Antibody-specified B-cell epitope prediction in line with the principle of context-awareness

Liang Zhao Limsoon Wong Jinyan Li*

Abstract—Context-awareness is a characteristic in the recognition between antigens and antibodies, highlighting the reconfiguration of epitope residues when an antigen interacts with a different antibody. A coarse binary classification of antigen regions into epitopes, or non-epitopes without specifying antibodies may not accurately reflect this biological reality. Therefore, we study an antibody-specified epitope prediction problem in line with this principle. This problem is new and challenging as we pinpoint a subset of the antigenic residues from an antigen when it binds to a specific antibody. We introduce two kinds of associations of the contextual awareness: (i) residues-residues pairing preference, and (ii) the dependence between sets of contact residue pairs. Preference plays a bridging role to link interacting paratope and epitope residues while dependence is used to extend the association from one-dimension to two-dimension. The paratope/epitope residues' relative composition, cooperativity ratios, and Markov properties are also utilized to enhance our method. A non-redundant data set containing 80 antibody-antigen complexes is compiled and used in the evaluation. The results show that our method yields a good performance on antibody-specified epitope prediction. On the traditional antibody-ignored epitope prediction problem, a simplified version of our method can produce a competitive, sometimes much better performance, in comparison with three structure-based predictors.

Index Terms—Epitope prediction, context dependence, antibody, antigen.

1 Introduction

A B-cell epitope is a small subset of residues in an antigen sequence which can be recognized by a specific antibody [1]. Epitopes can be categorized into two types: continuous and discontinuous [2]. A continuous epitope consists of residues that are contiguous in the primary antigen sequence, while a discontinuous epitope is formed by residues that are separated, sometimes far away, from each other in the primary sequence but they are close to each other in 3D space through polypeptide folding. For example, the epitope in the antigen chain E of PDB:1ZTX is constituted by five distant stretches: TYR302, SER306 LYS307 ALA308 PHE309, THR330 GLY331 THR332 ASP333, ALA365 THR366 ALA367 ASN368, and GLY389 GLU390 GLN391. All these short segments are required for the recognition by the antibody and thus they are not five epitopes on their own. Due to the importance of identifying epitopes in vaccine design, disease diagnosis and disease therapy [3], intensive efforts have been made to predict both continuous and discontinuous epitopes over the past few decades.

Context-awareness in the recognition between antigens and antibodies [4], [5], also known as context dependence, is the key idea that motivates the whole work in this paper. Context-awareness is a property highlighting that epitope is a dependent entity—An epitope cannot be formed without a corresponding antibody [5]. Therefore, epitope can be detected only at the point when an antigen interacts with a specific antibody area which is called the paratope of the antibody. A more intriguing facet of such binding mechanism is that the cluster of residues that acts as an epitope under one set of circumstances will not necessarily behave as epitope under another set [5]. In other words, an epitope of an antigen may be reconfigured with a different set of residues when a different antibody is presented in the interaction. For instance, the epitope of the prior protein as shown in PDB:1TQB consists of residues GLY127, LEU128, ASN156, TYR158, ARG159, ILE185, LYS188, GLN189, THR191, VAL192, THR193, THR194, THR195, THR196, LYS197, GLY198 and GLU199. However, the epitope of this antigen is reconfigured remarkably when the antibody is changed to another protein as shown in PDB:2W9E. The reconfigured epitope consists of residues GLY145, SER146, ASP147, TYR148, ASP150, ARG151, TYR152, ARG154, GLU155, ASN156, HIS158, ARG159, ASN200, THR202 and LYS207, having only two common residues (ASN156 and ARG159) overlapping with the first epitope. Therefore, a coarse binary classification of antigen regions into epitopes, or non-epitopes without specifying antibodies and paratopes may not accurately reflect biological reality [5].

The aim of this work is to predict the location of the epitope residues when we are given the sequence of an antigen and the heavy chain and light chain sequence of an antibody. Existing methods [6], [7], [8], [9], [10], [11], [12] generally overlooked the property

• *L. Zhao is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798. E-mail: s080011@e.ntu.edu.sg. L. Wong is with the School of Computing, National University of Singapore. E-mail: wongls@comp.nus.edu.sg. J. Li is with the School of Computer Engineering, Nanyang Technological University and with the School of Computing, National University of Singapore. E-mail: jinyan4@gmail.com. * Corresponding author.*

of context-awareness, and studied instead an antibody-ignored B cell epitope prediction problem. Given an antigen, these earlier methods just tell the union of all antigenic residues in this antigen. They were unable to specify which subset of the antigenic residues that can form an epitope binding to an antibody.

In light of this context-awareness, we introduce two kinds of associations: (i) the residues-residues pairing preference between epitopes and paratopes, and (ii) the dependence between two sets of contact residue pairs. Both of them are mined from antibody-antigen complexes for the training of our model. When an antigen-antibody complex without structural information is presented for epitope prediction, the preference association plays a bridging role to link the residues from the paratope to the epitope, and the dependence association is used to infer new set of contact paratope-epitope residue pairs. Our prediction begins by identifying antibody paratope residues. This idea is based on the fact that antibody paratopes are regularly structured and are easy to be identified. The residues of a paratope are usually located in six complementary determining regions (CDRs), namely L1, L2, L3, H1, H2 and H3 [13]. The first three CDRs are in antibody light chain, and the other three are in antibody heavy chain. The six CDRs can be identified by the Chothia CDR definition [14], and the paratope residues can be easily located by the residues' relative composition and cooperativity information. Figure 1 shows a diagram of our prediction method and illustrates how the preference association is used to predict epitope residues and how a dependence association is used to infer a new set of contact residue pairs. As there are still possibly many uncertain residues in the antigen, a semi-supervised Hidden Markov Model (HMM) is proposed to complete the process of epitope prediction.

Our technical ideas differ from traditional ones. Very often, physico-chemical properties, such as individual amino acid's hydrophilicity, backbone flexibility, residue's antigenic propensity, exposed surface area and so on, were used to tackle the problem of predicting continuous epitopes [6], [7], [8], [9]. Single or combinations of different properties including solvent-accessible surface area, spatial information, contact distance and amino acid statistics had been also proposed as features to study the discontinuous epitope prediction problem [10], [11], [12]. Some advanced machine learning approaches were also exploited [15], [16], [17], [18], [19]. Overall, the performance on the traditional antibody-ignored epitope prediction problem was not satisfactory for either continuous or discontinuous epitopes [20], [21], [22].

We name our method **ABepar** short for **A**ntibody-specified **B**-cell **e**pitope **p**rediction through **a**ssociation **r**ules. As mentioned, our method is trained on antibody-antigen PDB complexes, but it can be applied to any antibody-antigen sequence pairs that do not contain structural information. By simplifying the computational steps, our ABepar method can also be used to conduct the traditional epitope prediction when only antigen sequence is provided. The software and supplementary data are available at http://155.69.2.25/~s080011/index.html

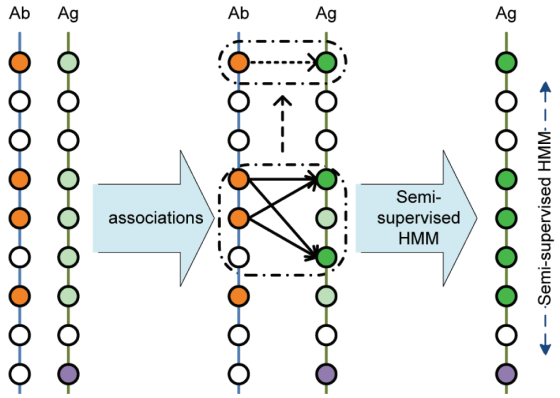

Fig. 1. Antibody-specified epitope identification by using associations and a semi-supervised Hidden Markov Model. Here, the unfilled circles represent non-paratope/epitope residues. The paratope residues are colored by orange, and the epitope residues are colored by light green and purple. We distinguish them by light green and purple because the light green circles are the epitope residues that interact with this specific antibody, while the purple circles are in a different epitope that interacts with another antibody when the environment changes. The solid arrows represent our epitope identification by the preference between paratope and epitope residues, while the dashed arrow and two boxes represent our epitope detection through dependence within two sets of paratope-epitope interacting residue pairs. The circles colored by green are the epitope residues that we can identify. The purple residues are not predicted as restricted by the context awareness.

