Identification of Unique Peptide Motifs as Linear epitopes with Pfam Domains and Families

Wei-Jun Zhung, Chih-Hong Liu, and Tun-Wen Pai*

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan, Republic of China.
*twp@mail.ntou.edu.tw

ABSTRACT

To identify the unique peptide motifs (UPMs) as specific linear epitopes from all family protein sequences, we have developed a database called linear epitope prediction database (LEPD) which applies reinforced merging techniques, background model analysis, and chemical property analysis to interpret its specific antigenicity. The UPMs extracted from more than 8,000 families defined by Pfam are regarded as linear epitopes providing significant information for the designs of antibodies and vaccines. They can be verified as epitope candidates according to their biological properties and compositions of the continuous stretch of amino acid residues. These potential epitopes predicted from each protein family can be revealed in a straightforward manner from the proposed database, and the corresponding chemical properties of each predicted epitopes are illustrated in graphical and tabular forms. Each identified UPM can be analyzed by scanning through the complete genomes of various organisms for guaranteeing the specificity of antigenic peptides. For any query protein possessing resolved three-dimensional structure, the contents in the proposed database also provide interactive visualization of protein structures for the allocation and comparison of the predicted linear epitopes. In terms of mapping a UPM as a linear epitope, the accuracy of the algorithm for linear epitope prediction is evaluated to be higher than 70% in comparison with known databases.

1: INTRODUCTIONS

1.1: Motivation

An epitope is defined as the sites of the antigen protein which is responsible for the specificity of the antigen in the Antigen-Antibody reactions. The epitope prediction is one of the most interesting issues in molecular immunology since the specialized immune protein, called antibody, is produced due to the introduction of an antigen into the body and possessing the remarkable ability to combine with the specific antigen that triggered the major function of the immune system. There are two major types of epitope: one is in a contiguous stretch of amino acid residues that are linked by peptide bonds and called as linear epitopes (LE), and the other type is in a non-contiguous format that are constructed by folding of polypeptide chain and called as conformational epitopes (CE). The characteristic of linear epitopes is determined by the sequential amino acids of a protein and the specificity of conformational epitopes depends on the spatial folding of the contributing individual sequential epitopes. In this paper, we will focus on the LE prediction and the predicted epitopes are computed in advance and collected into a database for further applications.

With respect to the variety from a set of protein family sequences, we found that several unique segments in a family are frequently located on the structure of loop. The reshaping feature of loop structure can be considered as a flexile signature of its specific functionality. Furthermore, the flexibility of loop structures certainly reflects the amino acids possessing high possibility of interacting with other protein structures. Therefore, the specific segments on loop structures can be considered as the high potential candidates for epitopes. A protein sequence family is usually classified by its chemical properties, functionalities, structures, and ancestors, etc. It is usually composed of several members with highly homologous sequences and similar biological functions. Up to now, the Pfam [1] is a well recognized protein family database which collects a large set of protein families covering many common protein domains and families. This database will be the main corpus and reference of our linear epitope predicting algorithms, and it will be introduced in the next section.

1.2: R_EMUS System

To extract the unique peptide motifs from a set of sequences, we employ the R_EMUS system to provide the major functions. The R_EMUS (**RE**inforced **M**erging techniques for **U**nique peptide **S**egments) is a web server for identification of the locations and compositions of unique peptide segments from a set of protein family sequences [2][3]. Different levels of uniqueness are determined according to substitutional relationship in the amino acids, frequency of appearance, and biological properties such as priority for serving as candidates for epitopes where antibodies recognize. The algorithm employs a sequence-based searching algorithm consisting of a three-phase operation including clustering, searching, and merging phases.

In the clustering phase, the module classifies 20 amino acids into different groups based on specified BLOSUM/PAM series of matrices or simply customized according to user's preference. The searching phase performs exact or approximate string matching to extract fundamental unique peptide segments, named as primary pattern. The length of a primary pattern is an important parameter for appropriate unique peptide extraction and strongly influences the final results. The rule of thumb for primary pattern lengths is that a shorter length setting for similar sequences and a longer length setting for dissimilar sequences. The searching module performs Boyer Moore algorithm to efficiently retrieve all primary patterns based on previous clustering definitions. Each searched fundamental unique peptide segment will be analyzed based on its frequencies of appearance and its representation level of uniqueness is calculated for the subsequent merging processes. In the last phase, merging algorithms initiate a bottom-up assembling to extract unique peptide segments. There are four different merging methods designed to combine the primary unique peptide segments and result in proper subset relationships that reflect the precision of their unique features. The R_EMUS web server is available at http://biotools.cs.ntou.edu.tw/REMUS. In this paper, we adopt the R_EMUS algorithm to extract the unique peptide motifs automatically according to the defined Pfam protein families.

