BIOINFORMATICS

Vol. 00 no. 00 2017
Pages 1–8

Novel overlapping subgraph clustering for the detection of epitopes from an antigen

Liang Zhao[1,2], Shaogui Wu[2], Jiawen Jiang[1], Wencui Li[1], Jie Luo[1,*], Jinyan Li[3, *]

[1]Taihe Hospital, Hubei University of Medicine, Hubei 442000, China.
[2]School of Computing and Electronic Information, Guangxi University, Nanning 530004, China.
[3] Advanced Analytics Institute and Centre for Health Technologies, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Antigens that contain overlapping epitopes have been reported occasionally. As current algorithms mainly take a one-antigen-one-epitope approach to the prediction of epitopes, they are not capable of detecting these multiple and overlapping epitopes accurately, or even those multiple and separated epitopes existing in some other antigens.

Results: We introduce a novel subgraph clustering algorithm for more accurate detection of epitopes. This algorithm takes graph partitions as seeds, and expands the seeds to merge overlapping subgraphs based on the term frequency-inverse document frequency (TF-IDF) featured similarity. Then, the merged subgraphs are each classified as an epitope or non-epitope. Tests of our algorithm were conducted on three newly collected datasets of antigens. In the first dataset, each antigen contains only a single epitope; in the second, each antigen contains only multiple and separated epitopes; and in the third, each antigen contains overlapping epitopes. The prediction performance of our algorithm is significantly better than the state-of-art methods. The lifts of the averaged f-scores on top of the best existing methods are 60%, 75%, and 22% for the single epitope detection, the multiple and separated epitopes detection, and the overlapping epitopes detection, respectively.

Availability The source code is available at `github.com/lzhlab/glep/`.

Contact: s080011@e.ntu.edu.sg

1 INTRODUCTION

A B-cell epitope is an antigenic determinant at the surface of an antigen that binds to an antibody, which is crucial for immune response (Abbas *et al.*, 2009). The relation between epitopes and antigens is not necessarily a one-to-one correspondence. One antigen may have multiple epitopes, even sometimes have multiple and overlapping ones; see Figure 1. Detection of all these epitopes from an antigen can be beneficial to effective vaccine design, disease diagnosis, disease therapy, and template-aware pathogenic-free reagent development (Sela-Culang *et al.*, 2014).

A number of experimental and computational methods have been developed for the detection of epitopes. The experimental approaches, such as X-ray crystallography, nuclear magnetic resonance and phage display, are laborious. They have also failed to detect the multiple and overlapping epitopes from an antigen (Sela-Culang *et al.*, 2014). On the other hand, the computational approaches are cheap and flexible, attracting intensive research recently (Esmaielbeiki *et al.*, 2016), including methods DiscoTope 2.0 (Kringelum *et al.*, 2012), ElliPro (Ponomarenko *et al.*, 2008), and SEPPA 2.0 (Qi *et al.*, 2014).

The core idea of these computational methods is to combine chemo-physical properties, e.g., hydrophobicity (Kyte and Doolittle, 1982), polarity (Cooper and Hausman, 2004), protrusion index (Ponomarenko *et al.*, 2008), with machine learning techniques, e.g., support vector machine, neural network, random forest, to identify epitopes. Depending on whether antibody information is required, these methods can be further categorized into antibody-specific methods and antibody-agnostic methods. The antibody-specific epitope prediction methods (Zhao and Li, 2010; Zhao *et al.*, 2011; Sela-Culang *et al.*, 2014, 2015; Krawczyk *et al.*, 2014; Zhao *et al.*, 2014; Sela-Culang *et al.*, 2015) require the associated antibody along with the antigen as input, and output a specific epitope that binds to the antibody but not others. For example, if the structures/sequences of the antigen contained in PDB 1A2Y are specified, these methods are intended to predict only the residues in part I+IV (of Figure 1(a)), but not other residues. The results obtained by these methods are trusty and biologically meaningful. However, the pairing rule of the input data markedly narrows down their applicability. For the antibody-agnostic approach, it has been intensively studied due to the large volume of available antigen data (Esmaielbeiki *et al.*, 2016). However, antibody-agnostic methods, to some extent, can be misleading as there may exist multiple and separated epitopes in an antigen, and even overlapping epitopes (Greenbaum *et al.*, 2007; Zhao *et al.*, 2012). See again the example in Figure 1. These methods can only detect partial or the whole set of *epitopic residues*, i.e., the parts labeled by I, II, III and IV in the panel (a), without the distinction of the residues belonging to the same or different epitopes.

[*]to whom correspondence should be addressed.

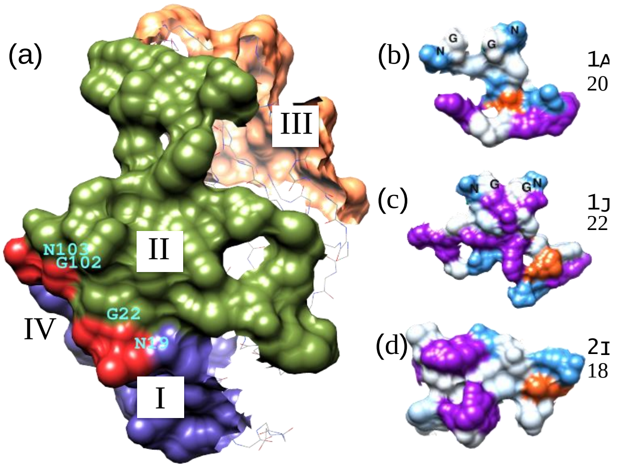

Fig. 1. The steric locations of the three representative epitopes located surface of the Hen egg white lysozyme. The panel (a) shows the lo c of the three epitopes, colored in dark slate blue (I), dark olive gre e and salmon (III) depending on which antibody it interacts with, i.e., 1A2Y, 1J1O and 2IFF. The four firebrick residues, N19, G22, G102 and N103 that are labeled as part IV, are the overlapping residues between the two epitopes I and II. Panel (b) to (d) are the front view of the three epitopes having hydrophobic residues shown in orange, the hydrophilic residues in dodge blue, the positive and negative charged residues in purple, and the residues with median to low hydrophobicity are shown in white. The images are produced by using Chimera (Pettersen *et al.*, 2004).

