

# Protein Science

## Prediction of residues in discontinuous B-cell epitopes using protein 3D structures

Pernille Haste Andersen, Morten Nielsen and Ole Lund

*Protein Sci.* 2006 15: 2558-2567; originally published online Sep 25, 2006;  
Access the most recent version at doi:[10.1110/ps.062405906](https://doi.org/10.1110/ps.062405906)

---

### References

This article cites 36 articles, 19 of which can be accessed free at:  
<http://www.proteinscience.org/cgi/content/full/15/11/2558#References>

Article cited in:  
<http://www.proteinscience.org/cgi/content/full/15/11/2558#otherarticles>

### Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

---

### Notes

---

To subscribe to *Protein Science* go to:  
<http://www.proteinscience.org/subscriptions/>

---

---

# Prediction of residues in discontinuous B-cell epitopes using protein 3D structures

---

PERNILLE HASTE ANDERSEN, MORTEN NIELSEN, AND OLE LUND

Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, DK-2800 Lyngby, Denmark

(RECEIVED June 16, 2006; FINAL REVISION August 9, 2006; ACCEPTED August 12, 2006)

## Abstract

Discovery of discontinuous B-cell epitopes is a major challenge in vaccine design. Previous epitope prediction methods have mostly been based on protein sequences and are not very effective. Here, we present DiscoTope, a novel method for discontinuous epitope prediction that uses protein three-dimensional structural data. The method is based on amino acid statistics, spatial information, and surface accessibility in a compiled data set of discontinuous epitopes determined by X-ray crystallography of antibody/antigen protein complexes. DiscoTope is the first method to focus explicitly on discontinuous epitopes. We show that the new structure-based method has a better performance for predicting residues of discontinuous epitopes than methods based solely on sequence information, and that it can successfully predict epitope residues that have been identified by different techniques. DiscoTope detects 15.5% of residues located in discontinuous epitopes with a specificity of 95%. At this level of specificity, the conventional Parker hydrophilicity scale for predicting linear B-cell epitopes identifies only 11.0% of residues located in discontinuous epitopes. Predictions by the DiscoTope method can guide experimental epitope mapping in both rational vaccine design and development of diagnostic tools, and may lead to more efficient epitope identification.

**Keywords:** discontinuous epitopes; B-cell epitope; antibody; vaccine design; protein structure; antigen; accessibility; hydrophilicity

A major task in vaccine design is to select and design proteins containing antibody-binding epitopes (B-cell epitopes) able to induce an efficient immune response. The selection can be aided by epitope prediction in relevant proteins or regions of proteins. In addition, prediction of B-cell epitopes may help to identify epitopes in proteins that have been analyzed using experimental techniques based on antibody affinity binding, e.g., Western blotting, immunohistochemistry, radioimmunoassay (RIA), and enzyme-linked immunosorbent assay (ELISA).

Most existing methods for prediction of B-cell epitopes exclusively use protein sequences as input, and are best

suitable to predict epitopes composed of a continuous stretch of amino acids (linear epitopes) (Hopp and Woods 1981; Parker et al. 1986; Jameson and Wolf 1988; Debelle et al. 1992; Maksyutov and Zagrebelnaya 1993; Alix 1999; Odorico and Pellequer 2003). In general, these methods are based on prediction of hydrophilicity, flexibility,  $\beta$ -turns, and surface accessibility using a number of amino acid propensity scales. A large amount of data exists on linear epitopes (Leitner et al. 2003; Saha et al. 2005; Toseland et al. 2005), since the annotation can be done by measuring the binding of antigen peptide fragments to antibodies. However, this method of annotation may lead to annotation errors, because a peptide can specifically bind an antibody even if some residues of the peptide are not interacting with the antibody. Predicting linear epitopes is still a nontrivial task, and the obtainable prediction accuracy is quite poor (Van Regenmortel and Pellequer 1994; Van Regenmortel 1996; Blythe and Flower 2005). However, combination of a hidden Markov

---

Reprint requests to: Ole Lund, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark; e-mail: [lund@cbs.dtu.dk](mailto:lund@cbs.dtu.dk); fax: 45-4593-1585.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062405906>.

model and a hydrophilicity scale constructed by Parker et al. (1986) has recently lead to some improvement in linear B-cell epitope prediction (Larsen et al. 2006).

It has been estimated that >90% of B-cell epitopes are discontinuous, i.e., consist of segments that are distantly separated in the pathogen protein sequence and brought into proximity by the folding of the protein (Barlow et al. 1986; Van Regenmortel 1996). Identification of discontinuous epitopes is difficult, since the complete analysis must be done in context of the native antigen structure. The most informative and accurate method for identification of discontinuous epitopes is determination of structures of antigen–antibody complexes by X-ray crystallography (Fleury et al. 2000; Mirza et al. 2000). The use of discontinuous epitopes derived from presently available X-ray structures is complicated by two major problems: First, the available data on discontinuous epitopes in different antigens is much reduced compared to linear epitopes; second, very few antigens have been studied to completely identify various discontinuous epitopes in the same antigen. The existence of undetected epitopes that are not identified in the data set can make it harder to develop good prediction algorithms because they influence the measured performance. However, detailed structural knowledge on antibody–antigen complexes is growing, and allows for broader analysis of discontinuous epitopes in various antigens and development of better prediction methods.

Correlation between surface exposure and B-cell epitopes has been known for many years (Novotny et al. 1986; Thornton et al. 1986). Recently, two new methods using protein structure and surface exposure for prediction of B-cell epitopes have been published (Kulkarni-Kale et al. 2005; Batori et al. 2006). However, none of these new methods using protein structure as input have the primary focus on discontinuous epitopes.

