Nucleic Acids Research, 2003, Vol. 31, No. 1 359–362
DOI: 10.1093/nar/gkg010

SDAP: database and computational tools for allergenic proteins

Ovidiu Ivanciuc, Catherine H. Schein and Werner Braun*

Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, 310 University Boulevard, Galveston, TX 77555-1157, USA

Received July 25, 2002; Revised and Accepted August 28, 2002

ABSTRACT

SDAP (Structural Database of Allergenic Proteins) is a web server that provides rapid, cross-referenced access to the sequences, structures and IgE epitopes of allergenic proteins. The SDAP core is a series of CGI scripts that process the user queries, interrogate the database, perform various computations related to protein allergenic determinants and prepare the output HTML pages. The database component of SDAP contains information about the allergen name, source, sequence, structure, IgE epitopes and literature references and easy links to the major protein (PDB, SWISS-PROT/TrEMBL, PIR-ALN, NCBI Taxonomy Browser) and literature (PubMed, MEDLINE) on-line servers. The computational component in SDAP uses an original algorithm based on conserved properties of amino acid side chains to identify regions of known allergens similar to user-supplied peptides or selected from the SDAP database of IgE epitopes. This and other bioinformatics tools can be used to rapidly determine potential cross-reactivities between allergens and to screen novel proteins for the presence of IgE epitopes they may share with known allergens. SDAP is available via the World Wide Web at http://fermi.utmb.edu/SDAP/.

INTRODUCTION

Allergic diseases, including allergic rhinitis, asthma and atopic dermatitis, are among the most common chronic health problems (1). As recombinant proteins, products of the genomic revolution are introduced into foods, medications and other products of our daily life, distinguishing allergens from other proteins becomes a pressing issue (2). This need was recently illustrated by the concern over possible allergenic effects of Starlink corn (3). The sequences and structures of many allergenic proteins have been determined. Most of these proteins can be grouped into relatively few families (4,5), suggesting that they share characteristics that contribute to their ability to bind IgE and trigger an allergic response (6,7). The web server SDAP (Structural Database of Allergenic Proteins) was created to aid in identifying these sequence commonalities. Computational and statistical tools to analyse sequences and structures, assessable within SDAP, have been designed to develop correlations that can be used to predict allergenicity of novel proteins and cross-reactivity between allergens.

The computational approach is briefly depicted in Figure 1. The sequences, 3D structures and IgE epitopes of known allergens, collected from databases and literature, are included in lists in the server. For allergens where no experimental structure has been determined, models will be computed with our MASIA/EXDIS/DIAMOD/FANTOM suite of programs (8–17).

Using this database and the computational tools from SDAP we will develop predictive models for IgE epitopes. This will allow a user to compare not just sequence or property similarity of a possible epitope homologue, but also factors such as local structure and relative surface exposure. SDAP is available at http://fermi.utmb.edu/sdap/.

DATABASE STRUCTURE

SDAP is designed as a web server (Fig. 2) controlled by a set of CGI scripts. These scripts mediate interaction with the user, the database and the computational tools. The database component of SDAP is implemented with MySQL under Linux. The information is collected in tables according to: allergen type; species; systematic name; brief description; sequence accession numbers from SWISS-PROT/TrEMBL, PIR-ALN, NCBI Taxonomy Browser and, where available, PDB. Sequences and IgE epitopes are collected into text files. The current lists of allergens were assembled from literature and from major sequence [SWISS-PROT/TrEMBL (18), PIR-ALN (19) and NCBI Taxonomy Browser (20)] and structure [PDB (21)] databases, guided by the list of allergen names from the IUIS website, http://www.allergen.org. As there is no public database summarizing information on known epitopes of allergenic proteins, the IgE epitope list in SDAP is based solely on primary literature sources. In its present version, SDAP allows searches restricted to the following fields: allergen name (according to the IUIS website listing), scientific and common name for the species and general source of the allergens. SDAP is the first allergen database that allows a user to retrieve IgE epitopes and identify similar regions in other allergenic proteins.

