

Structural bioinformatics

An attempt to define allergen-specific molecular surface features: a bioinformatic approachRuta Furmonaviciene¹, Brian J. Sutton⁴, Fabian Glaser⁵, Charlie A. Laughton², Nick Jones³, Herb F. Sewell¹ and Farouk Shakib^{1,*}¹Allergy Research Group, Institute of Infection, Immunity and Inflammation, ²Molecular Recognition Group, School of Pharmacy and ³Division of Otorhinolaryngology, School of Medical and Surgical Science, The University of Nottingham, Nottingham, UK, ⁴The Randall Division of Cell and Molecular Biophysics, King's College London, London, UK and ⁵European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UKReceived on July 11, 2005; revised and accepted September 28, 2005
Advance Access publication October 4, 2005**ABSTRACT**

Allergens are proteins that elicit T helper lymphocyte type 2 (Th2) responses culminating in IgE antibody production and allergic disease. However, we have no answer to the fundamental question of why certain proteins are allergens, while others are not. We hypothesized that analysis of the surface of diverse allergens may reveal common structural features which might enable them to be recognized as Th2-inducing antigens by cells of the innate immune system. We have therefore used the ConSurf server to search for allergen-specific motifs. This has enabled us to identify residue conservation patterns in the homologues of Ara t 8 (plant profilin), Act c 1 (actinidin), Bet v 1 (plant pathogenesis-related protein) and Ves v 5 (venom allergen). The results demonstrate the presence of allergen-specific patches consisting of an unusually high proportion of surface-exposed hydrophobic residues. The patches that have been identified may represent molecular patterns recognizable by cells of the innate immune system.

Contact: farouk.shakib@nottingham.ac.uk**Supplementary Information:** <http://www.nottingham.ac.uk/immunology/research/BI>**INTRODUCTION**

Allergens are proteins that elicit powerful T helper lymphocyte type 2 (Th2) responses, culminating in IgE antibody production and the development of allergic conditions such as asthma. However, we have no answer to the fundamental question of how allergens make the immune system respond in this way. Work done in the past few years has clearly established the central role of dendritic cells in the induction of Th2-mediated allergic diseases (Eisenbarth *et al.*, 2003), but given the innocuous nature of most allergens a major question remains: why do allergen-activated dendritic cells induce Th2, rather than Th1, responses? Thus, there is considerable interest in defining the nature of the molecular surface of allergenic proteins and the mechanisms involved in their initial recognition, and subsequent Th2 cell polarization, by cells of the innate immune system.

Innate immune responses against pathogens are thought to be triggered when a pathogen-associated molecular pattern (PAMP)

is recognized by a pattern recognition receptor (PRR) (Janeway, 1989), such as a toll-like receptor (TLR) (Janeway and Medzhitov, 1999; Reis e Sousa, 2001; Takeda *et al.*, 2003), leading to the development of protective Th1 responses. PAMPs represent conserved molecular patterns, also called 'molecular signatures', that are essential for the survival of microbes and are often shared by large groups of microorganisms. It is therefore conceivable that allergens too are endowed with a molecular pattern which, when recognized by a PRR, triggers a deleterious Th2 response, culminating in IgE antibody production and allergy.

A number of previous studies used different algorithms to try to predict allergenicity (WHO, 2001; Zorzet *et al.*, 2002; Stadler and Stadler, 2003), but these have only searched for sequence motifs, rather than common surface features. An earlier study (Aalberse, 2000) suggested that most allergens could be classified into four structural families, but again it fell short of demonstrating any common surface motifs.

In an attempt to define the molecular surface features of allergens that might enable them to be recognized as Th2-inducing antigens by cells of the innate immune system, we used the ConSurf server (Glaser *et al.*, 2003) to identify the conservation patterns and to search for allergen-specific motifs in diverse allergens. The ConSurf server enables the identification of functionally important regions on the surface of a protein or domain, based on the phylogenetic relations between its close sequence homologues, and projects the data onto a representative crystal structure. Here we report our results.