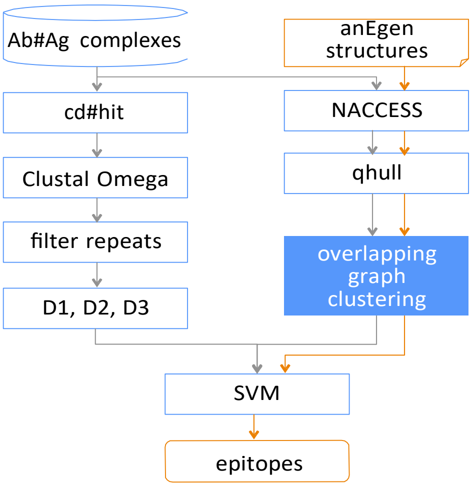

Fig. 2. The diagram of our data preparation, model construction and evaluation. The model construction is comprised by the solid lines connected components, while the model evaluation is illustrated by the dash lines connected components.

The existing computational approaches mainly focus on the single or multi-separated epitope prediction although the emerging of overlapping epitopes has been reported occasionally. Studies have shown that the immune response related proteins are rich of multiple and overlapping binding sites (Zhao *et al.*, 2012). Particularly, the Proteasome beta subunit protein has 7.6 binding sites on average, and the hen egg white lysozyme has up to 43 interacting partners accounting for 9 different epitopes, among which three are visually distinguishable (Figure 1). Another recent study on the D8 antigen has unveiled that it possesses four epitopes with two of them heavily overlapped (Sela-Culang *et al.*, 2014). These findings suggest that the existing antibody-agnostic methods may only detect antigenic residues (the constituent residues of epitopes) but not epitopes *per se*. Although a few methods can discover multiple epitopes (Zhao *et al.*, 2012; Ponomarenko *et al.*, 2008; EL-Manzalawy *et al.*, 2008), their performance is far from satisfactory. Furthermore, none of them is able to detect overlapping epitopes, particularly the sharing antigenic residues, e.g., the four residues colored in red in Figure 1.

We introduce a novel antibody-agnostic epitope prediction algorithm that is able to detect single, multi-separated, as well as overlapping epitopes from antigens, assuming the data of the corresponding antibodies are not given. Our algorithm, named Glep (short for overlapping graph clustering-based B-cell epitope predictor), achieves the goal by three major steps: construct a residue-level graph of an antigen; partition the graph into subgraphs, which are further expanded into overlapping ones using a new idea based on the term frequency-inverse document frequency (Rajaraman and Ullman, 2011) featured similarity; and classify each expanded subgraph as an epitope or a non-epitope by SVM (Chang and Lin, 2011). Figure 2 depicts the whole diagram of the data preparation, model construction and evaluation steps.

2 MATERIALS AND METHODS

2.1 Data preparation

Antibody-antigen complexes are retrieved from the protein data bank (PDB) (Berman *et al.*, 2000). Per the prevalent data preparation procedures (Zhao *et al.*, 2012), the complexes are selected via the following criteria: (i) the macromolecular type is protein only, i.e., those containing DNA or RNA are excluded; (ii) the length of an antigen sequence is greater or equal to 30 residues; (iii) the complex contains at least one asymmetric unit composed of antibody-antigen interacting quaternary structure; (iv) the X-ray resolution is not worse than 3Å; and (v) the complex title contains at least one of the terms: "antibody", "Fab", "Fv", and "VHH". Under these criteria, 808 antibody-antigen complexes are collected. Some illusive interactions are excluded by removing those having epitope size smaller than 5 (Ponomarenko and Bourne, 2007). The duplicate antigens are also removed using cd-hit (Li and Godzik, 2006) under the minimum sequence similarity of 0.9 and the minimum epitope similarity of 0.8 (Zhao *et al.*, 2012).

The epitope similarity is calculated in three steps: (i) determine the epitopic residues from each antibody-antigen complex by using the maximum Euclidian distance of 4Å as suggested by (Kringelum *et al.*, 2012; Sweredoski and Baldi, 2008; Zhao and Li, 2010); (ii) align the antigen sequences within each group generated by cd-hit through Clustal Omega (Sievers *et al.*, 2011); and (iii) map the epitopic residues to the consensus sequence (produced from the alignment of the multiple antigen sequences by voting), and calculate the pair-wise epitope similarity using $|e_1 \cap e_2|/min(|e_1|, |e_2|)$, where e_1 and e_2 are two epitopes having positions calibrated, and $|e|$ is the number of residues for epitope e.

As a result, 205 groups containing 258 PDB complexes are obtained, and they are further divided into three datasets: (i) D1: single epitope antigens—each antigen only has one epitope, which is used to evaluate the generic epitope prediction methods; (ii) D2: multi-separated epitope antigens—each antigen has more than one epitopes and none of them overlaps with another, which is used to evaluate traditional multi-epitope prediction methods; and (iii) D3: overlapping epitope antigens—each antigen contains two, or more

Table 1. Features used to factorize graphs.

feature group	metrics	citation
graph density		Barabasi A.L., et al. (Barabasi and Oltvai, 2004)
degree statistics	mean, variance, median, maximum	Barabasi A.L., et al. (Barabasi and Oltvai, 2004)
degree correlation statistics	mean, variance, and maximum	Stelzl U., et al. (Newman, 2002)
clustering coefficient statistics	mean, variance, and maximum	Barabasi A.L., et al. (Barabasi and Oltvai, 2004)
topological coefficient	mean, variance, and maximum	Stelzl U., et al. (Stelzl and et al., 2005)

epitopes having at least one overlapping epitope, which is used to assess the performance of the so called generalized epitope prediction method. D1 consists of 163 antigens, D2 consists of 21 antigens containing 42 separated epitopes, and D3 consists of 21 antigens containing 53 epitopes. More details of these datasets are shown in supplementary Table S1, S2 and Table 2.

2.2 Glep: steps for construction and prediction

Major steps in Glep include surface residue graph construction, overlapping subgraph clustering and subgraph classification. Details of each step are presented as follows.

