

PROCESS AND SYSTEM FOR ANALYZING THREE-DIMENSIONAL STRUCTURE SIMILARITIES OF PROTEINS AND IDENTIFYING THREE- DIMENSIONAL STRUCTURE MOTIFS OF PROTEINS AND THEIR APPLICATIONS

Field of the Invention

[0001] The present invention generally relates to bioinformatics, and more particularly to a process and system for analyzing three-dimensional (3D) structure similarities of proteins and identifying 3D structure motifs of proteins, and further to their applications including the prediction of the allergenicity of a query protein.

Background of the Invention

[0002] Proteins sharing similar functions usually have high homology of their amino acid sequences, at least in a portion of their amino acid sequences; it is assumed that similar amino acid sequences would have similar 3D structures, enabling different proteins to perform similar functions. A group of amino acid sequences with high homology is usually defined as a motif. For example, tyrosine kinases have a conserved motif for their kinase activity, even though different tyrosine kinases have different specificities in terms of substrates. Many amino acid sequence motifs are well known and very useful for predicting the functions of newly identified proteins. However, every functional protein has its own unique 3D structure; thus a 3D structure similarity between proteins would be more relevant to their similarity of functions.

[0003] The 3D structures of proteins are large, complex and constantly evolving; thus the comparison of the 3D structures of two proteins is more difficult than the alignment of the amino acid sequences. Certain approaches have been used for comparison of proteins' 3D structures. For example, Distance Matrix constructs a matrix of inter-atomic distance to represent the 3D structure of proteins; then the matrix is used for searching similarities between proteins. Furthermore, attempts have been made to reduce the complexity of proteins for comparison without introducing gross errors. For example,

one known method utilizes the coordinates of all Ca atoms to represent a protein's 3D structure, resulting in reduced complexity of proteins' 3D structures, performs the comparison of proteins' 3D structures using a set of linear Ca atoms for each protein, resulting in less requirements of computer power for performing the comparison of proteins' 3D structures. Nonetheless, all current approaches still require complex operations and high computer powers. Therefore, there is a need for a new scheme of searching 3D structure similarities between proteins.

Summary of the Invention

[0004] One embodiment of the present invention provides a process for analyzing three-dimensional (3D) structure similarities between a plurality of proteins. The method comprises the steps of providing 3D structure data of the plurality of proteins; calculating solvent accessible surface area for all atoms of each protein; extracting 3D positions data for every Ca atom of each protein; extracting 3D positions data for surface Ca atoms of each protein to represent its 3D structure; wherein the surface Ca atoms have a solvent accessible surface area value less than a predetermined threshold; reorienting the surface Ca atoms of each protein within a three dimension space so that all proteins share the same origin point to make them comparable; and calculating the SCaD value between two proteins; thereby the 3D structure similarities between proteins are classified on the basis of the SCaD value.

[0005] In another embodiment of the process for analyzing 3D structure similarities, after the step of reorienting, the process further comprises storing the reoriented proteins into a protein database for further use.

[0006] In another embodiment of the process for analyzing 3D structure similarities, the step of reorienting each protein in a three-dimension space comprises translating the protein so that the first surface Ca atom is positioned at the origin point; rotating the protein about y-axis and z-axis so that the second surface Ca atom is at the x-axis; and rotating the protein about the x-axis such that the third surface Ca atom is situated in the xy plane.

[0007] In another embodiment of the process for analyzing 3D structure similarities, in the step of calculating the SCaD value between two proteins, the SCaD values are calculated according to:

$$[0008] \quad SCaD = \sum_{i=1}^N dist_i$$

[0009] where N is the number of the amino acid in each protein and $dist_i$ is the euclidean distance between the i th Ca atom of the two proteins and is calculated according to:

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

[0011] where $(x1, y1, z1)$ and $(x2, y2, z2)$ are the position coordinates of the Ca atoms.

