IMMUNOME RESEARCH

REVIEW Open Access

Recent advances in B-cell epitope prediction methods

Yasser EL-Manzalawy[1,2*], Vasant Honavar[2*]

Abstract

Identification of epitopes that invoke strong responses from B-cells is one of the key steps in designing effective vaccines against pathogens. Because experimental determination of epitopes is expensive in terms of cost, time, and effort involved, there is an urgent need for computational methods for reliable identification of B-cell epitopes. Although several computational tools for predicting B-cell epitopes have become available in recent years, the predictive performance of existing tools remains far from ideal. We review recent advances in computational methods for B-cell epitope prediction, identify some gaps in the current state of the art, and outline some promising directions for improving the reliability of such methods.

Review

Antigen-antibody interactions play a pivotal role in the humoral immune response. Antibodies bind to antigens at specific sites which correspond to the antigenic determinants or B-cell epitopes. Identification and characterization of B-cell epitopes in target antigens is one of the key steps in epitope-driven vaccine design, immunodiagnostic tests, and antibody production. B-cell epitopes typically belong to one of two classes: linear (continuous or sequential) epitopes or conformational (discontinuous) epitopes. Linear epitopes are short peptides that correspond to a contiguous amino acid sequence fragment of a protein [1,2]. Linear epitopes are usually identified using assays such as PEPSCAN. Consequently, current experimental methods offer little direct evidence indicating that each residue in the epitope does in fact make contact with one or more residues in the paratope (the part in the antibody that binds to the antigen) [3]. Conformational epitopes are composed of amino acids that although not contiguous in primary sequence, are brought into close proximity within the folded 3-dimensional protein structure. Most B-cell epitopes, although they are composed of short linear peptides, appear to be conformational epitopes.

* Correspondence: yasser@iastate.edu; honavar@iastate.edu
[1]Department of Systems and Computer Engineering, Al-Azhar University, Egypt
[2]Department of Computer Science, Artificial Intelligence Research Laboratory, Center for Computational Intelligence, Learning, and Discovery, Iowa State University, USA
Full list of author information is available at the end of the article

Several experimental techniques are currently available for experimental mapping of B-cell epitopes [4]. However, the high cost and effort involved makes them impractical for application on a genomic scale. Computational techniques offer a fast, scalable, and cost-effective approach for predicting B-cell epitopes, for focusing experimental investigations and for improving our understanding of antigen-antibody interactions. Hence, there is a growing interest in the development of sophisticated computational tools for reliable prediction of B-cell epitopes.

Several computational methods for B-cell epitope prediction have been developed in recent years (e.g., [5-14]). However, the predictive performance of current methods is far from ideal [15]. To complicate matters, immunogenicity of proteins is poorly understood [16] and whether B-cell epitopes could be deciphered as an intrinsic features of the protein remains an open question [17]. Recent studies have pointed out some of the limitations of current epitope prediction methods [6,17,18]. Hence, increasing the reliability of computational methods for B-cell epitope prediction remains a major challenge in computational vaccinology [19]. In 2007, the National Institute of Allergy and Infectious Diseases (NIAID) sponsored a workshop and a meeting of a panel of immunologists and bioinformaticians in order to assess the current state of the art in epitope prediction and to identify some areas for further research [15]. One of the key goals of the workshop was to facilitate and expedite the development of improved

methods for B-cell epitope prediction. The report from the workshop recommended (among other things) developing benchmark datasets, standardizing the data formats, and identifying suitable performance metrics for comparing alternative methods.

Against this background, we review recent advances in computational methods for B-cell epitope prediction, identify some gaps in the current state of the art, and outline some promising directions for improving the reliability of such methods.

Predicting linear B-cell epitopes

Although it is believed that the majority of B-cell epitopes are conformational epitopes [20], experimental determination of epitopes has focused primarily on the identification of linear B-cell epitopes [21]. However, even in the case of linear B-cell epitopes, antibody-antigen interactions are often conformation-dependent. The conformation-dependent nature of antigen-antibody binding complicates the problem of B-cell epitope prediction. Hence, B-cell epitope prediction is less tractable than T-cell epitope prediction [22]. In what follows, we review the major approaches for predicting linear B-cell epitopes.

