METHOD AND SYSTEM FOR PREDICTION OF ALLERGENICITY OF PROTEINS USING MOTIF TABLE

Field of the Invention

[0001] The present invention generally relates to bioinformatics, and more particularly to a method and system that uses motif table to predict the allergenicity of a query protein.

Background of the Invention

[0002] 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.

[0003] 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.

16 MAR 2007

[0004] 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 sequence similarities with the known allergens. Thus, the potential allergenicity of query proteins can be predicted by examining their sequence similarities with known allergens.

[0005] Attempts have been made to predict allergenicity of a query protein by its amino acid sequences. 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. See, FAO/WHO: Allergenicity of Genetically Modified Foods. 2001 [http://www.who.int/foodsafety/publications/biotech/en/ec_jan2001.pdf]. Rome, Italy, FAO/WHO. However, these guidelines do not have much practical use because they predict more than two-thirds of proteins as allergens. See, Stadler and Stadler, The FASEB Journal, 17: 1141-3 (2003). Stadler and Stadler tried to improve the prediction efficiency by generating a minimal set of sequence motifs from the known allergens and predicting a novel protein as a potential allergen if the novel protein contains at least one sequence motif generated from the known allergens. While the improvement over the FAO/WHO guidelines is shown, their prediction is still not very useful (with ~1 true allergen in 10 predicted allergens. In a further improvement attempt, Li et al. clustered the sequences first, predicted conserved motifs for each of the clusters, built a profile for each predicted motif, and predicted the allergenicity of a query. See, Li et al., Bioinformatics, 20:2572-2578 (2004). While it improved the accuracy over the Stadler's method, it has its own shortcomings: too complex, and impractical in clustering allergens when the number is small and they are diversified.

Summary of the Invention

[0006] Therefore, there is an imperative need of a method and system that can predict the allergenicity of a query protein with high specificity with easy implementation. The present invention satisfies this need by providing a method and system that uses a

motif table to predict the allergenicity of a query protein. The method and system can be easily implemented.

[0007] In one aspect, the present invention provides a method of prediction of the allergenicity of a query protein. The method comprises the steps of providing a set of amino acid sequences of known allergens, identifying a pre-set number of motifs from the amino acid sequences of known allergens, constructing a motif table of the identified motifs, searching for any identified motifs within the query protein with unknown allergenicity, and predicting that the query protein is an allergen if it contains at least one positive match with the motif table. In various embodiments, the set of amino acid sequences of known allergens may be obtained from public resources or private databases; and the motifs may be identified by any known algorithms that can identify motifs in multiple amino acid sequences including expectation maximization and Gibbs Sampler.

[0008] In one embodiment, in the step of constructing the motif table, when the motif table comprises 1, 2, 3, ...n-1, n columns, and 1, 2, 3, ...m-1, m rows, wherein the n and m are integers, and n may be equal to m, and wherein the number of motifs may be equal to the numbers of columns and rows, all cells in the motif table are initially set as 0, and any cell is updated by adding 1 each time by the following rules: lining all motifs (i.e., x, y, z) found in one allergen in sequence, wherein x, y, z may be any number assigned to the motifs; finding all possible combinations of all motifs (i.e., xy, xz, yz); updating any cell by 1 every time when the motif combination matches the combination of the column and row numbers, wherein the first number of the motif combination may be assigned to be column number or row number, and when the first number of the motif combination is assigned, the second number of the motif combination must be assigned to the remaining.

[0009] In another embodiment, in the step of searching for any identified motifs within the query protein with unknown allergenicity, the searching may be done by any known alignment algorithm; wherein the alignment algorithm is the Mast algorithm.