[0012] Another embodiment of the present invention provides a process for generating a three-dimensional (3D) structure motif from a plurality of proteins. The process comprises the steps of providing 3D structure data of the plurality of proteins; calculating solvent accessible surface area for all atoms of each protein; extracting 3D positions data for every Ca atom of each protein; extracting 3D positions data for surface Ca atoms of each protein to represent its 3D structure; wherein the surface Ca atoms have a solvent accessible surface area value less than a predetermined threshold; generating a set of discontinuous surface peptides (DSP) for each of the plurality of proteins, wherein the set of DSP have the same number of amino acids; reorienting each DSP within a three dimension space so that all DSPs share the same origin point to make them comparable; calculating the SCaD values between two DSPs; and classifying at least two DSPs into a 3D structure motif if their SCaD values are lower than a predetermined threshold; thereby the 3D structure motif families are generated.

[0013] In another embodiment of the process for generating a three-dimensional (3D) structure motif, after the step of reorienting, the process further comprises storing the reoriented DSPs into a DSP database for further use.

[0014] In another embodiment of the process for generating a three-dimensional (3D) structure motif, in the step of generating a set of DSPs for each of the plurality of proteins, each DSP has 6-40 amino acids, and preferably 15-25 amino acids.

[0015] In another embodiment of the process for generating a three-dimensional (3D) structure motif, in the step of generating a set of DSPs for each of the plurality of proteins, each DSP is generated by nearest neighbor method.

[0016] In another embodiment of the process for generating a three-dimensional (3D) structure motif, the step of reorienting each peptide in a three-dimension space comprises translating the DSP so that the first surface Ca atom is positioned at the origin point; rotating the DSP about y-axis and z-axis so that the second surface Ca atom is at the x-axis; rotating the DSP about the x-axis such that the third surface Ca atom is situated in the xy plane.

[0017] In another embodiment of the process for generating a three-dimensional (3D) structure motif, in the step of calculating the SCaD value between two DSPs, the SCaD values are calculated according to:

$$[0018] \quad SCaD = \sum_{i=1}^N dist_i$$

[0019] where N is the number of the amino acid in each DSP and $dist_i$ is the euclidean distance between the i th Ca atom of the two DSPs and is calculated according to:

$$[0020] \quad dist_i = \sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2}$$

[0021] where (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Ca atoms.

[0022] Another embodiment of the present invention provides a process for predicting the allergenicity of a query protein. The process comprises the steps of providing 3D structure data of the plurality of known allergens; calculating solvent accessible surface area for all atoms of each allergen; extracting 3D positions data for every Ca atom of each allergen; extracting 3D positions data for surface Ca atoms of each allergen to represent its 3D structure; wherein the surface Ca atoms have a solvent accessible surface area value less than a predetermined threshold; generating a set of discontinuous surface peptides for each of the plurality of known allergens, wherein the set of DSPs have the same number of amino acids; reorienting each DSP within a three dimension space so that all DSPs share the same origin point to make them comparable; calculating the SCaD value between two DSPs; clustering the DSPs into a 3D structure motif family if the minimum distance value between one member and the rest fall below a predetermined threshold; and manipulating the query protein by the above relevant steps to get a set of DSPs, so that each DSP of the query protein is measured against each member of the 3D structure motif family; wherein the lowest SCaD value among all the DSPs of one protein will become the protein SCaD value; thereby the query protein is predicted as

an allergen if its protein SCaD value with all members of the 3D structure motif family fall below a predetermined threshold.

[0023] In another embodiment of the process for predicting the allergenicity, after the step of reorienting, the process further comprises storing the reoriented DSPs into a DSP database for further use.

[0024] In another embodiment of the process for predicting the allergenicity, in the step of generating a set of DSPs, each DSP has 6-40 amino acids, preferably 15-25 amino acids.

[0025] In another embodiment of the process for predicting the allergenicity, in the step of generating a set of DSPs, the DSPs are generated by nearest neighbor method.