Propensity scale methods

Propensity scale methods [23-26] assign a *propensity value* to each amino acid which measures the tendency of an amino acid to be part of a B-cell epitope (as compared to the background). To reduce fluctuations, the score for each target amino acid residue in a query sequence is computed as the average of the propensity values of the amino acids in a sliding window centered at the target residue. The propensity scores are then used as a basis of predicting whether a given amino acid sequence residue is likely to be part of a linear B-cell epitope. Propensity scale based methods rely on the observed correlations between specific physico-chemical properties of amino acids and the antigenic determinants in protein sequences to identify the location of the linear B-cell epitopes in the query protein sequence.

The first propensity scale method for predicting linear B-cell epitopes was introduced by Hopp and Woods [27] and utilized Levitt hydorophilicity scale [28] to assign a propensity value to each amino acid. This method is based on the assumption that antigenic determinants of protein sequences correspond typically to sequence windows that contain a large number of charged and polar residues and lack large hydrophilic residues.

Subsequently, several other propensity scales have been proposed for predicting linear B-cell epitopes. For example, Parker et al. [23], Karplus et al. [24], Pellequer et al. [29] and Emini et al. [26] have proposed propensity scale based methods that use hydrophilicity, flexibility, turns, or solvent accessibility propensity scales (respectively). PREDITOP [25], PEOPLE [30], BEPITOPE [31], and BcePred [32] predict linear B-cell epitopes based on combinations of physico-chemical properties as opposed to propensity measures that rely on individual properties.

Recently, Blythe and Flower [18] have performed an exhaustive assessment of 484 amino acid propensity scales to examine the correlation between propensity scale-based profiles and the location of linear B-cell epitopes in a dataset of 50 proteins. Their study found that even the best combinations of amino acid propensities yielded B-cell epitope predictions that were only marginally better than chance. The study concluded that the performance of propensity scale based methods reported in the literature is likely to have been overly optimistic, in part due to the small size of the datasets on which the methods had been evaluated. They suggested that more sophisticated approaches (i.e., machine learning approaches) for predicting linear B-cell epitopes need to be developed and rigorously evaluated in order to advance the state-of-the-art in linear B-cell epitope prediction.

Improved propensity scale methods

Several authors have explored methods for improving the predictive performance of propensity scale methods in predicting linear B-cell epitopes. BepiPred [11] combines the hydorophilicity scale constructed by Parker et al. [23] with a Hidden Markov Model (HMM) and demonstrates a slight but statistically significant improvement in performance over the propensity scale based methods of Parker et al. [23] and Levitt et al. [28] on a test dataset of 14 proteins and 83 epitopes. Chen et al. [8] have developed an amino acid pair (AAP) antigenicity scale that assigns to each possible pair of amino acids (i.e., dipeptides), a propensity value. The resulting AAP propensities are then used to represent each peptide sequence using 400 features. Chen et al. [8] trained and evaluated a support vector machine (SVM) classifier using this representation on a dataset of 872 unique epitopes and 872 non-epitopes. They found that the SVM classifiers trained using amino acid pair (AAP) propensity derived features outperform SVM classifiers trained using amino acid propensity derived features.

Machine learning methods

Motivated by the findings of Blythe and Flower [18] and the increasing numbers of experimentally characterized linear B-cell epitopes, several authors have explored machine learning based methods for predicting linear B-cell epitopes using amino acid sequence information. ABCPred [12] uses recurrent artificial neural networks for predicting linear B-cell epitopes and was evaluated on a dataset of 700 B-cell epitopes and 700 non-epitope peptides using 5-fold cross validation tests. Input

sequence windows ranging from 10 to 20 amino acids flanking the target residue, were tested and the best performance, 66% accuracy, was obtained using a window size of 16 amino acids. Söllner and Mayer [13] represent each peptide using a set of 1487 features derived from a variety of propensity scales, neighborhood matrices, and respective probability and likelihood values. Among the machine learning methods explored, they found that the best performing method, a nearest-neighbor classifier combined with feature selection, attained an accuracy of 72% on a dataset of 1211 B-cell epitopes and 1211 non-epitopes using a 5-fold cross validation test [13]. BCPred [5] and FBCPred [10] predicts linear B-cell epitopes and flexible length linear B-cell epitopes (respectively) using support vector machine (SVM) classifiers that use string kernels [33]. COBEpro [34] uses a two-step procedure for predicting linear B-cell epitopes. In the first step, an SVM classifier is used to assign scores to fragments of the query antigen. The input of the SVM is a vector of similarities between the input fragment and all training peptide fragments. In the second step, a prediction score is associated with each residue in the query antigen based on the SVM scores for the peptide fragments. Using several benchmark datasets, COBEpro has been shown to achieve a competitive performance with other linear B-cell epitope prediction methods.

