The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Prediction of Conformational Epitopes by Geometrical Affinity and Chemical Propensity Approaches

Cheng-Ying Tang[1] Wei-Kuo Wu[1] Yu-Ping Hsu[1] Hsin-Wei Wang[1]
Tun-Wen Pai[1,1] Hao-Teng Chang[2]

[1] Department of Computer Science and Engineering, National Taiwan Ocean University, No. 2, Pei-Ning Rd., Keelung, Taiwan 20224, R. O. C.
[2] Graduate Institute of Molecular Systems Biomedicine, China Medical University, No. 91, Hsueh-Shih Rd., Taichung, Taiwan 40402, R. O. C.

Abstract A conformational epitope is composed of several discontinuous segments as antigenic determinants which are spatially close to each other in the three dimensional structure. These segments form the antigen which may bind with a specific receptor of the immune system, and play an important role in vaccine designs and immuno-biological experiments. Though there are two major types of epitopes: linear and conformational epitopes, it has been estimated that more than 90% of B-cell epitopes depend on nonsequential amino acids and are geometrically clustered due to molecular folding. Therefore, prediction of conformational epitopes rather than linear ones becomes an important and challenging task for practical applications. In this paper, a novel conformational epitope prediction system was developed based on the characteristics of surface rate analysis of side chain atoms, distribution of surface curvature attribute, and physicochemical propensity of each surface residue. It is the first conformational epitope prediction system based on the combinatorial features of curvatures and surface rates of side chain atoms. In this paper, benchmark datasets were employed to train the optimal parameter settings, and thirty extra antigen-antibody complexes from three different data resources with verified conformational epitopes were adopted to evaluate the prediction accuracy. Comparing with those well-developed tools, our proposed method outperformed the others in both aspects of accuracy and efficiency. For this testing dataset, the proposed system achieved an average sensitivity of 39.4%, an average specificity of 91.2%, and an average AUC value of 0.69.

Keywords conformational epitope, side chain surface rate, surface curvature, physicochemical propensity

1 Introduction

B-cell epitopes, also known as antigenic determinants, are defined as a part of an

[1] twp@mail.ntou.edu.tw

antigen which is able to bind with either a specific antibody molecule or a particular B-cell receptor to elicit humoral immune response (HIR) [1]. The main purpose of predicting B-cell epitopes is to facilitate the synthetic peptide design that can replace an antigen in vaccine design to reduce injuries for researchers or experimental animals [2]. It is categorized into two major types: linear type, that contains a short continuous stretch of amino acid residues, and conformational type, that comprises several discontinuous peptides but close in three-dimensional structure.

In previous work, several prediction tools have focused on linear epitope prediction and widely provided, such as LEPD [3], BEPITOPE [4], and BEPIPRED [5], *etc*.. However, the number of continuous epitopes (linear epitopes) on native proteins had been estimated only 10% from all B-cell epiotpes in past surveys [6]. Most of B-cell epitopes are recognized and adopted to form a native conformation as a result of conformational epitopes. Therefore, to identify discontinuous epitopes becomes a practical and important task for performing synthetic peptide design, developing recombinant vaccines, and running specific diagnostic tests.

Only a few predictors have been designed for identifying discontinuous epitope sites in recent years. The conformational epitope predictor (CEP) is one of the first methods for identification of discontinuous epitopes which only adopted the attributes of geometrical information of protein structures for conformational epitope prediction [7]. Discotope predictor developed by Haste Andersen *et al*. applied amino acid statistics and structural surface information to obtain possible epitope sites [8]. PEPOP is another predicting method which identified fundamental segments composed of continuous surface accessible residues and clustered these segments according to their spatial vicinity for enumerating putative epitope candidates [9]. Finally, the newly proposed ElliPro utilized the 3-D structural information and calculated corresponding protrusion indices to acquire discontinuous epitope candidates [10].

To integrate the advantageous features from previous works and discover innovative and important characteristics from predicted conformational epitopes, we have proposed a novel algorithm which combined geometry affinity of conformational epitopes and physicochemical propensity to reveal the highly potential surface residues as conformation epitope sites. In this study, we have analyzed the characteristics of protein antigen surface curvature distribution, surface rate of side chains, and effectiveness of chemical propensity/affinity for antigen-antibody binding. From the experimental results, our proposed method outperformed those existing techniques and provided effective candidates on discontinuous epitope prediction.