Here, we present a prediction method for residues located in discontinuous B-cell epitopes. DiscoTope uses a combination of amino acid statistics, spatial information, and surface exposure. It is trained on a compiled data set of discontinuous epitopes from 76 X-ray structures of antibody/antigen protein complexes. We present the performance of DiscoTope compared to the Parker hydrophilicity scale (Parker et al. 1986) for a comparison to a classical, sequence-based method that has been shown recently to perform well for prediction of linear epitopes (Larsen et al. 2006). In addition, we compare the performance with predictions based on surface accessibility measured on antigen structures using the program NACCESS (Hubbard and Thornton 1993). We demonstrate that DiscoTope is generally the best performing of all methods described here. Finally, we present the delineation of epitopes in the malaria protein apical membrane antigen 1 (AMA1) where DiscoTope success-

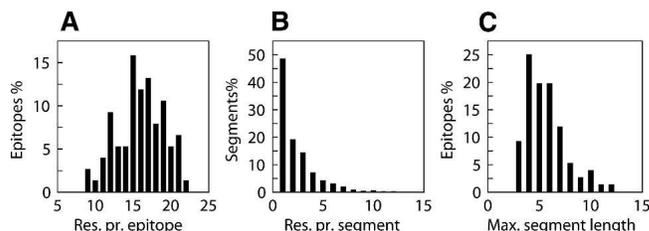
fully predicts epitope residues that have been identified using either various experimental or sequence analysis techniques.

## Results

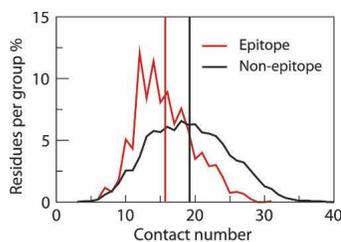
### *Properties of discontinuous B-cell epitopes*

In order to get a well-established basis for development and evaluation of the prediction method, we compiled a discontinuous epitope data set from 76 X-ray structures of complexes between antibodies and protein antigens. We analyzed the data set to find distributions for the number of residues per epitope, the number of residues per sequential stretch in epitopes, and the longest sequential stretch per epitope. These distributions are shown in Figure 1. The total number of residues per epitope ranged from 9 to 22, and >60% of the epitopes consisted of 14 to 19 residues (Fig. 1A). Segments with a single epitope residue represented >45% of the 528 segments in the data set (Fig. 1B). The longest sequential stretch of identified residues per epitope ranged between 3 and 12 residues, and >75% of epitopes comprised a sequential stretch of a maximum length of 4 to 7 residues (Fig. 1C). These findings confirm that most epitopes in the data set are indeed discontinuous, and composed by small parts of the antigen sequence forming a binding region for the antibody.

The data set was analyzed with respect to surface exposure by determining the number of intramolecular C $\alpha$  atom contacts for each residue (Fig. 2). A low contact number correlates with localization close to the surface or in protruding regions of antigen structures. A *t*-test showed that residues identified as part of epitopes in the data set had significantly lower numbers of contacts compared to the nonepitope residues ( $P < 10^{-5}$ ). The average number of contacts and standard error of mean for epitope residues was  $15.7 \pm 0.12$ , and for nonepitope residues the average contact number and standard error of mean was  $19.2 \pm 0.05$  (see Fig. 2, vertical lines). The



**Figure 1.** Analysis of the complete data set of discontinuous B-cell epitopes. (A) Distribution of the number of residues per epitope. (B) Distribution of the number of residues per sequential stretch of epitopes. (C) Distribution of the maximum length of a sequential stretch per epitope.



**Figure 2.** Contact numbers of epitope residues in the data set compared to non-epitope residues. The curves show the distribution of contact numbers for epitope residues (red curve) compared to non-epitope residues (black curve). The vertical lines represent the mean value of contact numbers for the epitope residues (red line) and for the non-epitope residues (black line).

finding that epitopes are in exposed or protruding regions is in agreement with previous analysis of B-cell epitopes (Novotny et al. 1986; Thornton et al. 1986). As shown in Figure 2, the two distributions are overlapping. This is most probably caused by the incomplete annotation of the data set or because other factors than contact numbers are important in defining an epitope.

For the development and evaluation of prediction methods, the 76 antigens in the data set were grouped into 25 nonhomologous groups (for more details, see Materials and Methods). From these 25 groups, five sets (of five groups each) were constructed and used for fivefold cross-validated training and evaluation, to avoid optimizing and evaluating on similar antigens.

#### Log-odds ratios calculated from the epitope data set

We analyzed the statistics of amino acids in epitopes and non-epitopes of the data set by calculation of log-odds ratios from peptides of the data set. A peptide-based approach of similarity reduction was chosen to avoid skewing log-odds ratios toward highly redundant epitopes in the data set. Peptides with high similarity in the data set were weighted lower than peptides with low similarity, and therefore, the length of the peptides played an important role in the derivation of log-odds ratios. We used raw log-odds ratios as epitope propensities for prediction of epitopes in the training sets and found a peptide length of nine residues to be optimal.

Table 1 shows epitope log-odds ratios calculated from homology-reduced peptides of the total data set of 76 proteins. Of the 20 amino acids, asparagine (N), arginine (R), proline (P), and lysine (K) had the highest log-odds ratios, meaning that they are overrepresented in epitopes compared to non-epitopes of the data set. Cysteine (C), alanine (A), leucine (L), valine (V), and phenylalanine (F) had very low log-odds ratios, and are correspondingly underrepresented in epitopes. Interestingly, we found several discrepancies between the Parker hydrophilicity

scale and the log-odds ratios (Table 1). For example, the most hydrophobic residue, tryptophan (W), did not have a particularly low log-odds ratio. The most hydrophilic residues, aspartate (D) and glutamate (E), had relatively moderate log-odds ratios. Arginine (R) and proline (P) had some of the highest log-odds ratios, but are ranked close to the middle of the Parker hydrophilicity scale. Cysteine (C) and alanine (A) are ranked close to the middle of the Parker scale, but had some of the lowest log-odds ratios.

