching, or identification of various properties of protein sequences. Homologous proteins often have similar 3-D structures, and similar shape may imply similar function. Proteins with similar function often display some level of similarity in their primary sequences, typically with more than 20–35% identity. The first step in the study of a novel protein is normally a search of public databases for homologous sequences. The most popular methods for searching sequence databases are the FASTA (29) and BLAST (30) programs. A number of specialized sequence alignment algorithms such as Smith–Waterman (31) (which is considered more accurate but slower) or PropSearch (32) (suitable for finding distant homologues) provide more detailed analyses. Another important set of bioinformatic tools uses pairwise alignment of sequences, aimed at exploring the similarity of a query sequence with a number of preselected target sequences. The task of producing an optimal local or global alignment of multiple sequences is mathematically difficult, but has a great utility in molecular biology. Multiple sequence alignment programs include those that focus on optimal global alignment [such as ClustalW (33)] or local alignment [such as PIMA (34)].

While DNA sequence analysis focuses largely on characterization of DNA sequences and genes, protein

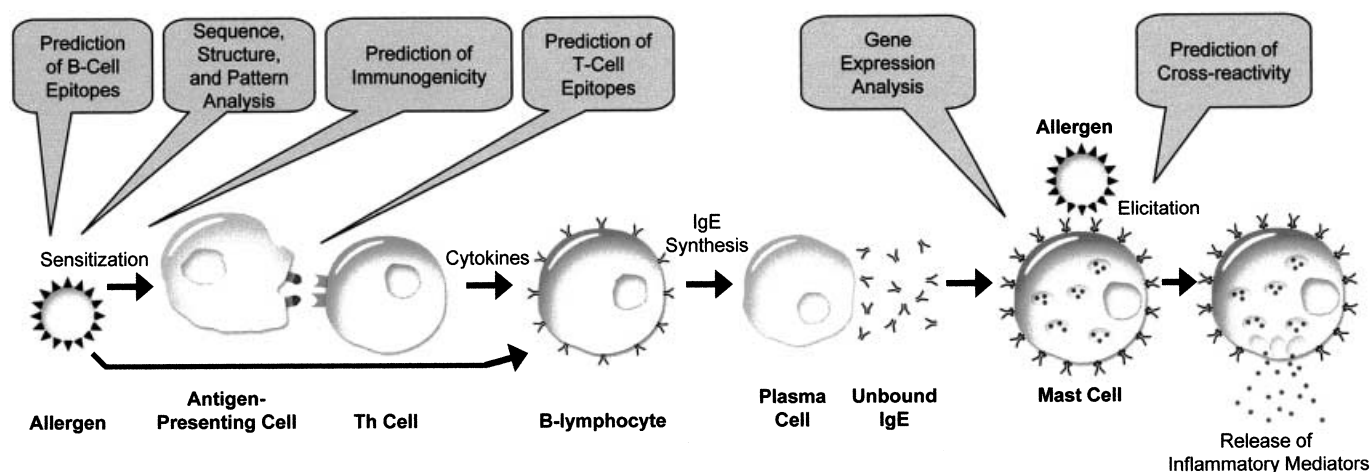


Figure 1. Bioinformatics tools and applications used in the study of specific steps in the allergic cascade.

sequence analysis focuses on solving the following problems (35):

- Finding conserved patterns of sequence in related proteins that may correspond to functional domains.
- Prediction of protein secondary structure by multiple alignment or from its primary sequence.
- Prediction of 3-D structure from sequence either by homology to a known molecular structure or *de novo*, from first principles, using knowledge of fundamental atomic interactions.
- Prediction of function of whole molecules or their domains, either by homology to proteins of known function or *de novo* analysis.

Some examples of sequence comparison and analysis in the study of allergens include molecular characterization of an allergen group from dust mites (36), characterization of a superfamily of proteins containing the allergenic lipid transfer proteins (37), structural characterization of a rice allergen (38), cross-reactivity studies of allergens (27, 39), analysis of mutations of pollen allergens (23), study of structural and molecular basis of allergenicity (25), and identification of characteristic motifs in cat (40) or cockroach (24) allergens.

Structural motifs and 3-D structure analysis

More advanced bioinformatic analyses of allergens require more specialized tools. A selection of representative, although not exhaustive, bioinformatic tools is available at the ExPASy (<http://www.expasy.ch>) server. Applications include identification of specific sequence motifs, structural motifs or characteristic patterns in allergen sequences. Advanced bioinformatic tools have been used for various predictions including secondary structure, allergenicity, sensitization, cross-reactivity, T- and B-cell epitopes and active sites in allergens. Secondary structure prediction of a selection of pollen, fruit and vegetable allergens using PredictProtein (41) software helped the identification of very similar structural elements and, in particular, the 'P-loop' region as a common domain of pollen and related food allergens (27). This group of sequences displayed strong allergenic cross-reactivity and the presence of common and specific epitopes. Third generation software such as PredictProtein or Jpred2 (41) accurately predict more than 75% of amino acids that form α -helices, β -sheets, or coils. Structural motif analysis can be useful in studies of specific protein properties, e.g. it successfully revealed the presence of coiled-coil helices in the group 5 allergen Der p 5 from the house dust mite (42). Bioinformatic analysis of this allergen explained the mechanism of allergen filament formation. In this study, the programs PAIRCOIL (43) and COILS (44) were used for prediction of coil structure, and PSORT (45) for prediction of protein localization sites in cells.

