PROCESS AND SYSTEM FOR PREDICTION OF PROTEIN ALLERGENICITY USING DISCONTINUOUS SURFACE PEPTIDE SEQUENCE AND VIRTUAL AMINO ACID

Field of the Invention

[0001] The present invention generally relates to bioinformatics, and more particularly to a process and system for prediction of the allergenicity of a query protein using discontinuous surface peptide sequence and virtual amino acid

Background of the Invention

[0002] Allergies affect approximately one-third of the human population and are the most common cause of chronic illness in developed countries. The current treatment of allergies relies chiefly on symptomatic therapies that include antihistamines, corticosteroids or, in severe cases, injectable epinephrine. However, it is well known that allergies can be prevented if an allergen is known to and avoided by an allergic individual, so the effective way to prevent allergy is allergen avoidance. It is even more effective when the allergens are proteins in our food. If any protein in the food is known as a potential allergen, any allergy induced by the allergic protein can be prevented by avoiding the consumption of the allergic protein.

[0003] The assessment of whether a protein is allergenic includes three main aspects: immunogenicity, cross-reactivity and clinical symptoms. It is well known that allergenic reactions are immune responses in nature; thus, immunogenicity of an allergenic protein refers to the likelihood of an immune response to the allergenic protein, normally in the form of either an immunoglobulin E (IgE) antibody or T-cell response. Because B-cell and T-cell epitopes are defined recognition sites on allergens, it is possible to evaluate the potential allergenicity of proteins by examining a protein or its gene source so as to determine its sequence similarity to known allergens.

[0004] One example of the sequence similarity approach to allergen detection is based on the " eight amino acids match (EAAM) approach". The current recommendations

for identification of potential allergens include the use of FASTA or BLAST searches for any eight or more contiguous amino acids that are identical to any segment of a known allergen (food, inhalant or contact allergen). Because the identity of eight amino acids between 2 proteins rarely occurs by chance, such an occurrence is thought to be a good indicator of potential allergenicity. Other methods for sequence search include: a minimum of 35% identity within an 80 amino acids segment; or any match of six contiguous amino acids, using motif or short peptides.

[0005] All the above allergen prediction methods use only the sequence information but not structural information despite the fact that it has been demonstrated that the structure of a protein has strong influence over the function of the protein. The inventors of the present invention have earlier filed a patent application with a title "Process and System For Analyzing Three-Dimensional Structure Similarities Of Proteins And Identifying Three-dimensional Structure Motifs Of Proteins And Their Applications", which discloses the use of three-dimensional information for allergen prediction. However, the massive information contained in the 3D data requires larger computational power than do traditional methods that employ only the sequence data. In order to utilize 3D data but reduce the requirement of computational power, the present invention provides processes and systems for prediction of allergens using discontinuous surface peptide sequence and virtual amino acid.

Summary of the Invention

[0006] One embodiment of the present invention provides a process for predicting the allergenicity of a query protein. The process comprises the steps of providing three-dimensional (3D) structure data of a plurality of known allergens, calculating solvent accessible surface area for all atoms of each of the plurality of known allergens, extracting 3D positions data for every Cα atom of each of the plurality of known allergens, extracting 3D positions data for surface Cα atoms of each of the plurality of known allergens to represent its 3D structure; wherein the surface Cα_atoms have a solvent accessible surface area value less than a predetermined threshold, generating a set of discontinuous surface peptides (DSPs) for each of the plurality of known allergens, wherein the set of DSPs have the same number of amino acids, calculating for each DSP the distance between two neighboring Cα atoms and inserting one or more virtual amino acids if the distance is larger

than a predetermined threshold, translating all DSPs into linear amino acid sequences including any inserted virtual amino acids, finally cutting all DSPs into a predetermined length and storing all the processed DSPs into the discontinuous surface peptide database (DSPD), manipulating the query protein by the above relevant steps to get a set of DSPs, calculating the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD and storing the maximum similarity score as the query protein's similarity score, and predicting the query protein as an allergen if the maximum similarity score of the query protein is bigger than a threshold value.

[0007] In another embodiment of the process, the predetermined length of all DSPs is 6-40 amino acids. In a further embodiment of the process, the predetermined length of all DSPs is 15-25 amino acids.