#### Evaluation of uncombined methods for B-cell epitope prediction

To test the predictive strength of contact numbers and the epitope propensity scale of log-odds ratios on discontinuous epitopes, we used the area under receiver operator curves (AUC) averages over different evaluation sets (see details in Materials and Methods). We additionally tested a sequential average of log-odds ratios as prediction score similar to the approach recommended for the hydrophilicity scale by Parker et al. (1986). The optimal window size for sequential averaging of log-odds ratios was found to be nine residues based on the predictive performance on the training sets (data not shown).

We found that the epitope log-odds ratios used with sequential averaging performed better than the sequentially averaged Parker hydrophilicity scale on the discontinuous epitopes (Fig. 3A). The raw epitope log-odds

**Table 1.** The Parker hydrophilicity scale and epitope log-odds ratios

| Amino acid | Parker | Log-odds ratios |
|------------|--------|-----------------|
| D          | 2.460  | 0.691           |
| E          | 1.860  | 0.346           |
| N          | 1.640  | 1.242           |
| S          | 1.500  | -0.145          |
| Q          | 1.370  | 1.082           |
| G          | 1.280  | 0.189           |
| K          | 1.260  | 1.136           |
| T          | 1.150  | -0.233          |
| R          | 0.870  | 1.180           |
| P          | 0.300  | 1.164           |
| H          | 0.300  | 1.098           |
| C          | 0.110  | -3.519          |
| A          | 0.030  | -1.522          |
| Y          | -0.780 | 0.030           |
| V          | -1.270 | -1.474          |
| M          | -1.410 | 0.273           |
| I          | -2.450 | -0.713          |
| F          | -2.780 | -1.147          |
| L          | -2.870 | -1.836          |
| W          | -3.000 | -0.064          |

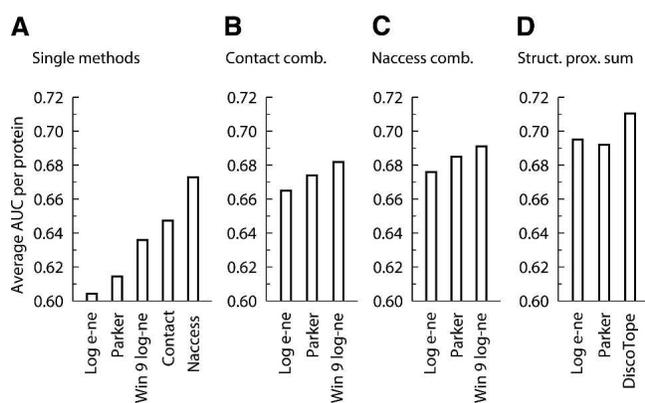
Amino acids are listed with descending hydrophilicity using the values of the Parker scale.

propensity scale gave an average performance of 0.604 on the evaluation sets. Smoothing of the log-odds ratios using a sequential average of nine residues improved the performance to 0.636. The Parker scale was used with a smoothing window of seven residues and had a performance of 0.614. Compared to the methods based on propensity scales, the methods based on contact numbers and NACCESS relative surface area (RSA) values had considerably higher performances of 0.647 and 0.673, respectively (Fig. 3A).

### Combination methods for epitope prediction

We additionally tested the prediction of epitope residues using surface localization values based on contact numbers or NACCESS RSAs in combination with epitope log-odds ratios or the Parker hydrophilicity scale. One combination approach was to use a sum of weighted prediction scores from surface localization measures and methods based on sequential information (log-odds ratios or hydrophilicity scores). A second approach was tested by summing log-odds ratios, sequentially averaged log-odds ratios or Parker scale scores of residues in spatial proximity and adding the contact numbers to give a prediction score. For each combination, we estimated the relative weight on the surface localization score by optimizing the predictive performance on the training sets measured in average AUC. The optimized weights are listed in Table 2.

The predictive performances of the combination methods were tested by calculating the average AUC from



**Figure 3.** Evaluation of B-cell epitope prediction methods. The average AUC of various methods on the five evaluation sets. “Log e-ne” denotes raw log-odds ratios; “Parker” denotes the Parker hydrophilicity scale; “Win9 log e-ne” denotes the log-odds ratios used with a smoothing window of nine residues; “Contact” denotes contact numbers; and “Naccess” denotes NACCESS RSA values. (A) The performance of single methods. (B) The performance of simple combination methods using contact numbers. (C) The performance of simple combination methods using NACCESS RSA values. (D) The performance of the structural proximity sum methods.

**Table 2.** Optimal weights on surface localization scores for combination methods

|          | Sequential average |                |        | Structural proximity sum |                |        |
|----------|--------------------|----------------|--------|--------------------------|----------------|--------|
|          | Log-odds           | Win 9 log-odds | Parker | Log-odds                 | Win 9 log-odds | Parker |
| Contacts | 1.5                | 0.25           | 1.75   | -1.5                     | -0.5           | -1.5   |
| NACCESS  | 2.0                | 0.25           | 1.75   | —                        | —              | —      |

Win 9 denotes epitope log-odds ratios smoothed by a window of nine sequential residues.

predictions on the evaluation data sets (Fig. 3B–D). Simple linear combinations of the Parker scale, raw log-odds ratios, and smoothed log-odds ratios with structure-based methods (contact numbers and NACCESS RSA values) in general improved the performance (Fig. 3B,C). Combination methods using raw log-odds ratios had a performance of 0.665 for the combinations with contact numbers and 0.676 for the combination with NACCESS RSA values. The linear combinations with the Parker method had performances of 0.674 for the contact number combination and 0.685 for the NACCESS RSA combination. Using a combination of smoothed log-odds ratios combined with contact numbers yielded a performance of 0.682. The best performing method of the simple linear combinations was the combination of smoothed log-odds ratios with NACCESS RSAs. This method had a performance of 0.691 on the evaluation sets.