1.3: Pfam

Pfam is a database of two parts, the first is the curated part of Pfam containing over 8296 protein families, and the updated version 20.0 is employed in this paper. Each family is manually curated and is represented by two multiple sequence alignments, two profile-Hidden Markov Models (profile-HMMs) and an annotation file. There are two types of multiple sequence alignment provided from Pfam including seed alignment and full alignment. The seed alignment is a multiple alignment of a representative set of sequences which is verified manually, and the full alignment is created from the HMM-profile by searching the sequences database for all detectable members. All these distinct members will be collected and aligned to the Hmm-profile. In the proposed LEPD database, we collect all sequences from both seed alignment and full alignment of each families, but the pre-performed UPM extraction are only focused on the seed part of Pfam families.

2: IMPLEMENTATION

2.1: Platform and Used Language

This database is implemented by PHP language and MySQL database server on the platform of Windows 2003 server version and IIS 6.0 http server. The web interface is designed employing Java script and dynamic html. The 3D display module applies the Jmol system (*http://jmol. sourceforge.net*) from Open Source software suite, athe. The web server is IBM xSerice 336, with one HyperThreading Technology CPU Xeon 3.0GHz, 2G bytes ECC DDR400 memory, and one 80G bytes SCSI harddrive.

2.2: Building Flow Chart

The flow chart of constructing the LEPD database is depicted in the Figure1. We build sequence families database by entering Pfam SQL script. The Pfam ftp offers SQL scripts to build Pfam sequence database. The current release version of Pfam is 20.0 containing 8296 families. After construction the database, we employ R_EMUS to search the unique segments from the classified families. The R_EMUS is registered as a COM service on the server. This allows us to use any language that supports COM techniques to convoke its services. Then we use PHP script to fetch sequences from the built Pfam database and apply R_EMUS functions to extract UPMs. Finally, the searched peptide motifs are aggregated to formulate the LEPD database. The user query interface is designed as web service version which locates at *http://biotools.cs.ntou.edu.tw/LEPD_Search.php*. This is written by dhtml and java script. The designed web query page allows users to browse Pfam family entries, the predicted linear epitopes, chemical properties in graphical representation, and interactive 3D structural displays in a very friendly interface.

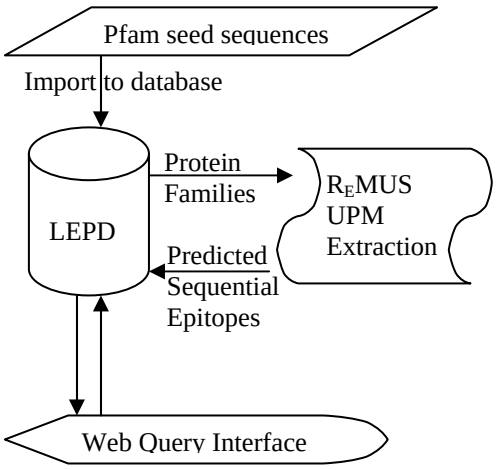

Figure 1. The flow chart for LEPD construction.

2.3: Mapping with PDB IDs

The Pfam sequences employ HMM and Blast tools to map similar PDB sequences. A Pfam sequence may coincide with multiple PDB IDs, and vice versa. In order to view the location of predicted linear epitopes from the coincidental 3D structures and to construct the possible conformational epitopes, the corresponding PDB files are required resources. In addition to the 3D visualization problems, the predicted linear epitopes in

LEPD are retrieved residue sites from Pfam family sequences, not from the sequence contents in its corresponding PDB file. In fact, the searched residues from Pfam family sequences are not exactly the same as the sequence contents in PDB files. Therefore, the corresponding matching and analysis is necessary for appropriate display for 3D visualization. Here we employ a simple sliding window algorithm to search the designated UPMs from its corresponding PDB file. The algorithm employed here allows one miss match due to the assumption that PDB sequence may contain missing sites. With the scanning processes, the list of found segments will be utilized for further 3D display and CE search.