2 Materials and Methods

2.1 Data set

We compiled a set of 80 non-redundant X-ray crystallographic antigen-antibody complexes with the resolution better than 3.0Å after an exhaustive search from PDB [23]. Each complex in this data set satisfies: (i) it contains three or more polypeptides each with a length $\geq$ 30 amino acids; (ii) it contains antibody variable regions, such as Fab, VHH or Fv fragments; and (iii) the paratope residues of each complex are mainly situated in the six CDRs. Besides, if a structure contains more than one asymmetric units but with no structural difference, only one unit is selected. To remove redundancy, an antigen structural pair-wise alignment was carried out by the combinatorial extension (CE) algorithm [24]. An antigen is redundant if the following criteria are met: (i) The alignment z-score is $\geq$ 4.0; (ii) The root mean square deviation (RMSD) of the alignment is $\leq$ 3Å; (iii) The

proportion of the matching residues against the aligned positions is $\geq$ 80%; (iv) The proportion of overlapping epitope residues is $\geq$ 80% against the epitope residues of the shorter sequence. The first criterion is based on the CE algorithm, while the middle two are suggested by [21].

2.2 Construction of our training model

Given an antibody-antigen PDB complex, the epitope and paratope residues are determined by using a distance threshold 4Å. (This threshold was recommended by [11] as it can capture the epitopes and paratopes with a high precision.) One residue is considered as an epitope residue if there exists an atom of this residue, except hydrogen, that is separated within a distance of 4Å from an atom of an antibody residue except hydrogen. This antibody residue is correspondingly called a paratope residue.

Taking all of the epitopes and paratopes from the training data, we mine the residues-residues pairing preference in these binding sites, as well as the dependency between sets of contact residue pairs. We also compute residue's relative composition and cooperativity values. These compositions and associations are then subsequently used when a pair of antibody-antigen sequences is given for epitope prediction. The detailed flowchart of training and testing is shown in Figure 2. The upper part of the flow chart shows our steps in model construction and the lower part depicts the steps for epitope prediction.

2.2.1 Associations: residues-residues pairing preference and dependence between two sets of contact residue pairs

A residues-residues pairing preference in paratope-epitope interacting complex is defined as a frequent biclique $G = \langle V_b, V_g, E\rangle$, where V_b is a set of vertices representing paratope residues, and V_g is a set of vertices representing epitope residues. E is a set of edges constituted by paratope-epitope contact residue pairs, satisfying the condition $|E| = |V_b| * |V_g|$, i.e. $\forall v_b \in V_b, \forall v_g \in V_g$, $\langle v_b, v_g\rangle \in E$. This concept of residue pairing preference underlies the pairing or full interaction between a group of paratope residues and a group of epitope residues. It extends the common pairing between two single residues, and it can expand the statistical and biological significance of residue pairing as demonstrated in our early work [25] where protein binding hotspots are studied.

Residues-residues pairing preferences are identified from a set of antibody-antigen PDB complexes through the following steps: (i) Determine all paratope and epitope residues of each antibody-antigen complex by using the distance threshold 4Å; (ii) Construct a bipartite graph from the paratope-epitope interacting residue pairs of each complex. The vertices are the paratope and epitope residues, and the edges stand for the contacts between paratope and epitope residues; (iii) Identify the maximal bicliques from every bipartite graph by using the LCM-MBC algorithm [26]; and (iv) select the frequent bicliques from all the antibody-antigen complexes with a minimum occurrence level of 7%.

A dependence between two sets of interacting residue pairs is described by an association rule. It highlights the extent to which one set of interacting residue pairs is dependent on the other set. In other words, if one set of interacting residue pairs is observed in a binding site, then what is the probability that (at which confidence level) the residue pairs in the other set also occur in this binding site. Formally, such an association rule is in the form $\{\langle p_1^o\text{-}e_1^o\rangle, \langle p_2^o\text{-}e_2^o\rangle, \ldots\} \rightarrow \{\langle p_1^i\text{-}e_1^i\rangle, \ldots\}$, where $\langle p\text{-}e\rangle$ represents a contact residue pair in the paratope-epitope binding site. The superscript o means observed paratope-epitope interacting residue pair, while the superscript i means the implied interaction.

Four steps are taken in detecting contact residue pair dependence from a set of antibody-antigen interacting complexes. (i) Identify all the paratope-epitope interacting residue pairs of each complex by using the distance threshold 4Å; (ii) Transform each interacting residue pair $\langle p\text{-}e\rangle$ into an item $I_{p,e}$ by using $I_{p,e} = I_p * 20 + I_e$, where I_p is the index of residue p, and I_e is the index of residue e. A residue index is the position number of this residue in the Kyte and Doolittle's increasing hydropathy index [27]; (iii) Form a transaction for each binding site of one complex with the transformed items; and (iv) Mine all frequent association rules from this transactional data set by an association rule mining software developed by [28]. The support and confidence level were set as 7% and 80% respectively during the dependence mining.

2.2.2 Residue's relative composition and cooperativity

Paratope residue's relative composition is defined as $RC_j^i = P_j^i * 2 * log_2(P_j^i/Q_j^i)$, where RC_j^i is the relative composition of residue j in CDR i, P_j^i is the composition of residue j over the paratope residues in CDR i, and Q_j^i represents the composition of residue j over all residues in CDR i. Epitope residue's relative composition is computed similarly, defined as $RC_j = P_j * 2 * log_2(P_j/Q_j)$, where RC_j is the relative composition of residue j, P_j is the composition of residue j over the epitope residues, and Q_j represents the composition of residue j over all antigen residues.

Regarding paratope residue's cooperativity, it is defined as a ratio $CO_{jk}^i = (P_{jk}^i/Q_{jk}^i)$, where CO_{jk}^i is the cooperativity of residue pair jk with respect to CDR i, P_{jk}^i is the composition of residue pair jk over the paratope residues in CDR i, and Q_{jk}^i is the composition of residue pair jk over all residues in CDR i. Epitope residues' cooperativity is calculated similarly by $CO_{jk} = (P_{jk}/Q_{jk})$, where CO_{jk} is the cooperativity of residue pair jk within antigens, P_{jk} is the composition of residue pair jk over the epitope residues and Q_{jk} is the composition of residue pair jk over all antigen residues.