Less than 50 allergen structures, representing approximately a dozen structural groups, are available in the

PDB database as of December 2002 (19). The deposited structures include those of ovalbumin, milk allergens (α -lactalbumin, β -lactoglobulin, and lactoferrin), lysozymes, lipocalins, profilins, other pollen allergens, insect venom allergens, dust mite, horse, and fungal allergens.

Protein allergenicity is determined by 3-D structure and structural knowledge can, therefore, help provide an insight into the molecular basis of allergenicity. For example, a common structural motif comprising a groove located inside an α - β motif has been identified within diverse allergens (dust mite allergen Der p 1, cysteine protease papain, lipocalin Mus m 1, and ragweed allergen Amb a 5) (25). The unknown 3-D structure of allergens can be deduced from homology to proteins whose structure is known. For example, the Der p 1 dust mite allergen was modelled upon the known structure of cysteine proteinases (46), and cat cystatin was modelled upon human cystatin structure (40). and the Blo t 13 allergen of the *Blomia tropicalis* mite was modelled upon cytosolic lipid transport proteins (47).

Although structural similarities and differences are important for establishing the basis for understanding immunogenicity and cross-reactivity of allergens, a study of known protein folds of 40 allergens suggested that a general structural feature that conferred allergenicity did not exist (8). The same study provided a description of four major structural families that potentially conferred allergenicity (Fig. 2): (A) antiparallel β -strands: the immunoglobulin-fold family (grass group 2, mite group 2), serine proteases (mite groups 3, 6, and 9), and soya bean-type trypsin inhibitor (Ole e 1, grass group 11); (B) antiparallel β -sheets intimately associated with one or more α -helices: tree group 1, lipocalin, profilin, aspartate protease (cockroach group 2); (C) (α + β) structures, in which the α - and β -structural elements are not intimately associated: mite group 1, lysozyme/lactalbumin, vespid group 5; and (D) α -helical: nonspecific lipid transfer protein, seed 2S protein, insect haemoglobin, fish parvalbumin, pollen calmodulin, mellitin from bee venom, Fel d 1 chain 1, and serum albumin. However, experimentally determined structures alone cannot sufficiently resolve fine details required for accurate prediction of the dynamics of protein interactions. Molecular dynamics simulation tools (48) provide atomic-level descriptions of molecular interactions. Study of antigen-antibody interactions using molecular dynamics simulations enabled reliable quantification of structural changes over long periods of simulation (49). The accumulation of allergen data, both structural and functional, will provide an increasingly important resource for detailed bioinformatic studies of allergens.

Genomics

'Genomics' refers to the investigation of the structure, function, regulation and interaction of multiple genes.

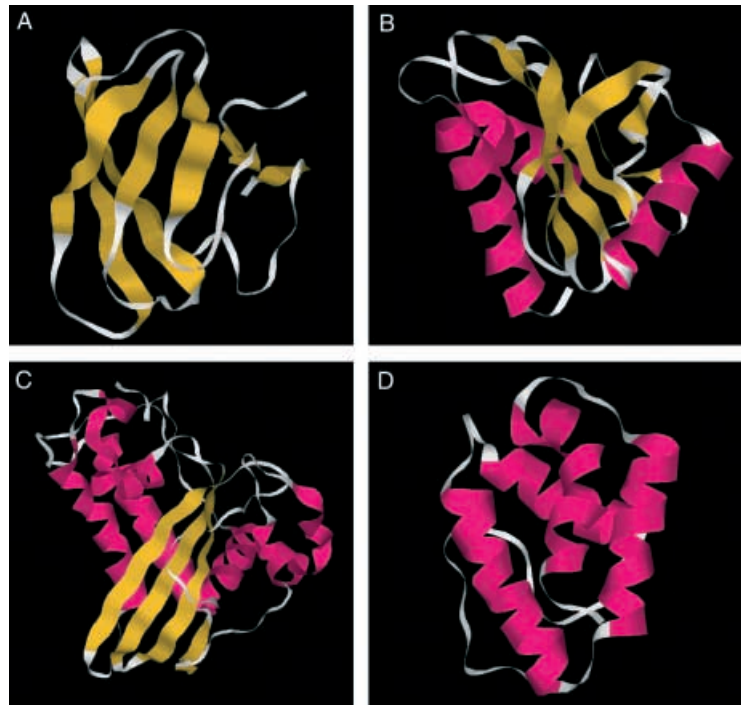


Figure 2. Structural groups of allergens as defined in (8): (A) Phl p 2, timothy grass pollen allergen group 2 (PDB code 1WHP), (B) Ara t 8, mouse-ear cress profilin I (PDB code 1A0K), (C) Ves v 5, wasp venom allergen 5 (PDB code 1QNX), (D) Zea m 14, maize lipid transfer protein (PDB code 1AFH).

The vast quantity of data produced by structural, functional, and comparative genomics cannot be fully interpreted without bioinformatic analyses. Study of the genetic factors that contribute to atopy and asthma has identified a high degree of sequence variation in the eotaxin gene (50). This study focused on classification of differentially expressed genes from activated T-helper (Th2) cells into biologically relevant functional categories (i.e. receptor, signal transduction pathways, transcription, and translation). Further studies have focussed on identification of differentially expressed genes and the comparison of gene expression in Th1 and Th2 clones (51, 52). Microarrays of multiple allergens have been used for detection of allergen-specific IgE in patient samples (53). Expression studies of pulmonary genes helped identify the complement factor 5 locus as a susceptibility gene for allergen-induced asthma in a mouse model (54). Oligonucleotide probe microarrays were used to study the complete transcript of human mast cells thereby identifying selective expression of 89 transcripts that may be important in allergic inflammation (55).