[0026] In another embodiment of the process for predicting the allergenicity, the step of reorienting each DSP in a three-dimension space comprises translating the DSP so that one end Ca atom is positioned at the origin point; rotating the DSP about y-axis and z-axis so that the second Ca atom is at the x-axis; and rotating the DSP about the x-axis such that the third Ca atom is situated in the xy plane.

[0027] In another embodiment of the process for predicting the allergenicity, in the step of calculating the SCaD value between two DSPs, the SCaD values are calculated according to:

[0028]
$$SCaD = \sum_{i=1}^N dist_i$$

[0029] where N is the number of the amino acid in each peptide and $dist_i$ is the euclidean distance between the ith Ca atom of the two DSPs and is calculated according to:

[0030]
$$dist_i = \sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2}$$

[0031] where (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Ca atoms.

[0032] 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 analyzing three-dimensional (3D) structure similarities between a plurality of proteins, the computer readable program code means in said article of manufacture comprising computer readable program code means for causing a computer to provide 3D structure data of the plurality of proteins; computer readable program code means for causing a computer to calculate solvent accessible

surface area for all atoms of each protein; computer readable program code means for causing a computer to extract 3D positions data for every Ca atom of each protein; computer readable program code means for causing a computer to extract 3D positions data for surface Ca atoms of each protein to represent its 3D structure; wherein the surface Ca atoms have a solvent accessible surface area value less than a predetermined threshold; computer readable program code means for causing a computer to reorient the surface Ca atoms of each protein within a three dimension space so that all proteins share the same origin point to make them comparable; and computer readable program code means for causing a computer to reorient each peptide within a three dimension space so that all peptides share the same origin point to make them comparable; and computer readable program code means for causing a computer to calculate the SCaD value between two proteins; thereby the 3D structure similarities between proteins are classified on the basis of the SCaD value.

[0033] One of advantages of the present invention is that no pre-alignment of amino acid sequences is required. In addition, the extracted Ca atoms make distance computation simple. Finally, the short peptides make wide range of proteins comparable.

[0034] 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

[0035] Preferred embodiments according to the present invention will now be described with reference to the Figures, in which like reference numerals denote like elements.

[0036] FIG 1 shows the 3D structure of a protein with forty amino acids with all the atoms present.

[0037] FIG 2 shows all of the forty Ca atoms that have been extracted from the same protein of FIG 1.

[0038] FIGS 3(a)-(d) show the sequential steps of reorientation in accordance with one embodiment of the present invention.

[0039] FIGS 4(a)-(c) illustrate the nearest neighbor method for generating DSPs in accordance with one embodiment of the present invention.

[0040] FIG 5 shows an exemplary process of obtaining 3D structure motifs from the training data in accordance with one embodiment of the present invention.

[0041] FIG 6 shows an exemplary process of prediction of allergenicity in accordance with one embodiment of the present invention.

[0042] FIG 7 shows the plot of sensitivity & specificity vs threshold value.

Detailed Description of the Invention

[0043] The present invention may be understood more readily by reference to the following detailed description of certain embodiments of the invention.

[0044] 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.

[0045] 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.

[0046] In one embodiment of the present invention, there is provided a process for analyzing 3D structure similarities of proteins. 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.

[0047] 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.

[0048] 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.

[0049] 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.

[0050] Referring to FIG 2, it presents all of the forty Ca atoms that have been extracted from the same protein of FIG 1. The extraction comprises obtaining the data of the 3D positions of the Ca atoms. Then, the extracted data of the 3D positions of the Ca atoms is saved into a file and used to represent the protein for further process.

[0051] Then, in combination with the solvent accessible surface data obtained from earlier calculation, extract the surface Ca 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. This operation will significantly reduce the number of Ca atoms to be included in the 3D similarity analysis, simplifying the analytic process.