2 Material and Method

2.1 Preparation of training materials and verified data sets

The training dataset for geometrical feature and chemical propensity analysis were collected from a benchmark dataset provided by Discotope [8], which consists of 75 antigen-antibody complexes. All these complex protein structures were

selected for training the features in this study. The epitope residues from the selected protein structures were defined and verified by evaluating each residue in the antigen chain within a 4 Å distance with respect to the correspondingly tied residue in the bound antibody structure.

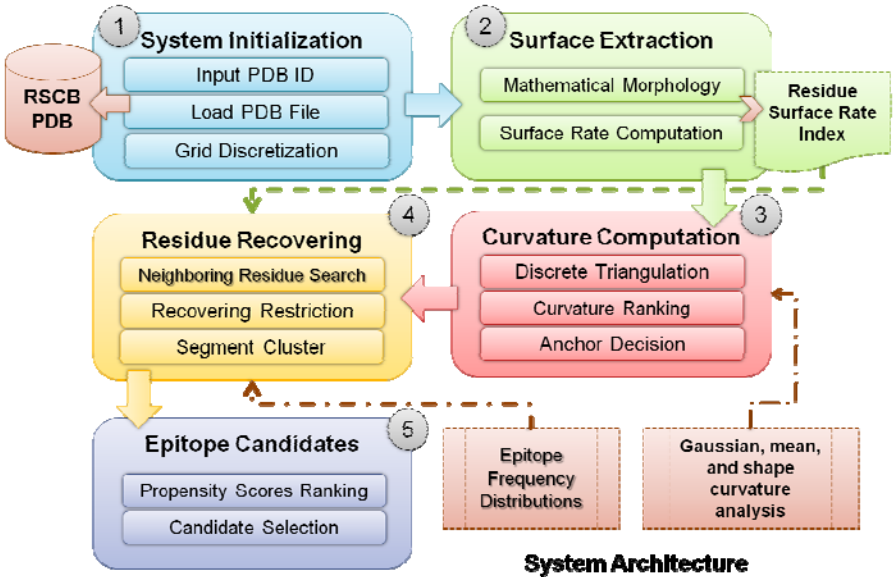

Figure 1: System configuration for the proposed CE prediction.

To evaluate system performance of the proposed method, we have randomly selected 30 antigen-antibody complexes from three different databases: CED [11], IEDB [12], and PEPOP [9] datasets. Positions of all detected conformation epitopes in these 30 structures depended on the description previously defined in their own literatures.

2.2 Methods

2.2.1 Algorithms of proposed method

As shown in Figure 1, in order to achieve the prediction of conformation epitopes from a protein structure, the proposed system was constructed by five main modules. Each module and its corresponding functions were described as the following:

Step 1: System Initialization:

The first module parses atom coordinates from a PDB file and performs the sampling processes in a 3-D grid space with a default molecular radius.

Step 2: Surface Extraction:

Mathematical morphology operations were applied to define the surface rate of each atom in a residue. The average surface rates of the set of side chain atoms and the set of backbone atoms were calculated individually. Only the side chain atoms appeared on the surface of a protein was considered as the possible candidates of

predicted conformational epitopes. In this module, a surface rate look-up table of epitope residues was constructed and a downsampled set of high surface rate residues was provided for further predicting processes. In the meantime, the predicting system also analyzed on epitope frequency distribution from the training dataset according to side chain surface rates.

Step 3: Curvature Computation:

Based on a surface rate filtering process in previous step, a set of downsampled amino acid residues was formulated as initial candidates for surface curvature analysis. The default settings were obtained according to the training processes on benchmark datasets and described in the next section in details. Here we adopted the Gaussian surface curvature to evaluate the shape curvatures of each atom on the surface. All Gaussian curvatures of initial candidate atoms were sorted and only the top 20% atoms with high Gaussian curvatures were selected as seeds for neighboring join and clustering processes.