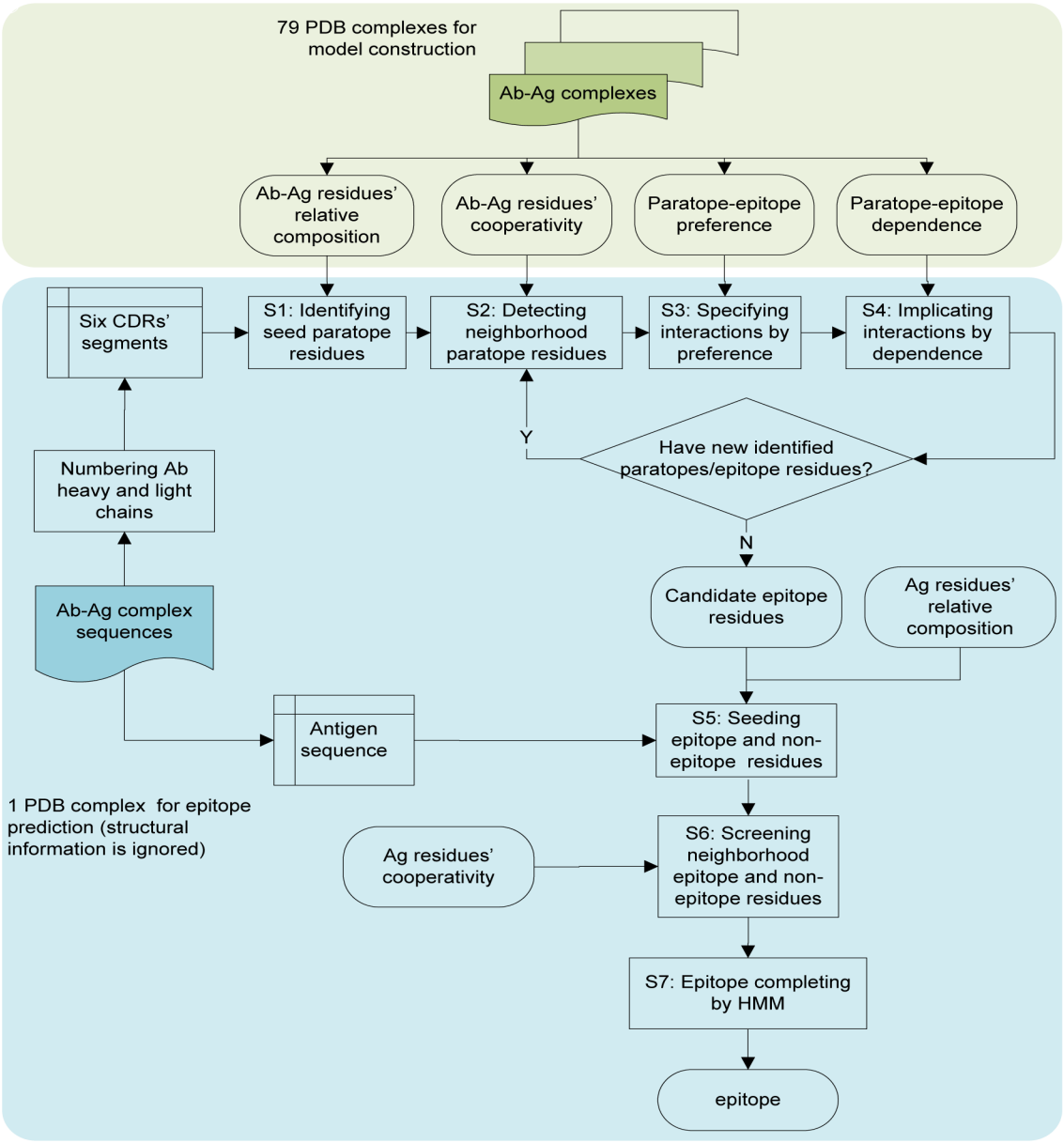

Fig. 2. Flow chart of model construction and epitope prediction.

Here residue pair specifies the paired neighborhood residues.

2.3 Sequence-based epitope prediction

Using the model components constructed from the training data of antibody-antigen PDB complexes as described in the above subsection, the epitope in a given antigen sequence can be predicted through the following steps when an antibody sequence is also given:

- Identifying seed paratope residues by using the paratope residues' relative composition;
- Detecting neighborhood paratope residues of the seeds by using the paratope residues' cooperativity ratios;
- Specifying candidate epitope residues by using the paratope-epitope interacting preference;
- Inferring new sets of paratope-epitope interacting residue pairs from the observed contact residue pairs by dependence;
- Seeding epitope residues from the candidate epitope residues by utilizing the epitope residues' relative composition;
- Screening out neighborhood residues of the seeds by using the epitope residues' cooperativity ratios;
- Completing epitope identification by using a semi-supervised Hidden Markov Model.

2.3.1 Identifying seed paratope residues

Given an antibody-antigen complex without structure information, the antibody heavy chain and light chain are numbered by the modified-Chothia numbering scheme

[29], then the six segments of CDRs are determined according to the Chothia CDR definition [14].

Every residue within the six CDRs is tested against a CDR-dependent paratope residue's relative composition threshold. Residues that meet the constraint are marked as seed paratope residues. Empirically, the thresholds are set to allow only the top three types of residues to pass the test.

2.3.2 Detecting neighborhood paratope residues of the seeds

Having identified the seed paratope residues, the neighborhood paratope residues of these seeds can be screened out by using the paratope residues' cooperativity ratios. More exactly, if a residue i has been marked as a paratope residue, then the cooperativity between its left neighbor i-1 and itself is tested against a CDR dependent threshold T_{co}. The left neighbor residue i-1 is marked as a paratope residue if and only if $CO_{i-1,i} \geq T_{co}$. This test is also applied to the right neighbor of i. Both the left and right neighbors are required to be in the same CDR as i's. Usually, the threshold is set to be such a proper number that only the top ten percent cooperativity ratios of the pair-wise neighborhood residues can pass the test.

2.3.3 Specifying candidate epitope residues by preference

A subset of paratope residues can be picked out from the first two steps, then the candidate epitope residues that could interact with these paratope residues are determined by residues-residues pairing preference which plays a subtle bridging role in linking paratope residues to epitope residues. To achieve this goal, all paratope residues of $V_b^{'}$ in each preference pattern $\langle V_b^{'}, V_g^{'}, E^{'} \rangle$ are checked against the already identified paratope residues, and then the epitope residues $V_g^{'}$ in this pattern are deemed as candidate epitope residues if all of the residues in $V_b^{'}$ are observed.

2.3.4 Inferring new sets of contact residue pairs by dependence

Based on the paratope-epitope interacting residue pairs specified by preference, new sets of paratope-epitope contact residue pairs can be implicated by dependence. For each dependence pattern $\{\langle p_1^o\text{-}e_1^o\rangle, \ldots, \langle p_m^o\text{-}e_m^o\rangle\} \rightarrow \{\langle p_1^i\text{-}e_1^i\rangle, \ldots, \langle p_n^i\text{-}e_n^i\rangle\}$, all the interacting pairs $\langle p_i^o\text{-}e_i^o\rangle$ from the observed set are checked against the identified interacting residue pairs. The implied interacting residue pairs $\langle p^i\text{-}e^i\rangle$ are considered as true interaction in this antibody-antigen binding site if all of the observed pairs $\langle p^o\text{-}e^o\rangle$ have been identified by the preference associations.

Preference combines paratope and epitope together, while dependence implicates new interactions between the paratope and epitope from the interactions produced by preference, and in return the result generated by dependence can provide new clues for preference. This reciprocal enhancement works well in identifying interactions between an antibody and an antigen. These two steps and the neighborhood residue detecting step are terminated until no more paratope/epitope residues can be identified from a given antibody-antigen sequence pair.

2.3.5 Seeding epitope residues

The seed epitope residues are identified through a similar strategy as seed paratope residues identification with only a slight difference. First, when determining the seed epitope residues, both epitope residues' relative composition and the candidate residues are considered. More exactly, a residue from the antigen sequence is marked as an epitope residue if and only if its relative composition exceeds the preset threshold T_{rc}^{eh} and this type of amino acid has already been fished out by the paratope-epitope associations (preference and dependence). Second, due to that the whole antigen sequence could be antigenic, the residues are checked one-by-one along the whole antigen sequence. Third, besides identifying the seed epitope residues, those residues with very low RC values with respect to a threshold T_{rc}^{el} are marked as non-epitope residues. Empirically, only the best four and the worst three types of residues can pass through these two thresholds respectively.

2.3.6 Screening out neighborhood epitope residues

Neighborhood epitope residues are detected by using exactly the same strategy of neighborhood paratope residue detection. One more thing is to screen out the non-epitope residues based on the non-epitope seeds. A neighborhood residue is marked as a non-epitope residue if the cooperativity between this neighbor and itself is within the lowest ten percent of all the observed cooperative residue pairs according to the epitope residues' cooperativity ratios.

2.3.7 Completing epitope identification

A stringent threshold based epitope identification strategy can identify a subset of epitope residues and non-epitope residues with high accuracy, but leaves a large number of antigen residues still hidden. Thus we introduce a semi-supervised Hidden Markov Model (HMM) to complete the prediction of epitope residues.