[0008] In another embodiment of the process, in the step of generating a set of DSPs, the DSPs are generated by a nearest neighbor method.

[0009] In another embodiment of the process, in the step of calculating for each DSP the distance between two neighboring Cα_atoms, the calculation is performed according to:

[0010] $$dis\tan ce = \sqrt{(x1-x2)^2+(y1-y2)^2+(z1-z2)^2}$$

[0011] where distance is the distance between the two neighboring Cα atoms; (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Cα atoms.

[0012] In another embodiment of the process, in the step of inserting one or more virtual amino acids, the threshold is determined by obtaining the median value of all Cα-Cα distances in all pdb files.

[0013] In another embodiment of the process, in the step of calculating the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD, the calculation is performed according to:

[0014] $$Similarity\ Score = \sum_{i=1}^{l} s(X_i Y_i)$$

[0015] where $s(X_i\ Y_i)$ is the substitution score between amino acid i in peptide X and the amino acid i in peptide Y , according to the substitution matrix employed. l is the length of the peptides.

[0016] In another embodiment of the process, in the step of predicting the query protein as an allergen if the maximum similarity score of the query protein is bigger than a

threshold value, the threshold value is an optimum threshold value from all maximum similarity scores from all positive and negative test data that could give best sensitivity and specificity results.

[0017] Another embodiment of the present invention provides an article of manufacture. The article comprises at least one computer usable medium having computer readable program code means embodied therein for predicting the allergenicty of a query protein. In one embodiment of the article, the computer readable program code means in said article of manufacture comprises computer readable program code means for causing a computer to provide three-dimensional (3D) structure data of a plurality of known allergens, computer readable program code means for causing a computer to calculate solvent accessible surface area for all atoms of each of the plurality of known allergens, computer readable program code means for causing a computer to extract 3D positions data for every Cα atom of each of the plurality of known allergens, computer readable program code means for causing a computer to extract 3D positions data for surface Cα atoms of each of the plurality of known allergens to represent its 3D structure; wherein the surface Cα atoms have a solvent accessible surface area value less than a predetermined threshold, computer readable program code means for causing a computer to generate a set of discontinuous surface peptides (DSPs) for each of the plurality of known allergens, wherein the set of DSPs have the same number of amino acids, computer readable program code means for causing a computer to calculate for each DSP the distance between two neighboring Cα atoms and insert one or more virtual amino acids if the distance is larger than a predetermined threshold, computer readable program code means for causing a computer to translate all DSPs into linear amino acid sequences including any inserted virtual amino acids, computer readable program code means for causing a computer to finally cut all DSPs into a predetermined length and store all the processed DSPs into the discontinuous surface peptide database (DSPD), computer readable program code means for causing a computer to manipulate the query protein by the above relevant steps to get a set of DSPs, computer readable program code means for causing a computer to calculate the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD and store the maximum similarity score as the query protein's similarity score, and computer readable program code means for causing a computer to predict the query protein as an allergen if the maximum similarity score of the query protein is bigger than a threshold value.

[0018] In another embodiment of the article, the predetermined length of all DSPs is 6-40 amino acids. In a further embodiment of the article, the predetermined length of all DSPs is 15-25 amino acids.

[0019] In another embodiment of the article, when generating a set of DSPs, the DSPs are generated by a nearest neighbor method.

[0020] In another embodiment of the article, when calculating for each DSP the distance between two neighboring Cα atoms, the calculation is performed according to:

[0021] $dis\tan ce = \sqrt{(x1-x2)^2+(y1-y2)^2+(z1-z2)^2}$

[0022] where distance is the distance between the two neighboring Cα atoms; (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the C α atoms.

[0023] In another embodiment of the article, when inserting one or more virtual amino acids, the threshold is determined by obtaining the median value of all Cα-Cα distances in all pdb files.

[0024] In another embodiment of the article, when calculating the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD, the calculation is performed according to:

[0025] $Similarity\ Score = \sum_{i=1}^{l} s(X_i Y_i)$

[0026] where $s(X_i\ Y_i)$ is the substitution score between amino acid i in peptide X and the amino acid i in peptide Y , according to the substitution matrix employed. l is the length of the peptides.