[0052] In order to compare the extracted Ca atoms from different proteins, they must be reoriented first. In one embodiment, the first three Ca atoms at either end (N-terminus or C-terminus) of one protein are subjected to sequential operations during the reorientation. In one comparison, the three Ca atoms of all compared proteins must be from the same terminus. Referring to FIGS 3(a)-(d), the sequential operations of the first three Ca atoms are illustrated with six Ca atoms. As will be seen hereinafter, the length of the protein shown is chosen for its simplicity for illustrating the principles of the present invention. In fact, the process of the present invention is applicable to all proteins, regardless of their lengths and amino acid compositions.

[0053] As shown in FIG 3(a), it is highly unlikely that any of the Ca atoms of one protein is located at the origin, making it impossible to compare two proteins. The reorientation starts by translating the Ca atoms of a protein so that the first Ca atom is located at the origin as shown in FIG 3(b). Then, the protein undergoes y-axis and z-axis rotation such that the second Ca atom is at the x-axis as shown in FIG 3(b). Then, the protein is rotated about the x-axis such that the third Ca atom is situated in the xy plane as

shown in FIG 3(c). When the reorientation is done, the protein as shown in FIG 3(d) is stored in the protein database.

[0054] The 3D structure similarities among reoriented proteins can be analyzed in any suitable ways. In one embodiment of the present invention, the 3D structure similarity between two proteins is based on the SCaD value. A brief description of calculating the SCaD value is provided herein.

[0055] The SCaD value between two proteins is defined as the Sum of the corresponding Ca atoms euclidean Distances between the two proteins (SCaD). The euclidean distance between the *i*th Ca atom of the two proteins is given by the following equation (1):

[0056]

$$dist_i = \sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2} \quad (1)$$

[0057] where $dist_i$ is the distance between the *i*th Ca atoms; (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Ca atoms. Then the SCaD value between the two proteins is calculated by the following equation (2):

[0058]
$$SCaD = \sum_{i=1}^N dist_i \quad (2)$$

[0059] where N is the number of the amino acid in each protein.

[0060] By calculating the SCaD values between one reference protein and any other proteins, the 3D structure similarities between the reference protein and other proteins can be classified into different categories based on the SCaD values. For example, in line with the tree structures of the amino acid sequence homologies of proteins, a similar tree for 3D structures can be built.

[0061] It is well known that proteins have different lengths of amino acid sequences. In order to use the SCaD value to evaluate the 3D structure similarities of different proteins, there are two concerns if the data of surface Ca atoms representing different proteins are not pre-processed. One is that different numbers of amino acids in two proteins will obviate the SCaD value; the other is that the similarity may be missed. For example, two proteins can perform one similar function (e.g., phosphorylation) 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 Ca atoms of each protein have been extracted, the smallest number of surface Ca atoms can be used as the reference number; whereby the proteins with bigger numbers of surface Ca atoms will be processed into a set of discontinuous surface peptides having the reference number of surface Ca atoms. Then, the discontinuous surface peptides with the same number of surface Ca atoms can be compared according to the processes described above. In addition, the proteins with different numbers of surface Ca atoms can be compared using 3D structure motifs as described hereinafter.

[0062] Another aspect of the present invention is to provide a process for obtaining 3D structure motifs from various proteins.

[0063] As is well known, proteins vary in their sizes, i.e., each protein has different number of amino acids, making it complicated to analyze the 3D structure similarities among proteins on the basis of SCAD values. It is still true when surface Ca atoms are used to represent the 3D structures of proteins. The inventors of the present invention discovered that by keeping the 3D positions of each surface Ca atom of a protein, a set of discontinuous surface peptides (DSPs) can be created for each protein, where the DSPs have less surface Ca atoms than any protein in the analysis. Thus, proteins with different numbers of surface Ca atoms can be represented by a set of DSPs with the same number of surface Ca atoms, and it is easy to reorient each DSP and calculate the SCaD value between DSPs following the process described above.

