

Letter to the editor

Basic immunology, allergen prediction, and bioinformatics

In two recent papers, Brusic et al. reviewed why and how bioinformatic methods could help to classify allergens and to predict the allergenicity of a given protein (1, 2). For instance, amino acid sequence comparisons are currently used during the allergenicity assessment of newly expressed proteins in bioengineered foods: any homology between the transgenic protein and known allergens is regarded as a risk for potential cross-reactivity and/or immunogenicity (i.e. IgE synthesis) (3).

But, as stressed by Brusic et al., general purpose molecular databases [e.g. SWISS-PROT (4)] or specialized databases of allergens are either incomplete or not suited to perform such bioinformatic computations. In an effort to set up a comprehensive 'Reference Allergen Database', Brusic et al. thus merged and studied the sequences retrieved from all public databases (2). No less than 3051 different sequences were finally kept in their file, a figure three times larger than those found in other databases devoted to allergens. Such a difference is rather surprising. It results quite exclusively from extra entries retrieved from Entrez NCBI's database (5). But Brusic's query of this database through the keywords '(iso)allergen(s)' clearly gave many false hits because these words often appears in annotation fields (e.g. literature references).

For example, the same query found 3042 sequences in Entrez database on October 29, 2003, a figure compatible with Brusic's data compiled 2 years before (2423 sequences). But 1463 out of these 3042 sequences did not correspond to allergens: 167 human or mouse proteins involved in immunology but not allergens, 172 proteins from lab-only species (*Arabidopsis*, *Caenorhabditis*, zebrafish), and 1124 unproven IgE-reactive sequences either ill-defined (89) or present in Entrez as US patent data (1035, among them 82% had less than 40 amino acids). After treating redundancy among the 1579 sequences left, a final score of only 1052 different sequences was found

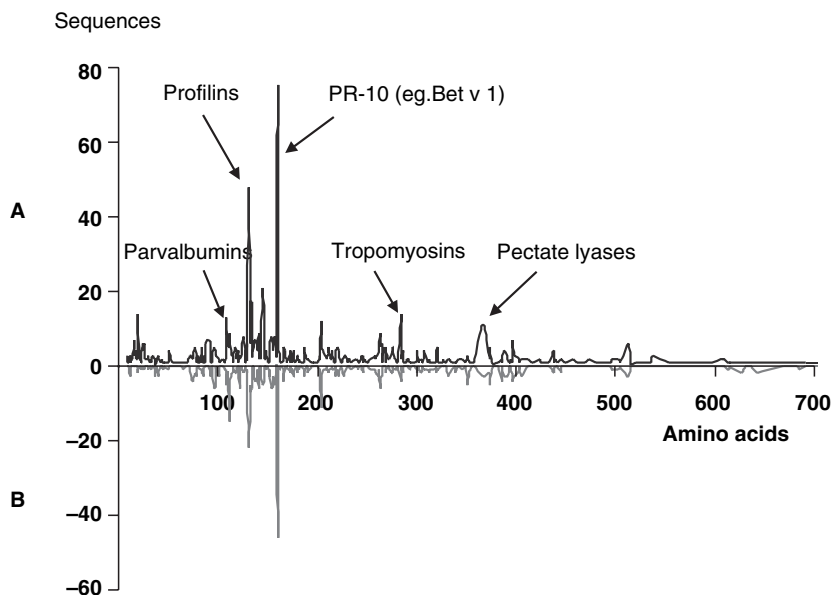


Figure 1. Number of allergen sequences as a function of protein length (source: NCBI Entrez database as of Oct 29, 2003): some allergens (eg. Bet v 1) or molecular families have been studied in great excess compared to others.

Panel A: non-redundant sequences ($n = 1052$). Not shown: 14 sequences longer than 700 amino acids.

Panel B: redundant sequences ($n = 527$).

(see Fig. 1), a result in line with the data stored in Allergome, today the most carefully updated allergen database (6).

The gap between bioinformatics and immunology is also underlined by a mistake in the allergic cascade depicted in Brusic's Fig. 1 (1): sensitization is not primed by molecular fragments such as B-cell epitopes. B-cell and even T-cell epitopes are only second-hand actors and cannot explain why a protein is sensed as an allergen by antigen presenting cells (APCs). That no characteristic sequence for a B- or T-epitope has been found so far is not surprising because APCs must detect whole molecules before cutting them into pieces. Thus recognizing an allergen depends on certain properties of the whole, genuine protein, these clues being lost in

T-cell epitopes. And recent studies in invertebrates and plants have shown that similar mechanisms are involved throughout evolution to sense danger: a panel of receptors [e.g. Toll-like (7)] are responsible for detecting some characteristic molecular patterns, often made of hybrid components (peptide, lipid, ionic, and/or carbohydrate moieties). This composite nature of many immunogenic patterns probably limits our ability to predict a sensitization risk with bioinformatic tools studying peptide backbones only.

Henri Malandain
Laboratoire de Biochimie, Centre
Hospitalier Chubert, Vannes, France

References

1. BRUSIC V, PETROVSKY N, GENDEL SM, MILLOT M, GIGONZAC O, STELMAN SJ. Computational tools for the study of allergens. *Allergy* 2003;**58**:1083–1092.
2. BRUSIC V, MILLOT M, PETROVSKY N, GENDEL SM, GIGONZAC O, STELMAN SJ. Allergen databases. *Allergy* 2003;**58**:1093–1100.
3. FAO/WHO. Evaluation of allergenicity of genetically modified foods. Report of a joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology (22–25/01/01). <http://www.fao.org/es/esn/food/allergygm.pdf>.
4. O'DONOVAN C, MARTIN MJ, GATTIKER A, GASTEIGER E, BAIROCH A, APWEILER R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 2002;**3**:275–284. <http://us.expasy.org/sprot/>.
5. NCBI PubMed. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein>.
6. MARI A. Allergome, a database of allergenic molecules. <http://www.allergome.org>.
7. TAKEDA K, KAISHO T, AKIRA S. Toll-like receptors. *Annual Rev Immunol* 2003;**21**:335–376.