Methods based on a combination of structural proximity sums of propensity scales with contact numbers gave the best performances on the evaluation sets (Fig. 3D). The performance of the structural proximity sum method based on Parker predictive values combined with contact numbers had a performance of 0.692. The corresponding structural proximity sum method using raw log-odds ratios had a performance of 0.695. The best performing method on the evaluation data sets was the structural proximity sum of sequentially smoothed epitope log-odds ratios combined with contact numbers. This method was shown to have a performance of 0.711, which is significantly better than the method based on structural averaging using raw log-odds ratios ( $P = 0.040$ ). The method is also significantly better than the Parker method ( $P = 0.007$ ) and marginally better than the NACCESS RSA method ( $P = 0.105$ ). We call this method DiscoTope.

### Analysis of the DiscoTope method for discontinuous B-cell epitope prediction

We decided to further analyze the Parker hydrophilicity, NACCESS RSA, and DiscoTope predictions to get a more detailed comparison of the performances of the methods.

**Table 3.** Sensitivity of methods corresponding to a number of selected specificity levels

| Specificity | DiscoTope   |           | NACCESS RSA |           | Parker      |           |
|-------------|-------------|-----------|-------------|-----------|-------------|-----------|
|             | Sensitivity | Threshold | Sensitivity | Threshold | Sensitivity | Threshold |
| 95%         | 15.5%       | -3.1      | 8.7%        | 88%       | 11.0%       | 1.07      |
| 90%         | 24.2%       | -4.7      | 18.9%       | 74%       | 19.6%       | 0.88      |
| 85%         | 32.3%       | -6.0      | 27.2%       | 67%       | 27.1%       | 0.74      |
| 80%         | 40.2%       | -6.9      | 37.3%       | 60%       | 36.0%       | 0.60      |
| 75%         | 47.3%       | -7.7      | 44.1%       | 55%       | 39.8%       | 0.50      |

Residues with a prediction value above the thresholds were predicted as parts of epitopes.

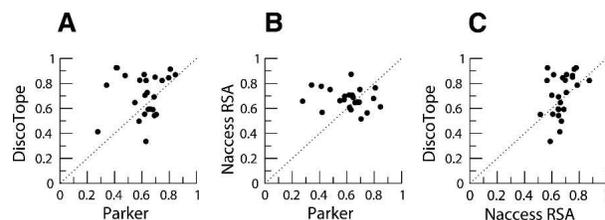
A comparison of the sensitivity of the three methods was done based on a number of selected specificities (Table 3). In Table 3, we have additionally listed prediction threshold values to facilitate general use of all three methods for B-cell epitope prediction. For all five specificity levels, DiscoTope had the highest sensitivity of the three methods. At a level of 95% specificity (which means only 5% false positive predictions) DiscoTope detected 15% of the epitopes. The Parker method had higher sensitivity than the NACCESS RSA method for the 95% and 90% specificity levels. This is in contrast to the averaged AUC value on the five evaluation sets, which was found to be higher for the NACCESS method than for the Parker method (Fig. 3A).

In order to analyze the performances of the three methods on different groups of antigens, we compared prediction AUC values for each of the 25 nonhomologous antigen groups (Fig. 4). For the majority of the groups of antigens in the data set, the DiscoTope method had a better performance than the Parker method (Fig. 4A). However, in eight groups of antigens the epitope residues were more accurately predicted using the Parker method. The same tendency was observed for the NACCESS RSA method, where the Parker method performed best for 12 groups (Fig. 4B). Comparison of the DiscoTope and NACCESS RSA methods showed that, even though the average AUC value for the 25 groups was highest for the DiscoTope method, the NACCESS RSA method performed best for 10 of the antigen groups (Fig. 4C). We found that the DiscoTope and the NACCESS RSA methods had six groups in common for which the Parker method performed best. These groups were represented by the PDB antigen entries 1JPS, 2JEL, 1TQB, 1AR1, 1OAZ, 1EO8. The fact that both surface accessibility based methods had lower performance than the Parker scale method suggests that the measured surface accessibility for single antigen chains is not sufficient for epitope prediction in all types of antigens. Three of the six groups (represented by antigens 1JPS, 1AR1, and 1EO8) contained antigens that have elongated structures. Furthermore, antigens of 1JPS, 1AR1, and 1EO8 are all known as subunits of larger biological complexes associated with

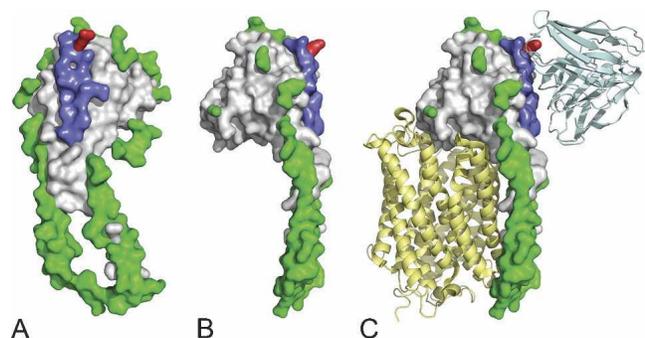
membranes (Ostermeier et al. 1997; Fleury et al. 2000; Faelber et al. 2001). Perhaps not surprising, the single antigen chain approach taken by the DiscoTope method clearly could not correctly measure the surface accessibility of all residues in such proteins. For an example, the structure of the antigen of 1AR1 is shown in Figure 5. On the plot, most of the residues in the antigen that had the lowest contact numbers are not in proximity of the epitope (Fig. 5A,B). In fact, only one residue of the epitope was among the 30% residues in the antigen with lowest contact numbers. The antigen of 1AR1 is a subunit of a membrane spanning cytochrome *c* oxidase (Fig. 5C) and the largest continuous region of residues with low contact numbers corresponds to a region of the protein that is described as membrane-spanning (Ostermeier et al. 1997).