An HMM is a statistical Markov model with unobserved states Π and observed variables X. In our case, the unobserved states are epitope and non-epitope, and the observed variables are the twenty standard amino acids. Figure 3 illustrates the diagram of the first-order hidden Markov model for epitope prediction which is quantified by parameters of transition probability and emission probability. The transition probability and emission probability are calculated by equation (1) and equation (2), respectively

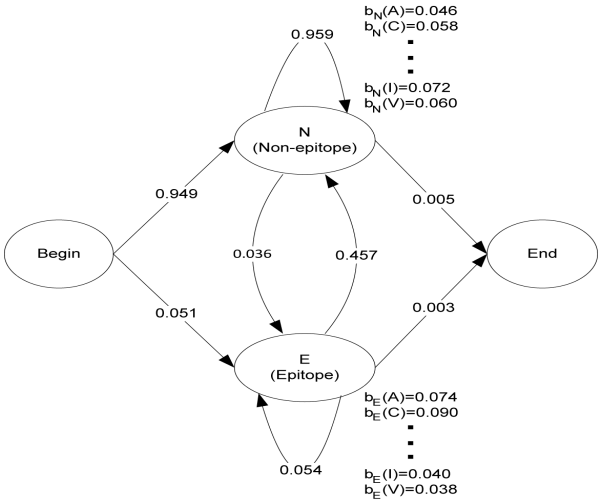

Fig. 3. First order hidden Markov model for epitope prediction. Arrows represent transition, and the numbers on arrows stand for transition probability from one state to another. Representative emission probabilities for state N and state E are shown at their right hand.

$$\hat{a}_{k,l} = \frac{A_{k,l}}{\sum_{l' \in \Pi} A_{k,l'}} \quad (1)$$

$$\hat{b}_k(\sigma) = \frac{B_k(\sigma)}{\sum_{\sigma' \in \Omega} B_k(\sigma')} \quad (2)$$

where $A_{k,l}$ is the number of transitions from state k to state l, Π is a set of states, $B_k(\sigma)$ is the number of emitting symbol σ at state k, and Ω is a set of symbols standing for the set of twenty standard amino acids.

In this work, the first-order HMM and second-order HMM are utilized together to build a better classifier. For the first-order HMM, $\Pi = \{begin, N, E, end\}$, and for the second-order HMM $\Pi = \{begin, NN, NE, EN, EE, end\}$. The symbol N represents the state of non-epitope residue, while E represents the state of epitope residue.

For a given sequence of antigen residues $x = (x_1, x_2, \cdots, x_m)$, the most probable sequence of states (or say labels) $\pi^* = (\pi_1^*, \pi_2^*, \cdots, \pi_m^*)$ is estimated by the Viterbi algorithm [30]. Suppose $\delta_k(i)$, which is calculated by equation (3), is the maximal probability of the path ending at sequence position i with state of k when observation is known, then the induction of $\delta_l(i+1)$ is carried out by equation (4). After determining the maximal probability for each site, then the most likely state for each site is determined by equation (5) if the state of site i has not been determined yet. If the state of site i has already been determined by the epitope seeding and neighborhood epitope residue screening, then it keeps the state unchangeable. We employ this semi-supervised learning method because of the significantly uneven distributed transition probability from non-epitope residue to epitope residue.

$$\delta_k(i) = \max_{\pi_1,\ldots,\pi_{i-1}} P(\pi_1, \ldots, \pi_{i-1}, \pi_i = k, x_1, \ldots, x_i) \quad (3)$$

$$\delta_l(i+1) = b_l(x_{i+1}) \max_k \{\delta_k(i) * a_{k,l}\} \quad (4)$$

$$\begin{cases} \pi_n^* = arg \max_k \{\delta_k(n)\} \\ \pi_i^* = arg \max_k \{\delta_k(i) * a_{k,\pi_{i+1}^*}\} \end{cases} \quad (5)$$

3 Experimental Results

3.1 Residues' relative composition and cooperativity in paratope and epitope

The paratope and epitope residues' relative compositions show that the same type of residue has diverse tendencies in antigen binding. Residue Y has a very high probability to become a paratope residue in CDR L1 and L2, while this tendency is weakened in other CDRs; residue D favors all the CDRs except H3 which is exactly a reverse tendency for residue G. Intriguingly, the same residue has very different relative composition in the six CDRs and antigen (the paratopes and epitopes comment: we intend to illustrate the diverse profiles of the same residue in six CDRs and antigen, but not the difference between different paratopes and epitopes). For example, residues Q, H and P are more likely to be epitope residues rather than paratope residues, while residues N, Y, W and S are in favor of paratope residues instead of epitope residues, in particular for residues Y and W. Our this finding on residues Y and W does not agree very well with the result reported by [31]. The reason should be that Y and W are rich in antigen sequence although they have a relatively large proportion in epitopes. In addition, our results clearly illustrate that epitopes prefer very much hydrophilic residues to hydrophobic residues. Detailed information is presented in supplementary Figure S1 and S2,

Our findings suggest that each CDR makes different contribution in antigen binding and that the same residue can exhibit diverse roles and profiles in six CDRs and antigen (epitopes). This confirms our idea of treating the six CDRs separately. The proportion of each CDR's contribution in antigen binding is presented in Table 1. The value is calculated based on the count of each paratope residue that interacts with epitope residues. For example, if one paratope residue interacts with two epitope residues then its contribution is counted twice instead of once. By this definition, H3 makes the highest contribution in antigen binding and L2 presents the lowest. Besides, the three CDRs from the antibody heavy chain take about two thirds contribution in antigen binding. Although a different and simple definition of calculating contribution is used here, our observations are consistent with the findings reported in [32].

The cooperativity ratios of epitope residue pairs and those in the paratope residue pairs in CDR H3 which

TABLE 1
Proportion of each CDR's contribution in antigen binding

CDR	L1	L2	L3	H1	H2	H3
Proportion	0.121	0.066	0.161	0.130	0.187	0.336

contributes about one third in the antigen binding are shown in supplementary Figure S3. The high values at the top-left corner of the epitope cooperativity picture emphasize that epitope residues prefer pairs of hydrophilic residues. With regard to the residue pairs within CDR H3, unexpectedly they cover only a few residues, e.g. Y, W, S, T and G. Although hydrophobic residues are scarce in the paratopes, they tend to cooperate with other residues once they occur in a paratope. This is interesting.

3.2 Associations in paratope-epitope interacting pairs

The key idea of this work is to use two kinds of associations, the preference and dependence, to address the problem of antibody-specified epitope prediction. As introduced, preference can reflect the contextual relation between paratope and epitope residues in a local manner, while dependence can span this contextual relation between two local sub-regions. The combination of these two rules, which extends association from one-dimension to two-dimension, can well capture the concept of context-awareness in the recognition between paratopes and epitopes.

The residues-residues pairing preferences with frequency (or, support) $\geq$ 12% are shown in Table 2. As an example, the second preference in Table 2 means that paratope residues {D,Y} interact frequently (15 out of 80) with epitope residue K. Overall, these most frequent paratope preference residues are enriched by Y, D, S and G, while the epitope residues are rich of K, R, Q and E. This observation applauds the findings given by residues' relative composition in another way. Interestingly, residue G does not have a very high relative composition, however G is rich in H3 which takes about one third contribution in antigen binding. Therefore residue G is also frequent in paratope. We take another example of preference to illustrate how the preference information is used in epitope prediction. Using $\langle\{G,Y\}\bigcup\{E\}\rangle$, we can say that once residues G and Y have been identified as paratope residues, then residue E can be marked as a candidate epitope residue.

The dependencies with support level $\geq$ 10% and confidence level $\geq$ 90% are listed in Table 3. Taking $\{\langle D\text{-}K\rangle,\langle W\text{-}K\rangle\}\rightarrow\{\langle Y\text{-}K\rangle\}$ as an example to demonstrate the usefulness of dependence, we can say that once paratope-epitope interacting residue pairs $\langle D\text{-}K\rangle$ and $\langle W\text{-}K\rangle$ are predicted or observed in the binding site, then $\langle Y\text{-}K\rangle$ is believed to occur in this site as well with 100% confidence. Compared with preference shown in Table 2, one can see that dependence involves a much broader interactions between paratope and epitope residues. Thus the reciprocal enhancement between these two associations facilitates the accurate antibody-specified epitope identification.

TABLE 2
Preference between paratope residues and epitope residues with frequency (or, support) $\geq$ 12%.