[0027] In another embodiment of the article, when predicting the query protein as an allergen if the maximum similarity score of the query protein is bigger than a threshold value, the threshold value is an optimum threshold value from all maximum similarity scores from all positive and negative test data that could give best sensitivity and specificity results.

[0028] One of advantages of the present invention is that no pre-alignment of amino acid sequences is required. In addition, the extracted Cα atoms make distance computation simple. Furthermore, the short peptides make wide range of proteins comparable. Finally, the employment of the DSPs with inserted virtual amino acids maintains structural information and at the same time requires less memory during the process than that of 3D data, thus making the process faster.

[0029] The objectives and advantages of the invention will become apparent from the following detailed description of preferred embodiments thereof in connection with the accompanying drawings.

Brief Description of the Drawings

[0030] Preferred embodiments according to the present invention will now be described with reference to the Figures, in which like reference numerals denote like elements.

[0031] FIG 1 shows the 3D structure of a protein with forty amino acids with all the atoms present.

[0032] FIG 2 shows all of the forty Cα atoms that have been extracted from the same protein of FIG 1.

[0033] FIGS 3(a)-(c) illustrate the nearest neighbor method for generating DSPs in accordance with one embodiment of the present invention.

[0034] FIG 4 shows the exemplary addition of one virtual amino acid (B) between R and K in a discontinuous surface peptide (DSP) in accordance with one embodiment of the present invention.

[0035] FIG 5 shows an exemplary process of obtaining the discontinuous surface peptide database (DSPD) from the training data in accordance with one embodiment of the present invention.

[0036] FIG 6 shows an exemplary process of testing positive and negative test data in accordance with one embodiment of the present invention.

[0037] FIG 7 shows the plot of sensitivity & specificity vs threshold value.

[0038] FIG 8 shows an exemplary blosum80 chart used for similarity score calculation in the present invention.

Detailed Description of the Invention

[0039] The present invention may be understood more readily by reference to the following detailed description of certain embodiments of the invention.

[0040] Throughout this application, where publications are referenced, the disclosures of these publications are hereby incorporated by reference, in their entireties, into this application in order to more fully describe the state of art to which this invention pertains.

[0041] In the following detailed description, specific details are set forth in order to provide a thorough understanding of the invention. However, it will be understood by those skilled in the relevant art that the present invention may be practiced without these specific details. In other instances, well-known methods, procedures, components, and materials have not been described in detail so as not to obscure the present invention.

[0042] In one embodiment of the present invention, there is provided a process for prediction of protein allergenicity using discontinuous surface peptides and virtual amino acids. Briefly, the process comprises the steps of generating DSPs and adding virtual amino acid(s) into the gaps between two neighboring amino acids if the distances of the gaps are larger than predetermined threshold. In this way, the comparison of DSPs between an allergen and a query protein is done on the basis of linear amino acid sequences so that it dramatically reduces the requirement of computer power. The term "protein" throughout this document refers to any substance having a linear amino acid sequence, including polypeptide and proteins in their ordinary meanings in biochemistry. In addition, for short polypeptide, peptide may be used. The process will be described in detail herein.

[0043] The 3D structure data of proteins may come from multiple sources. For example, many public and private database of proteins' 3D structure are available. When one protein only has its amino acid sequence available, current software enables us to obtain its 3D structure. Even though the 3D structure created by software alone is not verified by actual experiments, it is good enough for the process of the present invention.

[0044] Referring now to FIG 1, there is provided the 3D structure of a protein with forty amino acids (DDGLCYEGTNCGKVGKYCCSPIGKYCVCYDSKAICNKNCT) where all of the atoms are displayed. As shown in FIG 1, it is apparent that when all atoms are considered in comparing the 3D structures of two proteins, the comparison would be extremely complex and require high computer power.

[0045] For each 3D file, the solvent accessible surface area for all the atoms is calculated according to any of the known methods. For example, the Naccess software can be used to calculate the solvent accessible surface area.

[0046] Referring to FIG 2, it presents all of the forty Cα atoms that have been extracted from the same protein of FIG 1. The extraction comprises obtaining the data of the 3D positions of the Cα atoms. Then, the extracted data of the 3D positions of the Cα atoms is saved into a file and used to represent the protein for further process.