Step 4: Residue Recovering and Group Clustering:

According to previous analysis of epitope frequency distribution among epitopes, amino acid residues and their surface rates on training dataset, several neighboring residues around the seeds could be recovered since they did not possess either high surface rates or high Gaussian curvatures and were removed during the third step. After the recovering process, the proposed system grouped the geometrically closed surface residues into various sets of conformational epitope candidates. **Step 5**: Rearrangement and sorting of the predicted conformational epitopes:

All predicted sets of conformational epitopes were sorted according to it average surface rates, physico-chemical antigenicities, and Gaussian surface curvatures. Users can easily view these predicted results from a 3D model browser using the Jmol java molecular viewer package.

2.2.2 Definition of surface region:

Residues located on the surface regions were assumed as the first requirement to be considered as candidates of predicted conformational epitopes. Therefore, precise definition of surface characteristics is the first important issue to consider. In this paper, surface region identification was achieved by employing combinations of morphological operators including dilation and erosion operations. Mathematical morphology was initially devised as a rigorous theoretic framework for the shape and structure analysis of binary image [13]. Based on its superior characteristics in describing shape and structural attributes, an efficient and effective algorithm can be designed for approaching the precise surface rates of each residue in this study. Here, an antigen structure was denoted as X as an object in a 3-D grid:

$$X = \{\mathrm{v}: f(\mathrm{v}) = 1, \mathrm{v} = (x, y, z) \in \mathrm{Z}^3\}$$

where f was called as the characteristic function of X. On the other hand, the solvent elements were regarded as the background X^c which could be defined as follows:

$$X^c = \{\mathrm{v}: f(\mathrm{v}) = 0, \mathrm{v} = (x, y, z) \in \mathrm{Z}^3\}$$

And then, a sphere of radius of 1.4 Å was taken as a structure element *B*. The

symmetric of *B* with respect to the origin (0,0,0) was denoted by B^s and written as

$$B^s = \{-\mathrm{v}: \mathrm{v} \in B\}$$

The translation of *B* by a vector *d* was then denoted by B_d and performed as

$$B_d = \{\mathrm{v} + \mathrm{d}: \mathrm{v} \in B\}$$

Three elementary morphological operators were applied for surface region calculation and listed below:

Dilation: $X \oplus B^s = \{\mathrm{v} \in \mathrm{Z}^3: \mathrm{B_v} \cap X \neq \emptyset\}$

Erosion: $X \ominus B^s = \{\mathrm{v} \in \mathrm{Z}^3: \mathrm{B_v} \subset X\}$

Difference: $(X \oplus B^s) - (X \ominus B^s)$

The surface rate of each atom was obtained by calculating the ratio of intersected and un-intersected regions between the results of difference operation and the original protein surface atoms. Figure 2 depicts an example of surface rate calculation step by step.

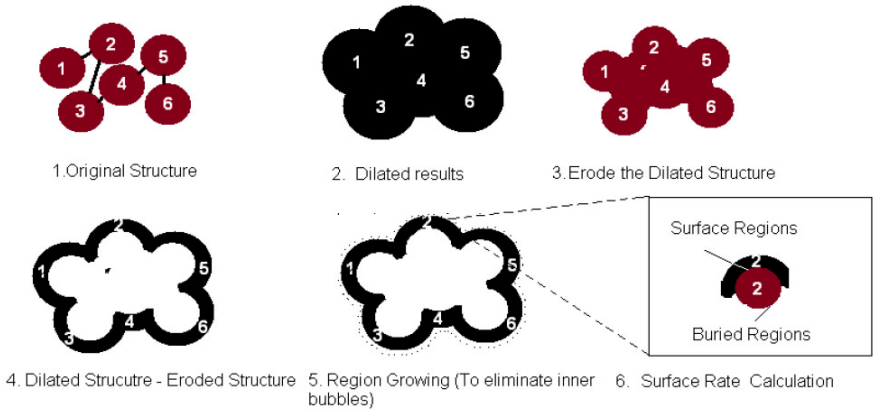

Figure 2: The mathematical morphology operations for surface rate calculation. The original structure, dilated, eroded, and differenced regions were displayed and utilized to find the surface regions of each residue.