Paratope(s)	Epitope(s)	Frequency	Redundancy
D	K	21.3%(17/80)	1.24(21/17)
D,Y	K	18.8%(15/80)	1.27(19/15)
S	E	16.3%(13/80)	1.15(15/13)
D	R	13.8%(11/80)	1.18(13/11)
D,Y	R	12.5%(10/80)	1.70(17/10)
G,Y	E	12.5%(10/80)	1.40(14/10)
S	K	12.5%(10/80)	1.00(10/10)

TABLE 3
Dependence in paratope-epitope interacting residue pairs with support $\geq$ 10% and confidence $\geq$ 90%

Obs. Pair(s)$^{|}$	Imp. Pair(s)$^{\rightarrow}$	confidence
$\langle$D-K$\rangle$, $\langle$W-K$\rangle$	$\langle$Y-K$\rangle$	100.0%
$\langle$T-K$\rangle$	$\langle$Y-K$\rangle$	100.0%
$\langle$Y-G$\rangle$, $\langle$T-K$\rangle$	$\langle$Y-K$\rangle$	100.0%
$\langle$D-K$\rangle$, $\langle$T-K$\rangle$	$\langle$Y-K$\rangle$	100.0%
$\langle$W-K$\rangle$	$\langle$Y-K$\rangle$	93.33%
$\langle$Y-G$\rangle$, $\langle$S-E$\rangle$	$\langle$Y-E$\rangle$	90.00%
$\langle$N-G$\rangle$	$\langle$Y-G$\rangle$	90.00%

$^{|}$: Observed paratope-epitope interacting pairs;
$^{\rightarrow}$: Implied paratope-epitope interacting pairs.

An example of the two-dimensional associations between an antigen and its antibody's six CDRs (in PDB:1EGJ) are partially shown in Figure 4. The prediction of the epitope residues is proceeded as follows: (i) residues TYR:32:H, TYR:30B:L, ASN:91:L, ASN:92:L, TRP:96:L are fished out in the step of identifying seed paratope residues, (ii) residue GLU:93:L is selected by using cooperativity detection, and (iii) epitope residues R, K and E are identified by utilizing these paratope information, residue's relative composition and the paratope-epitope association patterns, such as $\langle\{N\}\bigcup\{R\}\rangle$, $\langle\{Y\}\bigcup\{E\}\rangle$, and $\{\langle Y\text{-}R\rangle, \langle Y\text{-}E\rangle\}\rightarrow\{\langle Y\text{-}K\rangle\}$. Other epitope residues within this antigen sequence are identified by using the same strategy. To complete the identification, a semi-supervised HMM is used in the final step to label the undetermined residues.

3.3 Performance of ABepar for antibody-specified B-cell epitope prediction

Leave-one-out cross validation is carried out to evaluate the performance of our method. In every iteration, one complex without its structural information is tested against the model constructed by using the rest 79 PDB complexes in the training. Four metrics are adopted to quantify the performance which are sensitivity (sen.),

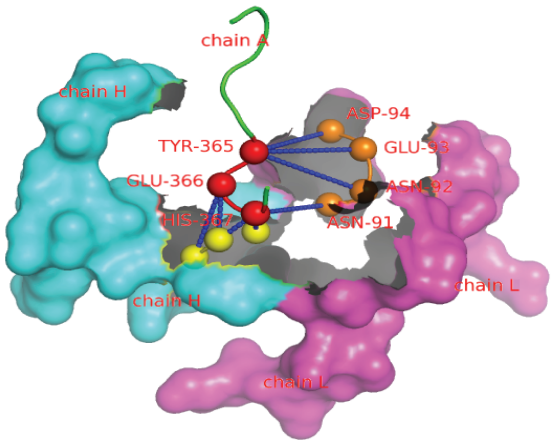

Fig. 4. Associations between the paratope and epitope residues within PDB:1EGJ. The epitope residues are rendered by red spheres, while the other antigen residues are colored by green. (Only a small segment of the antigen sequence is shown here.) The paratope residues are colored by yellow and orange for the antibody heavy and light chain respectively, meanwhile other residues within six CDRs are rendered by their surface. The dashed blue lines are used to illustrate the preference between the paratope and epitope residues. This picture is generated by PyMol [33].

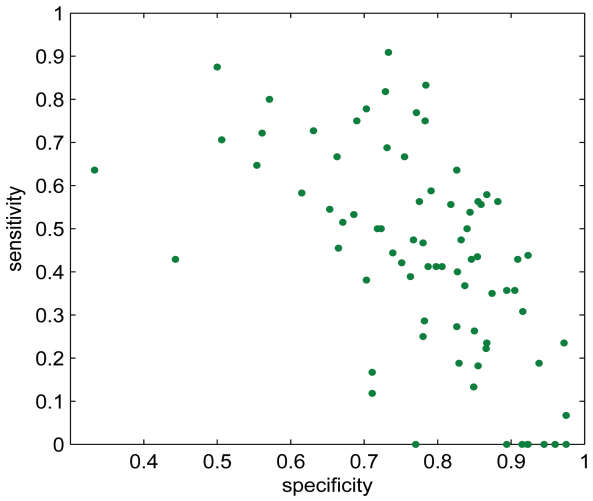

Fig. 5. Sensitivity versus specificity performance on the 80 complexes by ABepar.

specificity (spe.), accuracy (acc.) and F-measure (F1). Their definitions are given by: $sen. = TP/(TP + FN)$, $spe. = (TN)/(TN+FP)$, $acc. = (TP+TN)/(TP+TN+FP+FN)$, and $F1 = 2\times(precision\times recall)/(precision+recall)$, where TP is the number of correctly predicted epitope residues, TN is the number of correctly predicted non-epitope residues, FP is the number of incorrectly predicted epitope residues which should be non-epitope residues in reality, FN is the number of incorrectly predicted non-epitope residues which should be epitope residues in fact, precision is the proportion of correctly predicted epitope residues with respect to the total number of predicted epitope residues, and recall is the proportion of correctly identified epitope residues with respect to the total number of actual epitope residues.

The detailed performance on the 80 complexes are reported in (the) Table 4. Figure 5 plots a visualization for the performance of sensitivity versus specificity on all of the 80 complexes. The average sensitivity is 0.434±0.238, the average specificity is 0.781±0.126, and the overall accuracy is 0.749±0.103. We note that the distribution of the non-epitope and epitope residues for every antigen is extremely un-balanced with the ratio 12.65±10.15 on average. This nature of imbalance increases the complexity of the prediction problem and skews the competence of accuracy. Thus the F1-scores are also presented in Table 4 (columns 9 and 18) for a more meaningful measurement. To avoid inflated performance caused by the high similarity between antigen sequences, we further analyzed the antigen sequence similarity by using local pairwise sequence alignment with the BLOSUM62 substitution matrix. The pairwise sequence's (sequences) similarities of all the 80 antigen sequences are shown in Figure 6. We found that about 98.9% of the antigen sequence pairwise similarities are less than or equal to 0.25. That is, there is a faint over-fitting problem caused by the sequence similarity and the results well reflect the capability of our method. Although there are still a couple of antigen sequences having a very high similarity, their epitopes are quite different from each other. For instance, the epitope residues of PDB:1A2Y are 19N, 22G, 23Y, 24S, 27N, 102G, 103N, 116K, 117G, 118T, 119D, 120V, 121Q, 124I and 125R, while the epitope residues of PDB:1J1O are 15H, 16G, 19N, 20Y, 21R, 62W, 63W, 73R, 75L, 77N, 89T, 93N, 96K, 97K, 100S, 101D, 102G and 103N. Clearly there are only three overlapping residues between these two antigens, which are 19N, 102G and 103N. According to the performance on PDB:1A2Y and PDB:1J1O shown in Table 4, we can conclude that our method can well distinguish the antibody-specified epitopes.

The performance comparison between our method and those methods described by Ponomarenko and Bourne [21] shows that our method outperforms others according to the sensitivity and specificity values. Regarding the comparison with continuous epitope predictors, we note that there is no direct comparison because of the different problems addressed. We also note that the comparison with the methods in [21] is indirect as we are dealing with the new, antibody-specified epitope prediction problem rather than the traditional antibody-ignored epitope prediction problem. Therefore, a direct comparison under the strictly same platform is presented in the following subsection where a simplified version of ABepar is taken in the comparison.