*To whom correspondence should be addressed. Tel: +1 409 747 6810; Fax: +1 409 747 6850; Email: werner@newton.utmb.edu

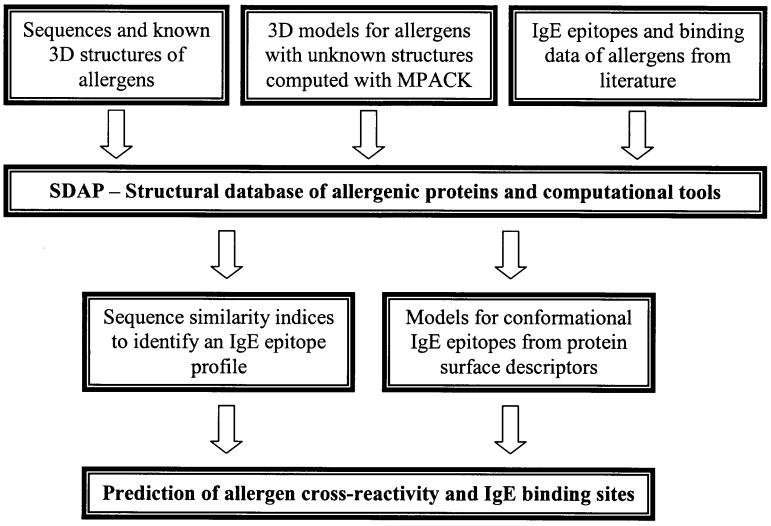

Figure 1. SDAP combines information from many sources and computational tools to rapidly determine potential cross-reactions among allergens and the allergenicity of novel proteins.

COMPUTATIONAL TOOLS

Special programs have been incorporated in SDAP to compare the sequences and structures in the database. In the present release, users can compare a given peptide sequence to all the sequences in SDAP, using either an exact match search or a similarity search based on physicochemical properties (22). The SDAP peptide exact match function is useful to identify allergens closely related to a peptide, for example when an IgE epitope may be responsible for clinically-defined cross-sensitivities to several allergens.

One use of SDAP is to quickly identify cross-reactivities between known allergens. For example, an exact match search using the Pen i 1 IgE epitope MQQLENDLDQVQESLLK from shrimp topomyosin identified the same sequence in the allergens Met e 1 (from another shrimp species) and Pan s 1 (lobster), consistent with the clinically observed cross-reactivity among these crustaceans (see Table 1). However, sequence identity, even among know cross-reactive allergens, is rare. To identify more distantly related sequences, the user can access a tool that uses the amino acids descriptors E_1–E_5 (22) to locate sequences with similar chemical properties. The vectors were derived by multidimensional scaling of 237 physical–chemical properties for all 20 naturally occurring amino acids. The mathematical procedure used ensures that the main variations of all 237 properties for the 20 amino acids are reflected by the five descriptors E_1–E_5. Using the E_1–E_5 descriptors, the similarity between two sequences A and B, each one consisting of N residues, is computed with the property distance function PD (23):

$$PD(A,B) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{5}\lambda_j(E_j(A_i) - E_j(B_i))^2\right]^{1/2}$$

where λ_j is the eigenvalue of the jth E component, $E_j(A_i)$ is the E_j value for the amino acid in the ith position from sequence A, and $E_j(B)_i$ is the E_j value for the amino acid in the ith position from sequence B.

The SDAP tool calculates the PD similarity index between the query sequence and each sequence-window with the same length from all allergens collected in the SDAP protein database. The search result is a list of similar sequences identified in allergenic proteins, presented in decreasing order of similarity (increasing PD) with the query sequence. Besides epitope identification, this tool can be used to find conserved regions in the allergens from the SDAP protein database.

The results of a sequence similarity search for the above Pen i 1 IgE epitope are shown in Table 1. Besides the two exact matches with Met e 1 and Pan s 1, similar regions were found in allergenic tropomyosins from insects (storage mite, American cockroach, American house dust mite and European house dust mite), lobster, crab, abalone and snail. Clinical tests with sera from sensitive patients indicate that the cross-reactivity between crustacean, mollusk and insect allergens is mediated by tropomyosin (24,25). Sensitization to house dust mites had been linked to oral allergic response to snails (7) and allergen immunotherapy with European house dust mite (*Dermatophagoides pteronyssinus*) may trigger severe reactions to mollusks and crustacea (26). These results demonstrate the utility of SDAP in identifying potential cross-reactivity among allergens.