[0047] Then, in combination with the solvent accessible surface data obtained from earlier calculation, extract the surface Cα atoms that have solvent accessible surface area lower than a predetermined value. For example, the predetermined value for the solvent accessible surface area is 1.0. The extracted surface Cα atoms will be used to generate the DSPs as described in detail below.

[0048] In order to solve the problem caused by the discontinuous nature of surface Cα atoms, one embodiment of the present invention provides a nearest neighbor method for generating discontinuous surface peptides (DSPs) from the surface Cα atoms with solvent accessible surface area less than a predetermined value. The nearest neighbor method will be described in detail hereinafter.

[0049] Now referring to FIGS 3(a)-(c), the nearest neighbor method is illustrated. First, choose one Cα atom to represent the first Cα atom of one discontinuous surface peptide (DSP); this can be done according to one's specification. It is noted that no surface Cα atom can be the first Cα atom in more than one DSP.

[0050] Then, find the nearest Cα atom to the first Cα atom; the nearest Cα atom will represent the second Cα atom of the DSP. Then, find the nearest Cα atom to the second Cα atom; the nearest Cα atom will represent the third Cα atom of the DSP. The process is repeated until the desired length of the DSP is achieved. It is noted that no Cα atom can appear twice in one DSP.

[0051] For a protein with n surface Cα atoms, it is possible to generate n DSPs with each Cα atom as the first Cα atom for each DSP. All DSPs form a DSP database.

[0052] Prior to the comparison of DSPs between allergens and query proteins, all DSPs are processed for possible addition of virtual amino acids. For each DSP, the distance between every two neighboring Cα atoms are calculated according the equation (1) below:

$$dis\tan ce = \sqrt{(x1-x2)^2+(y1-y2)^2+(z1-z2)^2} \qquad (1)$$

[0053]

[0054] where distance is the distance between the two neighboring Cα atoms; (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Cα atoms. If the distance is larger than the predetermined threshold, one or more virtual amino acids will be inserted between the two neighboring amino acids. As shown in FIG 4, one virtual amino acid (B) is inserted between two neighboring amino acids (R and K). When all the distances are calculated and all possible virtual amino acids are inserted, the DSPs will be translated into linear amino acid sequence information.

[0055] In one embodiment of the present invention, the predetermined threshold for inserting one or more virtual amino acids is obtained as follows. From an allergen data set, all of the neighboring Cα-Cα distances in all of the pdb files are calculated according to equation (1) described above; then the median value of all of the Cα-Cα distances is chosen to be the threshold. In the example below, the media value is 3.757A. Other values can also be used as the threshold. After the threshold is obtained, the insertion of virtual amino acids can be done in any suitable ways. For example, if the distance between two neighboring *Cα-Cα* is less than 2 times of the threshold, no virtual amino acid is inserted; if the distance is more than 2 times but less than 3 times of the threshold, one virtual amino acid is inserted; if the distance is more than 3 times but less than 4 times of the threshold, two virtual amino acids are inserted.

[0056] The virtual amino acid can be assigned any numerical value during similarity score calculation according to any specific situation. In one embodiment of the present invention, the virtual amino acid is treated as a gap and assigned with a constant gap penalty of -8 when the blosum80 chart as shown in FIG 8 is used for the similarity score calculation.

[0057] It is well known that proteins have different lengths of amino acid sequences, thus there are two concerns if the data of surface Cα atoms representing different proteins are not pre-processed. One is that different numbers of amino acids in two proteins will obviate the comparison; the other is that the similarity may be missed. For example, two proteins can perform one similar function (e.g., phosphoralation) but the first has 100 amino acids and the second 200 amino acids, and the function requires the whole amino acid sequence of the first protein but only the C-terminus of the second protein; if the comparison is done with the N-terminus 100 amino acids of the second protein, their apparent similarity

will be missed. Thus, after the surface Cα atoms of each protein have been extracted, the smallest number of surface Cα atoms can be used as the reference number; whereby the proteins with bigger numbers of surface Cα atoms will be processed into a set of discontinuous surface peptides having the reference number of surface Cα atoms. Then, the discontinuous surface peptides with the same number of surface Cα atoms can be compared.