TABLE 4
The performance for antibody-specified epitope prediction on 80 antibody-antigen complexes by ABepar

PDB ID	Ab	Ag	#Etp[1]	#NEtp[2]	sen.	spe.	acc.	F1	PDB ID	Ab	Ag	#Etp	#NEtp	sen.	spe.	acc.	F1
1A2Y	B A	C	15	114	0.600	0.711	0.698	0.316	1ZTX	H L	E	16	85	0.563	0.882	0.832	0.514
1AR1	C D	B	15	237	0.600	0.865	0.849	0.321	2ADF	H L	A	15	174	0.000	0.960	0.884	0.000
1BGX	H L	T	35	793	0.371	0.808	0.790	0.130	2AEP	H L	A	19	369	0.474	0.767	0.753	0.158
1BJ1	H L	W	16	78	0.438	0.923	0.840	0.483	2ARJ	H L	Q	18	97	0.222	0.866	0.765	0.229
1BQL	H L	Y	12	117	0.000	0.923	0.837	0.000	2B2X	H L	A	18	170	0.556	0.859	0.830	0.385
1EGJ	H L	A	11	90	0.909	0.733	0.752	0.444	2BDN	H L	A	16	52	0.688	0.731	0.721	0.537
1EO8	H L	A	15	304	0.133	0.849	0.815	0.063	2CMR	H L	A	17	178	0.235	0.972	0.908	0.308
1EZV	X Y	E	17	168	0.412	0.798	0.762	0.241	2DD8	H L	S	19	173	0.474	0.832	0.797	0.316
1FE8	H L	A	19	167	0.263	0.85	0.790	0.204	2FD6	H L	U	12	237	0.667	0.755	0.751	0.205
1FJ1	B A	F	17	234	0.588	0.791	0.777	0.263	2H9G	H L	S	12	52	0.583	0.615	0.609	0.359
1FNS	H L	A	12	184	0.750	0.783	0.781	0.295	2J4W	H L	D	11	23	0.636	0.826	0.765	0.636
1FSK	C B	A	17	142	0.118	0.711	0.648	0.067	2J5L	C B	A	11	23	0.273	0.826	0.647	0.333
1H0D	B A	C	16	106	0.875	0.500	0.549	0.337	2J88	H L	A	9	310	0.778	0.703	0.705	0.130
1IQD	B A	C	16	140	0.188	0.829	0.763	0.140	2JEL	H L	P	15	70	0.533	0.686	0.659	0.356
1J1O	H L	Y	18	111	0.444	0.739	0.698	0.291	2NR6	D C	A	14	315	0.357	0.905	0.881	0.204
1JHL	H L	A	11	118	0.818	0.729	0.736	0.346	2NY7	H L	G	22	268	0.500	0.840	0.814	0.289
1JPS	H L	T	19	181	0.579	0.867	0.840	0.407	2NYY	D C	A	20	1247	0.350	0.874	0.866	0.076
1JRH	H L	I	15	80	0.067	0.975	0.832	0.111	2Q8B	H L	A	18	274	0.500	0.723	0.709	0.175
1LK3	H L	A	18	118	0.389	0.763	0.713	0.264	2QQK	H L	A	12	529	0.167	0.711	0.699	0.024
1N8Z	B A	C	16	565	0.000	0.894	0.869	0.000	2QQN	H L	A	12	142	0.500	0.718	0.701	0.207
1NCA	H L	N	20	369	0.250	0.780	0.753	0.094	2R56	H L	A	21	138	0.381	0.703	0.660	0.229
1NFD	H G	D	12	227	0.833	0.784	0.787	0.282	2UZI	H L	R	18	148	0.722	0.561	0.578	0.271
1NMB	H L	N	19	369	0.421	0.751	0.735	0.134	2VXQ	H L	A	14	78	0.286	0.782	0.707	0.229
1NSN	H L	S	17	121	0.647	0.554	0.565	0.268	2VXS	I M	A	7	79	0.429	0.443	0.442	0.111
1OAZ	H L	A	17	98	0.235	0.867	0.774	0.235	2VXT	H L	I	17	139	0.412	0.806	0.763	0.275
1OB1	A B	C	13	83	0.769	0.771	0.771	0.476	2W9E	H L	A	15	84	0.800	0.571	0.606	0.381
1ORS	B A	C	10	122	0.000	0.975	0.902	0.000	2ZCH	H L	P	16	221	0.563	0.855	0.835	0.316
1OSP	H L	O	20	231	0.400	0.827	0.793	0.235	3B2U	H L	A	17	178	0.412	0.787	0.754	0.226
1OTS	C D	A	9	435	0.000	0.945	0.926	0.000	3B9K	L H	B	19	98	0.368	0.837	0.761	0.333
1P2C	E D	F	16	113	0.750	0.690	0.698	0.381	3BN9	F E	A	22	219	0.545	0.653	0.643	0.218
1R3J	B A	C	13	90	0.538	0.844	0.806	0.412	3CVH	H L	A	15	259	0.467	0.780	0.763	0.177
1RJL	B A	C	12	83	0.667	0.663	0.663	0.333	3D9A	H L	C	18	111	0.444	0.739	0.698	0.291
1SY6	H L	A	11	157	0.727	0.631	0.637	0.208	3D85	B A	C	16	117	0.000	0.915	0.805	0.000
1TQB	B C	A	17	85	0.706	0.506	0.539	0.338	3EOA	H L	I	14	165	0.429	0.909	0.872	0.343
1UJ3	B A	C	18	187	0.556	0.818	0.795	0.323	3G04	B A	C	23	205	0.435	0.854	0.811	0.317
1V7M	H L	V	16	129	0.563	0.775	0.752	0.333	3GBN	H L	B	12	161	0.000	0.770	0.717	0.000
1W72	H L	A	14	260	0.429	0.846	0.825	0.200	3GI9	H L	C	16	421	0.188	0.938	0.911	0.133
1WEJ	H L	F	11	93	0.636	0.333	0.365	0.175	3GRW	H L	A	33	173	0.515	0.671	0.646	0.318
1YJD	H L	C	14	104	0.357	0.894	0.831	0.333	3H42	H L	B	22	470	0.182	0.855	0.825	0.085
1YYM	R Q	P	13	286	0.308	0.916	0.890	0.195	3HI6	H L	A	22	158	0.455	0.665	0.639	0.235

[1]: number of epitope residues, and [2]: number of non-epitope residues.

3.4 Simplifying ABepar for antibody-*ignored* B-cell epitope prediction

To our best knowledge, all previous methods for B-cell epitope prediction do not use the context-dependence principle in the recognition between antigens and antibodies. Antigen is the only issue and data that is taken for the prediction. Given an antigen, the prediction just tells whether a residue is antigenic or not. The prediction result is an unordered union of the antigenic residues of the different epitopes that are formed when this antigen interacts with different antibodies. Such an approach can narrow down from the whole antigen sequence to the smaller antigenic region, but it cannot pinpoint the constituent residues of a specific epitope which is usually much shorter than the antigenic region of an antigen. An example is shown later. Therefore, such a generic epitope prediction is quite different from the problem we are addressing. To make a fair comparison between our method and the previous ones, we compile a new data set from PDB in which every antigen has multiple antibody partners. The purpose of constructing such a data set is aimed at alleviating the underestimation of current epitope predictors caused by incomplete epitope exploration [10], [11], [18]. This data set is shown in Table 5.

Three structure-based epitope prediction methods ElliPro [12], DiscoTope [11], and SEPPA [34] are used in the comparison. ElliPro predicts antigenic residues by means of the protrusion index (PI) which is calculated based on antigen structures. The performance here by ElliPro was calculated over all the predicted discontinuous epitopes with the default parameters. With regard to DiscoTope, the epitope residues from a testing antigen structure is predicted based on features that are extracted from a set of training antigen structures, such as residue solvent accessibility, residue spatial distribution and residue propensity. The performance of DiscoTope was achieved by using the default cutoff threshold -7.7. SEPPA identifies the antigenic residues by clustering the triangle residues of unit patches, and the result shown here was generated with the default parameter of 1.8.