The significance of any sequence match can be determined from a histogram of the distribution of PD values for all the sequences in SDAP to the shrimp epitope (Fig. 3). The sequences in Table 1 clearly have a lower PD value than the bulk of entries in SDAP. The histogram suggests that a significance cut-off value between 7.5 and 9 would be most appropriate for determining peptides with similar properties.

DATA SUBMISSION

Researchers in the allergy field are welcome to submit their published data by email to oiivanci@utmb.edu. Comments,

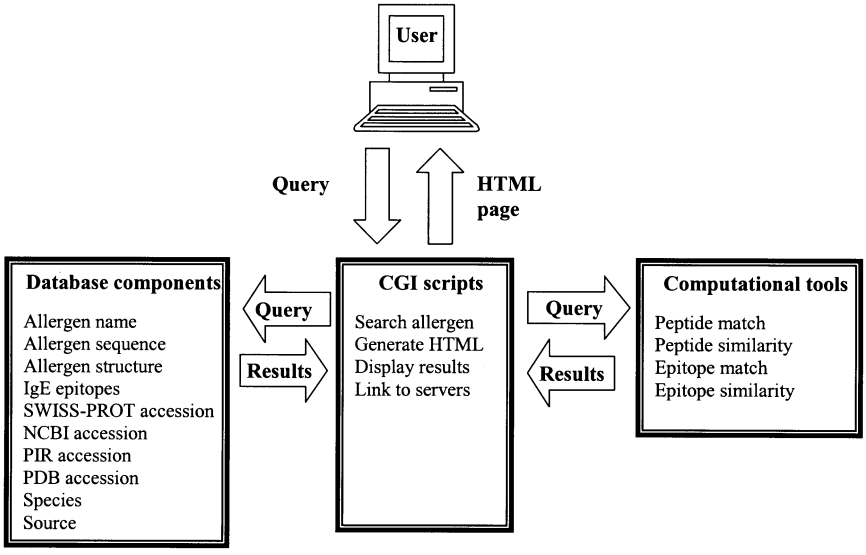

Figure 2. Structural blocks and main functions of SDAP.

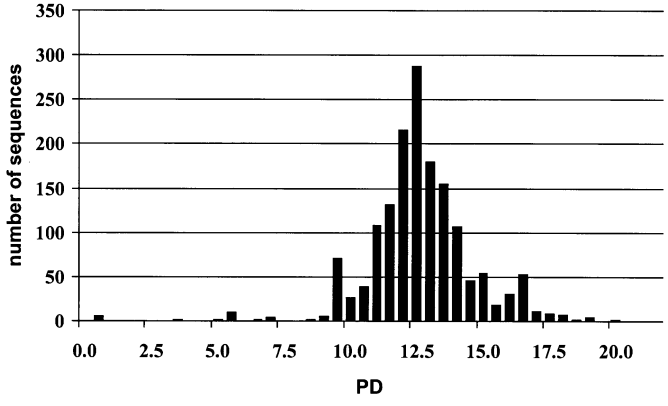

Figure 3. Histogram of *PD* values for an IgE epitope from shrimp for all SDAP entries. Only the most similar sequence region (lowest *PD* value) was recorded for each SDAP entry.

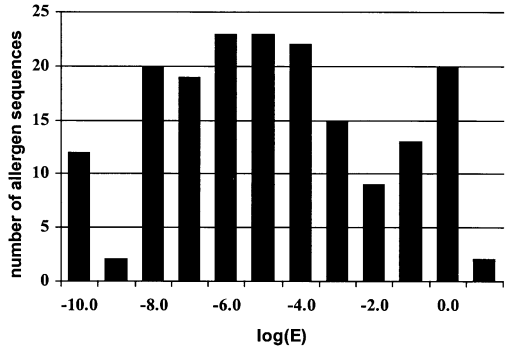

Figure 4. Distribution of log(E) values for PDB templates identified by 3D-PSSM for 180 allergen sequences. All structural alignments with $\log(E) < -1$ can be modeled with good precision.