Predicting conformational B-cell epitopes

Although more than 90% of B-cell epitopes are estimated to be conformational in nature [20], most experimental as well as computational methods focus on mapping linear B-cell epitopes. However, in the past few years, there is increasing interest in methods for predicting conformational B-cell epitopes. In what follows, we review three major approaches for predicting conformational B-cell epitopes.

Sequence-based prediction methods

Sequence-based methods in predicting conformational B-cell epitopes have the advantage that they do not require the structure of the target antigen to be known. The amino acid propensity scale methods that assign a prediction score to each residue in the antigen sequence can in principle be used to predict conformational B-cell epitopes [9]. Such methods provide a baseline for evaluation of more sophisticated conformational B-cell epitope prediction methods.

A large body of work using machine learning methods for predicting protein-protein [35,36], protein-DNA [37,38], and protein-RNA [39,40] interfaces using sequence-derived features has demonstrated the feasibility of using sequence-based classifiers in reliably identifying functionally important sites in proteins. It would be interesting to explore similar sequence-based machine learning methods for reliable prediction of conformational B-cell epitopes. The development of such B-cell epitope predictors would make it feasible to identify conformational B-cell epitopes in antigenic sequences for which no solved 3D structures are available.

Structure-based prediction methods

The most accurate experimental method for identifying conformational B-cell epitopes relies on determination of the structure of antigen-antibody complexes using X-ray crystallography [41,42]. Because the number of solved antigen-antibody complexes, or for that matter, the solved antigen structures, is small relative to the number of available antigenic sequences, there are only a small number of methods that utilize 3D structure-derived information in predicting conformational B-cell epitopes.

One of the first conformational B-cell epitope predictors is the conformational epitope predictor (CEP) [7]. Given an antigen with a known structure, CEP uses accessibility of residues and spatial distance cut-off to predict linear and conformational B-cell epitopes.

DiscoTope, a method developed by Andersen et al. [9], uses a combination of amino acid statistics, spatial context, and surface accessibility of amino acids to predict conformational B-cell epitopes. Dis coTope has been shown to outperform propensity scale methods on a dataset of 76 antigen-antibody complexes. The same study also showed that predictors that combine both sequence and structure-derived features of antigens are more accurate than those that rely on either sequence or structure derived features alone [9].

The B-cell conformational epitope predictor proposed by Rapberger et al. [43] works as follows (given the 3D structure of a query antigen): (i) Fast atomic density evaluation (FADE) [44] is applied to select an antibody among a library of 26 available antibodies showing best shape complementarity to the target antigen; (ii) FastContact algorithm [45] is used to identify the most likely interaction site between the selected antibody and the target antigen; (iii) Antigen residues that show a decrease in relative solvent accessible surface area (estimated using a probe size of $3\mathring{A}$) of at least 20% in the complex are predicted as belonging to a discontinuous epitope. This method was shown to outperform the CEP method [7] using a non-redundant dataset of 26 antigen-antibody complexes from Protein Data Bank (PDB) [46].

Ponomarenko et al. [17] introduced a benchmark data set of 62 antibody-antigen complexes extracted from PDB and used it to compare two conformational B-cell epitope prediction servers (CEP and DiscoTope) with six publicly available web servers for protein-protein binding site prediction using various approaches: i) protein-protein docking (ClusPro [47], DOT [48] and PatchDock [49]); ii) structure-based methods applying

different principals and trained on different datasets (PPI-PRED [50], PIER [51] and ProMate [52]); iii) residue conservation (ConSurf [53]). Their results suggest that docking methods outperform other methods when the top ten models and bound docking were considered. However, the overall performance was found to be relatively poor (with an average AUC no greater than 0.7) for all of the methods considered.