[0058] The length of each DSP can be determined by specific situations. For example, the length of each DSP may be from about 6 to 40 amino acid. It is preferable that the length of the DSP is about 15-25 amino acids. Of course, the set of DSPs can be generated in any other possible ways that suits one's need.

[0059] The following examples are provided to illustrate the applications of the principles of the present invention. They are by no means intended to limit the scope of the present invention.

[0060] EXAMPLES

[0061] 575 known allergens (positive allergen data) and 654 control proteins (negative data) were obtained from the list from Bjorklund et al. (Bioinformatics 201 (1): 39-50 (2005)). The input files were pdb data. For each protein in the positive and negative test set, if the structure is unknown, it will be created by Fugue homological modeling. They are used in the five-fold cross-validation to show the prediction accuracy of the algorithm.

[0062] Now there is provided a more detailed description of the five-fold cross-validation procedure to perform allergen prediction. Five-fold cross-validation was used to verify the effectiveness of the allergen prediction.

[0063] 1. Dividing the positive allergen data into 5 sets, each containing 115 allergen sequences, where the first set of the positive allergen data was used as the positive test data while the remaining 4 sets were combined to form the train data, and the negative data were used as the negative test data.

[0064] 2. Obtaining discontinuous surface peptides (DSPs) from the train data

[0065] Referring now to FIG 5, there is provided an exemplary process of obtaining discontinuous surface peptide database (DSPD) from the training data in accordance with one embodiment of the present invention. The process **10** starts with acquiring a training data set **11**; then calculating the solvent accessible surface area for all atoms **12**; then

extracting Cα atoms **13**; then further extracting surface Cα atoms **14** where the surface Cα atoms have a solvent accessible surface area less than a predetermined value (e.g., <1.0); then finding discontinuous surface peptides (DSPs) **15** as described earlier; then for each DSP, calculating the distance between two neighboring Cα atoms as described earlier and inserting one or more virtual amino acids if the distance is larger than a predetermined threshold **16**; then translating all DSPs into linear amino acid sequences **17**; and finally cutting all DSPs into a predetermined length (e.g., 24 amino acids including any inserted virtual amino acid(s)) and storing all the processed DSPs into the DSPD **18**.

[0066] 3. Testing positive and negative test data

[0067] Referring now to FIG 6, there is provided an exemplary process of testing positive and negative data in accordance with one embodiment of the present invention. The process **20** starts with providing a test data set **21**; then calculating the solvent accessible surface area of all atoms **22**; then extracting all Cα atoms **23**; then extracting surface Cα atoms **24**; then finding discontinuous surface peptides **25**; then for each DSP, calculating the distance between two neighboring Cα atoms as described earlier and inserting one or more virtual amino acids if the distance is larger than a predetermined threshold **26**; then translating all DSPs into linear amino acid sequences **27**; cutting all DSPs into a predetermined length (e.g., 24 amino acids including any inserted virtual amino acid(s)) **28**; calculating the similarity scores of the DSPs obtained from the test protein with all the DSPs in the DSPD as described below **29**; and storing the maximum similarity score as the test protein's similarity score **30**.

[0068] The similarity score between two DSPs were defined according to below equation (2) as:

[0069] $$Similarity\ Score = \sum_{i=1}^{l} s(X_i Y_i)$$

[0070] where $s(X_i\ Y_i)$ is the substitution score between amino acid i in peptide X and the amino acid i in peptide Y , according to the substitution matrix employed. l is the length of the peptides.

[0071] As the similarity score, any suitable scoring matrices can be employed in the present invention. For example, both PAM and BLOSUM scoring matrices can be used (e.g., BLOSUM 80 or PAM 30 etc). For example, BLOSUM: 30, 35, 40, 45, 50, 55, 60, 62, 65, 70, 75, 80, 85, 90, 100; PAM: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150,160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300,

310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410,420, 430, 440, 450, 460, 470, 480, 490, and 500. One exemplary blosum80 chart is shown in FIG 8.

[0072] 4. Steps 1-3 were repeated 5 times, using a different set of data from the 5 sets of positive allergen data as the positive test data. The same set of negative data is reused.