2.4: Chemical Properties Scoring

To obtain the corresponding graphical representation of the chemical properties of predicted epitopes, a specified protein sequence is scanned within a sliding window of a specified size defined by users. According to various types and scales of chemical properties of antigenicity, we assign different values and coefficient sets of corresponding chemical properties to each amino acid in the sliding window. The system calculates the mean value of the specified chemical property of the amino acids within the sliding window, and plots the figure according to the calculated value at the midpoint of the window[4]. The results of chemical property graphics will be discussed in the Section 4.

3: Usage of the LEPD

3.1: Search a Specific Family from Pfam

To query a specific family from the defined Pfam families, we can click on the "Alphabetical Indices of Protein Families" on the main searching webpage to find the appropriate results. For example, if we want to find the RnaseA family, we can simply click the capital "R" hyperlink directly. Then all Pfam sequence families with the leading character "R" will be displayed. The listed table for any query is ordered alphabetically for searched Pfam entries. If the inquiry family is found in the table, a user is able to click the hyperlink of "group name" field or "UPM/family number" field directly, he/she can browse the predicted linear epitopes for further analysis. The interface of the alphabetical indices of pre-defined protein families and an example of searched partial families are shown in the Figure 2. When an interested protein family is selected, the predicted linear epitopes are listed in a table according to the priority of antigenicity. The field of "site" represents the start location of the linear epitope in a specified sequence, and there are three graphic icons shown in the most right field representing the chemical property graph of the corresponding segments, the corresponding protein structure, and the advanced conformation epitope formulation for the selected linear epitope respectively.

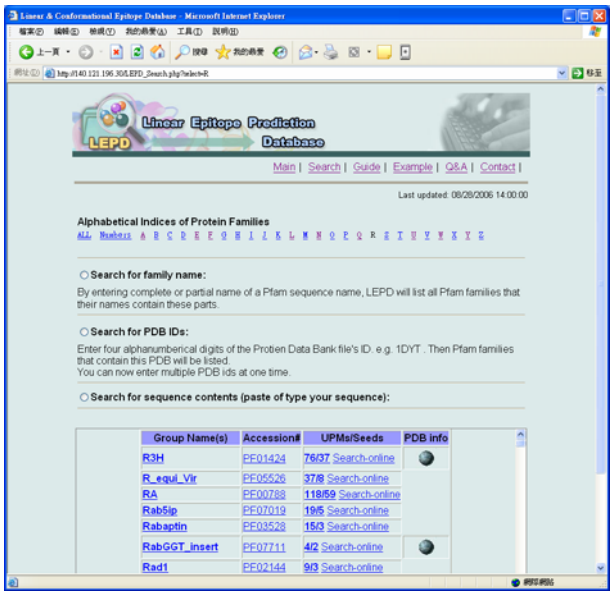

Figure 2. The LEPD interface for the alphabetical indices of protein families.

3.2 Display Chemical Properties Graphics

The first graphic icon connection provides the graphical representation of chemical properties of the selected segment including hydrophilicity, hydrophobicity, surface accessible, polarity, secondary structure, and other properties. There are various combinations of displaying the chemical properties of the selected UPMs that are optional to the users, such as only indicating the hydrophilicity and surface accessible properties at the same time. Furthermore, user can decide to show the chemical properties of the predicted epitopes one at a time or all at once. In this graph, the chemical score for entire sequence will be clearly revealed and the selected linear epitope will be highlighted in blue color for enhancing the contrast in graphics. The original sequence and all predicted epitopes in the corresponding positions are also displayed at the beginning of the webpage.

3.3 Display 3D Structure Viewer & CE Searching

On the interface of showing UPMs of a specified protein family, some of the entries may possess two extra icons. That means there existing corresponded PDB files for the sequence. By clicking on these icons, the 3D structure viewer and CE searching function will be displayed and initiated. The 3D viewer system is similar to the R_EMUS 3D structure display function except that this viewer shows only one structure at a time for the corresponding sequence. For the CE searching function, there are two parameters to be set in advance: one is the thresholding value of surface rate in percentage, and the other is the parameter of neighboring distance for the criteria of grouping multiple segments as a conformational epitope. The

CE prediction function can be considered as a completed program, but the performance is still under evaluation. Here we will only focus on the prediction of LE and demonstrate the efficiencies and effectivness of the database. Figure 3 shows an example of 3D visualization of predicted linear epitopes. In this example, there are 4 predicted linear epitopes from the Angi_Bovin/26-143 sequence in RNaseA protein family and users are able to select the specified epitopes and observe its structural position and properties.