#### Prediction of B-cell epitope residues in apical membrane antigen 1

To evaluate our method on B-cell epitopes that are mapped using other types of methods than X-ray crystallography we tested the predictions of DiscoTope on the structure of the ectodomain from AMA1 (Bai et al. 2005; Pizarro et al. 2005). No AMA1 epitopes are included in the data set of discontinuous epitopes derived from the PDB. However, two separate epitopes recognized by monoclonal antibodies Mab1F9 and Mab4G2



**Figure 4.** Dot plots showing comparisons of performances of the Parker method, the NACCESS RSA method, and the DiscoTope method. Circles indicate average AUC per group showed for the 25 groups of different antigens. The dotted lines indicate points where the methods perform equally.



**Figure 5.** Structure of the IAR1 antigen. The antigen is a subunit of the cytochrome *c* oxidase (Ostermeier et al. 1997). (A) The 30% of residues with lowest contact numbers are shown in green. In red is shown a residue that is part of the 30% with lowest contact numbers and of the epitope from the data set. The rest of the epitope is shown in blue. (B) The structure mentioned above and rotated 90 degrees. (C) The complex of the cytochrome *c* oxidase with antibody fragments. The IAR1 antigen is color coded as in A and B. Antibody fragments are shown in light blue. The other subunit of the cytochrome *c* oxidase is shown in yellow. Membrane spanning helices of the IAR1 antigen are part of the lower half of the structure.

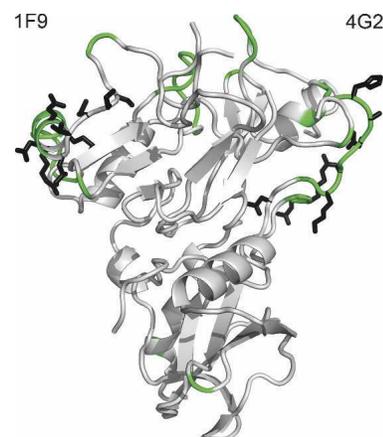
have been experimentally mapped on the AMA1 ectodomain: The Mab1F9 epitope was mapped using phage-display of peptides and point mutations of E197 (Coley et al. 2006); the discontinuous Mab4G2 epitope was mapped in detail by point mutation of nine residues (Pizarro et al. 2005). In addition, Bai et al. (2005) have classified five residues (including E197 and other residues in the same region of the structure) as highly polymorphic in *Plasmodium falciparum* AMA1 sequences. It has been suggested that the polymorphism is caused by selection pressure on the antigen to avoid the host immune system. We used a DiscoTope prediction threshold of  $-4.7$ , which corresponds to a specificity of 90% and 24% sensitivity (Table 3). In AMA1, 43 of 311 residues were predicted as epitope residues. Most of the predicted epitope residues cluster in three separate regions of the AMA1 structure (Fig. 6). DiscoTope successfully identified two of the eight residues in the 1F9 epitope that were mapped using phage-display (D196 and E197). In the discontinuous 4G2 epitope, all nine residues except D348 were predicted to be part of epitopes. All of the five highly polymorphic residues described by Bai et al. (2005) were predicted to be located in epitopes. Thus, DiscoTope successfully predicted epitope residues of AMA1 that have been mapped by using diverse methods.

## Discussion

In this paper, we present DiscoTope, a novel method for prediction of residues located in discontinuous B-cell epitopes. DiscoTope combines surface localization and spatial properties of a protein structure with a novel

epitope propensity scale. The combination is defined in terms of a simple weighted sum of the contact number and a sum of sequentially averaged epitope log-odds ratios of spatially proximate residues. We propose to use DiscoTope for prediction of discontinuous epitope residues for several reasons. First, we have shown on a data set of discontinuous epitopes that the average predictive performance of the DiscoTope is significantly higher than the Parker propensity scale and marginally higher than the surface localization score defined by the NACCESS RSA score. Second, we have shown that DiscoTope correctly predicts residues in epitopes that have been identified using different techniques such as phage-display, point mutation, and sequence analysis. Third, the DiscoTope prediction method is publicly available on [www.cbs.dtu.dk/services/DiscoTope](http://www.cbs.dtu.dk/services/DiscoTope), and the output of the method is easily interpreted.

The Parker hydrophilicity scale is often used for prediction of linear B-cell epitopes by smoothing values in a seven-residue window (Parker et al. 1986). Compared to the epitope log-odds ratios smoothed over a window of nine residues developed here, the Parker scale was not as accurate for prediction of discontinuous B-cell epitopes in the data set. The difference in ranking between the two scales suggests that our log-odds ratios represent more characteristics of the epitopes than only hydrophilicity. Possibly, this difference contributes to a better predictive performance on the data set since combinations of various propensity scales including hydrophilicity, flexibility, accessibility, and  $\beta$ -turn prediction are better than single propensity scales for epitope prediction (Pellequer et al. 1991). Our findings, that surface accessibility values improved the prediction of residues in B-cell epitopes, are in agreement with recently reported results by Batori et al. (2006). In addition, the combination of propensity



**Figure 6.** Predicted epitope residues of the AMA1 ectodomain. Backbone atoms of residues predicted by DiscoTope as parts of epitopes are highlighted in green. Side chains of the residues mapped to the monoclonal antibodies 1F9 and 4G2 are shown in black.

scale methods with structural information improved the performance considerably. This suggests that both accessibility and chemical characteristics are important in descriptors of discontinuous B-cell epitopes. Combination methods using a number of propensity scales have been used for B-cell epitopes for more than 15 years (Pellequer et al. 1991); however, DiscoTope is the first reported method combining a propensity scale with three-dimensional structural information, such as spatial proximity.

Van Regenmortel (1996) has addressed the problem of using protein sequences for prediction of B-cell epitopes, which are in reality multidimensional. He concluded that more input data, such as the antigen three-dimensional structure, is needed for accurate prediction. The requirement of structural input for B-cell epitope prediction is a limiting factor for the general use of the method. However, structural genomics projects help to increase the number of X-ray crystallography structures determined of proteins in general, and to cover larger areas of the structure space. Therefore, the requirement of protein structures as input for prediction methods will become a decreasing problem, because more structures will be determined and better homology models can be obtained.