[0073] 5. The maximum similarity scores from all the positive and negative test data in the earlier steps were combined and used to obtain an optimum threshold value that could give best sensitivity and specificity results. If the maximum similarity score of a protein is bigger than the threshold value, the protein is classified as an allergen.

[0074] The specificity and sensitivity were measured according to the following:

		Disease	
		+	-
Test Results	+	TP	FP
	-	FN	TN

TP = number of true positive Sensitivity =TP/(TP+FN)

FP = number of false positive

FN = number of false negative Specificity =TN/(TN+FP)

TN = number of true negative

[0075] Results

[0076] Table 1. The threshold values, sensitivity and specificity

Threshold	Sensitivity	Specificity
0	99.82609	0.614439
10	99.30435	3.533026
20	98.08696	9.953917
30	97.21739	14.65438
40	96.69565	19.7235
50	93.56522	31.95084
60	83.30435	66.14439
61	81.73913	69.0937
62	79.82609	73.33333
63	78.26087	76.52842
64	76.86957	79.84639
65	75.13043	82.5192
66	73.91304	83.96313
67	73.3913	87.46544
68	72.86957	88.90937

69	71.65217	90.16897
70	70.26087	91.58218
80	65.91304	97.45008
90	63.30435	98.4639
100	59.82609	98.95545
110	56	98.98618
120	51.47826	99.10906
130	47.82609	99.47773
140	44.86957	99.50845
150	38.95652	99.50845
160	33.73913	99.50845
170	27.47826	99.50845
180	25.21739	99.63134
190	21.56522	99.63134
200	15.82609	99.75422
210	9.913043	100

[0077] Table 1 shows the threshold values, sensitivity and specificity. FIG 7 shows the plot of sensitivity & specificity vs threshold value.

[0078] Another embodiment of the present invention provides a system for predicting allergenicity of proteins using the process described above. The system comprises a computer usable medium embedded with computer readable program coded means for performing the processes and storing programs and intermediate and final data.

[0079] The system may be any electronic device that has a computer readable medium embedded with computer programs. The electronic devices include PC, notebook, or the like.

[0080] While the present invention has been described with reference to particular embodiments, it will be understood that the embodiments are illustrative and that the invention scope is not so limited. Alternative embodiments of the present invention will become apparent to those having ordinary skill in the art to which the present invention pertains. Such alternate embodiments are considered to be encompassed within the spirit and scope of the present invention. Accordingly, the scope of the present invention is described by the appended claims and is supported by the foregoing description.

CLAIMS

What is claimed is:

1. A process for predicting the allergenicity of a query protein, said process comprising the steps of:

providing three-dimensional (3D) structure data of a plurality of known allergens;

calculating solvent accessible surface area for all atoms of each of the plurality of known allergens;

extracting 3D positions data for every Cα atom of each of the plurality of known allergens;

extracting 3D positions data for surface Cα atoms of each of the plurality of known allergens to represent its 3D structure; wherein the surface Cα atoms have a solvent accessible surface area value less than a predetermined threshold;

generating a set of discontinuous surface peptides (DSPs) for each of the plurality of known allergens, wherein the set of DSPs have the same number of amino acids;

calculating for each DSP the distance between two neighboring Cα atoms and inserting one or more virtual amino acids if the distance is larger than a predetermined threshold;

translating all DSPs into linear amino acid sequences including any inserted virtual amino acids;

finally cutting all DSPs into a predetermined length and storing all the processed DSPs into the discontinuous surface peptide database (DSPD);

manipulating the query protein by the above relevant steps to get a set of DSPs;

calculating the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD and storing the maximum similarity score as the query protein's similarity score; and

predicting the query protein as an allergen if the maximum similarity score of the query protein is bigger than a threshold value.

2. The process of claim 1, wherein the predetermined length of all DSPs is 6-40 amino acids.

3. The process of claim 2, wherein the predetermined length of all DSPs is 15-25 amino acids.

4. The process of claim 1, wherein in the step of generating a set of DSPs, the DSPs are generated by a nearest neighbor method.