[0010] In another aspect, the present invention provides a motif table for being used to predict allergenicity of a query protein. The motif table comprises at least two parameters forming a two-dimensional table; wherein the two-dimensional table comprises 1, 2, 3, ...n-1, n columns, and 1, 2, 3, ...m-1, m rows, wherein the n and m are integers, and n may be equal to m; wherein the number of motifs may be equal to the numbers of columns and rows; and wherein all cells in the motif table are initially set as 0, and any cell

is updated by adding 1 each time by the following rules: lining all motifs (i.e., x, y, z) found in one allergen in sequence, wherein x, y, z may be any number assigned to the motifs; finding all possible combinations of all motifs (i.e., xy, xz, yz); and updating any cell by 1 every time when the motif combination matches the combination of the column and row numbers, wherein the first number of the motif combination may be assigned to be column number or row number, and when the first number of the motif combination is assigned, the second number of the motif combination must be assigned to the remaining.

[0011] In yet another aspect, the present invention provides an article of manufacture comprising: at least one computer usable medium having computer readable program code means embodied therein for predicting allergenicity 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 a set of amino acid sequences of known allergens; computer readable program code means for causing a computer to identify a pre-set number of motifs from the amino acid sequences of known allergens; computer readable program code means for causing a computer to construct a motif table of the identified motifs; computer readable program code means for causing a computer to search for any identified motifs within the query protein with unknown allergenicity; and computer readable program code means for causing a computer to predict that the query protein is an allergen if it contains at least one positive match with the motif table. In various embodiments, the set of amino acid sequences of known allergens may be obtained from public resources or private databases; and the motifs may be identified by any known algorithms that can identify motifs in multiple amino acid sequences including expectation maximization and Gibbs Sampler.

[0012] In one embodiment, in the step of constructing the motif table, when the motif table comprises 1, 2, 3, ...n-1, n columns, and 1, 2, 3, ...m-1, m rows, wherein the n and m are integers, and n may be equal to m, and wherein the number of motifs may be equal to the numbers of columns and rows, all cells in the motif table are initially set as 0, and any cell is updated by adding 1 each time by the following rules: lining all motifs (i.e., x, y, z) found in one allergen in sequence, wherein x, y, z may be any number assigned to the motifs; finding all possible combinations of all motifs (i.e., xy, xz, yz); updating any cell by 1 every time when the motif combination matches the combination of the column and row numbers, wherein the first number of the motif combination may be assigned to be

column number or row number, and when the first number of the motif combination is assigned, the second number of the motif combination must be assigned to the remaining.

[0013] In another embodiment, in the step of searching for any identified motifs within the query protein with unknown allergenicity, the searching may be done by any known alignment algorithm; wherein the alignment algorithm is the Mast algorithm.

[0014] 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

[0015] Preferred embodiments according to the present invention will now be described with reference to the Figures, in which like reference numerals denote like elements.

[0016] FIGS 1a-1d illustrate the construction of a motif table in accordance with the present invention.

[0017] FIG 2 shows a plot of sensitivity vs motif length for different number of motifs.

[0018] FIG 3 shows a plot of specificity vs motif length for different number of motifs.

Detailed Description of the Invention

[0019] The present invention may be understood more readily by reference to the following detailed description of certain embodiments of the invention.

[0020] 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.

[0021] 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.

[0022] As discussed above, an allergen can elicit allergic immunoresponses within an individual because the allergen contains one or more epitopes, where the epitopes can be recognized by specific T- and B- cells. So far, the allergenic epitopes have been assumed as short contiguous amino acid sequences (motifs) in all the available methods of prediction of allergenicity of a query protein. This assumption may be the reason why the current algorithms for allergenicity prediction are dissatisfactory.