2.2.1 Step 1: Surface residue graph construction Since epitopic residues are located at the surface of an antigen, only these accessible ones are used to construct the graph G. The nodes of G are the accessible residues, and there is an edge between two nodes if the two residues have a Euclidean distance less than 6Å. Other than directly building the surface graph from the residues, we construct the graph using the accessible heavy atoms (carbon, nitrogen, or Oxygen) at the initial step, and then upgrade it into a residue-level graph by removing the edges within the same residues and those duplicate edges between two residues. An atom is considered as accessible if its accessible surface area (ASA) is greater than $10\mathring{A}^2$ (Zhao *et al.*, 2012). An ASA is computed using NACCESS (Hubbard and Thornton, 1992) with the default probe size. Then the graph G is constructed via Qhull (Barber *et al.*, 1996), a software tool which has implemented the Delaunay triangulation rule (Huan *et al.*, 2004). It is a tool widely used to construct protein surface graphs.

2.2.2 Step 2: overlapping subgraph clustering Overlapping subgraph clustering has been widely used to detect interacting communities, such as social networks (Goldberg *et al.*, 2010), protein-protein interaction networks, and metabolic pathways (Macropol *et al.*, 2009). These problems are intrinsically the same as the prediction of overlapping epitopes from an antigen. On top of these basic ideas, we introduce a novel subgraph clustering algorithm to detect single, multi-separated, and overlapping epitopes simultaneously for an antigen.

This algorithm has two important components:

1. *Seed detection*. Unlike existing approaches that use single node, a set of nodes, or a clique as a seed (Palla *et al.*, 2005; Ding *et al.*, 2016), we consider the partitions of a graph as the seeds. Graph partitioning is a well investigated area, and many renowned tools are available (Buluç *et al.*, 2013). To make our algorithm more flexible, we have designed a framework that can adapt to many graph partitioning approaches instead of just a single built-in method. This framework takes a graph and a partitioning approach as the input parameters, and outputs the partitions as the seeds. In this work, we use random walk (Lovász, 1996) to partition the whole surface residue graph into isolated subgraphs for expansion. The learned minimum weight of the edges for the graph partitioning is 0.035.
2. *Seed expansion*. Formally, let $G = \{G_1, G_2, \cdots, G_m\}$ be the set of seeds (partitioned isolated subgraphs) generated from the graph G. To expand a seed G_x, we first factorize it into a 14-dimensional feature vector denoted by $F = (f_1, f_2, \cdots, f_{14})$. The features include graph density, degree statistics, degree correlation statistics and cluster coefficient statistics (Table 1; see more details in the supplementary notes). Then the seed G_x can be factorized into a vector $F_{G_x} = (f_{G_x,1}, f_{G_x,2}, \cdots, f_{G_x,14})$. Suppose the set of nodes and edges of G_x are $V(G_x)$ and $E(G_x)$, respectively, then $f_{G_x,1}$, the graph density, is computed using

$$f_{G_x,1} = 2|E(G_x)|/(|V(G_x)| \cdot (|V(G_x)| - 1)).$$

The values of other features are calculated by the definitions in the supplementary notes. Since the values of the features have different ranges, they are further calibrated by

$$f^{I}_{G_i,j} = \frac{f_{G_i,j} - min(f_{G_x,j}) + E_j}{max(f_{G_x,j}) - min(f_{G_x,j}) + E_j},$$

where $x \in \{1, 2, \cdots, m\}$ having m number of seeds obtained from G, and E_j is the pseudo value for feature f_j, which is set as the minimum value of the feature by default. Based on the calibrated feature vectors, the TF-IDF (short for term frequency-inverse document frequency) score is computed for each feature using

$$\text{TF-IDF}(f_{G_i,j}) = \frac{f_{G_i,j}}{\sum_{x=1}^{14} f_{G_i,x}} \cdot \log \frac{\sum_{x=1}^{m} f_{G_x,j}}{f_{G_i,j}}.$$

Thus, the feature vector F_{G_x} can be transformed into

$$\text{TF-IDF}(F_{G_x}) = (\text{TF-IDF}(f_{G_x,j}) : j \in (1, \cdots, 14)).$$

From the TF-IDF-transformed feature vectors, each factorized seed can be expanded as follows. Let G^{I}_{x} be the new seed after a node v is added to the seed G_x, i.e.,

$$\begin{aligned} V(G^{I}_{x}) &= \{v\} \cup V(G_x) \\ E(G^{I}_{x}) &= E(G_x) \cup \{\text{edge}(v_i, v) : \\ &\quad v_i \in V(G_x) \wedge \\ &\quad edge(v_i, v) \in E(G)\}. \end{aligned}$$

To determine whether G^{I}_{x} is acceptable, i.e., whether v can be added to G_x, we calculate the similarity between G^{I}_{x} and G_x using

$$\text{sim}(G^{I}_{x}, G_x) = \frac{\text{TF-IDF}(F_{G^{I}_{x}}) \cdot \text{TF-IDF}(F_{G_x})}{||\text{TF-IDF}(F_{G^{I}_{x}})|| \cdot ||\text{TF-IDF}(F_{G_x})||}.$$

In case the similarity is greater than the predefined minimum value s_0, the seed G_x is expanded by v. The value s_0 is learned from the training data. The seed cluster is expanded layer-by-layer so long as the similarity score is satisfied.

Our subgraph clustering algorithm differs from the existing approaches at two important aspects: (i) the partition itself is considered as the seed; and (ii) the TF-IDF score is used to determine the seed expansion. Further details about overlapping subgraph clustering can be found at (Fortunato, 2010; Amelio and Pizzuti, 2014).

Note that the seed expansion is the key to detecting residues that are shared by multiple epitopes; see Figure 1. The existing approaches place each residue to an epitope, a non-epitope, or different epitopes exclusively.

However, our seed expansion allows a residue to be shared by multiple epitopes. More interestingly, we exploit the idea of TF-IDF to determine whether each residue should be assigned to an epitope or not in a local manner (layer by layer), which is in line with the specificity of antibody-antigen interaction.