TABLE 5
Performance evaluation for ABepar, ElliPro, DiscoTope, and SEPPA on an epitope relatively completely explored data set

antigen	corresponding PDB ID	#Ab†	F-measure			
			$F1^a$	$F1^e$	$F1^d$	$F1^s$
Hen egg white lysozyme	1A2Y 1BQL* 1BVK[C,F]‡ 1C08 1DQJ 1FDL 1G7H 1G7I 1G7J 1G7L 1G7M 1IC4 1IC5 1IC7 1J1O 1J1P 1J1X 1KIP 1KIQ 1KIR 1MLC[E,F] 1NDG 1NDM 1P2C[C,F] 1UA6 1UAC 1VFB 2DQC 2DQD 2DQE 2DQF[C,F] 2DQG 2DQH 2DQI 2DQJ 2EIZ 2EKS 2IFF 2YSS 3D9A	44	0.747	0.683	0.519	0.722
Vascular endothelial growth factor	1BJ1[V,W] 1CZ8[V,W] 2FJG[V,W]	6	0.464	0.571	0.540	0.567
Ubiquitin	3DVG[X,Y] 3DVN[U,V,X,Y]	6	0.367	0.375	0.456	0.430
Influenza virus neuraminidase	1NCD 1A14 1NMB 1NCB 1NCC 1NCA	6	0.244	0.336	0.387	0.378
Von willebrand factor	1FE8[A,B,C] 2ADF	4	0.343	0.343	0.150	0.483
Outer surface protein A	1FJ1[E,F] 1OSP	3	0.381	0.347	0.379	0.367
Integrin alpha-L	3HI6[A,B] 3EOA[I,J]	4	0.385	0.381	0.386	0.479
Prion protein	1TPX 1TQB 1TQC 2W9E	4	0.630	0.355	0.526	0.480
Tissue factor	1JPS 1UJ3	2	0.426	0.00	0.241	0.195

^{a}: ABepar, ^{e}: ElliPro, ^{d}: DiscoTope, and ^{s}: SEPPA.
†: number of antibodies interacting with this antigen. Some of the antibodies are mutants of its wild type.
‡: multi-antigen chains in the same PDB complex.
*: it is Bobwhite quail lysozyme, but it has very similar sequence and structure with hen egg white lysozyme.

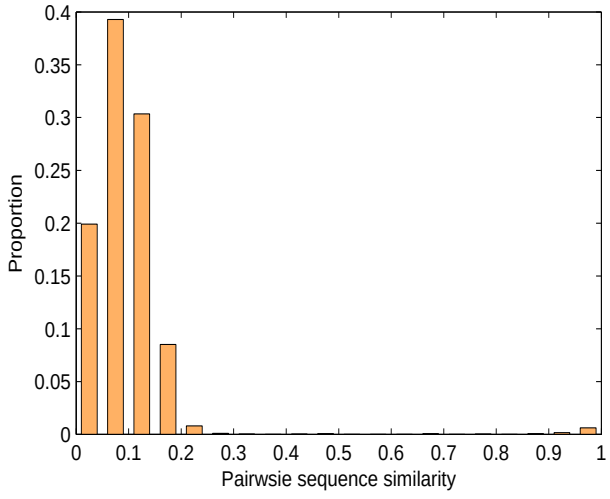

Fig. 6. Pairwise sequence similarity of all the 80 antigen sequences.

To identify antibody-ignored epitope residues by our method, we modified our computational process by skipping the first three steps which are specially for identifying paratope associated candidate epitope residues.

From the F1-scores shown in Table 5, we can see that our method is competitive to, sometimes much better than the three structure based methods even though we take only sequence information as input. In particular, our method can provide much better results on identifying epitope residues from those more completely explored antigen sequences. For example, we can correctly identify 37 of the 55 epitope residues and include only 7 non-epitope residues as epitope residues from hen egg white lysosyme, achieving an F1-score of 0.75. This performance is much better than 0.68 by ElliPro, 0.52 by DiscoTope, and 0.72 by SEPPA. All these demonstrate that even without 3D structural information, the simplified version of ABepar for antibody-ignored prediction is also powerful to identify epitope residues.

We once again stress the importance and complexity of the antibody-specified epitope prediction problem. We show some extreme pairwise dissimilarities between different epitopes of an antigen. Taking hen egg white lysozyme again, this antigen can interact with as many as 44 different antibodies (some are mutants of the wild type antibody, for instance 1G7H, 1G7I, 1G7J, 1G7L, 1G7M are mutants of 1VFB), and there are 55 antigenic residues in total in the 129 residues of this antigen. However, the average number of resides over these antibody-specified epitopes is only 16.3±2.0. The average pairwise epitope similarity of these 44 epitopes is 0.359. More remarkably, 58.4% of these pairwise epitope

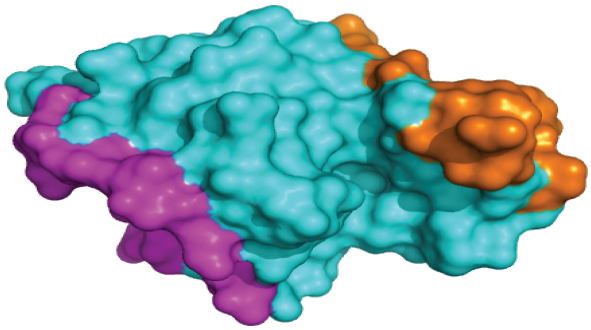

Fig. 7. Reconfiguration of epitope when hen egg white lysozyme interacts with different antibodies (PDB:1A2Y and PDB:2IFF). Epitope residues in PDB:1A2Y are colored by magenta while those epitope residues within PDB:2IFF are colored by orange. All the rest residues are rendered by cyan.

similarities are < 0.1, i.e. with less than 2 residues in common between two epitopes. For example, the epitope residues of PDB complex 1A2Y and 2IFF are distributed on two entirely non-overlapping locations on the surface of hen egg white lysozyme as shown in Figure 7. Here the similarity between two epitopes A and B is calculated by $|A \bigcap B|/(|A| * |B|)$, where $A \bigcap B$ represents the overlapping epitope residues, and $|X|$ means the residue number of X. These findings do manifest the extreme reconfiguration of epitope when an antigen binds to different antibodies. Therefore, it is important to specify the corresponding antibody when predicting epitope residues on a given antigen. We do not mean that identifying epitope regardless of antibody is less important, in fact we emphasize that identifying an epitope with the specification of its antibody in some applications is necessary as also stressed by [5].

4 Discussion

In this study, the antibody-specified epitope prediction problem is newly formulated. It is a more meaningful but more challenging problem than the traditional epitope prediction problem that does not consider antibody information in the prediction. There are two main differences. First, antibody-antigen PDB complexes are used in our model training, and antibody-antigen sequence pairs without structure information are required for our model testing. However, just antigen sequences or structures without antibodies are needed for epitope identification by the traditional epitope prediction models. Second, the output of our model is a set of antigenic residues that interact with the specific antibody excluding all other antigenic residues that interact with other antibodies. However, the output of the traditional model is a residue union of all possible epitopes without any clarity to label the constituent residues of any epitope.

Antibody-specified epitope prediction is first time proposed by this work, hence there is no direct comparison between the new problem and the traditional problem. Therefore a simplified version of our method is introduced for a fair comparison, and it is tested on a data set containing complexes which have been relatively completely explored for epitopes. The purpose of creating this data set is to avoid as much as possible the underestimation problem caused by incomplete epitope exploration which has been argued by many researchers [10], [11], [18]. The more complete epitope exploration of one antigen is, the more convincing of the result conducted on the data. Therefore hen egg white lysozyme, whose epitopes are relatively completely explored, is taken as example to demonstrate the prediction capability of different models.

In this study, we adopted a semi-supervised hidden Markov model to complete the process of epitope prediction. This approach can alleviate the wrong predictions caused by the stringent cutoff thresholds of residues' relative composition and cooperativity, and it also overcomes the thorny problem raised by the imbalanced distribution of non-epitope residues and epitope residues.

The proposed antibody-specified epitope prediction model is novel and promising, but it still surfers from the scarce of training data set. Thus the leave-one-out cross validation was used to evaluate this model. In every iteration, one sample was left out for testing and the remaining samples were taken for model construction. Along with the increasing available samples, we believe our model will show more powerful performance in antibody-specified epitope prediction.

5 Conclusion

We summarize the whole work as follows. In this paper, we have proposed a method called **ABepar** to predict antibody-specified epitopes with a high accuracy and broad applicability. This method is trained on a relatively small set of PDB complexes that contains antibody-antigen structures, but it can be applied to any antibody-antigen sequences without 3D structural information. The novel idea of this method is based on the context-awareness in the recognition between paratope and epitope which is implemented by two kinds of associations (preference and dependence). In addition, the semi-supervised HMM also plays an important role in completing the process of epitope identification. It can capture the sequential relationship within epitope residues at one hand, and also it overcomes the insurmountable obstacles caused by the imbalance distribution of the epitope and non-epitope residues at the other hand. Furthermore, we also assessed the performance of our model on epitope prediction when antibody information is not given. The comparison results show that our sequence-based epitope prediction method has a competitive performance compared with structure-based prediction methods which have much smaller applicability.