The comparison of specific cDNA with genomic sequences is useful for identification of variants (isoforms) of members of gene families. This type of analysis helped characterize a major peanut allergen Ara h 2 (56). Given the explosion of sequence data on a plethora of genomes, genomics applications will continue to grow in importance in allergy research.

Bioinformatic applications

Prediction of allergenicity

Several criteria can be used for bioinformatic assessment of allergenic potential of a protein. The molecular basis of allergenicity of a protein can be studied through the analysis of its sequence, structure, B- or T-cell epitopes (57). The recognition of allergens by IgE and IgG antibodies involves binding to linear or conformational epitopes on allergen surface (58, 59). Identification of B-cell epitopes is important for characterization of allergens and for better understanding of complex allergic reactions. Synthetic overlapping peptides have been used for identification of B-cell epitopes in several food allergens (60). Experimental epitope mapping requires large number of binding studies and has limited value for identification of discontinuous conformational B-cell epitopes. The complementarity of molecular shape and amino acid charge was observed in B-cell epitopes of a known Der p 1 epitope (28). Nuclear magnetic resonance studies indicate that only a small number of residues within an epitope are functionally important for antibody binding (9). Alanine replacements of single amino acids of B-cell epitopes showed that IgE binding can be reduced or eliminated by point mutation of a protein (61). Several bioinformatic methods for identification of potential B-cell epitope candidates have been reported. Early methods used hydrophilicity scans (62), properties of amino acids (63), or combination thereof (64) [e.g. in the identification of viral B-cell epitopes

(65, 66)]. More recent methods have used knowledge-based homology modelling and prediction of 3-D structure in addition to the earlier strategies. For example, a study of Japanese encephalitis virus showed that both conformational and linear B-cell epitopes can be predicted using a combined approach that also allowed for identification of critical residues (67). Knowledge of the structural requirements for effector cell activation may also be used for conversion of allergens into hypoallergenic immunotherapeutics by reorientation of IgE epitopes (68). A relatively small number of characterized 3-D structures and computationally intensive methodology currently prevent the use of these methods for systematic prediction of protein allergenicity.

Contrary to CD4⁺ Th1 cells, which produce cytokines that activate macrophages and are involved in delayed-type hypersensitivity, Th2 cells mediate strong antibody responses (69), including IgE-mediated allergic inflammatory responses. Allergen-specific T-cell clones have been studied in patients allergic to eggs, milk, nuts, apple, pollen, and latex, amongst others and T-cell epitopes have been identified for a variety of allergens (70). However, some of these data are of limited use because the human leucocyte antigen (HLA) specificity of these T-cell epitopes has not been identified. More than 500 variants of HLA-DR, -DP, and -DQ molecules have been characterized, to date (71). Because individuals have different combinations of HLA molecules, their T-cell responses will involve different T-cell epitopes. In addition, presence of certain HLA alleles may confer susceptibility or protection against specific allergies (72). Experimental T-cell epitope mapping involves binding or functional assays of multiple overlapping peptides. T-cell epitope mapping of an allergen requires a significant experimental effort even for a single HLA molecule and the systematic experimental characterization of all T-cell epitopes of a particular allergen is generally not practical. However, highly promiscuous HLA allele-specific T-cell epitopes have been identified in allergens (11). Promiscuous T-cell epitopes may have clinical applications in the design of immunotherapies for allergy.

Computational prediction methods can dramatically reduce the number of experiments required to identify relevant T-cell epitopes. Binding of peptide to a HLA molecule is a prerequisite for T-cell recognition and thus can be used for prediction of candidate T-cell epitopes (73, 74). Methods for prediction of HLA-binding peptides include those based on binding motifs (75), quantitative matrices (76) or sophisticated artificial intelligence methods [artificial neural networks (77) or hidden Markov models (78)]. Prediction of HLA class II binding peptides has successfully been used for identification and characterization of T-cell epitopes in autoimmunity (79), cancer (80), infectious diseases (81), and allergy (12, 82). Recent developments of predictive technologies allow the prediction of peptide binding to a wide range of HLA-DR molecules and identification of promiscuous or patient-

specific T-cell epitopes (83). Molecular modelling (84) or molecular dynamic simulations (85) have also been used for modelling and prediction of candidate HLA-class II binders, but being computationally intensive these methods cannot be routinely applied to large-scale screening.

A method for the assessment of the allergenic potential of proteins, using a decision tree that classifies the source of the protein, has been described (86). The decision components in this approach combine the measurements of *in vivo* and *in vitro* immunological responses and include a sequence comparison step. The FAO/WHO current recommendations (87) for identification of potential epitopes include searches for any eight or more contiguous amino acids that are identical to any segment of any known allergen (food, inhalant, or contact). The eight amino acid identity between two proteins rarely occurs by chance (88) and such an occurrence may have biological significance. However, it is not clear what level of sequence similarity is appropriate to account for possibly similar conformational epitopes. Two criteria have been proposed: 35% identity over an 80-amino acid segment (8) (for prediction of allergic cross-reactivity, described in the following section), or any match of six contiguous amino acids (unlikely to be used because of its low specificity). Current *in silico* analyses for the assessment of allergenicity are based only on sequence comparisons to known allergen proteins using FASTA or BLAST searches. This approach can be improved by the use of an appropriate bioinformatic infrastructure including a reference allergen database and custom-made advanced bioinformatic tools.