[0023] Allergenic immunoresponses involve both T- and B-cells. T cells are usually considered to recognize only contiguous epitopes that are formed by linear amino acid sequences, while B cells are usually considered to recognize conformational epitopes that can be formed by either linear amino acid sequences (contiguous epitopes) or amino acid sequences (discontiguous epitopes) that are not sequentially contiguous but brought close together in space by the folding of a polypeptide chain. The fact that the specific epitopes are formed by amino acid sequences provides the foundation for amino acid sequence-based prediction of allergenicity of novel proteins. However, simply using the occurrence of one linear epitope (motif) in a query protein to predict the protein's allergenicity has been investigated but found that it generates too high rate of false positives to be satisfactory. The present invention provides that the prediction of a query protein's allergenicity is drastically improved using a motif table. Without wish to be bound by any specific or particular theories, the inventors of the present invention reason that the motif table makes it possible to identify discontiguous epitopes contained in the query proteins so that the prediction in accordance with the present invention has a high success rate.

[0024] As used herein, a "motif" in a sequence context refers to a conserved pattern of amino acids that is found in two or more proteins. Motifs are represented as position-dependent scoring matrices that describe the score of each possible letter at each position in the pattern. Examples of sequence motif databases are the Prosite catalog (http://www.expasy.ch/prosite) and the Stanford Motifs Database (http://dna.stanford.edu/emotif/). In the Prosite catalog, a motif is an amino acid pattern that is found in a group of proteins that have similar biochemical activity, and that often is near the active site of the protein.

[0025] As used herein, a "motif table" means that motifs identified in a set of known protein allergens are arranged in a column and row manner. For example, as discussed in detail hereinafter, both the column and row represent the motifs. It is to be appreciated that the columns and rows are not necessarily equal and may have other parameters than the motifs.

[0026] As used herein, a "query protein" includes any polypeptide or protein that has an amino acid sequences, where the amino acids may be any of the natural ones or modified ones known to those skilled in the art.

[0027] The present invention provides a method for prediction of the allergenicity of a query protein using a motif table. The present invention also provides a system for executing the method in accordance with the principles of the present invention, where the system comprises a computer usable medium having computer readable program code means embodied there for predicting the allergenicity of a query protein. The computer readable program code means include means for causing a computer to procure a training set of known protein allergens; means for causing a computer to generate a profile of motifs from the known protein allergens; means for causing a computer to generate a motif table; and means for causing a computer to predict the allergenicity of a query protein using the motif table. The computer readable program code means further include means for causing a computer to store the original, intermediate, and final data, and input data; means for causing a computer to output the final data; and means for causing a computer to be trained with various known allergens. The computer usable medium can be a PC, a computer, a PDA or a notebook. The computer readable program code means may be any computer program that can perform the designated functions.

[0028] Prior to a more detailed description, the method of the present invention may be outlined briefly as follows: providing a set of known allergens; generating a set of motifs from the set of known allergens; generating a motif table based on the generated profile of motifs; and predicting the allergenicity of a query protein using the motif table. While specific terms and samples will be used to illustrate the principles of the present invention, they are in no means intended to limit the scope of the invention.

[0029] High-quality repositories of amino acid sequences of proteinaceous allergens and non-allergen counterparts provide a prerequisite for the generation of motif sets. Amino acid sequences of allergens and non-allergens may be obtained from a variety

16 MAR 2007

of sources including private and public databases, published articles, and in-house development. The public databases include WHO-IUIS allergen list (www.allergen.org/list.htm), Swiss-Prot's Allergen Index (www.expasy.org/cgi-bin/lists?allergen.txt), BIFS (www.iit.edu/~sgendel/fa.htm), FARRP (www.allergenonline.com), the Allergen Database (www.csl.gov.uk/allergen), the Protall Database (www.ifrn.bbsrc.ac.uk/protall/).

[0030] The amino acid sequences of known allergens and non-allergens may be refined before they are used to generate the motif profile. For example, the leading signals may be deleted because they are usually not present in the mature proteins. The leading signals may generate some undesirable motifs if they are used to generate the motif profile. In addition, the amino acid sequences from different sources may be cross-checked for any errors arising from sequencing and/or data entry.