Ellipro, a conformational B-cell epitope predictor developed by Ponomarenko et al. [54], implements a modified version of a method that was originally introduced by Thornton et al. [55] for predicting linear B-cell epitopes. Ellipro approximates a protein surface patch by an ellipsoid. Then, a protrusion index is assigned to each residue in the patch the residues are clustered according to their protrusion index values. The resulting clusters are predicted to be part of a conformational B-cell epitope. Ponomarenko et al. [54] reported that Ellipro outperforms six other structure-based predictors of protein-protein interfaces on a dataset of 39 PDB antibody-antigen complexes.

PEPITO, a method for predicting conformational B-cell epitopes introduced by Sweredoski and Baldi [56], uses a weighted linear combination of amino acid propensity scores and half sphere exposure values [57] which encode side chain orientation and solvent accessibility of amino acid residues. An improvement in performance over DiscoTope method has been reported [56].

Rubinstein et al. [58] explored the closely related problem of discriminating the antigenic determinant of an antigen from the rest of the antigen surface. They carried out an analysis of a non-redundant dataset of 53 antigen-antibody complexes. The results of their analysis suggest that epitopes can be discriminated from the rest of the antigen surface using features such as amino acid preferences, compositions of secondary structure, geometrical shape, and evolutionary conservation.

Mimotope analysis -based prediction methods

Pizzi et al. [59] have proposed an approach that combines both experimental and computational techniques for mapping B-cell epitopes. In this approach, a phage-display library of random peptides is scanned against an antibody of interest to obtain a panel of peptides (called mimotopes) that bind to the antibody with high affinity. It is assumed that this panel of mimotopes mimics the physico-chemical properties and spatial organization of the genuine epitopes. However, the precise identification of the epitope mimicked by the set of mimotopes is not straightforward since the epitope is often discontinuous (conformational) and the epitope and mimotopes do not necessarily share a high degree of sequence similarity. Moreover, some of the mimotopes may reflect noisy biological observations and should be filtered out in the analysis. Hence, several computational methods have been proposed for localizing the panel of affinity-selected peptides on the surface of a target antigen have been proposed in literature [60-65].

In general, mimotope analysis methods available in the literature differ from each other in terms of: i) how they represent the antigen structure/sequence; ii) how they align mimotopes with the target antigen structure/sequence; iii) how they cluster the mimotopes and rank the predicted epitopes. For example, PepSurf [63,64] represents the target antigen as a surface graph, wherein the nodes denote surface residues and an edge connects two nodes if the distance between the corresponding residues is lower than a specified threshold. Each mimotope is then aligned to the surface graph using a dynamic programming algorithm in order to obtain a highest scoring path in the graph. The set of highest scoring paths that are connected to each other correspond to the predicted conformational B-cell epitope. SiteLight [60] divides the antigen surface into overlapping patches and then aligns each mimotope with each patch based on a maximal bipartite matching algorithm. Mapitope [62] extracts amino acid pairs (AAPs) from mimotopes and the most statistically significant pairs (SSPs) are identified. These are then mapped on the surface of the antigen and the most elaborate and diverse clusters are identified. These are regarded as the predicted epitope candidates.

Current developments and promising directions

In this section, we highlight recent developments, ongoing efforts, and some promising directions for developing reliable B-cell epitope prediction methods.

Predicting protective linear B-cell epitopes

Söllner et al. [14] have recently investigated the utility of predicted antigenicity, sequence variability, and conservation of post-translational-modification motifs in predicting protective linear B-cell epitopes, i.e., linear B-cell epitopes associated with biological activity. Their analysis showed that focusing on a subset of domains in the query protein sequence (e.g., conserved regions and regions lacking post-translational modification sites) can potentially improve the predictive performance of linear B-cell epitope prediction methods. El-Manzalawy et al. [6] have recently shown that a Naive Bayes classifier trained using evolutionary information (e.g., Position specific scoring matrix (PSSM) profiles obtained using PSI-BLAST [66]) outperforms propensity scale based methods in predicting protective linear B-cell epitopes. These results suggest the possibility of improving the performance of B-cell epitope prediction methods by designing classifiers that are trained on specific subclasses of B-cell epitopes (e.g., protective or neutralizing epitopes).