[0064] As discussed above, the use of surface Ca atoms with solvent accessible surface area less than a predetermined value for 3D similarity analysis simplifies the comparison process, but the discontinuous nature of the surface Ca atoms creates a problem for generating the set of DSPs for establishing 3D motif family; this is because the surface Ca atoms with solvent accessible surface area less than a predetermined value are not a linear sequence anymore. In order to solve the problem caused by the discontinuous nature of surface Ca atoms, one embodiment of the present invention provides a nearest neighbor method for generating discontinuous surface peptide from the surface Ca atoms with solvent accessible surface area less than a predetermined value. The nearest neighbor method will be described in detail hereinafter.

[0065] Now referring to FIGS 4(a)-(c), the nearest neighbor method is illustrated. First, choose one Ca atom to represent the first Ca atom of one discontinuous surface peptide (DSP); this can be done according to one's specification. It is noted that no surface Ca atom can be the first Ca atom in more than one DSP.

[0066] Then, find the nearest Ca atom to the first Ca atom; the nearest Ca atom will represent the second Ca atom of the DSP. Then, find the nearest Ca atom to the second Ca atom; the nearest Ca atom will represent the third Ca atom of the DSP. The process is repeated until the desired length of the DSP is achieved. It is noted that no Ca atom can appear twice in one DSP.

[0067] For a protein with n surface Ca atoms, it is possible to generate n DSPs with each Ca atom as the first Ca atom for each DSP. Each DSP is reoriented so that they can be aligned and compared easily with other DSPs. The reorientation can be done according to the process described above. All reoriented DSPs form a DSP database.

[0068] 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. If a big 3D structure motif is desired, the DSP can be even longer. 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 suites one's need.

[0069] For a DSP database, the minimum SCaD value of each DSP against all the other DSPs is calculated. For each reference DSP in the DSP database, all other DSPs that are found with the SCaD value less than 120.0a (in the case when the DSP is 24 amino acids long) are classified to belong to the 3D motif family of the reference DSP. A 3D structure motif will be formed if there is at least one other DSP besides the reference DSP. If no family member is found, the DSP is discarded and will not be classified into any 3D structure motif family.

[0070] The 3D structure motif has many practical applications. It is assuming that if one unknown DSP can be classified as a member of one 3D structure motif family and the 3D structure motif family has a common functional feature, the unknown DSP is likely to have the same functional feature.

[0071] 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.

[0072] EXAMPLES

[0073] Allergens induce allergic responses including elicitation of immunoglobulin E (IgE) antibodies and causation of symptoms of allergy within an affected individual. Allergies affect approximately one-third of the human population and are the most common cause of chronic illness in developed countries. The current treatments of allergies include the administration into an infected individual of pharmaceutical chemicals such as 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.

[0074] The identification and validation of protein allergens becomes more acute nowadays because more and more transgenic proteins have been introduced into our food chains. Currently, there are three main aspects in the assessment of protein allergenic potential: immunogenicity, cross-reactivity and clinical symptoms. Immunogenicity is meant the immune response to a particular allergen, normally in the form of either an immunoglobulin E (IgE) antibody or T-cell response. While animal studies of immune responses provide the ultimate validation of an allergen, they are too expensive to be applied to every protein. Therefore, a reliable prediction of potential allergens is imperative for allocating the money to study the real potential allergens, rather than all proteins.

[0075] Attempts have been made to predict allergenicity of a query protein by its amino acid sequences. One example of the sequence similarity approach to allergen detection is based on the “eight amino acid 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 two proteins rarely occurs by chance, such an occurrence is considered to be a good indicator of potential allergenicity. In addition, according to the FAO/WHO guidelines for allergenicity evaluation of foods derived from biotechnology, a query protein is potentially allergenic if it either has an identity of at least six contiguous

amino acids (identity length $n=6$) or $>35\%$ sequence similarity over a window of 80 amino acids when compared with known allergens.