[0031] After a set of allergens is chosen, the motif set for the chosen allergens may be generated by any of current available algorithms that are capable of discovering motifs in a group of related sequences. The most commonly used algorithms are Multiple Expectation Maximization for Motif elicitation (MEME) and Gibbs Sampler. Both methods have been described in details in the literature so that both methods will only be discussed in brief. See, Bailey and Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, (1994); Bailey and Gribskov, "Combining evidence using p-values: application to sequence homology searches" Bioinformatics, Vol. 14, pp. 48-54, (1998).

[0032] MEME as a tool for discovering motifs in a group of related sequences is based on Expectation Maximization (EM) algorithm. The EM algorithm is used to identify conserved areas in unaligned DNA and proteins: assuming that a set of sequences is expected to have a common sequence pattern; making an initial guess as to location and size of the site of interest in each of the sequences; loosely aligning the interest sites; and providing from this alignment an estimate of base or aa composition of each column in the site. The EM algorithm consists of the two steps, which are repeated consecutively. Step 1, the expectation step, the column-by-column composition of the site is used to estimate the probability of finding the site at any position in each of the sequences. These probabilities are used to provide new information as to expected base or aa distribution for

each column in the site. Step 2, the maximization step, the new counts for bases or aa for each position in the site found in the step 1 are substituted for the previous set.

[0033] MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern (a profile). Individual MEME motifs do not contain gaps. MEME takes as input a group of DNA or protein sequences (the training set) and outputs as many motifs as requested. MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif.

[0034] The Gibbs sampler algorithm is slightly different from the EM approach. The Gibbs sampler also searches for the statistically most probable motifs and can find the optimal width and the number of motifs in each sequence. It iterates through two steps. In the first step a random start position for the motif is chosen for all sequences but for one. These sequences are then aligned and used to find an initial guess of the motif. The objective of the next step is to find the most probable pattern common to left out sequence (and on the next iterations to all of the sequences) by sliding them back and forth until the ratio of the motif probability to the background probability is a maximum. Then the next sequence is left out and the process is repeated until the residue frequencies in each motif do not change. The number of iterations may range from several hundred to several thousand.

[0035] After a pre-set number of motifs are generated, a motif table will be generated. The motif table may be designed according to the designer's rationales. In one embodiment, the columns and rows of a motif table represent all the motifs, and the cells within the table are updated each time when there is a combination of two motifs in one allergen.

[0036] For prediction of the allergenicity of a query protein using the constructed motif table, the query protein is first analyzed for any and every possible matches with the identified motifs from the known allergens. This pair alignment between the query protein and each of the motifs may be done by any sequence alignment algorithms. The common one is the Motif Alignment and Search Tool (MAST). MAST is a tool for searching biological sequence databases for sequences that contain one or more of a group of known motifs. MAST takes as input a file containing the descriptions of one or more motifs (generally the MEME output) and searches a query protein that match the motifs. The

16 MAR 2007

motif file can be the output of the MEME motif discovery tool or any file in the appropriate format. MAST outputs three things: The names of the high-scoring sequences sorted by the strength of the combined match of the sequence to all of the motifs in the group; Motif diagrams showing the order and spacing of the motifs within each matching sequence; Detailed annotation of each matching sequence showing the sequence and the locations and strengths of matches to the motifs.

[0037] After the motifs from the query protein are identified, they are compared with the motif table. If any combination of the motifs from the query protein matches a non-zero score in the corresponding cell of the motif table, the query protein is predicted as an allergen.

[0038] Examples

[0039] The following examples are provided for the sole purpose of illustration of the principles of the present invention. It by no means intends to limit the scope of the present invention.

[0040] A. Outline of the procedure of performing allergen prediction

[0041] 1. The positive allergen data is divided into 5 sets, each containing 115 allergen sequences.

[0042] 2. The first set of data is used as the positive test data while the remaining 4 sets are combined to form the train data. The protein sequences in the positive test data are randomly shuffled to form the negative test data.