Hybrid and consensus predictions of B-cell epitopes

Ensemble methods that combine the predictions of several predictors often outperform individual predictors in many biomolecular sequence and structure classification tasks [67-71]. Several strategies for combining a set of predictors, *S*, into a single consensus or meta predictor exist: (i) majority voting: the score for consensus prediction is obtained by the averaging the predicted scores of the individual predictors in *S*; (ii) weighted linear combination: the consensus prediction is obtained via a weighted sum of the predictions obtained from the predictors in *S*. The weights can be assigned based on the estimated performance of the predictors on a training dataset, or optimized to minimize the prediction error of the combined predictors on a training dataset; (iii) meta-learning: A meta-classifier is trained on a training dataset using the outputs of the predictors in *S* on each input sample as input to the classifier and the corresponding class label as the desired output of the classifier.

Recently, Söllner [72] introduced an approach for developing an ensemble of linear B-cell epitope classifiers. Initially, a large number of nearest neighbor and decision tree based classifies trained using different sets of training data features has been created. A strategy based on comparing Receiver Operating Characteristic (ROC) curves of classifiers was applied to select optimal performing classifiers. Finally, an ensemble of the optimal performing classifiers was developed based on a proposed majority voting strategy, positive unanimity voting. Simply, a positive prediction is accepted if and only if all the classifiers forming the ensemble returned a positive prediction.

Improved conformational B-cell epitope prediction tools

Antigen-antibody interactions constitute a subtype of protein-protein interactions. Therefore, the development of improved conformational B-cell epitope prediction tools may benefit from recent advances in developing protein-protein interface residue prediction methods. Hence, it would be interesting to explore the development of conformational B-cell epitope predictors that utilize or adapt sophisticated protein-protein interface predictors e.g., those that make use of sequence and structure-derived features, analyses of surface patches [51,73,74] shape descriptors [75,76] or docking [77].

Immune epitope database and analysis resources

The immune epitope database (IEDB) [78,79] is perhaps the most comprehensive database of experimentally characterized B-cell and T-cell epitopes. IEDB provides users with access to several epitope-related analysis and prediction tools including: (i) several methods for predicting linear and conformational B-cell epitopes; (ii) a tool for visualizing the predicted conformational epitopes on the 3D structure of an antigen; (iii) several tools for analyzing epitope data (e.g., computing epitope conservation and epitope population coverage). IEDB allows users to retrieve both intrinsic biochemical and extrinsic context dependent information about epitopes [78]. This makes it possible to easily assemble customized datasets (e.g., the protectivity data set [14]). Additionally, several researchers have utilized IEDB to conduct meta-analyses of pathogens of interest [80-82], thereby further enhancing the utility of IEDB in the analysis and prediction of B-cell epitopes.

Critical assessment of B-cell epitope prediction methods

Given the large number of B-cell epitope prediction methods available, there is an urgent need for systematic assessment of different methods on standard benchmark datasets [15]. In practice, it is not easy to compare different methods because of several factors: inadequate documentation of the datasets, prediction methods, or the evaluation methodology employed; the unavailability of the benchmark datasets used to evaluate the methods; the unavailability of the code that implements the method (especially in the case of predictors trained using machine learning) as opposed to a server that accepts an antigen sequence or structure as input and outputs the predicted epitopes (fair comparison of alternative machine learning methods or data representations needs to be based on the same training and test datasets); differences in data formats used for inputs and outputs of the predictors.

Rigorous comparative analyses of alternative methods are indispensible for improving our understanding of the strengths and weaknesses of different B-cell epitope prediction methods and for expediting the development of improved methods. Such critical assessment of alternative methods has proven quite valuable in other tasks e.g., protein structure prediction [83], protein-protein interaction site prediction [84].

The development of standardized data representations that would allow different prediction methods to be evaluated on standardized benchmark datasets would be extremely useful not only for comparing the methods but also for developing meta-servers combining the predictions of several prediction tools [15].

The Epitopes Toolkit (EpiT) [85] represents an attempt at standardizing the development and comparison of alternative epitope prediction methods. EpiT standardizes not only the data input and output formats for the predictors but also the encoding of the predictors themselves as serialized Java objects (model files) that can be executed within the EpiT environment.

EpiT consists of two main components: i) *model builder,* an application for building and evaluating epitope predictors and serializing these models in a binary format (model files). This application is an extension of WEKA [86], a widely used open source machine

learning toolkit that includes implementations of several machine learning algorithms. WEKA provides tools for data pre-processing, classification, regression, clustering, validation, and visualization. Furthermore, WEKA provides a framework for implementing new machine learning methods and data pre-processors; ii) *predictor,* an application for applying a model to test data (e.g., set of epitopes or protein sequences).