[0076] It is well known that IgE and T-cell responses are epitope-dependent. With the accumulation of hundreds of proteins as known allergens, it is reasonable to assume that a potential allergenic protein shares 3D structure similarities with the known allergens. Thus, the potential allergenicity of query proteins can be predicted by examining their 3D structure similarities with known allergens.

[0077] 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 Jackknife test to show the prediction accuracy of the algorithm.

[0078] Now there is provided a more detailed description of the Jackknife test procedure to perform allergen prediction. Jackknife test was used to verify the effectiveness of the allergen prediction.

[0079] 1. The positive allergen data was divided into 5 sets, each containing 115 allergen sequences.

[0080] 2. 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. The negative data were used as the negative test data.

[0081] 3. Obtaining 3D structure motifs from the train data

[0082] Referring now to FIG 5, there is provided an exemplary process of obtaining 3D structure motifs 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 Ca atoms **13**; then further extracting surface Ca atoms **14** where the surface Ca atoms have a solvent accessible surface area less than a predetermined value; then finding discontinuous surface peptides (DSPs) **15**; then reorienting the DSPs **16** where all the reoriented DSPs form a DSP database; then calculating ScaD values between any two DSPs **17**; and finally finding 3D motif families **18**. For example, all DSPs in one 3D motif family have a ScaD value less than 120a (in the case when the DSP is 24 amino acids long).

[0083] 4. Testing positive and negative test data

[0084] Referring now to FIG 6, there is provided an exemplary process of prediction of allergenicity 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 Ca atoms 23; then extracting surface Ca atoms 24; then finding discontinuous surface peptides 25; then reorienting the DSPs 26; then finding protein SCaD value against 3D motif families 27; and finally predicting the test data is an allergen if the protein SCaD value is less than a threshold 28.

[0085] 5. Steps 2-4 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.

[0086] 6. The protein SCaD values 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 a protein's SCaD value is below the threshold value, the protein is classified as an allergen.

[0087] The specificity and sensitivity were measured according to the following:

		Disease	
		+	-
Test Results	+	TP	FP
	-	FN	TN

$TP = \text{number of true positive}$ $\text{Sensitivity} = TP / (TP + FN)$

$FP = \text{number of false positive}$

$FN = \text{number of false negative}$ $\text{Specificity} = TN / (TN + FP)$

$TN = \text{number of true negative}$

[0088] Results

Threshold	Sensitivity	Specificity
0	0	100
0.1	11.13043	99.75422
0.2	21.21739	99.38556
0.3	22.08696	99.38556
0.4	24	99.35484
0.5	26.95652	99.35484
0.6	28.69565	99.35484
0.7	31.82609	99.35484
0.8	33.73913	99.35484

0.9	35.30435	99.35484
1	37.3913	99.35484
1.1	38.78261	99.35484
1.2	41.04348	99.35484
1.3	42.26087	99.35484
1.4	44.17391	99.23195
1.5	45.91304	99.10906
1.6	46.78261	99.07834
1.7	47.65217	98.92473
1.8	49.73913	98.80184
1.9	51.82609	98.43318
2	52.69565	98.43318
2.1	53.91304	98.06452
2.2	55.47826	97.88018
2.3	56.86957	97.51152
2.4	57.56522	96.89708
2.5	58.78261	95.97542
2.6	60.34783	95.54531
2.7	62.43478	93.76344
2.8	64.52174	91.12135
2.9	65.21739	88.90937
3	66.95652	86.91244
3.1	69.3913	84.20891
3.2	71.30435	82.45776
3.3	72.86957	78.77112
3.4	76	75.02304
3.5	76.17391	70.69124
3.6	78.26087	66.5745
3.7	80.17391	62.54992
3.8	82.26087	56.9278
3.9	84.86957	54.07066
4	85.56522	49.80031
4.1	86.95652	45.99078
4.2	88.69565	43.01075
4.3	89.56522	40.18433
4.4	89.91304	37.57296
4.5	90.6087	34.50077
4.6	90.78261	32.68817
4.7	91.47826	31.09063
4.8	91.82609	29.61598
4.9	92.34783	28.63287
5	92.86957	27.64977
5.1	93.04348	26.91244
5.2	93.21739	26.14439
5.3	93.3913	25.5914
5.4	93.56522	25.16129
5.5	93.56522	24.88479
5.6	93.56522	24.79263
5.7	93.56522	24.23963
5.8	93.56522	24.23963

