Prediction of Conformational Epitopes by Knowledge-based Energy Function and Geometrical Neighbouring Residue Contents

Ying-Tsang Lo[1], Tun-Wen Pai[1,2]*, Wei-Kuo Wu[1], Hao-Teng Chang[3,4]*

[1]Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan, R.O.C..

[2]Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan, R.O.C..

[3]Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung, Taiwan, R.O.C..

[4]China Medical University Hospital, Taichung, Taiwan, R.O.C..

*Corresponding author: Dr. Tun-Wen Pai, Department of Computer Science and Engineering & Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, No. 2, Peining Road, Keelung, 20224, Taiwan, R.O.C.. TEL: +886-2-24622192 ext. 6618, FAX: +886-2-24623249, E-mail: twp@mail.ntou.edu.tw

*Co-corresponding author: Dr. Hao-Teng Chang, Graduate Institute of Molecular Systems Biomedicine, College of Medicine, China Medical University, No. 91, Hsueh-Shih Road, Taichung, 40402, Taiwan, R.O.C.. TEL:+886-4-22052121 ext. 7721, FAX:+886-4-22333641, E-mail: htchang@mail.cmu.edu.tw

Keywords: antigenicity, immunogenicity, conformational epitope, solid angle, side chain surface rate, geometrical amino acid pair, knowledge-based energy function

Email addresses:

YTL: yt.lo@mail.ntou.edu.tw

TWP: twp@mail.ntou.edu.tw

WKW: yorkeway@gmail.com

HTC: htchang@mail.cmu.edu.tw

Abstract

Background

A conformational epitope (CE) is composed of several antigenic determinants which are spatially near to each other on structural surface of a protein. These segments form an antigen's epitope which may be bound by a specific paratope either from a B-cell receptor or an antibody within the immune system. The prediction of CEs plays an important role in vaccine designs and immuno-biological experiments.

Methods

The grid-based and mathematical morphological algorithms were applied for efficient detection and extraction of surface atoms, and initial surface resides of predicted CE candidates were exclusively selected according to the local average energy distribution. The novel CE prediction system was then developed based on the characteristics of surface rates, occurrence frequency of geometrical neighbouring residue combination, and knowledge-based energy functions. The trained and weighted combinatorial features of surface residue contents and potentials were integrated for a simple and effective CE prediction system.

Results

In this paper, three benchmark datasets were employed for evaluating the prediction performance. Compared to those well-developed tools, the proposed method performed well in both aspects of accuracy and efficiency. For these benchmark datasets, the proposed system achieved an average of 38.12% for sensitivities, 88.05% for specificities, and 82.79% for accuracy under a 10-fold verification mechanism.

Conclusions

The proposed method combined both features of energy profile of surface residues and occurrence frequency of geometrical amino acid pair to identify possible CEs for antigen structures, which facilitates biologists to achieve better solutions of immune-biological studies and to develop synthetic vaccines.

1. Introduction

B-cell epitopes, also known as antigenic determinants, are defined as a binding portion of an antigen which is able to interact with an antibody to elicit either cellular or humoral immune response [1, 2]. It indicates that epitopes are entities contributing with a specific recognition activity that can be recognized by a particular B-cell receptor within the immune system to generate antibody responses [3]. B-cell epitope recognition possesses huge potential for immune applications such as disease prevention, vaccine design, diagnosis and treatment. Though clinical and biological researchers depend on biochemical and biophysical experiments to identify epitope binding sites, these approaches are expensive and time-consuming but not always successful. Therefore, a reliable prediction of B-cell epitopes is an important task for computational immunology and vaccine design [4]. With the aids of accurate epitope prediction tools, immunologists are able to extract appropriate protein segments of

their interests, and it reduces experimental efforts for the design of vaccines and immunodiagnostics.

In general, epitopes can be categorized into linear (or continuous) and conformational (or discontinuous) types. A linear epitope (LE) is given to a short continuous fragment of amino acids. Though it doesn't retain information of the fold surface conformation, it reacts weakly with antibodies. The other type of a conformational epitope (CE) is composed of a patch of residues that doesn't require continuous in the protein sequence but preserving spatial vicinity [5]. In previous work, several tools focused on linear epitope predictions which required the content of protein sequences as the essentials, such as BEPITOPE [6], BCEPred [7], BepiPred [8], ABCpred [9], LEPD [10], LEPS [11] and BCPreds [12]. These tools utilized physical-chemical propensity of amino acids within a protein sequence, such as hydrophilicity, polar, charge or secondary structure, and applied quantitative matrices or machine learning algorithms, such as hidden Markov model (HMM), support vector machine (SVM) and artificial neural network (ANN) techniques, to predict the binding peptides. However, the number of LEs on native proteins had been estimated with a portion of 10% on B-cell epitopes in past analysis [13]. Most of B-cell epitopes were recognized and constructed to form the native conformation as CEs. Therefore, to identify discontinuous epitopes becomes a more practical and valuable task.

For CE prediction, several prediction tools based on the spatial information and combined with various epitope characteristics were proposed in the past decade, which include CEP [14], DiscoTope [15], PEPOP [16], ElliPro [17], PEPITO [18], and SEPPA [19]. All these prediction tools adopted various combinations of physical-chemical characteristics and trained statistical features from known antigen-antibody complexes to identify CE candidates.

A different approach based on phage display was utilized to discover relationship between protein-protein interaction from interested antigens. Phage display is one of widespread techniques applied to obtain peptide mimotopes that are selected by binding with a given monoclonal antibody in a similar way to a native epitope. The location of mimotope on the surface of the antigen can be considered as functional epitope mimics. Therefore, not only LEs but also CEs could be identified based on the mimotope analysis. The MIMOP is a hybrid computational tool that provided an epitope region prediction from information of a mimotope peptide sequence [20]. Similarly, Mapitope and Pep-3D-Search combined mimotopes and their own developed algorithms to search matched patterns on an antigen surface respectively. The algorithms identified discontinuous epitopes according to the Ant Colony Optimization (ACO) technique and several statistical thresholding parameters of amino acid pair affinity [21, 22].