[0043] 3. The Meme software program is used to find a predetermined number of motifs of fixed length w from the train data. The motifs are extracted from Meme result.

[0044] 4. The Meme result from step 3 is used to form a motif table as follows: (a) all cells in the motif table, MT, are initialized to 0; (b) for a sequence containing 2 motifs, m1, m2, cell[m1][m2] is increased by one; if a sequence contains 3 motifs: m1, m2, m3, cell[m1][m2] is increased by one; cell[m1][m3] is increased by one; cell[m2][m3] is increased by one; (c) in the same manner, the motif table is populated until the last sequence is processed.

[0045] 5. The test data and the results from the Meme program in step 3 are used as input for the Mast program to find similar motifs in the test data;

[0046] 6. From the Mast result in step 5, motifs from each protein sequence are tested to check if they are present in the motif table. The protein sequence is considered to be allergen if at least one entry is found in the motif table.

[0047] 7. Step 2 to step 6 are repeated with the second data set as test data and the rest of the data constituting the train data. The procedure is repeated until the last data set is used as the test set.

[0048] B. Construction of an exemplary motif table

[0049] A simple example will be used to illustrate the construction of the Motif table. Table 1 shows an exemplary set of five (5) motifs discovered in three hypothetical training proteins (P1, P2 and P3) by the Meme program from the train data.

[0050] Table 1. Example of the motifs found in proteins sequences in the train data

Protein	Motifs		
P1	1	3	
P2	2	4	5
P3	4	4	5

[0051] The Motif table is updated in the following manners:

[0052] 1. Initially, all the cells in the Motif table are set to 0 (FIG 1a).

[0053] 2. From protein sequence P1, cell [1,3] is set to 1 (FIG 1b);

[0054] 3. From protein sequence P2, cells [2,4], [2,5], [4,5] are set to 1 (FIG 1c);

[0055] 4. From protein sequence P3, cells [4,4] is set to 1, [4,5] is set to 3 (two counts of [4,5]) (FIG 1d).

[0056] C. Prediction of allergenicity of exemplary query proteins

[0057] Table 2 shows the motifs identified from the hypothetical test data with three proteins (P4, P5 and P6).

Protein	Motifs		
P4	2	5	
P5	1	1	3
P6	1	4	

[0058] Whether a test protein is an allergen is predicted by comparing the data in the motif table with the motifs discovered by the Mast program from the test data. The prediction for protein sequences P4, P5, and P6 are performed as follows:

[0059] 1. For P4 with motifs 2 and 5, cell[2,5] contains a value 1, P4 is predicted to be allergen;

[0060] 2. For P5 with motifs 1, 1, and 3, cell [1,1] contains 0 but cell[1,3] contains 1, P5 is predicted to be allergen;

[0061] 3. For P6 with motifs 1 and 4, cell[1,4] contains 0, P6 is predicted to be non-allergen.

[0062] D. Validation of the prediction of allergenicity using a motif table

[0063] 1. Data generation

[0064] The positive data set was obtained from the allergen list published in Bjorklund et al., Supervised Identification Of Allergen-Representative Peptides For In Silico Detection Of Potentially Allergenic Proteins, Bioinformatics, 21: 39-50 (2005); the negative data set was obtained by shuffling the positive allergen sequence randomly. There are 578 protein sequences each in the positive and negative data sets.

[0065] 2. Results

[0066] The motifs were identified using MEME and MAST. See, e.g., http://meme.sdsc.edu/meme/website/intro.html. The sensitivity results for different motif lengths are given in Table 3, where the sensitivity is defined as the percentage of true positive over the summation of true positive and false negative. For comparison, the prediction algorithm is repeated with randomly shuffled allergen sequences and the specificity results are shown in Table 4, where the specificity is defined as the percentage of true negative over the summation of true negative and false positive. Based on the results below, it can be seen that the best sensitivity result of 0.758 is obtained (25 amino acid length, 300 motifs). The corresponding specificity prediction result is 0.923. A plot of sensitivity versus motif length for different number of motifs is shown in Figure 2 while the plot for specificity versus motif length is shown in Figure 3.