5.9	93.73913	23.96313
6	93.73913	23.65591
6.1	93.91304	23.5023
6.2	94.08696	23.34869
6.3	94.08696	23.19508
6.4	94.08696	23.01075
6.5	94.08696	22.85714
6.6	94.26087	22.64209
6.7	94.26087	22.27343
6.8	94.26087	21.84332
6.9	94.26087	21.41321

[0089] Table 1 shows the threshold values, sensitivity and specificity. FIG 7 shows the plot of sensitivity & specificity vs threshold value.

[0090] Another embodiment of the present invention provides a system for analyzing 3D structure similarities of proteins, generating 3D structure motif families and predicting allergenicity of proteins. 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.

[0091] 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.

[0092] 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 analyzing three-dimensional (3D) structure similarities between a plurality of proteins, comprising the steps of:
 - providing 3D structure data of the plurality of proteins;
 - calculating solvent accessible surface area for all atoms of each protein;
 - extracting 3D positions data for every Ca atom of each protein;
 - extracting 3D positions data for surface Ca atoms of each protein to represent its 3D structure; wherein the surface Ca atoms have a solvent accessible surface area value less than a predetermined threshold;
 - reorienting the surface Ca atoms of each protein within a three dimension space so that all proteins share the same origin point to make them comparable; and
 - calculating the SCaD value between two proteins;
 - thereby the 3D structure similarities between proteins are classified on the basis of the SCaD value.
2. The process of claim 1, after the step of reorienting, further comprising:
 - storing the reoriented proteins into a protein database for further use.
3. The process of claim 1, wherein the step of reorienting each protein in a three-dimension space comprises:
 - translating the protein so that the first surface Ca atom is positioned at the origin point;
 - rotating the protein about y-axis and z-axis so that the second Ca atom is at the x-axis; and
 - rotating the protein about the x-axis such that the third Ca atom is situated in the xy plane.

4. The process of claim 1, wherein in the step of calculating the SCaD value between two proteins, the SCaD values are calculated according to:

$$SCaD = \sum_{i=1}^N dist_i$$

where N is the number of the amino acid in each protein and $dist_i$ is the euclidean distance between the i th Ca atom of the two proteins and is calculated according to:

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

where $(x1, y1, z1)$ and $(x2, y2, z2)$ are the position coordinates of the Ca atoms.

5. A process for generating a three-dimensional (3D) structure motif from a plurality of proteins, comprising the steps of:

providing 3D structure data of the plurality of proteins;

calculating solvent accessible surface area for all atoms of each protein;

extracting 3D positions data for every Ca atom of each protein;

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

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

reorienting each DSP within a three dimension space so that all DSPs share the same origin point to make them comparable;

calculating the SCaD values between two DSPs; and

classifying at least two DSPs into a 3D structure motif if their SCaD values are lower than a predetermined threshold;

thereby the 3D structure motif families are generated.

6. The process of claim 5, after the step of reorienting, further comprising:

storing the reoriented DSPs into a DSP database for further use.

7. The process of claim 5, wherein in the step of generating a set of DSPs for each of the plurality of proteins, each DSP has 6-40 amino acids.