Table 1. SDAP search results for allergens that contain regions similar to the Pen i 1 IgE epitope MQQLENDLDQVQESLLK

NCBI Taxonomy Browser Accession	Allergen	Species	Seq. Len.	Position	Sequence	*PD*
TPM1_METEN	Met e 1	*Metapenaeus ensis*, shrimp	274	40–56	MQQLENDLDQVQESLLK	0
TPM1_PANST	Pan s 1	*Panulirus stimpsoni*, lobster	274	40–56	MQQLENDLDQVQESLLK	0
TPM_LEPDS	Lep d 10	*Lepidoglyphus destructor*, storage mite	284	50–66	IQQIENELDQVQESLTQ	3.345
TPM1_HOMAM	Hom a 1	*Homarus americanus*, American lobster	284	50–66	MQQVENELDQVQEQLSL	4.731
TPM_PERAM	Per a 7	*Periplaneta Americana*, American cockroach	284	50–66	IQQIENDLDQTMEQLMQ	5.062
BAA04557	Der f 10	*Dermatophagoides farinae*, American house dust mite	299	65–81	IQQIENELDQVQEQLSA	5.074
TPM_DERPT	Der p 10	*Dermatophagoides pteronyssinus*, European house dust mite	284	50–66	IQQIENELDQVQEQLSA	5.074
TPM1_CHAFE	Cha f 1	*Charybdis feriatus*, crab	264	50–66	MQQVENELDQAQEQLSA	5.389
AF216518_1	Hal d 1	*Haliotis diversicolor*, abalone	284	50–66	CANLENDFDNVNEQLQE	6.490
TPMM_ANISI	Ani s 3	*Anisakis simplex*	284	50–66	MMQTENDLDKAQEDLST	6.661
JE0229	Tur c 1	*Batillus cornutus*	146	1–17	AANLENDFDNVNEQLQD	6.710

corrections and suggestions for new computational tools for allergenic determinants should be sent to the same address.

FUTURE DEVELOPMENTS

The SDAP server will be maintained on a regular basis. The database sequence, structure and IgE epitopes lists will be updated with information as it becomes available in other databases and the literature. The next major addition to SDAP will be data and software to compare the structures of epitopes. The 3D structure is known for only about 10% of the sequences in SDAP. For the sequences without structures, homology models will be prepared with our self-correcting distance geometry based EXDIS/DIAMOD/FANTOM suite. To determine whether suitable templates were available, the sequences of 180 allergens in SDAP with unknown 3D structure were submitted to the fold recognition server 3D-PSSM (27) (http://www.sbg.bio.ac.uk/~3dpssm) and a histogram of the distribution of the log(E) values for the best template found in the PDB (Fig. 4). For 150 sequences, a good template [$\log(E) < -1$] was identified. For the remaining allergens, eight have a log(E) between -1 and 0, meaning that the modeling may require combined information from other fold recognition servers. For these, we plan to use our MASIA program (8) (http://www.scsb.utmb.edu/masia/masia.html) to identify conserved motifs, which can be used to identify possible templates and improve alignments with distantly related proteins. Only 22 sequences, with $\log(E) > 0$, will require an alternative approach. Here, we will use secondary structure prediction methods to model the proteins *ab initio*.

ACKNOWLEDGEMENT

This work was supported by a Research Development Grant (#2535-01) from the John Sealy Memorial Endowment Fund for Biomedical Research.