2.2.3 Definition for curvature

Curvatures of a surface represent local measures of its shape. In this paper, for simplicity, analogous measures for discrete curves and surfaces were taken into consideration, and represented as polygonal curves and triangulated polyhedral surfaces.

Gaussian (K) and Mean (H) curvatures are the most widely used indicators for surface shape classification. Besl has defined the Gaussian and mean curvatures [14], and which are calculated from two principal curvatures k_1 and k_2. The Gaussian curvature is defined as the product of the principal curvatures, while the mean curvature equals to the arithmetic average of them:

$$\mathrm{K} = \mathrm{k_1} * \mathrm{k_2} \quad \text{and} \quad \mathrm{H} = \frac{\mathrm{k_1} + \mathrm{k_2}}{2}$$

To provide an even more compact description of local surface topology, Koenderink defined alternative curvature representation as the shape index (S) and

the magnitude of the curvedness (C) [15].

$$S = \frac{2}{\pi} * \arctan\left(\frac{k_1 + k_2}{k_1 - k_2}\right) \quad \text{and} \quad C = \sqrt{\frac{{k_1}^2 + {k_2}^2}{2}}$$

With these various measurements on shapes, we tried to analyze and discover the relationship between verified epitope residues and their corresponding curvatures.

Table 1 showed that the proportions of verified epitope residues located nearby the top 20% ranked curvature residues within a range distance. The results from training dataset revealed that the top 20% ranked residues in terms of Gaussian curvatures could be effective to be considered as the initial conformational seeds than the other two types of curvature calculation. There are about 66.7% of verified epitope residues can be retrieved based on the characteristics of high Gaussian curvatures. Figure 3(a) depicted the protein structure of 2JEL:P, in which the red elements illustrated the position of verified epitope sites, and the green elements represented the downsampled residues for initial analysis; Figure 3(b) showed the approximated version of the protein structure by considering the downsampled elements only; Figure 3(c) displayed the positions within the top 20% Gaussian curvatures in red. It can be clearly observed that the verified epiotpe residues indeed possessed high Gaussian curvature characteristics.

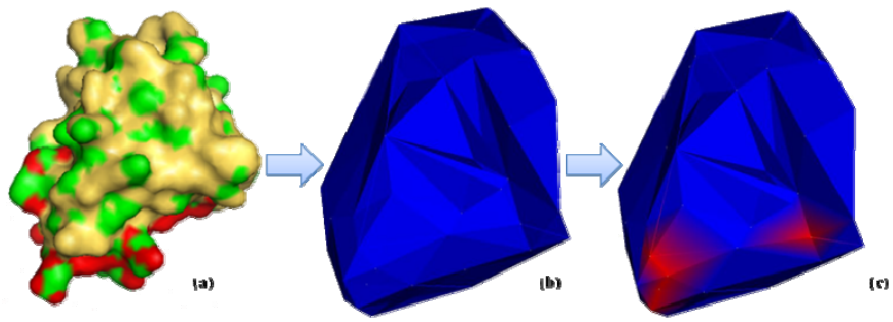

Figure 3: An example for conformational epiotpe anchor selection. (a) the protein structure 2JEL:P;(b) downsampled representation; (c) predicted seeds of conformational epitope by Gaussian curvature characteristics.

Table 1: Average percentages of three various curvatures of verified epitope residues in the top 20% of curvature features.

Curvature type	Average percentage of verified epitope residues
Gaussian Curvature	66.7%
SC Curvature	56.9%
Mean Curvature	54.9%

2.2.4 Residue Recovering and Group Clustering

Based on the analysis of distribution statistics of verified epitopes and the construction of a surface rate look-up table of residues obtained in module 2, the

proposed system recovered the neighboring residues which possessed high potentialities to be epitope candidates within a radius around the seeds. A default radius setting was obtained according to the training processes on benchmark datasets and the details were described in the next section. Subsequently, the proposed system performed the merging operations on the qualified groups into the final conformation epitope candidates by considering the amount of overlapped residues. The default setting was three overlapped residues for merging two groups into one. Finally, various scores of average surface rates, physicochemical antigenicities [18, 19], and Gaussian surface curvatures were considered as the sorting criteria for all possible candidates. The number of groups averagely ranged from 3 to 5 groups in all benchmark datasets. All predicted conformational epitopes can be easily selected and viewed from a 3D model browser employing the Jmol java molecular viewer package.