Prediction of allergic cross-reactivity

Patterns of allergic cross-reactivity, are important because they may reflect (a) the pattern of clinical sensitivities, (b) regulation of allergic sensitization via B- and T-cell epitope cross-reactivity and (c) the need for screening of novel foods for allergenicity (89). A large proportion of patients with pollen allergies show hypersensitivity after eating fresh fruit or vegetables. For example, the cross-reactivity of the major allergens of birch pollen (Bet v 1a), sweet cherry (Pru a 1), apple (Mal d 1), pear (Pyr c 1), celery tuber (Api g 1) and carrot (Dau c 1) were traced to high structural and amino acid sequence similarities (27). The cross-reactivity of IgE epitopes was traced down to a single amino acid and mutation studies showed that small changes in amino acid sequence can result in a significant reduction of allergenicity. Similar results were found in the study of cross-reactivity of profilins from birch and celery (90). Individual patients demonstrate variable B-cell epitope recognition, often responding against several epitopes from the same allergen. Studies with the latex allergen Hev b 5 indicate that amino acid substitutions in multiple epitopes are necessary for reduction in IgE binding (91). In some cases high sequence homology is related to

pan-allergenicity [as in lipid transfer proteins (92)], while in other cases high homology does not result in cross-reactivity [as in birch and carrot cyclophylin (93)]. Non-anaphylactic allergen fragments or variants of the IgE-reactive pollen allergens may serve as useful tools for new immunotherapies (27). The FAO/WHO expert recommendations provide guidelines for the assessment of potential cross-reactive allergenicity of food proteins (87). These include sequence comparison and several steps of homology search against protein databases and multiple sequence analysis. According to the recommendations, cross-reactivity has to be considered if there is >35% homology with any known allergen across any 80-amino acid window of a query protein, or if there is an identity of six or more contiguous amino acids. However, similarity in short stretches of a linear amino acid sequence is unlikely to result in cross-reactivity between two proteins unless they share similar protein folding (82). Searching for the presence of common structural motifs (25, 83, 94) is likely to improve the quality of assessment of cross-reactivity and allergenicity. More advanced methods of establishing the structure–activity relationship will provide further insight into the molecular basis of allergenicity.

Other applications

A number of specialized computational tools have been developed and applied in the field of allergy. One example is the allergen avoidance database tool (95) which, when queried, produces an extensive list of skin-care products that do not contain known allergens specific for a given patient. Another tool predicts sensitization to flour allergens in bakers using a stepwise logic regression method (96). That model correlates allergologic variables (such as measurements of total IgE or allergen-specific IgE) to sensitization. We can anticipate that more specialist bioinformatic tools will appear in the future to support research, clinical practice and the screening of novel synthetic or genetically manipulated foods and products.

Current needs

The growth of allergy-related data is entering an exponential phase because of advances in genomic, proteomic, and molecular biology techniques. Bioinformatics provides tools for managing this avalanche of data and is complementary to wet-lab research. The purpose of allergy-related databases is to facilitate the collection, retrieval, and analysis of such data. An ideal allergy-related molecular database – a ‘reference database’ – should have a comprehensive, accurate and expert-annotated content. It should also contain a set of specialist tools that provide for extraction of complex information. Such databases and corresponding search tools must be accurate and robust.

Bioinformatic tools of higher sophistication will be needed for deeper analysis of the allergic properties of proteins. Tools for the analysis of allergens will be based on both relatively simple sequence comparison and motif extraction techniques and advanced statistical and artificial intelligence methods. The tools that will facilitate the use of the allergen databases may include:

- Clustering tools for grouping allergens into specific structure–function groups; these tools will use multiple sequence analysis, motif identification, and advanced clustering techniques (97, 98).
- Structural analysis tools for B-cell epitope identification (99).
- Advanced predictive tools for T-cell epitope identification (76–78).
- Sophisticated pattern recognition tools for prediction of allergenicity, using machine learning tools (100, 101), for search, classification, and modelling relationships.

Conclusion

New technologies enable the improvement of agronomic characteristics (such as resistance to insects, herbicides, or diseases) or quality (such as improved taste and texture, shelf life, and nutritional value) and an increasing number of foods are being developed and produced using biotechnology approaches. Genetically manipulated staples such as soya beans, canola, corn, or potatoes, are already reaching mass-market consumers (102). Novel foods, as well as nutraceuticals, cosmetics, and other products need to be carefully assessed for safety before they reach consumers to ensure they are not more allergenic than traditional products. The requirements for the safety assessment of such products are growing increasingly complex.

Rapidly increasing amounts of data and the accumulating knowledge in the field of allergy provide rich sources of data for aiding safety assessment of novel foods as well as data to support basic and applied allergy research. A significant amount of work has already taken place in allergy-related data management. In applying bioinformatic tools to this data it is critical that we understand the strengths and drawbacks of individual methods. The obvious next step is the integration of allergy data resources and the creation of a bioinformatic infrastructure to allow the systematic management of allergen data. A reference allergen database integrated with specialist search and analysis tools could provide an ideal core resource for the study of allergens. This would, in turn, enhance the growth and development of other allergy-related databases that focus on specialized aspects of allergens.

Combined with experimental methodology validated on animal models that has been extensively used for

allergenicity testing (103), a bioinformatics approach will help define a more efficient and comprehensive approach to both the assessment of allergenicity and assist towards better understanding of the biological and medical basis of allergy.