In general, methods based on structural information were shown to predict residues in discontinuous B-cell epitopes with a higher performance measured in average AUC than propensity scale methods, which only used sequential information. In all methods of evaluation, the DiscoTope method was shown to have the highest performance. However, we found that the Parker hydrophilicity scale had a higher sensitivity than the NACCESS RSA method on the 95% and 90% specificity levels. These results illustrate the importance of using other measures of performance for evaluation in addition to the AUC.

We found that for antigen groups that contain antigens that are part of larger biological complexes, the performances of both the NACCESS RSA method and the DiscoTope method were relatively low. The low performances were due to an incorrect measure of surface accessibility of regions that are part of protein–protein interaction sites or are embedded in a membrane. Therefore, we believe that the outcome of prediction methods for B-cell epitopes should be combined with additional information about properties such as biological complex formation, membrane interaction, and glycosylation.

The accuracy of the described methods for B-cell epitope prediction was still relatively moderate. This may partly be caused by the incomplete identification of epitopes in the antigens of the data set. If the methods correctly predicted an epitope that was not bound by the antibody in the corresponding complex PDB file it counted as a false positive. However, since the same data

set was used in the evaluation of all methods described here, we assumed that incomplete identification had the same influence on the predictive performance of all methods, and hence, negligible influence on their relative ranking. The predictive performance of the method developed by Batori et al. (2006) was evaluated on six epitopes of one single antigen. This evaluation approach, using an antigen where all epitopes are more completely identified, possibly had the effect that the false positive proportion was lower and the measured performance was higher. In our approach of evaluation, we chose to include as much variation as possible and thereby avoid biasing the method toward a certain type of antigen or epitope. However, a future evaluation of our DiscoTope method using a data set of antigens with more completely identified epitopes would be of interest.

Recently, Schlessinger et al. (2006) have developed a sophisticated method for identification of epitopes in antibody/antigen complex structures. The method is based on an analysis and identification of complementarity determining regions (CDRs) of the antibody and a subsequent identification of epitopes by mapping residues in the antigen in proximity to CDRs. The identification described in this paper was simply based on antigen residues in proximity to antibody residues in general, and it is plausible that a future application of the identification method developed by Schlessinger et al. could improve the DiscoTope method.

Because of their nonlinearity, discontinuous epitopes pose other problems than linear epitopes in vaccine design. Not only must the new vaccine contain the amino acids or atoms that are necessary for binding and eliciting specific antibodies, but a conservation of the correct spatial conformation is also needed. DiscoTope can predict residues that are likely to be part of discontinuous epitopes. Subsequently, antibody binding studies and site-directed mutagenesis may help to group predicted epitope residues into epitopes and validate binding. Analysis of the local conformations of epitope residues in the antigen structure may also aid the design of vaccines, because a vaccine based on a discontinuous epitope must have these conformations preserved. The preservation may be obtained using native proteins, subdomains of a protein, redesigned proteins carrying the epitope, or mimotope peptides in vaccines. Therefore, we consider discontinuous epitopes useful for rational vaccine design.

## Materials and methods

### *Preparation of the data set*

A list of experimentally determined protein antigen–antibody structures was obtained from the SACS database of antibody

crystal structure information (Allcorn and Martin 2002). The list was filtered to contain only structures determined to a resolution  $<3 \text{ \AA}$  with protein antigens of  $>25$  amino acids. Coordinate files corresponding to the filtered list were downloaded from the Protein Data Bank (PDB, <http://www.rcsb.org/pdb>). The final data set contained 76 complexes of antibody–antigen pairs. Epitope residues in the data set were defined as antigen amino acids having atoms within a  $4 \text{ \AA}$  distance from antibody atoms. Comparisons based on a subset of five identified epitopes with residues reported as antibody interacting (Padlan et al. 1989; Muller et al. 1998; Fleury et al. 2000; Mirza et al. 2000; Romijn et al. 2003) showed that a distance threshold of  $4 \text{ \AA}$  gave an annotation corresponding well to that made by human experts (92% of the epitope residues were correctly identified, and only 1% of the nonpeptide residues were identified as epitope residues). Only a single epitope was represented in each PDB file. All other epitopes that might exist in a given antigen were treated as nonpeptides in our analysis. Certain antigens were represented multiple times in the data set (29 antigens are variants of lysozyme). Therefore, we grouped the data set according to antigen homology. Homology in the data set of 76 proteins was determined using a BLAST search (Altschul et al. 1997) with the BLOSUM80 matrix against all other antigens in the data set combined with a homology threshold as described by Lund et al. (1997). Antigens were then split into 25 groups with low homology between the groups (BLAST E-values  $>0.30$  between groups). The data set annotations and the groups of antigens are publicly available at <http://www.cbs.dtu.dk/suppl/immunology/DiscoTope>. Finally, the 25 nonhomologous groups of antigens were divided into five data sets used for cross-validated training and evaluation.

#### *Use of the Parker hydrophilicity scale*

The average Parker scale value over a window of seven residues was used for the per-residue epitope prediction value as proposed by Parker et al. (1986).

#### *Definition of surface residues*

A combined measure of amino acid surface localization and structural protrusion was obtained by using residue contact numbers. The residue contact number is the number of  $C\alpha$  atoms in the antigen within a distance of  $10 \text{ \AA}$  of the residue  $C\alpha$  atom (Nishikawa and Ooi 1980). For a more direct measure of residue solvent accessibility, the relative solvent-accessible surface area per residue was calculated for antigen chains extracted from each PDB file using the NACCESS program (Hubbard and Thornton 1993). NACCESS default options were used with a probe radius of  $1.4 \text{ \AA}$ .

#### *Performance measures*

The area under a receiver operator characteristics curve (AUC) (Swets 1988) was used as performance measure. A receiver operator characteristics curve is constructed by varying the prediction threshold and plotting the false-positive proportion, or 1-specificity, on the  $X$ -axis against the true positive proportion, or sensitivity, on the  $Y$ -axis (Swets 1988; Lund et al. 2005). We calculate the AUC on a per protein basis. This ensures that a prediction where all residues in a protein are predicted as only epitopes or only nonpeptides has an AUC of

0.5 corresponding to a random prediction. The performance of each method was measured as the average AUC, average specificity, and average sensitivity for the 25 antigen groups.