[0067] Table 3. Sensitivity results for different motif lengths

	5 aa length	15 aa length	25 aa length	35 aa length
100 motifs	0.372174	0.6	0.648696	0.633043
200 motifs	0.335652	0.697391	0.735652	0.707826
300 motifs	0.311304	0.704348	0.758261	0.753043
400 motifs	0.318261	0.723478	0.747826	0.754783

[0068] Table 4. Specificity results for different motif lengths

	5 aa length	15 aa length	25 aa length	35 aa length

100 motifs	0.937391	0.961739	0.913043	0.921739
200 motifs	0.911304	0.94087	0.937391	0.946087
300 motifs	0.89913	0.923478	0.923478	0.93913
400 motifs	0.890435	0.895652	0.930435	0.961739

[0069] 3. Comparison of predicted results with experimental results

[0070] The method and system of the present invention has been validated with its applications to allergens with known epitopes. The experimental data was obtained from the SDAP database (http://fermi.utmb.edu/SDAP/). In the case of the Par j 1 allergen, experimental data shows that the allergen contains 4 discontinuous epitopes. The proposed algorithm predicted 4 motifs which cover the epitope regions of the experimental data using parameters of motif length = 25, motif number =200. Similar result is obtained using parameters of motif length = 25, motif number =300. The predicted and the experimental results are shown in Tables 6 and 7.

[0071] Table 6: Comparing experimental data of Par j 1 allergen with predicted epitopes using parameters motif length = 25 bp, motif number = 200

[0072]

Predicted Result (pos 1:70)	QETCGTMVRALMPCLPFVQGKEKEPSKGCCSGAKRLDGETKTGPQRVHACECIQTAMKTYSDIDGKLVSE
Expt Data (pos 1:70)	QETCGTMVRALMPCLPFVQGKEKEPSKGCCSGAKRLDGETKTGPQRVHACECIQTAMKTYSDIDGKLVSE
Predicted Result (71:139)	VPKHCGIVDSKLPPIDVNMDCKTVGVVPRQPQLPVSLRHGPVTGPSDPAHKARLERPQIRVPPPAPEKA
Expt Data (71:139)	VPKHCGIVDSKLPPIDVNMDCKTVGVVPRQPQLPVSLRHGPVTGPSDPAHKARLERPQIRVPPPAPEKA

[0073] Table 7: Comparing experimental data of Par j 1 allergen with predicted epitopes using parameters motif length = 25 bp, motif number = 300

Predicted Result (pos 1:70)	QETCGTMVRALMPCLPFVQGKEKEPSKGCCSGAKRLDGETKTGPQRVHACECIQTAMKTYSDIDGKLVSE
Expt Data (pos 1:70)	QETCGTMVRALMPCLPFVQGKEKEPSKGCCSGAKRLDGETKTGPQRVHACECIQTAMKTYSDIDGKLVSE
Predicted Result (71:139)	VPKHCGIVDSKLPPIDVNMDCKTVGVVPRQPQLPVSLRHGPVTGPSDPAHKARLERPQIRVPPPAPEKA
Expt Data (71:139)	VPKHCGIVDSKLPPIDVNMDCKTVGVVPRQPQLPVSLRHGPVTGPSDPAHKARLERPQIRVPPPAPEKA

[0074] For the allergen Jun a 3, experimental data shows that it contains 4 epitopes. Using parameters of motif length=25bp, motif number =200, the algorithm predicted 2 motifs which coincides with 3 of the 4 epitopes. Using parameters of motif length=25bp, motif number =300, the algorithm predicted 5 motifs which covers most of the regions of

the 3 epitopes. However, the first and the fifth motifs are false positives. The predicted and the experimental results are shown in Tables 8 and 9.