The complementarily bounded surface between an antigen and an antibody could be observed from a structural complex, and the binding specificity could be determined according to hydrogen bonds, van der Walls contacts, electrostatics hydrophobic interactions. It was also experimentally verified that only a few energetic residues located within the total contact area contributed sufficient binding affinity and could be defined as true antigenic epitopes [23]. Hence, an intuitive concept to extract energetic residues from structural surface was proposed in this study. Based on thermodynamic hypothesis, we assumed that a native unbound antigen structure is at the lowest free energy state, but relatively, the most active residues located on this antigen's surface would possess higher potential energies for binding with an antibody under various physiological conditions. There are two different types of potential energy functions are currently estimated in recent study. One is the physical-based potential which focuses on the fundamental analysis of forces between atoms,

and the other is the knowledge-based potential which extracts parameters from experimentally solved protein structures[24]. Due to heavy computational complexities required by the former approach, the second way to estimate energy by distance-dependent atomic knowledge-based potential was adopted in this study, and the formulated energy functions of all surface residues were mainly provided by the Protein Structure Analysis web system (ProSA) [25].

In addition to the energy function of surface residues, according to Chen's study in linear epitope prediction [26], it showed that the occurrence frequencies of some amino acid pairs in CE epitope datasets are significantly higher than those in non-epitope datasets. This statistical feature might be reasonably applied to enhance the performance of a CE prediction system. Hence, both the advantages of featuring statistical distribution of verified CE epitopes and preserving high energy function of candidate surface residues are considered simultaneously in this study. Here, surface residues with higher energy function and located within a constrained radius were initially and exclusively assigned as initial anchors, and followed by extending neighbouring residues to formulate predicted CE clusters. To our best knowledge, the combination of energy function of surface residues and the occurring frequency of neighbouring residues was not proposed yet. Hence, in this study the distributions of energy function and appeared combination of geometrical paired residues from true epitopes were discussed and analyzed, and the adoption of these information was applied to train the best parameters for CE prediction. From the experimental results, our proposed method provided outstanding performance for extracting effective candidates on discontinuous epitope prediction. All the details will be presented in the following sections.

2. Material and Methods

2.1 System Architecture

The proposed system for predicting conformational epitope was constructed by four main stages including grid-based surface structure analysis, energy profile computation, anchor assignment and CE clustering and ranking. The system flow is depicted in Figure 1.

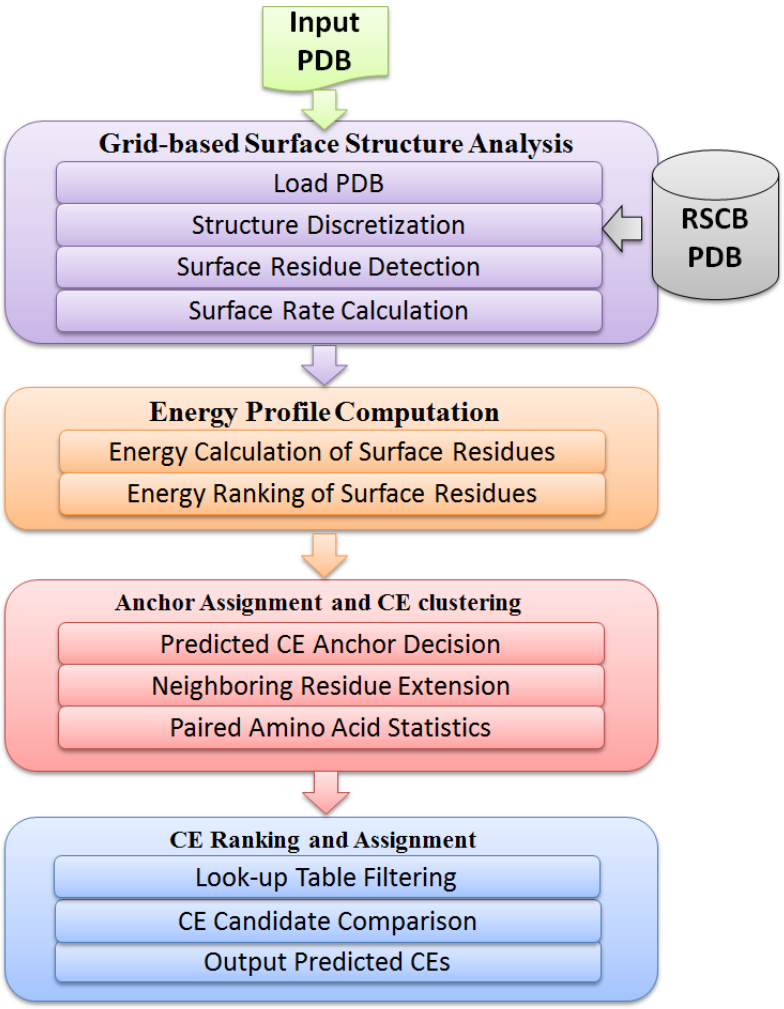


Figure 1 System configuration for the proposed CE prediction system.

The first module of "Grid-based Surface Structure Analysis" receives a PDB ID or PDB file from RCSB Protein Data Bank [27] and performs the protein data sampling processes for extracting surface information. Subsequently, 3D

mathematical morphology techniques were applied to extract the solvent accessible surface from a protein antigen in “Surface Extraction” [28], and surface rates of atoms were calculated by evaluating the exposure ratio contacted with solvent molecules. Then the proposed system would sum up all side chain atoms of each residue as the residue surface rates and exported to a look-up table. The next module of “Energy Profile Computation“ utilized the results from ProSA web system to rank the energy profiles of each residue on antigen surface. The energy profiles were ranked for the following CE anchor selection. Surface residues with higher ranked values and located within mutually exclusive positions will be considered as initial conformational epitope anchors. The third module performed CE neighbouring residue extension from initial CE anchors to retrieve neighbouring residues based on energy indices and surface residue distance. Besides, the pairwise amino acid statistics were calculated for selecting suitable predicted CE clusters. At the last module, the values of knowledge-based energy propensity and occurrence frequencies of geometrical amino acid pairs were combined with weighted coefficients to provide final predicted CEs.