2.2.3 Step 3: subgraph classification Our Step 2 described above can divide the antigen surface graph into expanded subgraphs possibly containing both epitopic ones and non-epitopic ones. It remains to determine which ones are epitopic. Following the study (Zhao *et al.*, 2012), the graphlets having size ≤ 3 are used as features to perform protein-domain graph classification, including single residue, residue pair, and residue triplet. That is, there are a total of 1770 (=$C_{20}^1 + C_{21}^2 + C_{22}^3$) features. Per the study, a subset of 144 features that maximize the validation f-score are selected for classification by using LIBSVM (Chang and Lin, 2011).

2.3 Glep: summary for epitope prediction

Four major steps are taken by Glep to detect epitopes from an antigen with structure information available: (i) select accessible heavy atoms from the antigen via NACCESS (Hubbard and Thornton, 1992) under the parameter ASA≥10Å^2; (ii) construct an atom-level graph for the selected atoms by Qhull (Barber *et al.*, 1996) and upgrade it into a residue-level graph; (iii) conduct overlapping subgraph clustering for the graph using our newly designed algorithm; and (iv) classify each subgraph as epitope or non-epitope using LIBSVM (Chang and Lin, 2011).

3 RESULTS AND DISCUSSIONS

3.1 Evaluation metrics

Measures *f-score*, *recall* and *precision* are used to assess the prediction performance, which are defined as

$$\text{recall} = TP/(TP + FN)$$

$$\text{precision} = TP/(TP + FP)$$

$$\text{f-score} = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}},$$

where TP is the number of epitopic residues that have been correctly called, FN is the number of epitopeic residues that have missed out, and FP is the number of non-epitopic residues that are wrongly called.

Note that, specificity and accuracy are not included as TN, the number of non-epitopic residues that are correctly identified, is usually 12 times greater than TP on average (Zhao *et al.*, 2011). Thus, specificity and accuracy are not very informative to understand the prediction performance. Instead, f-score is the most informative and meaningful measurement. In addition, all the performance is computed based on the whole antigen sequence instead of the antigen surface residues.

3.2 Superior performance on the single-epitope detection

Glep was applied to all the antigens in D1 to predict the epitopes. On average, the f-score is 0.579±0.127, recall is 0.518±0.169, and precision is 0.716±0.153. Detailed results are shown in Table S1 of the supplementary notes.

To compare the performance between Glep and the state-of-art methods, we carried out epitope prediction on D1 using BCPREDS (EL-Manzalawy *et al.*, 2008), DiscoTope 2.0 (Kringelum *et al.*, 2012), Ellipro (Ponomarenko *et al.*, 2008), EpiPred (Krawczyk *et al.*, 2014) and BepiPred (Jespersen *et al.*, 2017), under their default parameters if applicable. The f-scores of Glep are better than all these literature methods for 122 of the total 163 epitopes. Figure 3 and Figure S1 present the distribution of the f-scores, recalls and precisions for all the predictors. Clearly, the f-score of Glep is remarkably higher than that of other approaches. We have observed (from Figure S1) that the performance superiority of Glep is mainly attributed to its very high precision. That is, the epitopes determined by Glep contains very small portion of non-epitopic residues, sometimes no common residues. Although the averaged recall generated by Glep is not so striking, it is often better than, at least competitive to, the literature methods. In fact, the averaged f-score produced by BCPREDS, BepiPred, DiscoTope 2.0, Ellipro, EpiPrd and SEPPA 2.0 are are 0.234±0.199, 0.090±0.117, 0.167±0.179, 0.361±0.182, 0.353±0.197, 0.269±0.157; the averaged recalls generated by these approaches are 0.221±0.200, 0.201±0.231, 0.271±0.315, 0.533±0.267, 0.484±0.281, 0.62±0.3; and the averaged precisions obtained from these approaches are 0.264±0.222, 0.066±0.094, 0.178±0.224, 0.372±0.27, 0.291±0.172, 0.185±0.12. On average, the *lift* of f-score by Glep on top of BCPREDS, BepiPred, DiscoTope 2.0, Ellipro, EpiPred and SEPPA 2.0 is 147%, 543%, 247%, 60%, 64% and 115%, respectively. Here, *lift* is given by $(x - y)/y$, where x and y are the two values to be compared. More detailed comparison results are presented in Table S1.

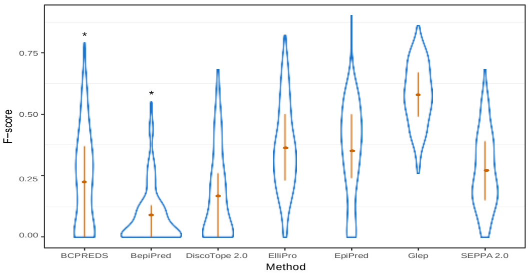

Fig. 3. The f-score comparison of epitope prediction by applying Glep, BCPREDS, BepiPred, DiscoTope 2.0, Ellipro, EpiPred and SEPPA 2.0 on dataset D1. The lower and upper end of the red line within each plot indicate the 25th and 75th percentile, respectively. The plots marked by "*" are obtained solely from antigen sequence.

3.3 Accurate detection of multi-separated epitopes from an antigen

Although predicting multiple epitopes is just a small step ahead from the single epitope prediction, it has much more biological significance. Firstly, it is in line with the principle of context-awareness of epitope binding (Zhao *et al.*, 2011). Secondly, it can, to a certain extent, identify epitopes rather than just antigenic residues, which can be used to guide practical applications, such as vaccine design.

To assess the performance of our algorithm for the detection of multi-separated epitopes from an antigen, we applied Glep to dataset D2. Glep achieved an excellent performance. It can detect all the epitopes from the 21 antigens in D2, with an averaged f-score 0.560±0.114, recall 0.523±0.145, and precision 0.655±0.161. Detailed performance is presented in Table S2. Some highlights of the performance are as follows. The pair-wise sequence similarity of

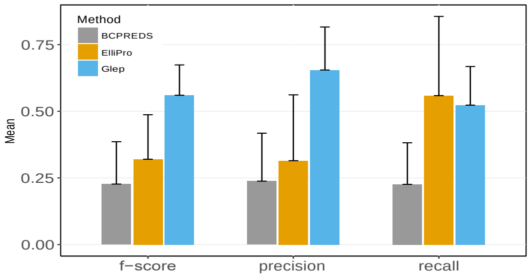

Fig. 4. The performance comparison of separated multi-epitope prediction by applying Glep, BCPREDS, and Ellipro on dataset D2.