8. The process of claim 7, wherein each DSP has 15-25 amino acids.
9. The process of claim 5, wherein in the step of generating a set of DSPs for each of the plurality of proteins, each DSP is generated by nearest neighbor method.
10. The process of claim 5, wherein the step of reorienting each peptide in a three-dimension space comprises:
 - translating the DSP so that the first surface Ca atom is positioned at the origin point;
 - rotating the DSP about y-axis and z-axis so that the second surface Ca atom is at the x-axis; and
 - rotating the DSP about the x-axis such that the third surface Ca atom is situated in the xy plane.
11. The process of claim 5, wherein in the step of calculating the SCaD value between two DSPs, the SCaD values are calculated according to:

$$SCaD = \sum_{i=1}^N dist_i$$

where N is the number of the amino acid in each DSP and $dist_i$ is the euclidean distance between the i th Ca atom of the two DSPs and is calculated according to:

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

where $(x1, y1, z1)$ and $(x2, y2, z2)$ are the position coordinates of the Ca atoms.

12. A process for predicting the allergenicity of a query protein; comprising the steps of:
 - providing 3D structure data of the plurality of known allergens;
 - calculating solvent accessible surface area for all atoms of each allergen;
 - extracting 3D positions data for every Ca atom of each allergen;
 - extracting 3D positions data for surface Ca atoms of each allergen to represent its 3D structure; wherein the surface Ca atoms have a solvent accessible surface area value less than a predetermined threshold;

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

reorienting each DSP within a three dimension space so that all DSPs share the same origin point to make them comparable;

calculating the SCaD value between two DSPs;

clustering the DSPs into a 3D structure motif family if the minimum distance value between one member and the rest fall below a predetermined threshold; and

manipulating the query protein by the above relevant steps to get a set of DSPs, so that each DSP of the query protein is measured against each member of the 3D structure motif family; wherein the lowest SCaD value among all the DSPs of one protein will become the protein SCaD value;

thereby the query protein is predicted as an allergen if its protein SCaD value with all members of the 3D structure motif family fall below a predetermined threshold.

13. The process of claim 12, after the step of reorienting, further comprising:
storing the reoriented DSPs into a DSP database for further use.
14. The process of claim 12, wherein in the step of generating a set of DSPs, each DSP has 6-40 amino acids.
15. The process of claim 14, wherein each DSP has 15-25 amino acids.
16. The process of claim 12, wherein in the step of generating a set of DSPs, the DSPs are generated by nearest neighbor method.
17. The process of claim 12, wherein the step of reorienting each DSP in a three-dimension space comprises:
translating the DSP so that one end Ca atom is positioned at the origin point;
rotating the DSP about y-axis and z-axis so that the second Ca atom is at the x-axis;
and
rotating the DSP about the x-axis such that the third Ca atom is situated in the xy plane.

18. The process of claim 12, wherein in the step of calculating the SCaD value between two DSPs, the SCaD values are calculated according to:

$$SCaD = \sum_{i=1}^N dist_i$$

where N is the number of the amino acid in each peptide and $dist_i$ is the euclidean distance between the i th Ca atom of the two DSPs and is calculated according to:

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

where (x1,y1,z1) and (x2,y2,z2) are the position coordinates of the Ca atoms.

19. An article of manufacture, comprising:

at least one computer usable medium having computer readable program code means embodied therein for analyzing three-dimensional (3D) structure similarities between a plurality of proteins, the computer readable program code means in said article of manufacture comprising:

computer readable program code means for causing a computer to provide 3D structure data of the plurality of proteins;

computer readable program code means for causing a computer to calculate solvent accessible surface area for all atoms of each protein;

computer readable program code means for causing a computer to extract 3D positions data for every Ca atom of each protein;

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

computer readable program code means for causing a computer to reorient the surface Ca atoms of each protein within a three dimension space so that all proteins share the same origin point to make them comparable; and

computer readable program code means for causing a computer to reorient each peptide within a three dimension space so that all peptides share the same origin point to make them comparable; and

computer readable program code means for causing a computer to calculate the SCaD value between two proteins;

thereby the 3D structure similarities between proteins are classified on the basis of the SCaD value.