2.2 Preparation of testing datasets

In this study, the DiscoTope, Epitome database, and IEDB (Immune Epitope Database) were used to verify the prediction results. DiscoTope provided a benchmark dataset consisted of 70 antigen-antibody complexes which were obtained from the SACS database [29] with only structures determined to a resolution less than 3 Å and with protein antigens of greater than 25 amino acid residues. The estimated epitope residues from DiscoTope dataset were defined and verified by evaluating each residue in the antigen chain within a 4 Å distance with respect to the correspondingly tied residues in the bound antibody structure. Epitome dataset contained 134 protein

chains which were inferred by the distance between protein antigens and Complementary Determining Regions (CDRs) of the correspondingly tied antibodies. They labelled residues as interaction sites if any of their particular atoms were measured within a distance of less than 6 Å from CDRs of the antibody. The IEDB dataset collected in this study was composed of 43 protein antigen chains from IEDB website (www.immuneepitope.org). This dataset contained only for protein antigens with complex structure annotation in the field of "ComplexPdbId" from "iedb_export" zip file. Since there were 11 protein chains with total number of residues less than 35, here we only selected 45 antigen-antibody complexes to represent the IEDB benchmark dataset. A non-redundant dataset of 163 antigen structures from the previous datasets was also constructed for the final verification processes.

2.3 Surface Structure Analysis

Interaction between an antigen and an antibody was usually induced by the atoms located on surface areas. The definitions of protein surface including solvent accessible surface and molecular surface were first implied by [30] as shown in Figure 2. Then, Richards introduced molecular surface constructed by "contact surface" and "re-entrant surface". The contact surface represents the part of the van der Waals surface which can be directly touched by solvent. The re-entrant surface consists of the inward-facing part of the probe sphere when it is in contact with more than one molecular atoms [31]. In 1983, Connolly employed the Gauss-Bonnet approach to calculate molecular surface, who defined a probe with small size to roll over a whole protein structure and obtained the molecular surface. Based on these definitions described above, we proposed a novel algorithm to efficiently retrieve surface regions.

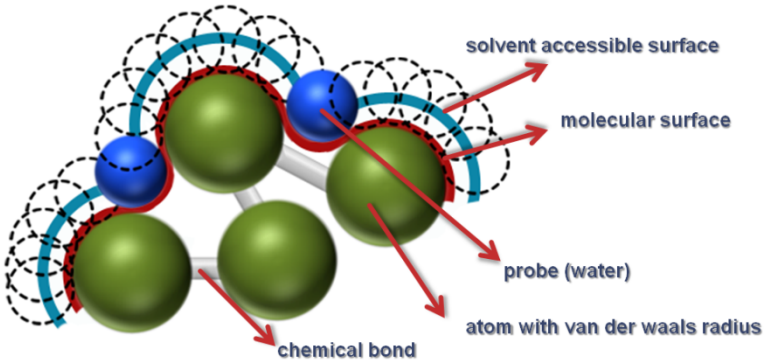


Figure 2 A cartoon illustration of protein surface definitions.

2.3.1 Three-dimension Mathematical Morphology

In this study, surface region identification was achieved by employing combinatorial morphological operators including dilation and erosion operations. Mathematical morphology was initially devised as a rigorous theoretic framework for shape and structural analysis of binary images. Based on its superior characteristics in describing shape and structural attributes, an efficient and effective algorithm can be designed for detecting precise surface rates from each residue. Here, an antigen structure was denoted as X as an object in a 3-D grid:

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

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:

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

A sphere with pre-defined radius of 1.5 Å was defined 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 then applied for surface region calculation and listed below:

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

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

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

The surface rate of each atom could be obtained by calculating the ratio of intersected and un-intersected regions with respect to the overlapping areas between the results of morphological difference operation and the original protein atoms. Figure 3 depicts an example step by step for extracting the surface regions and calculating the surface rate of an atom.

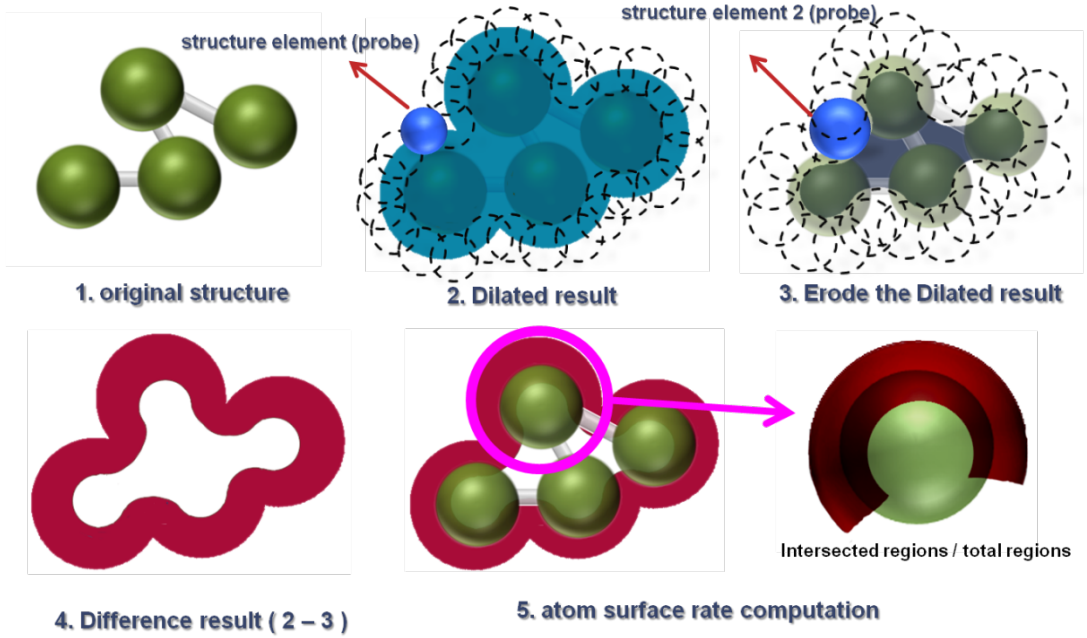


Figure 3 Procedures of mathematical morphology operations for surface rate calculation: including the original structure, dilated, eroded, difference regions, and surface regions of each atom.