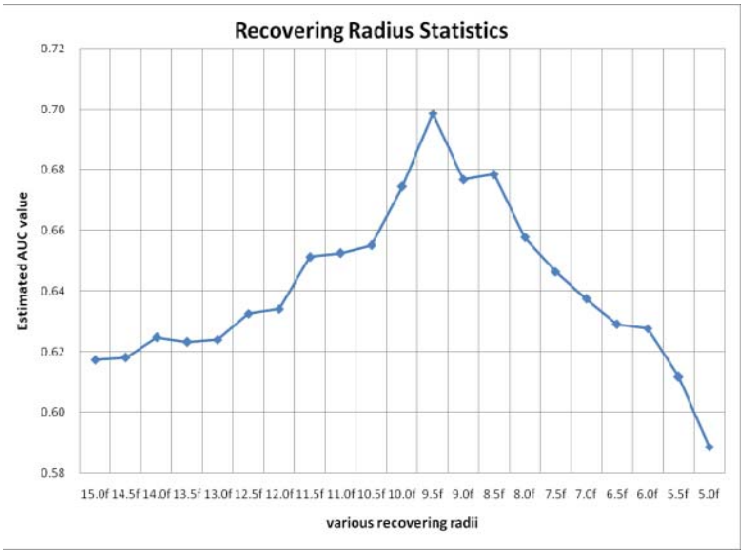

Figure 4: The estimated AUC measurement for various recovering radii. The threshold of 9.5 Å provided the best prediction by applying the benchmark data sets from Discotope.

3 Result and Discussion

In this paper, a novel algorithm was proposed to predict conformational epitopes from a protein structure. To verify the performance of the developed system, we have employed 30 protein structures with known conformation epitopes as our testing dataset. Each predicted conformational epitope from the query protein, we have calculated the number of residues of correctly predicted epitopes (TP), the number of non-epitope residues incorrectly predicted as epitope residues (FP), the number of not predicted as epitopes and indeed non-epitope residues (TN), and the number of verified epitope residues not predicted by the system (FN). The following parameters were calculated in each prediction for comparison:

Sensitivity = TP (true positive) / [TP (true positive) + FN (false negative)]
Specificity = TN (true negative) / [TN (true negative) + FP (false positive)]
AUC - Area under the ROC Curve: a summary measure of the receiver operating characteristic (ROC) curve

To achieve the best performance, we have applied the benchmark dataset from Discotope [8] as our training dataset. First of all, these 55 protein structures were applied to the prediction system with various recovering radii thresholds ranging from 5.0 Å to 15 Å. The estimated AUC measure, (sensitivity + specificity)/2, was utilized to each recovering radius setting and shown in Figure 4. It is obvious that the threshold of a recovering radius of 9.5 Å for each seed residue providing the best performance with respect to the estimated AUC measurement. Hence, the default value of the recovering radius of a seed at the forth step was set as 9.5 Å in this study. The default curvature threshold was set to 20% because it provided a better performance of determining the achors from the training dataset. Once all default settings were decided, all other steps were executed and the performance could be verified. As we mentioned in the dataset preparation section, 30 protein structures from three different databases were adopted for comparison. The ROC curves of the proposed system, Discotope and random distribution were shown in Figure 5. From the ROC curves, it was clearly demonstrated that our proposed system provided better prediction results than the Discotope prediction system.