ACKNOWLEDGMENTS

This research work was supported by two Singapore MOE (Ministry Of Education) ARC (Academic Research Council) funding projects (Tier-2 grant T208B2203 and Tier-2 grant MOE2009-T2-2-004), and by a Nanyang Technological University Tier-1 grant RG66/07. We are also thankful to Dr. Hoi Chu-Hong, Steven for his advice.

REFERENCES

[1] N. K. Jerne, "Immunological speculations," *Annu. Rev. Microbiol.*, vol. 14, pp. 341–358, 1960.
[2] M. H. V. Van Regenmortel, "Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity," *Methods.*, vol. 9, no. 3, pp. 465–472, 1996.
[3] M. B. Irving, O. Pan, and J. K. Scott, "Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics," *Curr. Opin. Chem. Biol.*, vol. 5, no. 3, pp. 314–324, 2001.
[4] M. H. V. Van Regenmortel, "Structural and functional approaches to the study of protein antigenicity," *Immunol. Today*, vol. 10, no. 8, pp. 266–272, 1989.
[5] J. A. Greenbaum, P. H. Andersen, M. Blythe, H.-H. Bui, R. E. Cachau, J. Crowe, M. Davies, A. Kolaskar, O. Lund, S. Morrison, B. Mumey, Y. Ofran, J.-L. Pellequer, C. Pinilla, J. V. Ponomarenko, G. P. S. Raghava, M. H. V. 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, no. 2, pp. 75–82, 2007.
[6] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *Proc. Natl. Acad. Sci. USA*, vol. 78, no. 6, pp. 3824–3828, 1981.
[7] P. Karplus and G. Schulz, "Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen," *Naturwissenschaften*, vol. 72, no. 4, pp. 212–213, 1985.
[8] J. Parker, D. Guo, and R. Hodges, "New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites," *Biochemistry*, vol. 25, no. 19, pp. 5425–5432, 1986.
[9] A. Kolaskar and P. C. Tongaonkar, "A semi-empirical method for prediction of antigenic determinants on protein antigens," *FEBS Lett.*, vol. 276, pp. 172–174, 1990.
[10] U. Kulkarni-Kale, S. Bhosle, and A. S. Kolaskar, "CEP: a conformational epitope prediction server," *Nucleic Acids Res.*, vol. 33, pp. 168–171, 2005.
[11] P. H. Andersen, N. Morten, and L. Ole, "Prediction of residues in discontinuous B-cell epitopes using protein 3D structures," *Protein Sci.*, vol. 15, no. 11, pp. 2558–2567, 2006.
[12] J. Ponomarenko, H.-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, pp. 514+, 2008.
[13] T. T. Wu and E. A. Kabat, "An analysis of the sequences of the variable regions of bence jones proteins and myeloma light chains and their implications for antibody complementarity," *J. Exp. Med.*, vol. 132, pp. 211–250, 1970.
[14] C. Chothia and A. M. Lesk, "Canonical structures for the hypervariable regions of immunoglobulins," *J. Mol. Biol.*, vol. 196, no. 4, pp. 901–917, 1987.
[15] J. Söllner and B. Mayer, "Machine learning approaches for prediction of linear b-cell epitopes on proteins," *J. Mol. Recognit.*, vol. 19, pp. 200–208, 2006.
[16] S. Saha and G. P. S. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 40–48, 2006.
[17] 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: Structure, Function, and Bioinformatics*, vol. 68, no. 1, pp. 294–304, 2007.
[18] N. D. Rubinstein, I. Mayrose, and T. Pupko, "A machine-learning approach for predicting B-cell epitopes," *Mol. Immunol.*, vol. 46, no. 5, pp. 840–847, 2008.
[19] M. J. Sweredoski and P. Baldi, "Cobepro: a novel system for predicting continuous b-cell epitopes," *Protein Engineering, Design and Selection*, vol. 22, no. 3, pp. 113–120, 2009.
[20] M. J. Blythe and D. R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Sci.*, vol. 14, no. 1, pp. 246–8, 2005.
[21] J. V. Ponomarenko and P. E. Bourne, "Antibody-protein interactions: benchmark datasets and prediction tools evaluation," *BMC Struct. Biol.*, vol. 7, p. 64, 2007.
[22] S. P. Singh, S. Tyagi, K. Feroz, and M. B.N, "Benchmarking the Propensity Scales for the Prediction of Linear B-Cell Epitopes," *J. Comput. Intell. Bioinformatics*, vol. 1, no. 1, pp. 45–53, 2008.
[23] 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, no. 1, pp. 235–242, 2000.
[24] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein eng.*, vol. 11, no. 9, pp. 739–747, 1998.
[25] J. Li and Q. Liu, "'Double water exclusion': a hypothesis refining the O-ring theory for the hot spots at protein interfaces," *Bioinformatics*, vol. 25, no. 6, pp. 743–750, 2009.
[26] J. Li, G. Liu, H. Li, and L. Wong, "Maximal Biclique Subgraphs and Closed Pattern Pairs of the Adjacency Matrix: A One-to-one Correspondence and Mining Algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 12, pp. 1625–1637, 2007.
[27] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, 1982.
[28] F. Coenen, G. Goulbourne, and P. Leng, "Tree Structures for Mining Association Rules," *Data Min. Knowl. Discov.*, vol. 8, no. 1, pp. 25–51, 2004.
[29] K. Abhinandan and A. C. Martin, "Analysis and improvements to kabat and structurally correct numbering of antibody variable domains," *Mol. Immunol.*, vol. 45, no. 14, pp. 3832–3839, 2008.
[30] G. D. Forney, "The Viterbi algorithm," *Proceedings of The IEEE*, vol. 61, no. 3, pp. 268–278, 1973.
[31] N. D. Rubinstein, I. Mayrose, D. Halperin, D. Yekutieli, J. M. Gershoni, and T. Pupko, "Computational characterization of B-cell epitopes," *Mol. Immunol.*, vol. 45, no. 12, pp. 3477–3489, 2008.
[32] J. C. Almagro, "Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires," *J. Mol. Recognit.*, vol. 17, no. 2, pp. 132–143, 2004.
[33] W. L. Delano, "The PyMOL Molecular Graphics System," DeLano Scientific, Palo Alto, CA, USA, 2002. [Online]. Available: http://www.pymol.org
[34] 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 Research*, vol. 37, no. suppl 2, pp. W612–W616, 2009.

PLACE PHOTO HERE

Liang Zhao received the BS degree from Wuhan University, P.R. China, in 2007. Since August 2008, he has been a PhD student in the school of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include computational biology and machine learning.

PLACE PHOTO HERE

Limsoon Wong Limsoon Wong is Provost's Chair Professor of Computer Science and Professor of Pathology at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. Limsoon has written about 150 research papers, a few of which are among the best cited of their respective fields. He has/had served on the editorial boards of Information Systems (Elsevier), Journal of Bioinformatics and Computational Biology (ICP), Bioinformatics (OUP), IEEE/ACM Transactions on Computational Biology and Bioinformatics, and Drug Discovery Today (Elsevier). He is a co-founder and chairman of Molecular Connections in India.

PLACE PHOTO HERE

Jinyan Li Jinyan Li received his PhD degree in computer science from the University of Melbourne in 2001. He was an associate professor in the School of Computer Engineering, Nanyang Technological University, Singapore. His research is focused on protein structural bioinformatics, statistically important discriminative patterns, interaction subgraphs, and classification methods. Jinyan has published over 100 research articles. One of his most interesting work was a cancer diagnosis technique for childhood leukemia disease through the discovery of emerging patterns from the gene expression data, and currently he is very interested in infectious disease studies and water bioinformatics by exploring graph theories and biological water exclusion principles. One of his data mining articles is widely cited over 500 times, and another paper on bioinformatics is cited over 1000 times, according to google scholar.