2.3.2 Surface Rate Computation

The properties of side chains of amino acid residues were considered as an important factor in protein-protein interaction. There were numerous literatures dealing with influences of side chains on protein binding issues. Antigen-antibody binding might bring conformational change of protein structures, and amino acids with flexible side chains were considered as potentially useful in this situation. Moreover, hydrophobic and polar side chains were also regarded as major binding affinities formed protein-protein interface in experiments [32-37]. Therefore, side chains of residue affections were considered as main inspection in the proposed algorithm. Through 3-D mathematical morphology operations, the rate of each molecular atom, *AR(r)*, can be precisely acquired. Here, only side-chain atoms were involved on surface rate computation, and the surface rate of each residue was denoted as *SR*(*r*). It was calculated by the following formula:

$$SR(r) = \left\{ i \in R : \frac{1}{N} \sum_{i=1}^{N} AR(i) \right\}$$

where i represents the i^{th} atom on a specific amino acids, R is all atom types of selected residue, and N is the total number of atoms of the residue "r".

According to the definition, statistics of surface rate of verified epitope residues and all residues of benchmark datasets were analysed and illustrated in **Error! Reference source not found.**. From the statistics, it indicated that true epitopes often possessed higher surface rates to bind with surface residues of antibodies. After surface rate analysis, the proposed system set a look-up table and minimum thresholding of surface rate for further prediction processes.

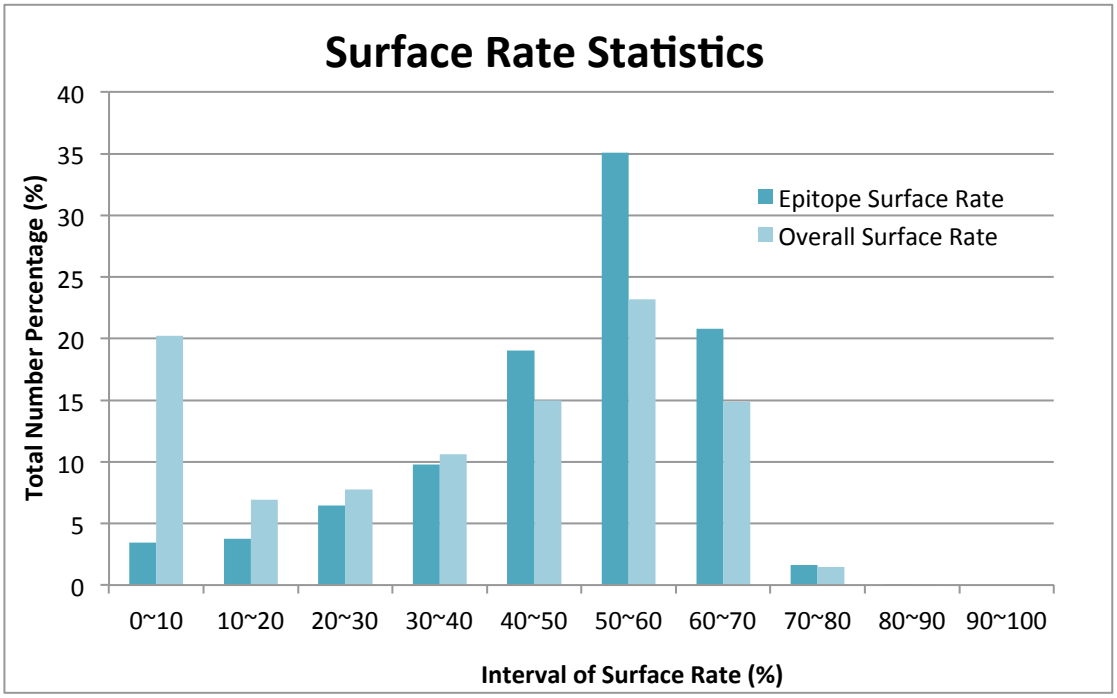


Figure 4 The distribution of surface rate in true CE Residues and overall residues.

2.4 Energy Profile Computation

The knowledge-based potential was adopted for representing the energy of each surface residue, which was obtained from the distribution of pairwise distances to extract effective potentials between residues. The potential of each residue was usually constructed from an all-heavy-atom representation, and the heavy atoms in a protein were previously categorized according to the specific types of residues or atoms. Basically, the potential for any pair of two heavy atoms is calculated according to the observed and expected number of contacts within a certain distance. The potential between two atoms indicated the level of attractive interaction between two different residues. Though the main application of the knowledge-based potential was used for improving the folding recognition, structure prediction and refinement, here we adopted the advantages of calculated energy function from each surface residue to distinguish various statuses of active conditions. To understand the differences of knowledge-based potential between the residues of true CE epitopes

and non-CE epitopes, we calculated the surface energy profiles of various parameter settings from all 247 known antigens. The results have shown that the true CE residues possessed higher energy functions than non-epitope residues. If the window size was set as 8 from ProSA system, the average energy of each verified CE residue cluster of an antigen from Epitome, DiscoTope, and IEDB datasets were 69.4%, 75% and 51.2% higher than average energy of non-CE residues of that antigen, respectively. It was also observed that at least one of the CE residues ranked in the top 20% of energy function of all surface residues, and most of the highest energy values of all CE residues ranked in the top 3% for all antigens. Therefore, in this study we selected the first 20% residues with higher energy as our initial CE anchors. Besides, the selected initial seeds should possess surface rates within the range of 20% to 50% according previous statistics. All satisfied seed residues would be mutually examined with a shortest distance of 12 Å to eliminate possible CE candidate groups. Once the initial seeds were decided, the neighbouring residues will be included within the radius of 10 Å.

2.5 Occurrence Frequency Analysis of Geometrical Amino Acid Pairs

The proposed filtering mechanism is adopted from Chen's idea as statistical features for CE verification. However, the amino acid pairs were formulated from geometrical neighbouring relationship instead of continuing sequence sense. **Error! Reference source not found.** defines the required variables in statistical analysis for query amino acid pair. Since there are 20 different amino acids and neglect order relationship within a pair of residues, a total of 210 possible combinations of surface residue pair were analyzed for their occurence frequencies within the true CE epitope and non-CE epitope datasets. Higher occurence frequencies of geometrical amino acid

pair within various radii (range from 2 Å to 6 Å) were analysed and their corresponding CE indices for each pair were also calculated.

Table 1: Required variables in statistical analysis for geometrical amino acid pairs (GAAP).