[0075] Table 8: Comparing experimental data of Jun a 3 allergen with predicted epitopes using parameters motif length = 25 bp, motif number = 200

Predicted Result (pos 1:70)	MARVSELAFLLAATLAISLHMQEAGVVKFDIKNQCGYTVWAAGLPGGGKRLDQGQTWTVNLAAGTASARF
Expt Data (pos 1:70)	MARVSELAFLLAATLAISLHMQEAGVVKFDIKNQCGYTVWAAGLPGGGKRLDQGQTWTVNLAAGTASARF
Predicted Result (71:140)	WGRTGCTFDASGKGSCQTGDCGGQLSCTVSGAVPATLAEYTQSDQDYYDVSLVDGFNIPLAINPTNAQCT
Expt Data (71:140)	WGRTGCTFDASGKGSCQTGDCGGQLSCTVSGAVPATLAEYTQSDQDYYDVSLVDGFNIPLAINPTNAQCT
Predicted Result (1411:210)	APACKADINAVCPSELKVDGGCNSACNVFKTDQYCCRNAYVDNCPATNYSKIFKNQCPQAYSYAKDDTAT
Expt Data (141:210)	APACKADINAVCPSELKVDGGCNSACNVFKTDQYCCRNAYVDNCPATNYSKIFKNQCPQAYSYAKDDTAT
Predicted Result (71:139)	FACASGTDYSIVFCP
Expt Data (71:139)	FACASGTDYSIVFCP

[0076] Table 9: Comparing experimental data of Jun a 3 allergen with predicted epitopes using parameters motif length = 25 bp, motif number = 300

Predicted Result (pos 1:70)	MARVSELAFLLAATLAISLHMQEAGVVKFDIKNQCGYTVWAAGLPGGGKRLDQGQTWTVNLAAGTASARF
Expt Data (pos 1:70)	MARVSELAFLLAATLAISLHMQEAGVVKFDIKNQCGYTVWAAGLPGGGKRLDQGQTWTVNLAAGTASARF
Predicted Result (71:140)	WGRTGCTFDASGKGSCQTGDCGGQLSCTVSGAVPATLAEYTQSDQDYYDVSLVDGFNIPLAINPTNAQCT
Expt Data (71:140)	WGRTGCTFDASGKGSCQTGDCGGQLSCTVSGAVPATLAEYTQSDQDYYDVSLVDGFNIPLAINPTNAQCT
Predicted Result (1411:210)	APACKADINAVCPSELKVDGGCNSACNVFKTDQYCCRNAYVDNCPATNYSKIFKNQCPQAYSYAKDDTAT
Expt Data (141:210)	APACKADINAVCPSELKVDGGCNSACNVFKTDQYCCRNAYVDNCPATNYSKIFKNQCPQAYSYAKDDTAT
Predicted Result (71:139)	FACASGTDYSIVFCP
Expt Data (71:139)	FACASGTDYSIVFCP

[0077] In summary, the motifs predicted by the algorithm have been found to agree with real experimental epitopes data for both allergens Par j1 and jun a 3.

[0078] 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 method for predicting the allergenicity of a query protein, comprising the steps of:

providing a set of amino acid sequences of known allergens;

identifying a pre-set number of motifs from the amino acid sequences of known allergens;

constructing a motif table of the identified motifs;

searching for any identified motifs within the query protein; and

predicting that the query protein is an allergen if the query protein contains at least one positive match with the motif table.

2. The method of claim 1, wherein, in the step of providing the amino acid sequences of known allergens, the set of amino acid sequences of known allergens may be obtained from public resources or private databases.

3. The method of claim 1, wherein, in the step of identifying motifs, the motifs may be identified by any known algorithms that can identify motifs in multiple amino acid sequences.

4. The method of claim 3, wherein the motif identification algorithms include expectation maximization and Gibbs Sampler.