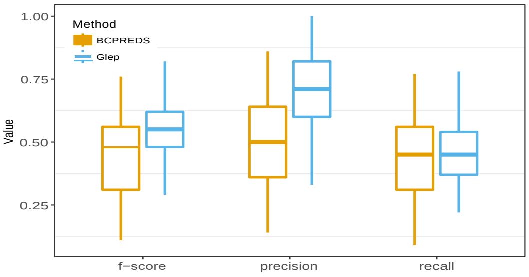

Fig. 5. The performance comparison of overlapping multi-epitope prediction by applying Glep and BCPREDS on dataset D3.

1CZ8 chain W and 2FJG chain W is 100%. However, their epitopes are quite different. The epitope residues of 1CZ8 chain W are TYR45, LYS48, GLN79, ILE80, MET81, ARG82, ILE83, HIS86, GLN87, GLY88, GLN89, HIS90, ILE91, GLY92, GLU93, MET94, while the epitope residues of 2FJG chain W that are mapped to 1CZ8 chain W are PHE17, MET18, TYR21, GLN22, TYR25, ASN62, ASP63, GLU64, LEU66, PRO106. Glep can detect 11 of the 16 residues for the epitope on 1CZ8 chain W, introducing only 2 non-epitopic residues. For 2FJG, Glep can detect 6 of the 10 epitopic residues, introducing only 3 non-epitopic residues.

We have also conducted experiments using BCPREDS and ElliPro on D2 for multi-epitope prediction. For BCPREDS, the parameters of "flexible epitope length" and "non-overlapping epitope prediction" are set; while for ElliPro the default parameters are used as it can predict multiple epitopes *per se*. BCPREDS has an empty result on 6 epitopes, while ElliPro is unable to detect 2 epitopes. In contrast, Glep can detect all of them. In addition, the f-score on 37 of the 42 epitopes generated by Glep are much higher than BCPREDS and ElliPro. On average, the lift of f-score by Glep from BCPREDS and ElliPro is 147% and 75%, respectively. Figure 4 presents the summarized performance of recall and precision for all the approaches besides f-score. Obviously, the performance of Glep is markedly better than that of the other two. More detailed results are reported in Table S2.

Experiments are not carried out using BepiPred, DiscoTope 2.0, EpiPrd and SEPPA 2.0, as they can only predict single epitopes from an antigen.

3.4 Accurate detection of overlapping epitopes from an antigene

Existence of overlapping epitopes in antigens has been known for a long time, but reported in detail only recently (Narayan *et al.*, 2011; Zhao *et al.*, 2012; Faleri *et al.*, 2014; Zhao *et al.*, 2012; Abdiche *et al.*, 2017). BCPREDS is the only computational method which has an option to predict overlapping epitopes from an antigen, but has exclusive parameter settings comparing with separated epitope prediction. Our method has a better performance than BCPREDS. Experiments are carried out on dataset D3, and the detailed results are presented in Table 2. Glep can successfully detect all the epitopes from the 21 antigens, including the overlapping epitopes. The averaged f-score, recall and precision are 0.549±0.123, 0.483±0.166, and 0.700±0.163, respectively. BCPREDS has a much lower performance with an averaged f-score at only 0.449±0.156, recall at 0.433±0.170, and precision at 0.496±0.175. The lift of f-score, recall or precision by Glep from BCPREDS is 22.3%, 11.5%, and 41.1%, respectively. Comprehensive comparison of the performances between these two approaches is shown in Figure 5. The paired t-test p-value of f-score, recall and precision between Glep and BCPREDS are 1.7e-4, 1.1e-1 and 7.1e-8, statistically indicating that Glep significantly outperforms BCPREDS from all the performance measurements.

3.5 Simultaneous detection of both overlapping and separated epitopes from an antigen

There are cases of antigens which contain epitopes separated from the cluster of overlapping epitopes (Zhao *et al.*, 2012; Faleri *et al.*, 2014). Figure 6 shows an example. None of existing approaches can detect both separated and overlapping epitopes simultaneously. Although BCPREDS can make predictions for non-overlapping and overlapping epitopes, it achieves this goal by different built-in procedures having different parameters. That is, it can only make predictions for overlapping epitopes or non-overlapping epitopes independently, other than simultaneously. Our Glep works very well to deal with these challenging situations.

We use the antigen contained in 1JHL chain A to demonstrate the effectiveness of Glep in the simultaneous prediction of overlapping epitopes as well as separated epitopes. Table 3 shows that the four antigen sequences are highly similar (in fact three of them—1p2c, 1bvk and 1dqj—are the same, while the other is slightly different), while their epitopes are distinguishable. Particularly, the epitope of 1p2c chain C is totally different from the others, while the rest three overlap with each other; see Figure 6.

Results in Table 2 suggest that Glep can successfully detect the four epitopes with very high accuracy (the f-score of detecting epitopes within 1p2c_BA_C, 1jhl_HL_A, 1dqj_BA_C and 1bvk_BA_C are 0.79, 0.67, 0.61 and 0.38, respectively). Specifically, Glep can detect 13 of the 18 residues of the separated epitope 1p2c_BA_C with only introducing 2 non-epitopic residues; detect 8 of the 11 residues of the epitope 1jhl_HL_A with introducing 5 non-epitopic residues; detect 10 of the 21 residues of the epitope 1dqj_BA_C with introducing 2 non-epitopic residues; and detect 5 of the 17 residues of the epitope 1bvk_BA_C with introducing 4 non-epitopic residues. More importantly, the overlapping epitope residues can be fished out for different epitopes. For instance, there are 4 overlapping residues between epitope 1dqj_BA_C and 1bvk_BA_C (positions are 18, 19, 102 and 103), and Glep can detect 3 of them (18, 102 and 103) for both epitopes.

Table 2. The performance of overlapping multi-epitope prediction by Glep and BCPREDS on dataset D3.