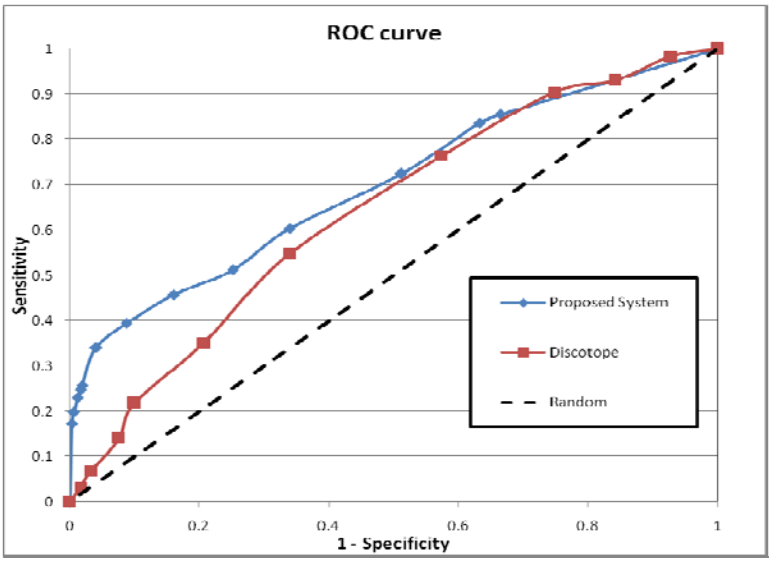

Figure 5: ROC curves of the proposed system and Discotope.

Acknowledgements

This work is supported by the Center for Marine Bioscience and Biotechnology (CMBB) in National Taiwan Ocean University, Keelung, Taiwan, and the National Science Council in Taiwan, R. O. C. (NSC97-2627-B-019-003 to T.-W. Pai).

References

[1] Yang, X., & Yu, X. (2009). An introduction to epitope prediction methods and software. *Rev Med Virol, 19*(2), 77-96.

[2] Ponomarenko, J. V., & Bourne, P. E. (2007). Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct Biol, 7*, 64.

[3] Chang, H. T., Liu, C. H., & Pai, T. W. (2008). Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches. *J Mol Recognit, 21*(6), 431-441.

[4] Odorico, M., & Pellequer, J. L. (2003). BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J Mol Recognit, 16*(1), 20-22.

[5] Larsen, J. E., Lund, O., & Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Res, 2*, 2.

[6] Van Regenmortel, M. H. V. (1996). Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. *Methods, 9*(3), 465-472.

[7] Kulkarni-Kale, U., Bhosle, S., & Kolaskar, A. S. (2005). CEP: a conformational epitope prediction server. *Nucleic Acids Res, 33*(Web Server issue), W168-171.

[8] Haste Andersen, P., Nielsen, M., & Lund, O. (2006). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci, 15*(11), 2558-2567.

[9] Moreau, V., Fleury, C., Piquer, D., Nguyen, C., Novali, N., Villard, S., et al. (2008). PEPOP: computational design of immunogenic peptides. *BMC Bioinformatics, 9*, 71.

[10] Ponomarenko, J., Bui, H. H., Li, W., Fusseder, N., Bourne, P. E., Sette, A., et al. (2008). ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics, 9*, 514.

[11] Huang, J., & Honda, W. (2006). CED: a conformational epitope database. *BMC Immunol, 7*, 7.

[12] Peters, B., Sidney, J., Bourne, P., Bui, H. H., Buus, S., Doh, G., et al. (2005). The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol, 3*(3), e91.

[13] Nikolaidis, N., & Pitas, I. (2000). *3-D Image Processing Algorithms*: John Wiley & Sons, Inc.

[14] Besl, P. J., & Jain, R. C. (1986). Invariant surface characteristics for 3D object recognition in range images. *Comput. Vision Graph. Image Process., 33*(1), 33-80.

[15] Koenderink, J. J., & Doorn, A. J. v. (1992). Surface shape and curvature scales. *Image Vision Comput., 10*(8), 557-565.

[16] Van Regenmortel, M. H. (2001). Antigenicity and immunogenicity of synthetic peptides. *Biologicals, 29*(3-4), 209-213.

[17] Wu, W. K., Chung, W. C., Chang, H. T., Cheng, R. S., & Pai, T. W. (2009). *B-Cell Conformational Epitope Prediction based on Structural Relationship and Antigenic Characteristics* Paper presented at the Intelligent Informatics in Biology and Medicine.

[18] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res, 36*(Database issue), D202-205.

[19] Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol, 157*(1), 105-132.