References

- FEIJEN M, GERRITSEN J, POSTMA DS. Genetics of allergic disease. *Br Med Bull* 2000;**56**:894–907.
- WALLEY AJ, WILTSHIRE S, ELLIS CM, COOKSON WO. Linkage and allelic association of chromosome 5 cytokine cluster genetic markers with atopy and asthma associated traits. *Genomics* 2001;**72**:15–20.
- DAVIES DE, DJUKANOVIC R, HOLGATE ST. Application of functional genomics to study of inflammatory airways disease. *Thorax* 1999;**54**:79–81.
- WESTERGREN-THORSSON G, MALMSTROM J, MARKO-VARGA G. Proteomics – the protein expression technology to study connective tissue biology. *J Pharm Biomed Anal* 2001;**24**:815–824.
- KING TP, HOFFMAN D, LOWENSTEIN H, MARSH DG, PLATTS-MILLS TAE, THOMAS W. Allergen nomenclature. *Allergy* 1995;**50**:765–774. (<http://www.allergen.org/List.htm>)
- SWOBODA I, JILEK A, FERREIRA F, ENGEL E, HOFFMANN-SOMMERGRUBER K, SCHEINER O et al. Isoforms of Bet v 1, the major birch pollen allergen, analyzed by liquid chromatography, mass spectrometry, and cDNA cloning. *J Biol Chem* 1995;**270**:2607–2613.
- KETTNER A, HUGHES GJ, FRUTIGER S, ASTORI M, ROGGERO M, SPERTINI F et al. Api m 6: a new bee venom allergen. *J Allergy Clin Immunol* 2001;**107**:914–920.
- AALBERSE RC. Structural biology of allergens. *J Allergy Clin Immunol* 2000;**106**:228–238.
- MUELLER GA, SMITH AM, CHAPMAN MD, RULE GS, BENJAMIN DC. Hydrogen exchange nuclear magnetic resonance spectroscopy mapping of antibody epitopes on the house dust mite allergen Der p 2. *J Biol Chem* 2001;**276**:9359–9365.
- SELO I, CLEMENT G, BERNARD H, CHATEL J, CREMINON C, PELTRE G et al. Allergy to bovine beta-lactoglobulin: specificity of human IgE to tryptic peptides. *Clin Exp Allergy* 1999;**29**:1055–1063.
- FRIEDL-HAJEK R, SPANGFORT MD, SCHOU C, BREITENEDER H, YSSEL H, JOOST VAN NEERVEN RJ. 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* 1999;**29**:478–487.
- DE LALLA C, STURNIOLO T, ABBRUZZESE L, HAMMER J, SIDOLI A, SINIGAGLIA F et al. Cutting edge: identification of novel T cell epitopes in Lol p5a by computational prediction. *J Immunol* 1999;**163**:1725–1729.
- KING TP, SPANGFORT MD. Structure and biology of stinging insect venom allergens. *Int Arch Allergy Immunol* 2000;**123**:99–106.
- SANTIAGO ML, HAFALLA JC, KURTIS JD, ALIGUI GL, WIEST PM, OLVEDA RM et al. Identification of the *Schistosoma japonicum* 22.6-kDa antigen as a major target of the human IgE response: similarity of IgE-binding epitopes to allergen peptides. *Int Arch Allergy Immunol* 1998;**117**:94–104.
- TODA M, SATO H, TAKEBE Y, TANIGUCHI Y, SAITO S, INOUE S et al. Inhibition of immunoglobulin E response to Japanese cedar pollen allergen (Cry j 1) in mice by DNA immunization: different outcomes dependent on the plasmid DNA inoculation method. *Immunology* 2000;**99**:179–186.
- MIRZA O, HENRIKSEN A, IPSSEN H, LARSEN JN, WISSENBAACH M, SPANGFORT MD et al. Dominant epitopes and allergic cross-reactivity: complex formation between a Fab fragment of a monoclonal murine IgG antibody and the major allergen from birch pollen Bet v 1. *J Immunol* 2000;**165**:331–338.
- BENSON DA, KARSCH-MIZRACHI I, LIPMAN DJ, OSTELL J, RAPP BA, WHEELER DL. Gen Bank. *Nucleic Acids Res* 2000;**28**:15–18.
- BAIROCH A, APWEILER R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;**28**:45–48.
- BERMAN HM, WESTBROOK J, FENG Z, GILLILAND G, BHAT TN, WEISSIG H et al. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–242.
- BRUSIC V, RUDY G, HARRISON LC. MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res* 1998;**26**:368–371.
- SCHÖNBACH C, KOH, JL, FLOWER DR, WONG L, BRUSIC V. FIMM, a database of functional molecular immunology. *Nucleic Acids Res* 2002;**30**:226–229.
- BRUSIC V, MILLOT M, PETROVSKY N, GENDEL SM, GIGONZAC O, STELMAN SJ. Allergen databases. *Allergy* 2003;**58**:1093–1100.
- MIDORO-HORIUTI T, GOLDBLUM RM, BROOKS EG. Identification of mutations in the genes for the pollen allergens of eastern red cedar (*Juniperus virginiana*). *Clin Exp Allergy* 2001;**31**:771–778.
- YANG CY, WU JD, WU CH. Sequence analysis of the first complete cDNA clone encoding an American cockroach Per a 1 allergen. *Biochim Biophys Acta* 2000;**1517**:153–158.
- FURMONAVICIENE R, SHAKIB F. The molecular basis of allergenicity: comparative analysis of the three dimensional structures of diverse allergens reveals a common structural motif. *Mol Pathol* 2001;**54**:155–159.
- SMITH AM, BENJAMIN DC, HOZIC N, DEREWENDA U, SMITH WA, THOMAS WR et al. The molecular basis of antigenic cross-reactivity between the group 2 mite allergens. *J Allergy Clin Immunol* 2001;**107**:977–984.
- SCHREURER S, SON DY, BOEHM M., KARAMLOO F, FRANKE S, HOFFMANN A et al. Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen. *Mol Immunol* 1999;**36**:155–167.
- FURTADO PB, FURMONAVICIENE R, McELVEEN J, SEWELL HF, SHAKIB F. Adaptations of the helix-grip fold for ligand prediction of the interacting surfaces in a trimolecular complex formed between the major dust mite allergen Der p 1, a mouse monoclonal anti-Der p 1 antibody, and its anti-idiotypic. *Proteins* 2001;**43**:134–144.
- PEARSON WR, LIPMAN DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;**85**:2444–2448.