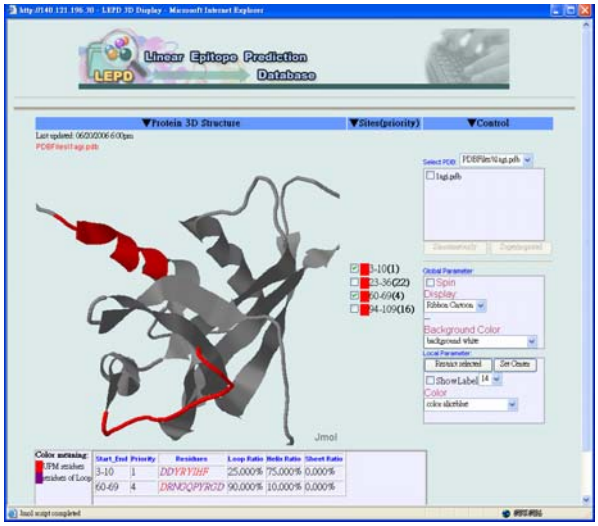

Figure 3. An example of showing 3D structural visualization of predicted epitopes.

4: EVALUATION

4.1: Evaluation On Chemical Properties

Antigenic peptides possess trivial characteristics of some chemical properties from previous experiments and reports [5~8]. For example, hydrophilicity, accessibility, secondary structure: beta-turn, etc. Therefore, we supply the graphics of these chemical properties of overall sequence and highlight the corresponding positions of predicted linear epitopes for further experiments. In these chemical properties calculation, several scales and weighting coefficient sets that produced by experts can be individually selected by users, and length of peptide chain can be specified as well. Excepting the chemical properties discussed before, the system also supports some other chemical properties. For example, there are Polaritym, Hydrophpbicity, data, and other properties. In addition to the expected chemical properties, the system also supplies some viewpoints from other properties. The chemical properties employed here can be referred to the website of Prot Scale of ExPASY (*http://www.expasy.org/tools/protscale.html*). In Figure 4, we select the P53 protein family to show a complete example of predicted epitopes. In this case, the highest rank of antigenicity of the predicted epitope is selected to show its chemical properties. From the figure, we can observe that the predicted epitope is a unique peptide motif possessing high scores of antigenicity, surface accessibility. Polarity, and low scales of huydrophobicity.

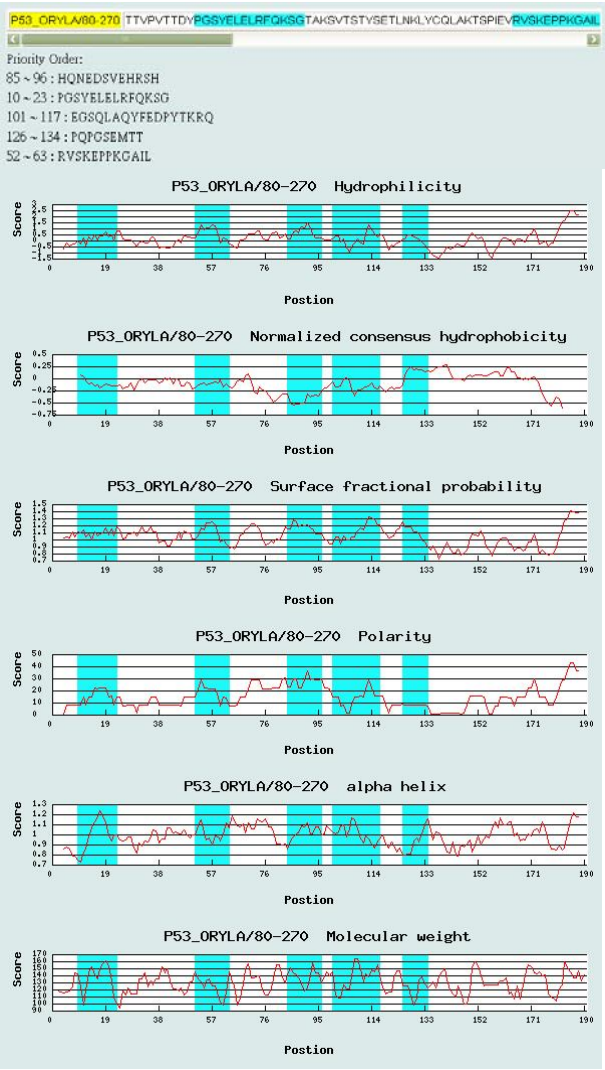

Figure 4. An example of P53 protein family and the chemical properties of the predicted epitope with the highest rank of antigenicity from the P53_ORYLA (80-270) sequence.