Variables	**Description**
N_{GAAP}^{+}	The occurrence times of a geometrical amino acid pair in the true CE epitope dataset.
N_{GAAP}^{-}	The occurrence times of a geometrical amino acid pair in the non-CE epitope dataset.
f_{GAAP}^{+}	The occurrence frequencies of a geometrical amino acid pair in the true CE epitope dataset.
f_{GAAP}^{-}	The occurrence frequencies of a geometrical amino acid pair in the non-CE epitope dataset.
$Total_{GAAP}^{+}$	The total occurrence times of all geometrical amino acid pairs in the true CE epitope dataset.
$Total_{GAAP}^{-}$	The total occurrence times of all geometrical amino acid pairs in the non-CE epitope dataset.
CEI_{GAAP}	CE Index of geometrical amino acid pair.

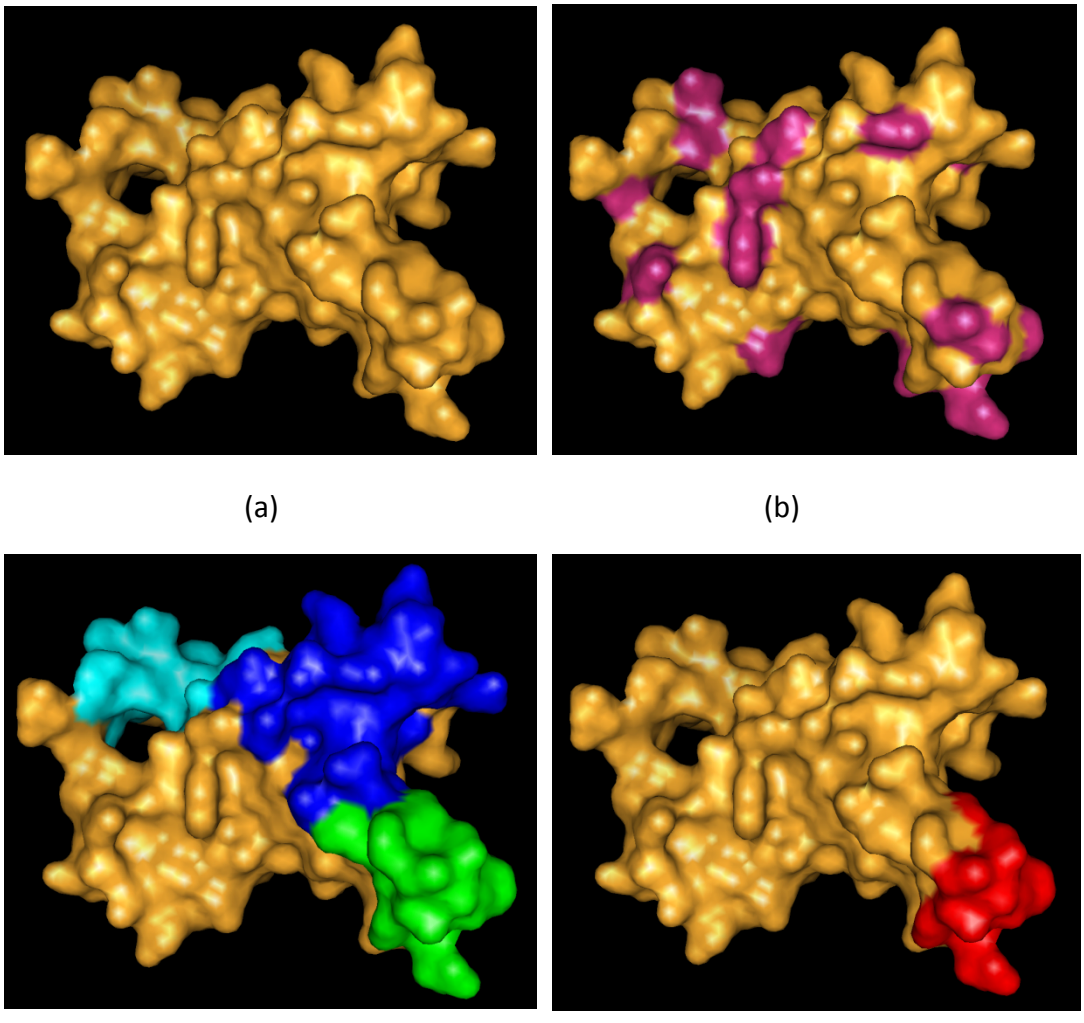
The CE Index (CEI_{GAAP}) of a geometrical amino acid pair was obtained by the following equations, which calculated the frequency of occurrence of a particular pair in the CE dataset divided by the frequency of occurrence of the same pair in the non-CE epitope dataset, and then took logarithm of the ratio to base 10. The final CE Index was normalized within the range of [0,1]. In this study, the total occurrence times of all geometrical amino acid pairs in the true CE epitope dataset are 2834 pairs, and the total occurrence times of all geometrical amino acid pairs in the non-CE epitope dataset are 36,118 under the radius of 2 Å. For example, the highest two CEIs were “HQ” of 0.921 and “EH” of 0.706 calculated from 247 antigens.

$$f_{GAAP}^{+} = \frac{N_{GAAP}^{+}}{Total_{GAAP}^{+}}$$

$$f_{GAAP}^{-} = \frac{N_{GAAP}^{-}}{Total_{GAAP}^{-}}$$

$$\mathrm{CEI}_{GAAP} = \log\left[\frac{f^{+}_{GAAP}}{f^{-}_{GAAP}}\right]$$

Individual CE index in a predicted CE cluster will be summed up and divided by total number of CE pairs within the cluster to obtain a corresponding average CE Index of the predicted CE patch. Finally, the CEI_{GAAP} will multiply with a weighting coefficient and combined with the average energy function to obtain a final CE ranking index. Based on the combined index, the proposed CE prediction system will provide three best CE candidate groups for users. One example of protein 1ORS:C is shown in the Figure 5. Protein surface detection (Figure 5a), energy thresholding (Figure 5b), three predicted CE clusters (Figure 5c), and the true-CE residue of protein 1ORS:C (Figure 5d) were shown for demonstration purposes.



(a) (b)

(c) (d)

Figure 5 (a) Protein surface of 1ORS:C; (b) surface seed residues possessing energy function within top 20%; (c) top three CE predicted groups by removing neighbouring seeds located within 12 Å and extended neighbouring residues with 10 Å; (d) the true CE residues.