PDB†	Glep			BCPRED S			PDB	Glep			BCPRED S		
	f-score	recall	precision	f-score	recall	precision		f-score	recall	precision	f-score	recall	precision
3sqo_HL_A	**0.64**	0.50	0.88	0.50	0.50	0.50	2adf_HL_A	0.50	0.40	0.67	**0.55**	0.53	0.57
2xtj_DB_A	**0.59**	0.47	0.80	0.58	0.53	0.64	1fe8_JN_A	**0.40**	0.36	0.44	0.24	0.27	0.21
3h42_HL_B	**0.48**	0.45	0.50	0.28	0.23	0.36	1fe8_IM_B	**0.55**	0.42	0.80	**0.55**	0.47	0.64
1a14_HL_N	**0.62**	0.53	0.75	0.45	0.41	0.50	1fe8_IM_C	**0.77**	1.00	0.62	0.21	0.40	0.14
1nca_HL_N	**0.65**	0.48	1.00	0.29	0.24	0.36	3hi6_HL_A	**0.40**	0.27	0.75	0.28	0.23	0.36
4lsu_HL_G	**0.46**	0.34	0.71	0.41	0.29	0.71	3eoa_HL_I	0.55	0.43	0.75	**0.57**	0.57	0.57
4h8w_HL_G	0.39	0.37	0.41	**0.61**	0.53	0.71	4g6m_HL_A	**0.52**	0.42	0.67	0.48	0.42	0.57
4om1_HL_G	0.49	0.38	0.71	**0.52**	0.38	0.86	4g6j_HL_A	**0.74**	0.67	0.82	0.17	0.14	0.21
3gbn_HL_A	**0.43**	0.50	0.38	0.40	0.67	0.29	1p2c_BA_C	**0.79**	0.72	0.87	0.00	0.00	0.00
3lzf_HL_A	**0.44**	0.32	0.75	0.36	0.32	0.43	1bvk_BA_C	0.38	0.29	0.56	**0.39**	0.35	0.43
4py8_IJ_A	**0.50**	0.50	0.50	0.42	0.50	0.36	1dqj_BA_C	**0.61**	0.48	0.83	0.29	0.24	0.36
3hi1_HL_G	**0.55**	0.45	0.69	0.41	0.35	0.50	1jhl_HL_A	**0.67**	0.73	0.62	0.32	0.36	0.29
2i5y_HL_G	**0.67**	0.67	0.67	0.34	0.33	0.36	3nfp_HL_I	**0.56**	0.50	0.64	0.38	0.33	0.43
2ny7_HL_G	0.41	0.26	1.00	**0.43**	0.35	0.57	3iu3_HL_I	0.48	0.35	0.80	**0.49**	0.39	0.64
3idx_HL_G	0.58	0.44	0.88	**0.60**	0.56	0.64	1tqb_BC_A	**0.82**	0.78	0.88	0.56	0.50	0.64
2ny5_HL_G	**0.48**	0.54	0.44	0.30	0.31	0.29	2w9e_HL_A	0.69	0.60	0.82	**0.76**	0.73	0.79
2xwt_AB_C	**0.55**	0.48	0.65	0.15	0.11	0.21	4h88_HL_A	0.59	0.42	1.00	**0.69**	0.75	0.64
3g04_BA_C	**0.34**	0.22	0.83	0.11	0.09	0.14	4bz2_HL_A	**0.59**	0.47	0.80	0.58	0.53	0.64
3bn9_DC_B	**0.48**	0.33	0.89	0.26	0.21	0.36	4bz1_HL_A	0.42	0.33	0.57	**0.54**	0.58	0.50
3nps_BC_A	**0.50**	0.44	0.58	0.26	0.20	0.36	4l5f_HL_E	**0.50**	0.46	0.55	0.44	0.46	0.43
3so3_CB_A	**0.54**	0.44	0.69	0.31	0.24	0.43	4al8_HL_C	**0.56**	0.64	0.50	**0.56**	0.64	0.50
3ma9_HL_A	0.44	0.32	0.73	**0.51**	0.40	0.71	3u2s_HL_G	**0.75**	0.90	0.64	0.50	0.60	0.43
2cmr_HL_A	0.40	0.28	0.71	**0.63**	0.56	0.71	4dqo_HL_C	**0.63**	0.56	0.71	0.52	0.67	0.43
4hc1_HL_A	**0.80**	0.75	0.86	0.67	0.62	0.71	4lu5_IM_A	0.54	0.47	0.64	**0.55**	0.53	0.57
4hcr_HL_A	**0.62**	0.44	1.00	0.25	0.22	0.29	4m1g_HL_A	0.59	0.62	0.57	**0.74**	0.77	0.71
3bgf_HL_S	0.48	0.40	0.60	**0.55**	0.53	0.57	4m1g_HL_B	**0.67**	0.73	0.62	0.40	0.45	0.36
2dd8_HL_S	0.29	0.26	0.33	**0.61**	0.53	0.71							

† The first part separated by "_" is the PDB code, the middle part contains the antibody heavy and light chain name, and the last part is the antigen chain name. The antibody-antigen complexes within the same block having the same antigen but different epitopes, i.e., the antigen sequence similarity is no less than 0.9, while the epitope similarity is no larger than 0.8.

Table 3. The details of overlapping and separated epitopes in antigen 1JHL chain A.

PDB[a]	antigen size	antigen similarity[b]	epitope similarity[c]	position of epitope residues[d]
1jhl_HL_A	129	92.25%	63.64%	21 23 103 106 112 113 116 117 118 119 121
1p2c_BA_C	129	92.25%	0	41 43 45 46 47 48 49 50 51 53 65 66 67 68 70 79 81 84
1bvk_BA_C	129	92.25%	63.64%	18 19 22 23 24 27 102 103 116 117 118 119 120 121 124 125 129
1dqj_BA_C	129	92.25%	23.53%	14 15 16 18 19 20 21 62 63 73 75 77 89 93 96 97 98 100 101 102 103

[a] The first part separated by "_" is the PDB code, the middle part contains the antibody heavy and light chain name, and the last part is the antigen chain name. [b] Antigen pair-wise sequence similarity, computed by using cd-hit (Li and Godzik, 2006). Here we only show the *minimum* similarity between the interest antigen and the rest. In fact, the chain "C" of 1p2c, 1bvk and 1dqj are the same. [c] Epitope similarity, determined by $|e_1 \cap e_2|/min(|e_1|, |e_2|)$, where $|e|$ is the size of e. We only show the *maximum* similarity between the interest epitope and the rest. [d] For ease of understanding, the position is the *calibrated* position other than the raw position of each complex, i.e., the four antigen sequences are aligned by using Clustal Omega (Sievers *et al.*, 2011) at first, then the position of epitope residues are mapped to the alignment later.