4.2: Performance of Linear Epitope Prediction

The linear epitopes can be validated by a variety of biochemical and biophysical methods such as Western blotting, Enzyme-Linked Immunosorbent Assay (ELISA), immunopricipitation, and X-ray crystallography, and the predicted results can further provide researchers valuable information on peptide antigen design and structural determination of heterocomplexes formed by specific antigen-antibody interactions. In this paper, the performance of the bottom-up merging algorithms for prediction of linear epitopes is evaluated by adopting the epitopes of human monoclonal antibodies retrieved from the website of Santa Cruz Biotechnology, Inc.(http://www.scbt.com/) which focuses on the ongoing development of research

antibodies. Among 8398 entries in the database, 83 monoclonal antibodies with specified epitope length less than 300 amino acid residues are generated against human antigens, which are classified into 63 human protein families containing a total number of 264 protein sequences. Each set of the family protein sequences is collected from GenBank and analyzed by the bottom-up merging algorithms. It is found that 275 UPMs are located within the antibody-antigen recognition sites of 66 monoclonal antibodies, indicating that the average accuracy of matching at least one of the UPM with the reported epitopes of the selected antibodies is 79.52% (66/83). As the lengths of the detected epitopes are decreased from 300, 200, to 100 amino acid residues, the accuracy of correlating a UPM with an epitope decreased from 94.12% (24/34), 81.25% (26/32), to 70.59% (16/17), respectively. Our results reveal that the accuracy is length-dependent and an overall accuracy higher than 70% is successfully achieved.

To further evaluate the specificity of the searched UPMs, we provide the background model analysis by scanning the identified UPMs over the completed genome species. Since all UPMs listed in our database are derived from discriminating the common sequence stretches during multiple sequence comparison of closely related protein families, it is expected that the identified UPMs are quite rare in the rest areas of the genome. Taking the RNase A-like families as an example, there are 30 seed sequences clustered in the family set and 47 UPMs are allocated by default settings. As expected, when each predicted epitope is scanned through the completed genome provided by NCBI database (ftp://ftp.ncbi.nih.gov/genomes/), only 6 out of 47 candidates are found to appear in more than one species. This advanced analysis provides an additional filter for improving the specificity of antigenic epitope selection. Taken together, we have developed a novel database for prediction of linear epitopes located in all protein families defined by Pfam. Antigens designed based on the identified UPMs have the advantages in differentiating one member of a large protein family due to the nature of specific molecular recognition.

REFERENCE

[1] Chang HT, Pai TW, Fan T, Su BH, Wu PC, Tang CY, Chang CT, Liu SH, and Chang M.D. (2006), A reinforced merging methodology for mapping unique peptide motifs in members of protein families, BMC Bioinformatics, 7:38.

[2] Pai, T.W., Chang, M.D.T., Tzou, W.S., Su, S.H., Wu, P.C., Chang, H.T., and Chou, W.I. (2006), REMUS: a tool for identification of unique peptide segments as epitopes, Nucleic Acids Res. 34, Web Server Issue. W198-201.

[3] Robert D. Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer and Alex Bateman(2006), Pfam: clans, web tools and services, Nucleic Acids Res., Database Issue 34:D247-D251.

[4] Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2005), Protein Identification and Analysis Tools on the ExPASy Server;(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press , pp. 571-607.

[5] Hopp TP, Woods KR. (1981), Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci, USA 1981; 78:3824±8.

[6] Janin J. (1979), Surface and inside volumes in globular proteins. Nature 277:491±2.

[7] Levitt, M. (1998). Normalized frequency for beta turn. Biochemistry 17, 4277-4285.

[8] Alix, A.J. (1999) Predictive estimation of protein linear epitope by using the program PEOPLE, Vaccine, 18, pp. 311-314.