3. Experimental Results

A novel algorithm based on surface energy and occurrence frequency of neighbouring residue was proposed to predict CEs. To verify the performance of the developed system, we have employed 247 protein structures and 163 non-redundant structures collected from three benchmark datasets under 10-fold verification mechanism. All these verified CEs on protein structure were obtained either from experimental observation or inferred from computational analysis. For each predicted CE of the query protein, we have calculated the number of epitope residues correctly predicted as epitope residues (TP), the number of non-epitope residues incorrectly predicted as epitope residues (FP), the number of not predicted as epitope residues and indeed non-epitope residues (TN), and the number of verified epitope residues not correctly predicted by the system (FN). The following parameters were calculated in each prediction for comparison:

$$\text{Sensitivity (SE)} = \text{TP} \div [\text{TP} + \text{FN}]$$

$$\text{Specificity (SP)} = \text{TN} \div [\text{TN} + \text{FP}]$$

$$\text{Positive Prediction Value (PPV)} = \text{TP} \div [\text{TP} + \text{FP}]$$

$$\text{Accuracy (ACC)} = [\text{TP} + \text{TN}] \div [\text{TP} + \text{TN} + \text{FN} + \text{FP}]$$

Error! Reference source not found. and **Error! Reference source not found.** provide the evaluation of weighting coefficient combination for both energy function and occurrence frequency of pairwise amino acid features. **Error! Reference source not found.** is the result for average energy function of residues located within a radius of 6 Å, and **Error! Reference source not found.** is the case for considering energy function of individual residue. The results have shown that average energy function provided a better performance than considering single residue. However, both approaches resulted in a quite stable performance for sensitivity, specificity, positive prediction value, and accuracy. The best combination of weighting coefficients for sensitivity is 10% for average patch energy function and 90% for occurrence frequency. This is mainly due to the energy function criteria had been applied in the previous step for CE anchor selection. Therefore, the feature of energy function would not affect the prediction results with obvious influences. In this case, the initial parameter settings for new target antigen protein and the following 10-fold verification will apply with these trained combinations.

Table 2: Average performance of CE prediction for various weighting coefficient combinations between average energy (Avg. EG) within a 6 Å-radius and pairwise residue occurrence rate (PR). Each antigen was predicted with three CE candidates.

Weighting Combinations	SE	SP	PPV	ACC
0%EG+100%PR	0.38174909	0.88026912	0.28948427	0.82762314
10%EG+90%PR	**0.41375626**	0.88491713	0.318401513	0.83550329
20%EG+80%PR	0.40411907	0.88339643	0.310372011	0.83364651
30%EG+70%PR	0.40071021	0.88472985	0.308931260	0.83462812
40%EG+60%PR	0.40235963	0.88500477	0.308956909	0.83484050
50%EG+50%PR	0.40032410	0.88526988	0.308866524	0.83494350
60%EG+40%PR	0.39826932	0.88709592	0.310329851	0.83674728
70%EG+30%PR	0.39788531	0.88708866	0.310057838	0.83681763
80%EG+20%PR	0.39440495	0.88639840	0.307165993	0.83575056
90%EG+10%PR	0.39315133	0.88647102	0.307463589	0.83588749
100%EG+0%PR	0.39477960	0.88665173	0.307860654	0.83606191

Table 3: Average performance of CE prediction for various weighting coefficient combinations between individual energy (Ind. EG) and pairwise residue occurrence rate (PR). Each antigen was predicted with three CE candidates.

Weighting Combinations	SE	SP	PPV	ACC
0%EG+100%PR	0.38904213	0.88545484	0.297620232	0.83316720
10%EG+90%PR	0.38730979	0.88374611	0.295145236	0.83109301
20%EG+80%PR	**0.40874497**	0.88785200	0.315718499	0.83729001
30%EG+70%PR	0.39293810	0.88612791	0.305437883	0.83393131
40%EG+60%PR	0.40530435	0.88759054	0.313223041	0.83635800
50%EG+50%PR	0.40110938	0.88624436	0.314452191	0.83427900
60%EG+40%PR	0.38267268	0.88614126	0.306830027	0.83289012
70%EG+30%PR	0.36904261	0.88510455	0.297330839	0.83028217
80%EG+20%PR	0.35784993	0.88327931	0.287382221	0.82740505
90%EG+10%PR	0.35565826	0.88242811	0.283611851	0.82639348
100%EG+0%PR	0.349151010	0.88206203	0.281820846	0.82577874