#### *Statistical analysis*

Mean values of contact numbers for epitope residues and nonpeptide residues were analyzed using a double-sided  $t$ -test (standard deviation = 0.121,  $n = 1202$  for epitope residues, and standard deviation = 0.050,  $n = 13,242$  for nonpeptide residues.) A bootstrapping approach was used for pairwise comparisons of the average AUC values to determine the significance of the performances (Efron and Tibshirani 1993). For each method, the 25 values of average AUC value per antigen group were resampled 100,000 times in order to obtain a robust estimate of the  $P$ -values.

#### *Derivation of epitope log-odds ratios*

Four of the five data sets (the training sets) were used for derivation of epitope log-odds ratios. A series of peptides were produced by sliding an odd-sized window through the sequences of antigens in the training sets. The peptides were then sorted into an epitope group and a nonpeptide group, depending on the identification of the residue in middle position as epitope residue or as nonpeptide residue. Weight matrices were calculated from the peptides in each group using the method described by Nielsen et al. (2004), including sequence clustering, sequence weighting, and pseudo counts with a weight of 200. Finally, the log-odds ratios at the central matrix position for each of the 20 amino acids in the epitope group relative to the nonpeptide group were calculated in half bits and used as an epitope propensity scale.

#### *Using log-odds ratios for epitope prediction*

For prediction of epitope residues, the raw log-odds ratios were used alone or in combination with a smoothing window calculating the sequential average of the epitope propensity scale values. The optimal length of peptides used for the derivation of log-odds ratios and the optimal size of the smoothing window were determined with respect to the predictive performance on the training sets used for calculating the log-odds ratios. The performance reported is the fivefold cross-validated performance on the data set. This reduces the risk of overestimating the performance, since the calculation of the log-odds ratios and optimization of other parameters, such as the peptide length and the smoothing window size, are estimated on the training set, and hence are not biased by the evaluation set data.

#### *Simple combinations of propensity scales with structure-based methods*

Contact numbers, NACCESS RSAs, and Parker hydrophilicity values were normalized by subtracting the mean and dividing with the standard deviation. The normalized contact numbers were multiplied by  $-1$  in order to correlate high values with surface localization. Subsequently, the different propensity scales were combined with contact numbers or NACCESS RSAs using a linear combination with a weight on the surface measure ranging from 0.001 to 100. Optimal weights were

determined using the training sets. Finally, the performance was evaluated on evaluation sets.

### Structural proximity sum of epitope log-odds ratios

Alternatively, the epitope log-odds ratios or the Parker hydrophilicity scale were used by summing values for all residues with C $\alpha$  atoms within a 10 Å distance of each residue. We tested a number of weighting schemes for the proximity sums, for instance, based on the distance to the central residue, the contact number for the residue, and a combination of the two. However, the simple approach where all residues carry equal weight gave the highest performance on the training sets (data not shown).

### Prediction of epitopes in AMA1

Chain A of the AMA1 ectodomain from *P. falciparum* (PDB code 1Z40) was used for DiscoTope epitope prediction. We chose to use 1Z40 instead of a full-length AMA1 ectodomain structure (1W8K) because the main part of the residues in the 4G2 epitope was not observed in the latter. Residues 348, 351, 352, 354–356, 385, and 388–389 were counted as residues in the 4G2 epitope (Pizarro et al. 2005); residues 191–199 were counted as part of the 1F9 epitope (Coley et al. 2006); and residues 187, 197, 200, 230, and 243 were counted as highly polymorphic residues (Bai et al. 2005).

### Acknowledgments

We thank Claus Lundegaard for helpful comments. This work was funded in part by NIH Contract No. HHSN266200400083C.

### References

Alix, A.J. 1999. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine* **18**: 311–314.

Allcorn, L.C. and Martin, A.C. 2002. SACS–Self-maintaining database of antibody crystal structure information. *Bioinformatics* **18**: 175–181.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.

Bai, T., Becker, M., Gupta, A., Strike, P., Murphy, V.J., Anders, R.F., and Batchelor, A.H. 2005. Structure of AMA1 from *Plasmodium falciparum* reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc. Natl. Acad. Sci.* **102**: 12736–12741.

Barlow, D.J., Edwards, M.S., and Thornton, J.M. 1986. Continuous and discontinuous protein antigenic determinants. *Nature* **322**: 747–748.

Batori, V., Friis, E.P., Nielsen, H., and Roggen, E.L. 2006. An in silico method using an epitope motif database for predicting the location of antigenic determinants on proteins in a structural context. *J. Mol. Recognit.* **19**: 21–29.

Blythe, M.J. and Flower, D.R. 2005. Benchmarking B cell epitope prediction: Underperformance of existing methods. *Protein Sci.* **14**: 246–248.

Coley, A.M., Parisi, K., Masciantonio, R., Hoeck, J., Casey, J.L., Murphy, V.J., Harris, K.S., Batchelor, A.H., Anders, R.F., and Foley, M. 2006. The most polymorphic residue on *Plasmodium falciparum* apical membrane antigen 1 determines binding of an invasion-inhibitory antibody. *Infect. Immun.* **74**: 2628–2636.

Debelle, L., Wei, S.M., Jacob, M.P., Hornebeck, W., and Alix, A.J. 1992. Predictions of the secondary structure and antigenicity of human and bovine tropoelastins. *Eur. Biophys. J.* **21**: 321–329.

Efron, B. and Tibshirani, R.J. 1993. *An introduction to the bootstrap*, 1st ed. Chapman and Hall, London.

Faelber, K., Kirchhofer, D., Presta, L., Kelley, R.F., and Muller, Y.A. 2001. The 1.85 Å resolution crystal structures of tissue factor in complex

with humanized Fab D3h44 and of free humanized Fab D3h44: Revisiting the solvation of antigen combining sites. *J. Mol. Biol.* **313**: 83–97.