METHODS

The ConSurf web server (Glaser *et al.*, 2003) facilitates the identification of patterns of conserved and variable residues in a protein by estimating the degree of conservation of amino acids within its close sequence homologues, and subsequently mapping the conservation scores onto a representative crystal structure. ConSurf requires a minimum number (five) of close homologous sequences of similar length and a representative crystal structure. Using the three-dimensional (3D) structure of a protein as an input, the ConSurf server obtains the sequence from the PDB (Berman *et al.*, 2000) file and carries out a search for close homologous sequences of the protein using PSI-BLAST (Altschul *et al.*, 1997). This search is based on using the Swiss-Prot database (Bairoch and Apweiler, 1999) and a default single

*To whom correspondence should be addressed at Division of Immunology, Queen's Medical Centre, Nottingham NG7 2UH, UK

iteration of PSI-BLAST with an E-value cutoff of 0.001. The sequences obtained are then aligned using CLUSTALW (Thompson *et al.*, 1994) (with default parameters) and a phylogenetic tree is built using the neighbour joining algorithm (Saitou and Nei, 1987), as implemented in the Rate4Site program (Pupko *et al.*, 2002). Conservation scores corresponding to the site's evolutionary rate are calculated using the Bayesian method (Mayrose *et al.*, 2004) (ConSurf Version 3.0), which is a significant improvement over previous methods, particularly when only a small number of sequences are available. The Bayesian method also assigns a confidence interval to each of the inferred evolutionary conservation scores (Susko *et al.*, 2002). The proteins, with their conservation scores colour-coded on their surface, can finally be visualized online using the Protein Explorer engine (Martz, 2002).

The conservation scale used for visualization is obtained through ConSurf by processing the original continuous conservation scores obtained from Rate4Site. The colour grades (from 1 to 9) are assigned as follows: conservation scores below average (i.e. negative values, which are indicative of slowly evolving, conserved sites) are divided into 4.5 equal intervals. The same 4.5 intervals are used for scores above average (i.e. positive values, which are indicative of rapidly evolving, variable sites). Thus, nine equally sized categories of conservation or grades are obtained. Using this procedure, the 'width' (i.e. the maximum and minimum scores) of each colour grade would vary for different polypeptide chains. Thus, the colouring results of a ConSurf calculation do not indicate the absolute magnitudes of evolutionary distances, but rather the relative degrees of conservation for each residue.

RESULTS

Out of the SDAP (Ivanciuc *et al.*, 2003) list of allergens (<http://fermi.utmb.edu/SDAP>), only four allergen groups (95 allergens in total) met the ConSurf utility criteria of having a minimum number of five homologous allergen sequences (E-value cutoff of 0.001) of similar length and a representative structure.

All known homologous sequences for each representative crystal structure [Ara t 8/3NUL (1.60 Å) (Thorn *et al.*, 1997), Act c 1/2ACT (1.70 Å) (Baker, 1980), Bet v 1/1BV1 (2.00 Å) (Gajhede *et al.*, 1996) and Ves v 5/1QNX (1.90 Å) (Henriksen *et al.*, 2001)] were searched using the ConSurf server (Glaser *et al.*, 2003). The sequences were then divided into allergens and non-allergens. The protein was considered allergenic if referred to as such in the Swiss-Prot/TrEMBL database (ExPaSy site <http://ca.expasy.org/sprot/>) or in SDAP. Admittedly, however, some of the proteins that were considered non-allergens might be allergenic in some predisposed individuals and this may introduce some noise into the comparisons, but the fact that we are sure about the allergenic group means that we can be certain when looking for conservation patterns in allergens.

The list of allergens and non-allergens used as input for the ConSurf search is shown in Supplementary Table I. There were 37 allergens and 46 non-allergens in the 3NUL group, 9 allergens and 162 non-allergens in the 2ACT group, 27 allergens and 19 non-allergens in the 1BV1 group, and 22 allergens and 51 non-allergens in the 1QNX group. Although, the number of homologues in the 2ACT allergen group is rather small, however it is still acceptable by the ConSurf server and was included in our analysis since a smaller number of non-allergen homologues did not alter the conservation pattern.

The selected protein sequences were examined for common residue patterns using the ConSurf server, searching for common allergen-specific patterns not present in the 'control', non-

allergen group. All allergen-specific and accessible residues that are either highly conserved (red colour, levels 8–9) or highly variable (blue colour, levels 1–2), but which have only average scores (white) in the non-allergen sequence, are underlined in Figure 1. Also identified (underlined) are residues that are highly conserved (red) in the allergen sequence but highly variable (blue) in the non-allergen sequence, or vice versa, provided that they are accessible.

Figure 2 shows that allergen-specific surface residues were frequently hydrophobic: 55% in 3NUL, 48% in 2ACT, 39% in 1BV1 and 44% in 1QNX (chain A).

DISCUSSION

Allergic reactions consist of a series of events that start with recognition of the native allergen structure by antigen presenting cells, such as dendritic cells, and culminates in IgE antibody production and mast cell sensitization and triggering. Our hypothesis is that allergens display molecular patterns that are recognized by PRRs on antigen presenting cells. This encounter between the native allergen and the PRRs provides instructive signals for the immune system to mount an IgE antibody response leading to allergy.