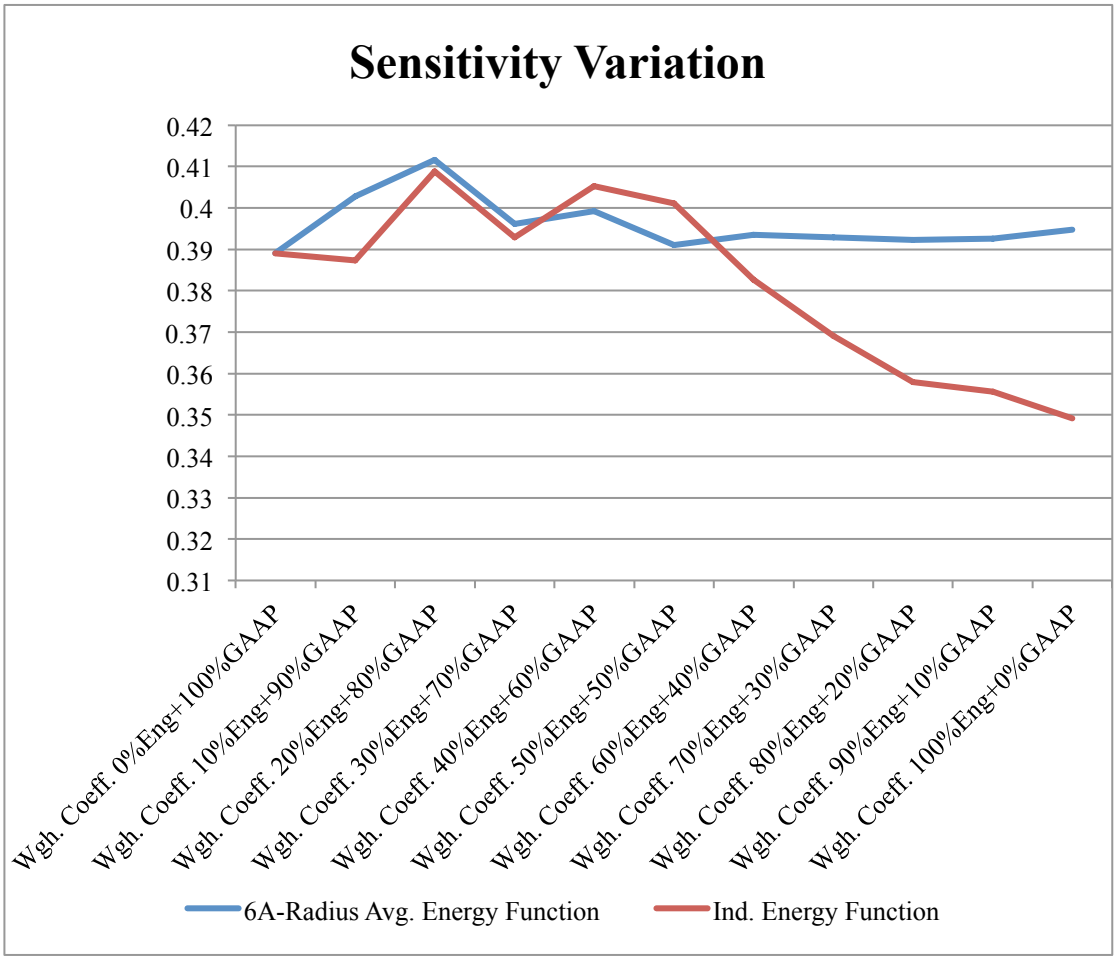


Figure 6 To observe the performance of different combinations of energy function and the occurrence frequency rate of pairwise amino acids. The results have shown that both average and individual energy indices maintained with a quite stable performance in sensitivities. The average energy profiles provided a better performance than individual residue consideration.

To evaluate the CE predicted system, we adopted a 10-fold cross validation mechanism. The total 247 protein antigens from DiscoTope, Epitome, and IEDB datasets and the 163 non-redundant antigens were applied as two individual testing datasets. For the first set of 247 antigens, the results indicated that the proposed system achieved an average sensitivity of 38.12%, an average specificity of 88.06%, an average positive prediction value of 29.18%, and an average accuracy of 82.79%.

For the second set of non-redundant 163 antigens, an average sensitivity of 34.59%, an average specificity of 88.70%, an average positive prediction value of 29.24%, and an average accuracy of 82.74% were obtained. For these two testing datasets, the number of CE clusters was limited to 3 predicted CE sets.

With rapidly increasing number of solved protein structures, CE prediction has been more and more desirable for biomedical and immunological scientists to obtain the ultimate capacity in immune applications. In this paper, a novel method combined characteristics of surface rate, knowledge-based energy function, and occurrence frequency of geometrical amino acid pairs was proposed for predicting CE residues located in discontinuous B cell antigenic determinates. Compared to those well-developed tools, the proposed method performed well in both aspects of accuracy and efficiency.

Discussion and Conclusion

With rapidly increasing number of solved protein structures, CE prediction has been more and more desirable for biomedical and immunological scientists to obtain the ultimate capacity in immune applications. In this paper, a novel method combined characteristics of surface rate, energy function, and geometrical amino acid pairs was proposed for predicting CE residues located in discontinuous B cell antigenic determinates. Since some existing systems do not allow users to evaluate the AUC values through parameter settings, there exists another approximated evaluation for AUC measurement by taking the average of specificity and sensitivity from system's default settings [19]. To compare the prediction performance with DiscoTope system with respect to the DiscoTope's testing dataset, our proposed system provided a higher average specificity of 89.1% than DiscoTope's 75%, and a higher average

sensitivity of 56.5% than DiscoTope's 47.3%. Hence, the estimated AUC value of 0.728 of our proposed method is superior to an estimated AUC value of 0.612 of DiscoTope. For comparing with the PEPTIO (BEPro) system, we have utilized both Epitome and DiscoTope datasets, and the PEPITO system produced averaged AUC values of 0.683 and 0.753, respectively. In comparison with our proposed system, we have achieved comparable AUC values of 0.694 and 0.728 for Epitome and DiscoTope datasets, respectively. The average number of predicted CE groups from our proposed system was about 6 CE candidates, and the best predicted CE cluster was ranked at an average of 2.9. This is also the reason for providing three CE candidates initially from our proposed system. Besides, due to the distance limitation for extending neighbouring residues, our proposed system generally predicted CEs with limited residues and it performed better than other system in terms of specificity. However, this will lower down the quality of sensitivity simultaneously. Perhaps future research can enhance on examining the distributions of various propensities between epitopes and non-epitopes, specific geometrical shape of query antigens, and unique corresponding relationship between antigens and antibodies. The clustered information should be able to facilitate appropriate selection of initial CE anchors and provide precise CE candidates for immunologists.

Competing interests

No competing interests.

Authors' contributions

YTL and WKW designed the algorithms and performed the experimental data analysis. TWP and HTC conceived of the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work is supported by the Center of Excellence for Marine Bioenvironment and Biotechnology of National Ocean University and National Science Council, Taiwan, R.O.C. (NSC 101-2321-B-019-001 and NSC 100-2627-B-019-006 to T.-W. Pai), and also supported in part by Taiwan Department of Health Clinical Trial and Research Center of Excellence (DOH101-TD-B-111-004).

References

[1] X. Yang and X. Yu, "An introduction to epitope prediction methods and software," *Rev Med Virol,* vol. 19, pp. 77-96, Mar 2009.

[2] N. S. Greenspan and E. Di Cera, "Defining epitopes: It's not as easy as it seems," *Nat Biotechnol,* vol. 17, pp. 936-7, Oct 1999.

[3] D. R. Davies and G. H. Cohen, "Interactions of protein antigens with antibodies," *Proc Natl Acad Sci U S A,* vol. 93, pp. 7-12, Jan 9 1996.

[4] J. A. Greenbaum, P. H. Andersen, M. Blythe, H. H. Bui, R. E. Cachau, J. Crowe, M. Davies, A. S. Kolaskar, O. Lund, S. Morrison, B. Mumey, Y. Ofran, J. L. Pellequer, C. Pinilla, J. V. Ponomarenko, G. P. Raghava, M. H. van Regenmortel, E. L. Roggen, A. Sette, A. Schlessinger, J. Sollner, M. Zand, and B. Peters, "Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools," *J Mol Recognit,* vol. 20, pp. 75-82, Mar-Apr 2007.