Fleury, D., Daniels, R.S., Skehel, J.J., Knossow, M., and Bizebard, T. 2000. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins* **40**: 572–578.

Hopp, T.P. and Woods, K.R. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci.* **78**: 3824–3828.

Hubbard, S.J. and Thornton, J.M. 1993. *NACCESS computer program*. Department of Biochemistry and Molecular Biology, University College of London, UK.

Jameson, B.A. and Wolf, H. 1988. The antigenic index: A novel algorithm for predicting antigenic determinants. *Comput. Appl. Biosci.* **4**: 181–186.

Kulkarni-Kale, U., Bhosle, S., and Kolaskar, A.S. 2005. CEP: A conformational epitope prediction server. *Nucleic Acids Res.* **33**: W168–W171.

Larsen, J.E., Lund, O., and Nielsen, M. 2006. Improved method for predicting linear B-cell epitopes. *Immunome Res* **2**: 2.

Leitner, T., Foley, B., Hahn, B., Marx, P., McCutchan, F., Mellors, J., Wolinsky, S., and Korber, B. 2003. *HIV sequence compendium 2003*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR04–7420.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak, S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* **10**: 1241–1248.

Lund, O., Nielsen, M., Lundegaard, C., Kesmir, C., and Brunak, S. 2005. *Immunological Bioinformatics*, 1st ed., pp. 100–101. MIT Press, Cambridge, MA.

Maksyutov, A.Z. and Zagrebelnaya, E.S. 1993. ADEPT: A computer program for prediction of protein antigenic determinants. *Comput. Appl. Biosci.* **9**: 291–297.

Mirza, O., Henriksen, A., Ipsen, H., Larsen, J.N., Wissenbach, M., Spangfort, M.D., and Gajhede, M. 2000. Dominant epitopes and allergic cross-reactivity: Complex formation between a Fab fragment of a monoclonal murine IgG antibody and the major allergen from birch pollen Bet v. 1. *J. Immunol.* **165**: 331–338.

Muller, Y.A., Chen, Y., Christinger, H.W., Li, B., Cunningham, B.C., Lowman, H.B., and de Vos, A.M. 1998. VEGF and the Fab fragment of a humanizing neutralizing antibody: Crystal structure of the complex at 2.4 Å resolution and mutational analysis of the interface. *Structure* **6**: 1153–1167.

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. 2004. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* **20**: 1388–1397.

Nishikawa, K. and Ooi, T. 1980. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. Pept. Protein Res.* **16**: 19–32.

Novotny, J., Handschumacher, M., Haber, E., Bruccoleri, R.E., Carlson, W.B., Fanning, D.W., Smith, J.A., and Rose, G.D. 1986. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl. Acad. Sci.* **83**: 226–230.

Odorico, M. and Pellequer, J.L. 2003. BEPITOPE: Predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recognit.* **16**: 20–22.

Ostermeier, C., Harrenga, A., Ermler, U., and Michel, H. 1997. Structure at 2.7 Å resolution of the *Paracoccus denitrificans* two-subunit cytochrome *c* oxidase complexed with an antibody FV fragment. *Proc. Natl. Acad. Sci.* **94**: 10547–10553.

Padlan, E.A., Silverton, E.W., Sheriff, S., Cohen, G.H., Smith-Gill, S.J., and Davies, D.R. 1989. Structure of an antibody–antigen complex: Crystal structure of the HyHEL-10 Fab–lysozyme complex. *Proc. Natl. Acad. Sci.* **86**: 5938–5942.

Parker, J.M., Guo, D., and Hodges, R.S. 1986. 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* **25**: 5425–5432.

Pellequer, J.L., Westhof, E., and Van Regenmortel, M.H. 1991. Predicting location of continuous epitopes in proteins from their primary structures. *Methods Enzymol.* **203**: 176–201.

Pizarro, J.C., Vulliez-Le Normand, B., Chesne-Seck, M.L., Collins, C.R., Withers-Martinez, C., Hackett, F., Blackman, M.J., Faber, B.W., Remarque, E.J., Kocken, C.H., et al. 2005. Crystal structure of the malaria vaccine candidate apical membrane antigen 1. *Science* **308**: 408–411.

Romijn, R.A., Westein, E., Bouma, B., Schiphorst, M.E., Sixma, J.J., Lenting, P.J., and Huizinga, E.G. 2003. Mapping the collagen-binding

- site in the von Willebrand factor-A3 domain. *J. Biol. Chem.* **278**: 15035–15039.
- Saha, S., Bhasin, M., and Raghava, G.P. 2005. Beipep: A database of B-cell epitopes. *BMC Genomics* **6**: 79.
- Schlessinger, A., Ofran, Y., Yachdav, G., and Rost, B. 2006. Epitome: Database of structure-inferred antigenic epitopes. *Nucleic Acids Res.* **34**: D777–D780.
- Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science* **240**: 1285–1293.
- Thornton, J.M., Edwards, M.S., Taylor, W.R., and Barlow, D.J. 1986. Location of “continuous” antigenic determinants in the protruding regions of proteins. *EMBO J.* **5**: 409–413.
- Toseland, C.P., Clayton, D.J., McSparron, H., Hemsley, S.L., Blythe, M.J., Paine, K., Doytchinova, I.A., Guan, P., Hattotuagama, C.K., and Flower, D.R. 2005. AntiJen: A quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res* **1**: 4.
- Van Regenmortel, M.H.V. 1996. Mapping epitope structure and activity: From one-dimensional prediction to four-dimensional description of antigenic specificity. *Methods* **9**: 465–472.
- Van Regenmortel, M.H. and Pellequer, J.L. 1994. Predicting antigenic determinants in proteins: Looking for unidimensional solutions to a three-dimensional problem? *Pept. Res.* **7**: 224–228.