EpiT is implemented in Java (and modulo the choice of the Java platform, platform-independent). EpiT can be freely downloaded from the project web site at http://ailab.cs.iastate.edu/epit. In addition, the project web site offers a rich resource for the developers of epitope prediction tools and for EpiT users. Some examples of the useful resources available at the EpiT project web site include:

- EpiT documentation: A tutorial and an API documentation of EpiT components that includes several examples of how to build an epitope predictor and how to build an ensemble or a consensus predictor.
- An expanding Repository of Epitope Predictors (REP): Ready to use models for predicting linear B-cell epitopes using BCPred [5], FBCPred [10], and AAP [8] methods. Other researchers can contribute new epitope predictors to this repository.
- A Repository of Epitope Datasets (RED): A repository of epitope benchmark datasets made available by the authors as well as several other publicly available datasets (in WEKA format) that can be used by EpiT users to build their own customized epitope prediction tools. Researchers can contribute additional benchmark datasets to the repository.

EpiT toolkit is available under the GNU General Public License (GPL) which allows others to freely extend or modify the software so long as the modified software is also made available under the GNU GPL.

Conclusions

Developing improved methods for predicting B-cell epitopes requires large datasets of experimentally well-characterized B-cell epitopes, and especially, antigen-antibody complexes and protective epitopes. Special care must be exercised to ensure that the datasets used to train and evaluate the predictors are of high quality:

- Constructing Non-Redundant datasets: Redundant antibody-antigen complexes should be eliminated from the datasets of B-cell epitopes used to train and evaluate B-cell epitope predictors. Otherwise, the estimated performance of predictors can be overly optimistic due to presence of similar complexes in both the training and test datasets. Most of existing linear B-cell epitope datasets [11,13,87] consist of *unique* epitopes. However, the uniqueness of epitope sequences is not a sufficient condition for non-redundancy of the dataset because a pair of *unique* linear B-cell epitopes can share a high degree of pairwise sequence similarity. Unless additional similarity reduction steps are taken to eliminate similar epitopes from the dataset, the predictive performance of B-cell epitope predictors estimated using cross-validation on such datasets can be overly optimistic; and in some cases, lead to false conclusions regarding the performance of different predictors (or the machine learning algorithms used to train the corresponding predictors) relative to each other [10].
- Identifying Non-epitope data: In the currently available conformational B-cell epitope datasets [9,17], the antigen residues that are not part of an antibody-antigen interface in the complex are treated as "non-epitope" residues. However, because each antigen can potentially bind to multiple antibodies, it is possible that some of the "non-epitope" residues may in fact be epitope residues in the complex(s) formed by the same antigen with other antibodies. In currently available linear B-cell epitope datasets comprised of whole antigen sequences [14,88], the antigen residues that are not covered by any of the *reported* epitope are treated as "non-epitope" residues. Because the experimental data available are necessarily incomplete, some of these "non-epitope" residues might in fact be epitope residues. Training an epitope predictor using a dataset in which some epitope residues are incorrectly labeled as non-epitope residues is tantamount to training the predictor on a noisy dataset. Performance estimates of the predictors on such a dataset tend to exaggerate the number of false positives (and hence the estimates of predictor performance that are based on the numbers of false positives, true positives, false negatives, and true negatives). Some authors [5,8,10,12,13] have tried to alleviate this problem by using non-epitope residues extracted from a random sample of the protein sequences in SwissProt [89]. However, a recent study [34] has shown that the resulting datasets may yield somewhat biased performance estimates. This underscores the importance for large experimentally well-characterized datasets in this field.

The use of large, non-redundant, and experimentally well-characterized datasets can help increase the accuracy of the cross-validation based estimates of the performance of B-cell epitope predictors trained on such datasets. However, it does not necessarily inform the