[5] M. H. Van Regenmortel, "Antigenicity and immunogenicity of synthetic peptides," *Biologicals,* vol. 29, pp. 209-13, Sep-Dec 2001.

[6] M. Odorico and J. L. Pellequer, "BEPITOPE: predicting the location of continuous epitopes and patterns in proteins," *J Mol Recognit,* vol. 16, pp. 20-2, Jan-Feb 2003.

[7] S. Saha and G. P. S. Raghava, "BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physical-chemical properties," *LNCS,* vol. 3239, pp. 197-204, 2004.

[8] J. E. Larsen, O. Lund, and M. Nielsen, "Improved method for predicting linear B-cell epitopes," *Immunome Res,* vol. 2, p. 2, 2006.

[9] S. Saha and G. P. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins,* vol. 65, pp. 40-8, Oct 1 2006.

[10] H. T. Chang, C. H. Liu, and T. W. Pai, "Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches," *J Mol Recognit,* vol. 21, pp. 431-41, Nov-Dec 2008.

[11] H. W. Wang, Y. C. Lin, T. W. Pai, and H. T. Chang, "Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification," *J Biomed Biotechnol,* vol. 2011, p. 432830, 2011.

[12] Y. El-Manzalawy, D. Dobbs, and V. Honavar, "Predicting linear B-cell epitopes using string kernels," *J Mol Recognit,* vol. 21, pp. 243-55, Jul-Aug 2008.

[13] M. H. V. Van Regenmortel, "Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity," *Methods,* vol. 9, pp. 465-72, Jun 1996.

[14] U. Kulkarni-Kale, S. Bhosle, and A. S. Kolaskar, "CEP: a conformational epitope prediction server," *Nucleic Acids Res,* vol. 33, pp. W168-71, Jul 1 2005.

[15] P. Haste Andersen, M. Nielsen, and O. Lund, "Prediction of residues in discontinuous B-cell epitopes using protein 3D structures," *Protein Sci,* vol. 15, pp. 2558-67, Nov 2006.

[16] V. Moreau, C. Fleury, D. Piquer, C. Nguyen, N. Novali, S. Villard, D. Laune, C. Granier, and F. Molina, "PEPOP: computational design of immunogenic peptides," *BMC Bioinformatics,* vol. 9, p. 71, 2008.

[17] J. Ponomarenko, H. H. Bui, W. Li, N. Fusseder, P. E. Bourne, A. Sette, and B. Peters, "ElliPro: a new structure-based tool for the prediction of antibody epitopes," *BMC Bioinformatics,* vol. 9, p. 514, 2008.

[18] M. J. Sweredoski and P. Baldi, "PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure," *Bioinformatics,* vol. 24, pp. 1459-60, Jun 15 2008.

[19] J. Sun, D. Wu, T. Xu, X. Wang, X. Xu, L. Tao, Y. X. Li, and Z. W. Cao, "SEPPA: a computational server for spatial epitope prediction of protein antigens," *Nucleic Acids Res,* vol. 37, pp. W612-6, Jul 2009.

[20] V. Moreau, C. Granier, S. Villard, D. Laune, and F. Molina, "Discontinuous epitope prediction based on mimotope analysis," *Bioinformatics,* vol. 22, pp. 1088-95, May 1 2006.

[21] E. M. Bublil, N. T. Freund, I. Mayrose, O. Penn, A. Roitburd-Berman, N. D. Rubinstein, T. Pupko, and J. M. Gershoni, "Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm," *Proteins,* vol. 68, pp. 294-304, Jul 1 2007.

[22] Y. X. Huang, Y. L. Bao, S. Y. Guo, Y. Wang, C. G. Zhou, and Y. X. Li, "Pep-3D-Search: a method for B-cell epitope prediction based on mimotope analysis," *BMC Bioinformatics,* vol. 9, p. 538, 2008.

[23] J. Novotny, R. E. Bruccoleri, and F. A. Saul, "On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5," *Biochemistry,* vol. 28, pp. 4735-49, May 30 1989.

[24] H. Lu and J. Skolnick, "A distance-dependent atomic knowledge-based potential for improved protein structure selection," *Proteins,* vol. 44, pp. 223-32, Aug 15 2001.

[25] M. Wiederstein and M. J. Sippl, "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins," *Nucleic Acids Res,* vol. 35, pp. W407-10, Jul 2007.

[26] J. Chen, H. Liu, J. Yang, and K. C. Chou, "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale," *Amino Acids,* vol. 33, pp. 423-8, Sep 2007.

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res,* vol. 28, pp. 235-42, Jan 1 2000.

[28] M. L. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *Science,* vol. 221, pp. 709-13, Aug 19 1983.

[29] L. C. Allcorn and A. C. Martin, "SACS--self-maintaining database of antibody crystal structure information," *Bioinformatics,* vol. 18, pp. 175-81, Jan 2002.

[30] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *J Mol Biol,* vol. 55, pp. 379-400, Feb 14 1971.

[31] F. M. Richards, "Areas, volumes, packing and protein structure," *Annu Rev Biophys Bioeng,* vol. 6, pp. 151-76, 1977.

[32] W. I. Chou, T. W. Pai, S. H. Liu, B. K. Hsiung, and M. D. Chang, "The family 21 carbohydrate-binding module of glucoamylase from Rhizopus oryzae consists of two sites playing distinct roles in ligand binding," *Biochem J,* vol. 396, pp. 469-77, Jun 15 2006.

[33] P. Herion and J. L. De Coen, "Production and characterization of a rat monoclonal antibody against leu5 enkephalin," *Mol Immunol,* vol. 23, pp. 209-15, Feb 1986.

[34] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: an open source platform for ligand pocket detection," *BMC Bioinformatics,* vol. 10, p. 168, 2009.

[35] I. S. Mian, A. R. Bradwell, and A. J. Olson, "Structure, function and properties of antibody binding sites," *J Mol Biol,* vol. 217, pp. 133-51, Jan 5 1991.

[36] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman, "Side-chain flexibility in proteins upon ligand binding," *Proteins,* vol. 39, pp. 261-8, May 15 2000.

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