3.6 Marginal impact of unbound antigen on epitopes detection by Glep

Glep is much better than existing approaches on epitope prediction that has been validated on bound antigens, i.e., D1, D2 and D3, while its performance on unbound antigens is unknown. To unveil its power of epitope prediction on unbound antigens, we have collected another data set that is in accordance with the bound antigens in terms of sequence similarity having minimum threshold of 95%, where the similarity is automatically computed via the tool provided by PDB (Berman *et al.*, 2000) . That is, we use the bound antigen sequence to query the PDB to retrieve the unbound antigens satisfying the similarity threshold. As a result, 110 unbound antigens are obtained.

Based on the unbound antigen, we carried out epitope prediction by using Glep. The overall f-score, recall and precision is 0.551±0.154, 0.483±0.179, and 0.696±0.182, respectively. The detailed performance is shown in the supplementary notes. Comparing with the performance produced from the bound antigens, the one obtained from the unbound antigens is slightly smaller. The difference between the two is 0.028 in terms of mean f-score. When it breaks down into the three types, we found that the performance of unbound antigen on D3 is slightly better than

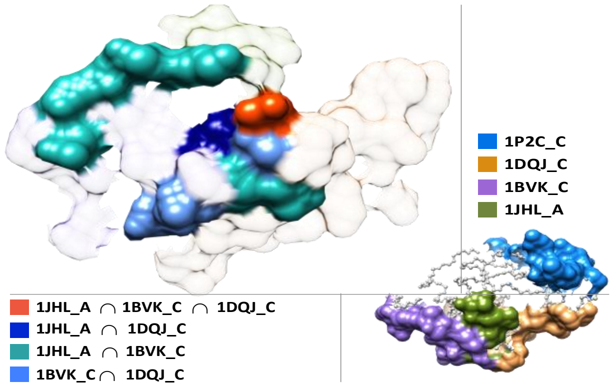

Fig. 6. Overlapping and separated epitopes resided in antigen 1JHL chain A. The whole picture of the four epitopes is shown in the lower right panel with partial covered by each other; the three overlapping epitopes are shown in the top left panel with overlapping residues highlighted in different colors.

that of the bound antigen; see figure 7. We have also calculated the correlation coefficient as well as the p-value between the f-score obtained from bound antigen and unbound antigen. It is clear that there has no significant difference; see figure 7. From the figure we can also see that the larger root mean square deviation (rmsd) between a bound antigen and the corresponding unbound antigen results in higher discrepancy of performance, which is in accordance with the intuition.

4 CONCLUDING REMARKS

Epitope prediction possesses broad significance, attracting many global research teams working on this area. However, they mainly take a one-antigen-one-epitope approach to the prediction of epitopes. Now that multiple, even overlapping epitopes in an antigen have been reported recently, this one-to-one prediction scenario is limited with unsatisfactory performance. To detect epitopes from antigens accurately as well as to reflect the biological facts, we have introduced a novel subgraph clustering algorithm especially for the prediction of overlapping epitopes. This algorithm takes surface residue graph partitions as seeds, and expands all the seeds to cluster overlapping subgraphs through term frequency-inverse document frequency (TF-IDF) featured similarity. Then, the subgraphs are each classified as an epitope or a non-epitope using SVM on delicately selected features. Experiments are conducted on three newly collected antigen datasets. The results have demonstrated that the performance of our newly proposed approach is significantly better than the state-of-art approaches. The lift of averaged f-score of our algorithm from that of the literature methods ranges from 60% to 543% on single epitope prediction. Regarding the averaged f-score of multiple separate epitope prediction, the lift of our algorithm from the second best is as high as 75%. For multiple overlapping epitope prediction, the lift is 22%. Furthermore, the proposed algorithm can detect multiple, both separated and overlapping, epitopes simultaneously with excellent performance.

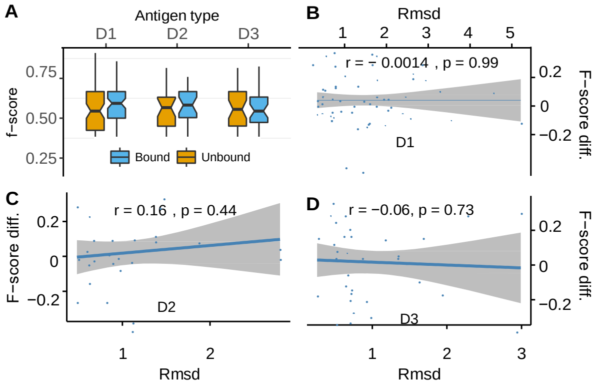

Fig. 7. The performance comparison of Glep between bound antigens and unbound antigens. Panel A is overall performance distribution, while B, C, and D are the detailed distribution on the three datasets. The x-axis of the panel B, C and D is the root mean square deviation (rmsd) between a bound antigen and the corresponding unbound antigen generated by structure alignment, while the y-axis is the difference of f-score obtained from the two.

Funding: This work is collectively supported by the National Natural Science Foundation for Young Scientists of China (Grant No. 31501070), the Natural Science Foundation of Guangxi Province, China (Grant No. 2016GXNSFCA380006), and the Scientific Research Foundation of GuangXi University (Grant No. XGZ150316).

REFERENCES

Abbas,A.K., Lichtman,A.H. and Pillai,S. (2009) *Cellular and Molecular Immunology*. 6th edition,, W.B. Saunders Company.

Abdiche,Y.N., Yeung,A.Y., Ni,I. and et al. (2017) Antibodies targeting closely adjacent or minimally overlapping epitopes can displace one another. *PLOS ONE*, **12** (1), 1–22.

Amelio,A. and Pizzuti,C. (2014) *Overlapping Community Discovery Methods: A Survey*. Vienna: Springer Vienna pp. 105–125.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**, 101–103.

Barber,C.B., Dobkin,D.P. and Huhdanpaa,H. (1996) The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, **22** (4), 469–483.