The ConSurf server enables the identification of functionally important regions of the surface of a protein of known 3D structure based on the phylogenetic relations between its close sequence homologues. Four groups of allergens (95 allergens in total) were found with sufficient homologous sequences and at least one crystal structure to which this method could be applied. In each group, the sequence of the allergen of known structure was used in the PSI-BLAST search to identify all known allergenic and non-allergenic homologues. The allergen-specific motifs (i.e. clusters of residues in the allergen but not the non-allergen group) that were identified consisted of both highly conserved and highly variable residues. Interestingly, a previous study which identified functionally important surface patches in the MHC class I peptide-binding groove and in the antigenic surfaces of the influenza haemagglutinin has shown that these patches also consisted of both highly conserved and highly variable residues (ConSurf Gallery: <http://consurf.tau.ac.il/gallery.html>).

The hypothesis behind the present approach is that if there is a set of receptors that have a rather limited repertoire that recognize allergen-associated molecular patterns (by analogy with recognition molecules of innate immune cells such as the PRR) then we might expect to find common features among allergens that consist of highly conserved and also highly variable residues within defined patches. Our search revealed that there were indeed allergen-specific motifs formed by adjacent conserved and variable residues, which, unusually for surface residues, were mostly hydrophobic. Since the four groups studied here constitute only a small sample of allergens, it may be premature to generalize this result. However, this is the first demonstration of an allergen-specific surface feature displayed by structurally and functionally diverse groups of allergens.

The above findings are in line with the recent notion that the innate immune system might have evolved to detect hydrophobic portions of immunogenic proteins (Seong and Matzinger, 2004) consisting of a string of hydrophobic amino acids, rather than being dependent on the exact amino acid sequence (Berezovsky and Trifono, 2000). Those authors have therefore argued that what



Fig. 1. Residue conservation amongst allergen (upper line) and non-allergen (lower line) sequences belonging to the four allergen groups that meet the ConSurf utility criteria, namely the homologues of Ara t 8, Act c 1, Bet v 1 and Ves v 5. All allergen-specific residues that are either highly conserved (red) or highly variable (blue), but which have only average scores (white) in the non-allergen sequence, are underlined. Also underlined are residues that are highly conserved (red) in the allergen sequence but highly variable (blue) in the non-allergen sequence, or vice-versa, provided that they are accessible. Shown in green are residues (in the 1QNX chain A sequences) that could not be assessed for conservation because of lack of homologous motifs (i.e. present in <10% of the sequences). Residues that are assigned confidence intervals too large to be trustworthy are coloured in yellow.

makes a protein recognizable as an allergen (i.e. Th2-inducing protein) by antigen presenting cells is its hydrophobic nature, as illustrated by allergens such as lipocalins, lipid transfer proteins and seed storage proteins (Seong and Matzinger, 2004). The hydrophobic residues of such proteins are normally buried, but they may become exposed upon unfolding, such as following the loss of a specific ligand that is integrated into their 3D structure. Examples of such ligands include metal ions (e.g. Ca in parvalbumins) and lipids

(e.g. MD-2 proteins, lipocalins and lipid transfer proteins) (Breiteneder and Mills, 2005).

The relevance to allergen recognition by dendritic cells of the residues identified here will clearly need to be validated. This could be done by mutagenesis experiments and the mutant proteins tested for allergenicity, or in the longer term by virtual screening for ligands (PRRs) that bind to the putative molecular patterns identified.

3NUL allergens (profilins)

(Hydrophobic 55%, uncharged hydrophilic 18%, acidic 11% and basic 16%)

.SWQSYVDDH L.CDVEGNHL TAAAILGQDG SVWAQSAKFP
QLKPQEIDGI KKDFEEPGFL APTGLFLGGE KY.VIQGEQG
AVIRGKKGPG GVTIKKTNQA LVFGFYDEP. TGQCNLVVE
RLGDYLIESE L

2ACT allergens (cysteine proteases)

(Hydrophobic 48%, uncharged hydrophilic 33%, acidic 19% and basic 0%)

LPSYVDWRS GAVVDIKSQG ECGGCWAFSA IATVEGINKI
TSGSLISLSE QELIDCGRTQ NTRGCDGGYI TDGFQFIIND
GGINTEENYP YTAQDGDCDV ALQDQKYVTI DTYENVPYNN
EWALQTAVTY QPVSVALDAA GDAFQYASG IFTGPGTAV
DHAIVIVGYG TEGGVDYWIV KNSWDTTWGE EGYMRILRN
GGAGTCGIAT MPSYPVKY

1BV1 allergens (plant pathogenesis-related proteins)

(Hydrophobic 39%, uncharged hydrophilic 8%, acidic 23% and basic 31%)