REFERENCES

1. Malone,D.C., Lawson,K.A., Smith,D.H., Arrighi,H.M. and Battista,C. (1997) A cost of illness study of allergic rhinitis in the United States. *J. Allergy Clin. Immunol.*, **99**, 22–27.
2. Gendel,S.M. (1998) Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food Nutrition Res.*, **42**, 63–92.
3. FIFRA (2000) A set of scientific issues being considered by the environmental protection agency regarding: assessment of scientific information concerning StarLink Corn. Rep. No. SAP Report No. 2000-06. FIFRA.
4. Aalberse,R.C. (2000) Structural biology of allergens. *J. Allergy Clin. Immunol.*, **106**, 228–238.
5. Breiteneder,H. and Ebner,C. (2000) Molecular and biochemical classification of plant-derived food allergens. *J. Allergy Clin. Immunol.*, **106**, 27–36.
6. Ipsen,H. and Lowenstein,H. (1997) Basic features of crossreactivity in tree and grass pollen allergy. *Clin. Rev. Allergy Immunol.*, **15**, 389–396.
7. Sicherer,S.H. (2001) Clinical implications of cross-reactive food allergens. *J. Allergy Clin. Immunol.*, **108**, 881–890.
8. Zhu,H., Schein,C.H. and Braun,W. (2000) MASIA: recognition of common patterns and properties in multiple aligned protein sequences. *Bioinformatics*, **16**, 950–951.
9. Schaumann,T., Braun,W. and Wuthrich,K. (1990) The program FANTOM for energy refinement of polypeptides and proteins using a Newton–Raphson minimizer in torsion angle space. *Biopolymers*, **29**, 679–694.
10. Mumenthaler,C. and Braun,W. (1995) Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.*, **4**, 863–871.
11. Fraczkiewicz,R. and Braun,W. (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.*, **19**, 319–333.
12. Zhu,H. and Braun,W. (1999) Sequence specificity, statistical potentials and 3D structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. *Protein Sci.*, **8**, 326–342.
13. Zhu,H., Schein,C.H. and Braun,W. (1999) Homology modeling and molecular dynamics simulations of PBCV-1 glycosylase complexed with UV-damaged DNA. *J. Mol. Model.*, **5**, 302–316.
14. Soman,K., Midoro-Horiuti,T., Ferreon,J., Goldblum,R., Brooks,E., Kurosky,A., Braun,W. and Schein,C.H. (2000) Homology modeling and characterization of IgE epitopes of mountain cedar allergen Jun a 3. *Biophys. J.*, **79**, 1601–1609.
15. Soman,K.V., Schein,C.H., Zhu,H. and Braun,W. (2001) Homology modeling and simulations of nuclease structures. In Schein,C.H. (ed.), *Nuclease Methods and Protocols, Methods in Molecular Biology.* Humana Press, Totowa, N.J., Vol. 160, pp. 263–286.
16. Schein,C.H., Nagle,G.T., Page,J.S., Sweedler,J.V., Xu,Y., Painter,S.D. and Braun,W. (2001) Aplysia attractin: Biophysical characterization and modeling of a water-borne pheromone. *Biophys. J.*, **81**, 463–472.
17. Murtazina,D., Puchkaev,A.V., Schein,C.H., Oezguen,N., Braun,W., Nanavati,A. and Pikuleva,I.A. (2002) Membrane protein interactions contribute to efficient 27-hydroxylation of cholesterol by mitochondrial cytochrome P450-27a1. *J. Biol. Chem.*, **277**, 37582–37589.
18. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
19. Wu,C.H., Huang,H.Z., Arminski,L., Castro-Alvear,J., Chen,Y.X., Hu,Z.Z., Ledley,R.S., Lewis,K.C., Mewes,H.W., Orcutt,B.C., Suzek,B.E., Tsugita,A., Vinayaka,C.R., Yeh,L.S.L., Zhang,J. and Barker,W.C. (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.
20. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16.
21. Westbrook,J., Feng,Z.K., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S., Bourne,P.E. and Berman,H.M. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
22. Venkatarajan,M.S. and Braun,W. (2001) New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *J. Mol. Model.*, **7**, 445–453.
23. Ivanciuc,O., Schein,C.H. and Braun,W. (2002) Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics*, **18**, 1358–1364.
24. Leung,P.S.C. and Chu,K.H. (2001) cDNA cloning and molecular identification of the major oyster allergen from the Pacific oyster *Crassostrea gigas*. *Clin. Exp. Allergy*, **31**, 1287–1294.
25. Santos,A.B.R., Chapman,M.D., Aalberse,R.C., Vailes,L.D., Ferriani,V.P.L., Oliver,C., Rizzo,M.C., Naspitz,C.K. and Arruda,L.K. (1999) Cockroach allergens and asthma in Brazil: Identification of tropomyosin as a major allergen with potential cross-reactivity with mite and shrimp allergens. *J. Allergy Clin. Immunol.*, **104**, 329–337.
26. vanRee,R., Antonicelli,L., Akkerdaas,J.H., Garritani,M.S., Aalberse,R.C. and Bonifazi,F. (1996) Possible induction of food allergy during mite immunotherapy. *Allergy*, **51**, 108–113.
27. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J.E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.