Berman,H.M., Westbrook,J., Feng,Z. and et al. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28** (1), 235–242.

Buluç,A., Meyerhenke,H., Safro,I., Sanders,P. and Schulz,C. (2013) Recent advances in graph partitioning. *CoRR*, **abs/1311.3144**.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27.

Cooper,G.M. and Hausman,R.E. (2004) *The Cell: A Molecular Approach*. ASM Press.

Ding,Z., Zhang,X., Sun,D. and Luo,B. (2016) Overlapping community detection based on network decomposition. *Scientific Reports*, **6**, 24115.

EL-Manzalawy,Y., Dobbs,D. and Honavar,V. (2008) Predicting flexible length linear B-cell epitopes. *Computational Systems Bioinformatics*, **7**, 121–132.

Esmaielbeiki,R., Krawczyk,K., Knapp,B., Nebel,J.C. and Deane,C.M. (2016) Progress and challenges in predicting protein interfaces. *Brief Bioinform*, **17** (1), 117–131.

Faleri,A., Santini,L., Brier,S. and et al. (2014) Two cross-reactive monoclonal antibodies recognize overlapping epitopes on neisseria meningitidis factor h binding protein but have different functional properties. *FASEB J*, **28** (4), 1644–1653.

Fortunato,S. (2010) Community detection in graphs. *Physics Reports*, **486** (35), 75–174.

Goldberg,M., Kelley,S., Magdon-Ismail,M., Mertsalov,K. and Wallace,A. (2010) Finding overlapping communities in social networks. In *2010 IEEE Second International Conference on Social Computing* pp. 104–113.

Greenbaum,J.A., Andersen,P.H., Blythe,M. and *et. al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recognit*, **20** (2), 75–82.

Huan,J., Wang,W., Bandyopadhyay,D., Snoeyink,J., Prins,J. and Tropsha,A. (2004) Mining Protein Family Specific Residue Packing Patterns from Protein Structure. In *Eighth Annual International Conference on Research in Computational Molecular*

Biology (RECOMB) pp. 308–315.

Hubbard,S.J. and Thornton,J.M. (1992). Naccess V2.1.1 - Solvent accessible area calculations. http://www.bioinf.manchester.ac.uk/naccess/.

Jespersen,M.C., Peters,B., Nielsen,M. and Marcatili,P. (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res,* .

Krawczyk,K., Liu,X., Baker,T., Shi,J. and Deane,C.M. (2014) Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics,* **30** (16), 2288–94.

Kringelum,J.V., Lundegaard,C., Lund,O. and Nielsen,M. (2012) Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLOS Computational Biology,* **8** (12), 1–10.

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

Li,W.L. and Godzik,A. (2006) cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* **22**, 1658–1659.

Lovász,L. (1996) Random walks on graphs: a survey. In *Combinatorics, Paul Erdős is Eighty* vol. 2,. Budapest pp. 353–398.

Macropol,K., Can,T. and Singh,A.K. (2009) RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics,* **10** (1), 283.

Narayan,V., Halada,P., Hernychova,L. and et al. (2011) A multi-protein binding interface in an intrinsically disordered region of the tumour suppressor protein interferon regulatory factor-1. *J Biol Chem,* **286** (16), 14291–303.

Newman,M.E. (2002) Assortative mixing in networks. *Phys Rev Lett,* **89** (20), 208701.

Palla,G., Derenyi,I., Farkas,I. and Vicsek,T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature,* **435** (7043), 814–818.

Pettersen,E.F., Goddard,T.D., Huang,C.C. and et al. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem,* **25** (13), 1605–12.

Ponomarenko,J., Bui,H.H., Li,W., Fusseder,N., Bourne,P.E., Sette,A. and Peters,B. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics,* **9** (1), 514.

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

Qi,T., Qiu,T., Zhang,Q., Tang,K., Fan,Y., Qiu,J., Wu,D., Zhang,W., Chen,Y., Gao,J., Zhu,R. and Cao,Z. (2014) SEPPA 2.0more refined server to predict spatial epitope considering species of immune host and subcellular localization of protein antigen. *Nucleic Acids Research,* **42** (W1), W59.

Rajaraman,A. and Ullman,J.D. (2011) *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA.

Sela-Culang,I., Ashkenazi,S., Peters,B. and Ofran,Y. (2015) PEASE: predicting B-cell epitopes utilizing antibody sequence. *Bioinformatics,* **31** (8), 1313–5.

Sela-Culang,I., Benhnia,M.R., Matho,M.H. and et. al. (2014) Using a combined computational-experimental approach to predict antibody-specific B cell epitopes. *Structure,* **22** (4), 646–657.

Sela-Culang,I., Ofran,Y. and Peters,B. (2015) Antibody specific epitope prediction - emergence of a new paradigm. *Current Opinion in Virology,* **11**, 98–102.

Sievers,F., Wilm,A., Dineen,D. and et. al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol,* **7** (539).

Stelzl,U. and et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell,* **122** (6), 957–968.

Sweredoski,M.J. and Baldi,P. (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics,* **24** (12), 1459–1460.

Zhao,L., Hoi,S.C., Li,Z., Wong,L., Nguyen,H. and Li,J. (2014) Coupling graphs, efficient algorithms and B-cell epitope prediction. *IEEE/ACM Trans Comput Biol Bioinform,* **11** (1), 7–16.

Zhao,L., Hoi,S.C.H., Wong,L., Hamp,T. and Li,J. (2012) Structural and functional analysis of multi-interface domains. *PLOS ONE,* **7** (12), 1–13.

Zhao,L. and Li,J. (2010) Mining for the antibody-antigen interacting associations that predict the B cell epitopes. *BMC Struct Biol,* **10** (Suppl 1), S6.

Zhao,L., Wong,L., Hoi,S.C., Lu,L. and Li,J. (2012) B-cell epitope prediction through a graph model. *BMC Bioinformatics,* **13** (17), S20.

Zhao,L., Wong,L. and Li,J. (2011) Antibody-Specified B-Cell Epitope Prediction in Line with the Principle of Context-Awareness. *IEEE/ACM Trans Comput Biol Bioinf,* **8** (6), 1483–1494.