Acknowledgments

Dr Gendel's work was partially supported by Cooperative Agreement FD-000431 between the U.S. Food and Drug Administration and the National Center for Food Safety and Technology.

30. ALTSCHUL SF, GISH W, MILLER W, MYERS EW, LIPMAN DJ. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–410.
31. SMITH TF, WATERMAN MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–197.
32. HOBOMH U, SANDER C. A sequence property approach to searching protein databases. *J Mol Biol* 1995;**251**:390–399.
33. THOMPSON JD, HIGGINS DG, GIBSON TJ. CLUSTALW, improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* 1994;**22**:4673–4680.
34. SMITH, RF, SMITH, TF. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng* 1992;**5**:35–41.
35. FASMAN KH, SALZBERG SL. An introduction to biological sequence analysis. In: SALZBERG SL, SEARLS DB, KASIF S, eds. *Computational methods in molecular biology*. Amsterdam: Elsevier, 1998:29–42.
36. MILLS KL, HART BJ, LYNCH NR, THOMAS WR, SMITH W. Molecular characterization of the group 4 house dust mite allergen from *Dermatophagoides pteronyssinus* and its amylase homologue from *Euroglyphus maynei*. *Int Arch Allergy Immunol* 1999;**120**:100–107.
37. IYER LM, KOONIN EV, ARAVIND L. Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. *Proteins* 2001;**43**:134–144.
38. IZUMI H, SUGIYAMA M, MATSUDA T, NAKAMURA R. Structural characterization of the 16-kDa allergen, RA17, in rice seeds. Prediction of the secondary structure and identification of intramolecular disulfide bridges. *Biosci Biotechnol Biochem* 1999;**63**:2059–2063.
39. DÍAZ-PERALES A, LOMBARDO M, SÁNCHEZ-MONGE R, GARCÍA-SELLÉS FJ, PERNAS M, FERNÁNDEZ-RIVAS M et al. Lipid-transfer proteins as potential plant panallergens: cross-reactivity among proteins of *Artemisia* pollen, Castanea nut and Rosaceae fruits, with different IgE-binding capacities. *Clin Exp Allergy* 2000;**30**:1403–1410.
40. ICHIKAWA K, VAILES LD, POMES A, CHAPMAN MD. Identification of a novel cat allergen – cystatin. *Int Arch Allergy Immunol* 2001;**124**:55–56.
41. CUFF JA, BARTON GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;**40**:502–511.
42. LIAW SH, CHEN HZ, LIU GG, CHUA KY. Acid-induced polymerization of the group 5 mite allergen from *Dermatophagoides pteronyssinus*. *Biochem Biophys Res Commun* 2001;**13**:308–312.
43. BERGER B, WILSON DB, WOLF E, TONCHEV T, MILLA M, KIM PS. Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci USA* 1995;**92**:8259–8263. (<http://www.expasy.ch/tools>)
44. LUPAS A. Prediction and the analysis of coiled-coil structures. *Methods Enzymol* 1996;**266**:513–525. (<http://www.expasy.ch/tools>)
45. NAKAI K, KANEHISA M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 1992;**14**:897–911. (<http://www.expasy.ch/tools>)
46. TOPHAM CM, SRINIVASAN N, THORPE CJ, OVERINGTON JP, KALSHEKER NA. Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng* 1994;**7**:869–894.
47. PUERTA L, KENNEDY MW, JIMENEZ S, CARABALLO L. Structural and ligand binding analysis of recombinant blo t 13 allergen from *Blomia tropicalis* mite, a fatty acid binding protein. *Int Arch Allergy Immunol* 1999;**119**:181–184.
48. WANG W, DONINI O, REYES CM, KOLLMAN PA. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct* 2001;**30**:211–243.
49. ISRAILEWITZ B, BAUDRY J, GULLINGSRUD J, KOSZTIN D, SCHULTEN K. Steered molecular dynamics investigations of protein function. *J Mol Graphics Model* 2001;**19**:13–25.
50. NICKEL RG, SAIITA FP, FREIDHOFF LR, YU XY, EHRLICH E, BARNES KC et al. Positional candidate gene approach and functional genomics strategy in atopy gene discovery. *Int Arch Allergy Immunol* 1999;**118**:282–284.
51. LI XD, ESSAYAN DM, LIU MC, BEATY TH, HUANG SK. Profiling of differential gene expression in activated, allergen-specific human Th2 cells. *Genes Immun* 2001;**2**:88–98.
52. CHUNG KF, ADCOCK IM. Pathophysiological mechanisms of asthma. Application of cell and molecular biology techniques. *Mol Biotechnol* 2001;**8**:213–232.
53. WILTSHIRE S, O'MALLEY S, LAMBERT J, KUKANSKIS K, EDGAR D, KINGSMORE SF et al. Detection of multiple allergen-specific IgEs on microarrays by immunoassay with rolling circle amplification. *Clin Chem* 2000;**46**:1990–1993.
54. KARP CL, GRUPE A, SCHATZ E, EWART SL, KEANE-MOORE M, CUOMO PJ et al. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat Immunol* 2000;**1**:221–226.
55. SAITO H, NAKAJIMA T, MATSUMOTO K. Human mast cell transcriptome project. *Int. Arch Allergy Immunol* 2001;**125**:1–8.
56. VIQUEZ OM, SUMMER CG, DODO HW. Isolation and molecular characterization of the first genomic clone of a major peanut allergen, Ara h 2. *J Allergy Clin Immunol* 2001;**107**:713–717.
57. BREDEHORST R, DAVID K. What establishes a protein as an allergen? *J Chromatogr B Biomed Sci Appl* 2000;**756**:33–40.
58. VAN REGENMORTEL MH, PELLEQUER JL. Predicting antigenic determinants in proteins: looking for unidimensional solutions to a three-dimensional problem? *Pept Res* 1994;**7**:224–228.
59. VAN KAMPEN V, LIEBERS V, SANDER I, CHEN Z, BAUR X, RAULF-HEIMSOOTH M et al. B-cell epitopes of the allergen Chi t 1.01: peptide mapping of epitopes recognized by rabbit, murine, and human antibodies. *Allergy* 2001;**56**:118–125.
60. WILLIAMS SC, BADLEY RA, DAVIS PJ, PUJIK WC, MELOEN RH. Identification of epitopes within beta lactoglobulin recognised by polyclonal antibodies using phage display and PEPSCAN. *J Immunol Methods* 1998;**213**:1–17.
61. HELM RM, COCKRELL G, CONNAUGHTON C, WEST CM, HERMAN E, SAMPSON HA et al. Mutational analysis of the IgE-binding epitopes of P34/Gly m Bd30 K. *J Allergy Clin Immunol* 2000;**105**:378–384.
62. PARKER JM, GUO D, HODGES RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* 1986;**25**:5425–5432.