5. The method of claim 1, wherein, in the step of constructing the motif table, when the motif table comprises 1, 2, 3, ...n-1, n columns, and 1, 2, 3, ...m-1, m rows, wherein the n and m are integers, and n may be equal to m, and wherein the number of motifs may be equal to the numbers of columns and rows, all cells in the motif table are initially set as 0, and any cell is updated by adding 1 each time by the following rules:

16 MAR 2007

lining all motifs (e.g., x, y, z) found in one allergen in sequence, wherein x, y, z may be any number assigned to the motifs;

finding all possible combinations of all motifs (e.g., xy, xz, yz);

updating any cell by 1 every time when the motif combination matches the combination of the column and row numbers, wherein the first number of the motif combination may be assigned to be column number or row number, and when the first number of the motif combination is assigned, the second number of the motif combination must be assigned to the remaining.

6. The method of claim 1, wherein, in the step of searching for any identified motifs within the query protein, the searching may be done by any known alignment algorithm.

7. The method of claim 6, wherein the alignment algorithm is the Mast algorithm.

8. A motif table for being used to predict allergenicity of a query protein, comprising at least two motifs forming a two-dimensional table; wherein the two-dimensional table comprises 1, 2, 3, ...n-1, n columns, and 1, 2, 3, ...m-1, m rows, wherein the n and m are integers, and n may be equal to m; wherein the number of motifs may be equal to the numbers of columns and rows; and wherein all cells in the motif table are initially set as 0, and any cell is updated by adding 1 each time by the following rules:

lining all motifs (e.g., x, y, z) found in one allergen in sequence, wherein x, y, z may be any number assigned to the motifs;

finding all possible combinations of all motifs (e.g., xy, xz, yz);

updating any cell by 1 every time when the motif combination matches the combination of the column and row numbers, wherein the first number of the motif combination may be assigned to be column number or row number, and when the first number of the motif combination is assigned, the second number of the motif combination must be assigned to the remaining.

9. An article of manufacture, comprising:

at least one computer usable medium having computer readable program code means embodied therein for predicting allergenicity 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 a set of amino acid sequences of known allergens;

computer readable program code means for causing a computer to identify a pre-set number of motifs from the amino acid sequences of known allergens;

computer readable program code means for causing a computer to construct a motif table of the identified motifs;

computer readable program code means for causing a computer to search for any identified motifs within the query protein; and

computer readable program code means for causing a computer to predict that the query protein is an allergen if the query protein contains at least one positive match with the motif table.

10. The article of manufacture of claim 9, wherein the set of amino acid sequences of known allergens may be obtained from public resources or private databases.

11. The article of manufacture of claim 9, wherein the motifs may be identified by any known algorithms that can identify motifs in multiple amino acid sequences.

12. The article of manufacture of claim 11, wherein the motif identification algorithms include expectation maximization and Gibbs Sampling.

13. The article of manufacture of claim 9, wherein the motif table comprises 1, 2, 3, ...n-1, n columns, and 1, 2, 3, ...m-1, m rows, wherein the n and m are integers, and n may be equal to m; wherein the number of motifs may be equal to the numbers of columns and rows; and wherein all cells in the motif table are initially set as 0, and any cell is updated by adding 1 each time by the following rules:

lining all motifs (e.g., x, y, z) found in one allergen in sequence, wherein x, y, z may be any number assigned to the motifs;

finding all possible combinations of all motifs (e.g., xy, xz, yz);

16 MAR 2007

updating any cell by 1 every time when the motif combination matches the combination of the column and row numbers, wherein the first number of the motif combination may be assigned to be column number or row number, and when the first number of the motif combination is assigned, the second number of the motif combination must be assigned to the remaining.

14. The article of manufacture of claim 9, wherein the searching of motifs in the query protein may be done by any known alignment algorithm.

15. The article of manufacture of claim 14, wherein the alignment algorithm is the Mast algorithm.