5. The process of claim 1, wherein in the step of calculating for each DSP the distance between two neighboring Cα atoms, the calculation is performed according to:

$$dis\tan ce = \sqrt{(x1-x2)^2 + (y1-y2)^2 + (z1-z2)^2}$$

where distance is the distance between the two neighboring Cα atoms; (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Cα atoms.

6. The process of claim 1, wherein in the step of inserting one or more virtual amino acids, the threshold is determined by obtaining the median value of all Cα-Cα distances in all pdb files.

7. The process of claim 1, wherein in the step of calculating the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD, the calculation is performed according to:

$$Similarity\ Score = \sum_{i=1}^{l} s(X_i Y_i)$$

where $s(X_i\ Y_i)$ is the substitution score between amino acid i in peptide X and the amino acid i in peptide Y , according to the substitution matrix employed. l is the length of the peptides.

8. The process of claim 1, wherein in the step of predicting the query protein as an allergen if the maximum similarity score of the query protein is bigger than a threshold value, the threshold value is an optimum threshold value from all maximum similarity

scores from all positive and negative test data that could give best sensitivity and specificity results.

9. An article of manufacture, comprising:

at least one computer usable medium having computer readable program code means embodied therein for predicting the allergenicty of a query protein, the computer readable program code means in said article of manufacture comprising:

computer readable program code means for causing a computer to provide three-dimensional (3D) structure data of a plurality of known allergens;

computer readable program code means for causing a computer to calculate solvent accessible surface area for all atoms of each of the plurality of known allergens;

computer readable program code means for causing a computer to extract 3D positions data for every Cα atom of each of the plurality of known allergens;

computer readable program code means for causing a computer to extract 3D positions data for surface Cα atoms of each of the plurality of known allergens to represent its 3D structure; wherein the surface Cα atoms have a solvent accessible surface area value less than a predetermined threshold;

computer readable program code means for causing a computer to generate a set of discontinuous surface peptides (DSPs) for each of the plurality of known allergens, wherein the set of DSPs have the same number of amino acids;

computer readable program code means for causing a computer to calculate for each DSP the distance between two neighboring Cα atoms and insert one or more virtual amino acids if the distance is larger than a predetermined threshold;

computer readable program code means for causing a computer to translate all DSPs into linear amino acid sequences including any inserted virtual amino acids;

computer readable program code means for causing a computer to finally cut all DSPs into a predetermined length and store all the processed DSPs into the discontinuous surface peptide database (DSPD);

computer readable program code means for causing a computer to manipulate the query protein by the above relevant steps to get a set of DSPs;

computer readable program code means for causing a computer to calculate the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD and store the maximum similarity score as the query protein's similarity score; and

computer readable program code means for causing a computer to predict the query protein as an allergen if the maximum similarity score of the query protein is bigger than a threshold value.

10. The article of claim 9, wherein the predetermined length of all DSPs is 6-40 amino acids.

11. The article of claim 10, wherein the predetermined length of all DSPs is 15-25 amino acids.

12. The article of claim 9, wherein when generating a set of DSPs, the DSPs are generated by a nearest neighbor method.

13. The article of claim 9, wherein when calculating for each DSP the distance between two neighboring Cα atoms, the calculation is performed according to:

$$dis\tan ce = \sqrt{(x1-x2)^2+(y1-y2)^2+(z1-z2)^2}$$

where distance is the distance between the two neighboring Cα atoms; (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Cα atoms.

14. The article of claim 9, wherein when inserting one or more virtual amino acids, the threshold is determined by obtaining the median value of all Cα-Cα distances in all pdb files.

15. The article of claim 9, wherein when calculating the similarity scores of the DSPs of the query protein with all the DSPs in the DSPD, the calculation is performed according to:

$$Similarity\ Score = \sum_{i=1}^{l} s(X_i Y_i)$$

where $s(X_i\ Y_i)$ is the substitution score between amino acid i in peptide X and the amino acid i in peptide Y , according to the substitution matrix employed. l is the length of the peptides.

16. The article of claim 9, wherein when predicting the query protein as an allergen if the maximum similarity score of the query protein is bigger than a threshold value, the threshold value is an optimum threshold value from all maximum similarity scores from all positive and negative test data that could give best sensitivity and specificity results.