63. WELLING GW, WEIJER WJ, VAN DER ZEE R, WELLING-WESTER S. Prediction of sequential antigenic regions in proteins. *FEBS Lett* 1985;**188**:215–218. (http://antheprot-pbil.ibcp.fr/ns_sommaire.html)
64. KOLASKAR AS, TONGAONKAR PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett* 1990;**276**:172–174.
65. TONIUTTO P, FABRIS C, BARBONE F, TISMINETZKY SG, LIANI D, GALAI T et al. Immunoreactivity to putative B-cell epitopes of hepatitis G virus polyprotein in viremic and nonviremic subjects. *Clin Diagn Lab Immunol* 1999;**6**:573–576.
66. DEWASTHALY S, AYACHIT VM, SARTHI SA, GORE MM. Monoclonal antibody raised against envelope glycoprotein peptide neutralizes Japanese encephalitis virus. *Arch Virol* 2001;**146**:1427–1435.
67. KOLASKAR AS, KULKARNI-KALE U. Prediction of three-dimensional structure and mapping of conformational epitopes of envelope glycoprotein of Japanese encephalitis virus. *Virology* 1999;**261**:31–42.
68. VALENTA R, KRAFT D. Recombinant allergen molecules: tools to study effector cell activation. *Immunol Rev* 2001;**179**:119–127.
69. ROMAGNANI S. The Th1/Th2 paradigm. *Immunol Today* 1997;**18**:263–266.
70. VAN NEERVEN RJ, EBNER C, YSSEL H, KAPSENBERG ML, LAMB JR. T-cell responses to allergens: epitope-specificity and clinical relevance. *Immunol Today* 1996;**17**:526–532.
71. ROBINSON J, MALIK A, PARHAM P, BODMER JG, MARSH SG. IMGT/HLA database – a sequence database for the human major histocompatibility complex. *Tissue Antigens* 2000;**55**:280–287.
72. KIM YK, OH HB, OH SY, CHO SH, KIM YY, MIN KU. HLA-DRB1*07 may have a susceptibility and DRB1*04 a protective effect upon the development of a sensitization to house dust mite *Dermatophagoides pteronyssinus*. *Clin Exp Allergy* 2001;**31**:110–115.
73. NELSON CA, FREMONT DH. Structural principles of MHC class II antigen presentation. *Rev Immunogenet* 1999;**1**:47–59.
74. MAFFEI A, PAPADOPOULOS K, HARRIS PE. MHC class I antigen processing pathways. *Hum Immunol* 1997;**54**:91–103.
75. RAMMENSEE HG, BACHMANN J, EMME- RICH NP, BACHOR OA, STEVANOVIC S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999;**50**:213–219.
76. HAMMER J, BONO E, GALLAZZI F, BELUNIS C, NAGY Z, SINIGAGLIA F. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J Exp Med* 1994;**180**:2353–2358.
77. BRUSIC V, RUDY G, HONEYMAN G, HAMMER J, HARRISON LC. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics* 1998;**14**:121–130.
78. BRUSIC V, PETROVSKY N, ZHANG G, BAJIC VB. Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol Cell Biol* 2002;**80**:280–285.
79. HONEYMAN MC, BRUSIC V, STONE NL, HARRISON LC. Neural network-based prediction of candidate T-cell epitopes. *Nat Biotechnol* 1998;**16**:966–969.
80. ZAROUR HM, KIRKWOOD JM, KIERSTEAD LS, HERR W, BRUSIC V, SLINGLUFF CL Jr et al. Melan-A/MART-1(51–73) represents an immunogenic HLA-DR4-restricted epitope recognized by melanoma-reactive CD4(+) T cells. *Proc Natl Acad Sci USA* 2000;**97**:400–405.
81. BORRAS-CUESTA F, GOLVANO J, GARCIA-GRANERO M, SAROBE P, RIEZU-BOJ J, HUARTE E et al. Specific and general HLA-DR binding motifs: comparison of algorithms. *Hum Immunol* 2000;**61**:266–278.
82. RAULF-HEIMSOETH M, CHEN Z, RIHS HP, KALBACHER H, LIEBERS V, BAUR X. Analysis of T-cell reactive regions and HLA-DR4 binding motifs on the latex allergen Hev b 1 (rubber elongation factor). *Clin Exp Allergy* 1998;**28**:339–348.