GVFNYETETT SVIPAARLFK AFILDGDNLF PKVAPQAIS
VENIEGNGGP GTIKKISFPE GLPFKYVKDR VDEVDTNFK
YNSVIEGGP IGDTEKISN EIKIVATPDG GSILKISNKY
HTKGDEHVKA EQVKASKEMG ETLLRAVESY LLAHSDAYN

1QNX(A) allergens (venom allergens)

(Hydrophobic 44%, uncharged hydrophilic 23%, acidic 9% and basic 24%)

.....AEAEF NNYCKIKCLK GGVHTACKYG SLKPNCGNKV
VVSYGLTKQE KQDILKEHND FRQKIARGLE TRGNPGPQPP
AKNMKNLVWN DELAYVAQVW ANQCQYGHDT CRDVAKYQVG
QNVALTGSTA AKYDDPVKLV KMWEDEVKDY NPKKKFSGND
FLKTGHYTQM VWANTKEVGC GSIKYIQEKW HKHYLVCNYG
PSGNFKNEEL YQTK

Colour-coding:

VALMGIPFC	RED	Hydrophobic
DE	BLUE	Acidic
RHK	MAGENTA	Basic
STYNQ	GREEN	Uncharged hydrophilic

Fig. 2. The nature of residues constituting allergen-specific surface patches (underlined) found on the crystal structures representing the four allergen groups that met the ConSurf utility criteria. Allergen-specific surface residues are mostly hydrophobic (shown in red).

ACKNOWLEDGEMENTS

This work was primarily funded by Asthma UK (London; Grant ID 02/005) and partially supported by the Nasal Research Fund (Nottingham; Grant ID 7365).

Conflict of Interest: none declared.

REFERENCES

- Aalberse, R.C. (2000) Structural biology of allergens. *J. Allergy Clin. Immunol.*, **106**, 228–238.
- Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bairoch, A. and Apweiler, R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.
- Baker, E.N. Structure of actinidin, after refinement at 1.7 Å resolution. *J. Mol. Biol.*, **141**, 441–484.
- Berezovsky, I.N. and Trifonov, E. (2000) Protein structure and folding: a new start. *J. Biomol. Struct. Dyn.*, **19**, 397–403.
- Berman, H.M. et al. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Breiteneder, H. and Mills, E.N. (2005) Molecular properties of food allergens. *J. Allergy Clin. Immunol.*, **115**, 14–23.
- Eisenbarth, S.C. et al. (2003) The master regulators of allergic inflammation: dendritic cells in Th2 sensitization. *Curr. Opin. Immunol.*, **15**, 620–626.
- Gajhede, M. et al. (1996) X-ray and NMR structure of Bet v 1, the origin of birch pollen allergy. *Nat. Struct. Biol.*, **3**, 1040–1045.
- Glaser, F. et al. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
- Henriksen, A. et al. (2001) Major venom allergen of yellow jackets, Ves v 5: structural characterization of a pathogenesis-related protein superfamily. *Proteins*, **45**, 438–448.
- Ivanciu, O. et al. (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.*, **31**, 359–362.
- Janeway, C.A., Jr (1989) Approaching the asymptote? Evolution and revolution in immunology *Cold Spring Harb. Symp. Quant. Biol.*, **54**, 1–13.
- Janeway, C.A., Jr and Medzhitov, R. (1999) Lipoproteins take their toll on the host. *Curr. Biol.*, **9**, 879–882.
- Martiz, E. (2002) Protein explorer: easy yet powerful macromolecular visualization. *Trends Biochem. Sci.*, **27**, 107–109.
- Mayrose, I. et al. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.
- Pupko, T. et al. Rate4Site: an algorithmic tool for the identification of functional regions on proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** (Suppl.), S71–77.
- Reis e Sousa, C. (2001) Dendritic cells as sensors of infection. *Immunity*, **14**, 495–498.
- Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Seong, S.Y. and Matzinger, P. (2004) Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses. *Nat. Rev. Immunol.*, **4**, 469–478.
- Stadler, M.B. and Stadler, B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J.*, **17**, 1141–1143.
- Susko, E. et al. (2002) Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Mol. Biol. Evol.*, **19**, 1514–1523.
- Takeda, K. et al. (2003) Toll-like receptors. *Annu. Rev. Immunol.*, **21**, 335–376.
- Thompson, J.D. et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Thorn, K.S. et al. (1997) The crystal structure of a major allergen from plants. *Structure*, **15**, 19–32.
- WHO (2001), Evaluation of allergenicity of genetically modified foods. *Report of a Joint FAO/WHO Expert Consultation. World Health Organization. Geneva.*
- Zorzet, A. et al. (2002) Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol.*, **2**, 1–10.