83. STURNIOLO T, BONO E, DING J, RADDRIZZANI L, TUERECI O, SAHIN U et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 1999;**17**:555–561.
84. JURCEVIC S, TRAVERS PJ, HILLS A, AGREWALA JN, MORENO C, IVANYI J. Distinct conformations of a peptide bound to HLA-DR1 or DRB5*0101 suggested by molecular modelling. *Int Immunol* 1996;**8**:1807–1814.
85. TOH H, KAMIKAWAJI N, TANA T, MUTA S, SASAZUKI T, KUHARA S. 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* 2000;**13**:423–429.
86. METCALFE DD, ASTWOOD JD, TOWNSEND R, SAMPSON HA, TAYLOR SL, FUCHS RL. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit Rev Food Sci Nutr* 1996;**36**:S165–S186.
87. FAO/WHO. Evaluation of allergenicity of genetically modified foods. Joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology. Rome: FAO/WHO, 2001;22–25.
88. GENDEL SM. The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv Food Nutr Res* 1998;**42**:45–62.
89. AALBERSE RC, AKKERDAAS J, VAN REE R. Cross-reactivity of IgE antibodies to allergens. *Allergy* 2001;**56**:478–490.
90. SCHEURER S, WANGORSCH A, NERKAMP J, SKOV PS, BALLMER-WEBER B, WUTHRICH B et al. Cross-reactivity within the profilin panallergen family investigated by comparison of recombinant profilins from pear (Pyr c 4), cherry (Pru av 4) and celery (Api g 4) with birch pollen profilin Bet v 2. *J Chromatogr B Biomed Sci Appl* 2001;**756**:315–325.
91. BEEZHOLD DH, HICKEY VL, SUSSMAN GL. Mutational analysis of the IgE epitopes in the latex allergen Hev b 5. *J Allergy Clin Immunol* 2001;**107**:1069–1076.
92. ASERO R, MISTRELLO G, RONCAROLO D, DE VRIES SC, GAUTIER MF, CIURANA CL et al. Lipid transfer protein: a pan-allergen in plant-derived foods that is highly resistant to pepsin digestion. *Int Arch Allergy Immunol* 2001;**124**:67–69.
93. FUJITA C, MORIYAMA T, OGAWA T. Identification of cyclophilin as an IgE-binding protein from carrots. *Int Arch Allergy Immunol* 2001;**125**:44–50.
94. NEUDECKER P, SCHWEIMER K, NERKAMP J, SCHEURER S, VIETHS S, STICHT H et al. Allergic cross-reactivity made visible: solution structure of the major cherry allergen Pru av 1. *J Biol Chem* 2001;**276**:22756–22763.
95. YIANNIAS JA, EL-AZHARY RA. Contact allergen avoidance program: a topical skin care product database. *Am J Contact Dermat* 2000;**11**:243–247.
96. POPP W, WAGNER C, KISS D, ZWICK H, SERTL K. Prediction of sensitization to flour allergens. *Allergy* 1994;**49**:376–379.
97. KRIVENTSEVA EV, BISWAS M, APWEILER R. Clustering and analysis of protein families. *Curr Opin Struct Biol* 2001;**11**:334–339.
98. MYERS G. Optimally separating sequences. *Genome Inform Ser Workshop Genome Inform* 2001;**12**:165–174.

99. YOSHIMORI A, DEL CARPIO CA. Automatic epitope recognition in proteins oriented to the system for macromolecular interaction assessment MIAX. *Genome Inform Ser Workshop Genome Inform* 2001;**12**:113–122.
100. MJOSLESNESS E, DECOSTE D. Machine learning for science: state of the art and future prospects. *Science* 2001;**293**:2051–2055.
101. BRUSIC V, ZELEZNIKOW J. Knowledge discovery and data mining in biological databases. *Know Eng Rev* 1999;**14**:257–277.
102. TAYLOR SL, HEFLE SL. Will genetically modified foods be allergenic? *J Allergy Clin Immunol* 2001;**107**:765–771.
103. ILSI Health and Environmental Sciences Institute. Allergenicity assessment for foods derived from genetically modified crops. Protein Allergenicity Subcommittee. ILSI Health and Environmental Sciences Institute, July 2001. <http://www.ilsa.org>.