choice of performance measures to use for comparing different predictors (or the machine learning algorithms used to train the predictors). Accuracy, sensitivity, specificity, and correlation coefficients are widely used metrics for evaluating the performance of prediction methods [90]. None of these measures when used alone provides a complete picture of the performance of the predictor. Each of these metrics is *threshold-dependent* because it describes the performance of the predictor for a given choice of the classification threshold. Moreover, it is possible to trade off one measure (e.g., sensitivity) against another (e.g., specificity). To get a more comprehensive picture of the performance of the predictor, it is useful to consider the Receiver Operating Characteristic (ROC) curve which describes the performance of the classifier over all possible choices of the classification threshold. The ROC curve is obtained by plotting the true positive rate as a function of the false positive rate or, equivalently, sensitivity versus (1-specificity) as the discrimination threshold of the binary classifier is varied. Each point on the ROC curve corresponds to a specific choice of the classification threshold, and hence a particular choice of the tradeoff between true positive rate and false positive rate. The area under ROC curve (AUC) is a useful threshold-independent summary statistic for comparing two ROC curves. The AUC is defined as the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative sample. An AUC = 1 corresponds to a perfect predictor, whereas an AUC = 0.5 corresponds to a predictor that classifies each input sample by randomly guessing its class label. Any predictor with a performance that is better than random will have an AUC value that lies between 0.5 and 1.0. AUC is one of the most widely used metrics for comparing the performance of different B-cell epitope predictors. It is also the recommended metric for assessing the performance of B-cell epitope predictors [15]. However, a comparison of predictors based on AUC can yield misleading conclusions when the corresponding ROC curves cross. A recent study [91] has shown that AUC has a more severe limitation: Using AUC to compare different predictors is tantamount to using different misclassification cost distributions and hence different metrics to evaluate different predictors. Because the cost of misclassifying a sample is a property of the prediction problem, and not a property of the classifier, there is a need for better metrics for fair comparison of the performance of different predictors.

List of abbreviations used

AAP: Amino Acid Pair; AUC: Area Uunder ROC Curve; CEP: Conformational Epitope Predictor; EpiT: Epitopes Toolkit; FADE: Fast Atomic Density Evaluation; GPL: General Public License; HMM: Hidden Markov Model; IEDB: Immune Epitope Database; NIAID: National Institute of Allergy and Infectious Diseases: PDB: Protein Data Bank; PSSM: Position Specific Scoring Matrix; RED: Repository of Epitope Datasets; REP: Repository of Epitope Predictors; ROC: Receiver Operating Characteristic; SSP: Statistically Significant Pair; SVM: Support Vector Machine.

Acknowledgements

This research was supported in part by a grant from the National Institutes of Health (GM066387) to Vasant Honavar.
This article has been published as part of *Immunome Research* Volume 6 Supplement 2, 2010: Computational Vaccinology: State-of-the-art Assessments. The full contents of the supplement are available online at http://www.immunome-research.com/supplements/6/S2.

Author details

[1]Department of Systems and Computer Engineering, Al-Azhar University, Egypt. [2]Department of Computer Science, Artificial Intelligence Research Laboratory, Center for Computational Intelligence, Learning, and Discovery, Iowa State University, USA.

Authors contributions

YE and VH conceived and wrote this review article.

Competing interests

The author declare no competing interests.

Published: 3 November 2010

References

1. Barlow D, Edwards M, Thornton J, *et al*: **Continuous and discontinuous protein antigenic determinants.** *Nature* 1986, **322**:747-748.
2. Langeveld J, martinez Torrecuadrada J, boshuizen R, Meloen R, Ignacio C: **Characterisation of a protective linear B cell epitope against feline parvoviruses.** *Vaccine* 2001, **19**:2352-2360.
3. Van Regenmortel M: **What is a B-cell epitope?** *Methods in molecular biology (Clifton, NJ)* 2009, **524**:3.
4. Reineke U, Schutkowski M: **.** *Epitope mapping protocols. Preface* Humana Press, 2 2009.
5. EL-Manzalawy Y, Dobbs D, Honavar V: **Predicting linear B-cell epitopes using string kernels.** *J. Mol. Recognit.* 2008, **21**:243-255.
6. EL-Manzalawy Y, Dobbs D, Honavar V: **Predicting linear B-cell epitopes using evolutionary information.** *IEEE International, Conference on Bioinformatics and Biomedicine* 2008.
7. Kulkarni-Kale U, Bhosle S, Kolaskar A: **CEP: a conformational epitope prediction server.** *Nucleic Acids Res.* 2005, **33**:W168.
8. Chen J, Liu H, Yang J, Chou K: **Prediction of linear B-cell epitopes using amino acid pair antigenicity scale.** *Amino Acids* 2007, **33**:423-428.
9. Haste Andersen P, Nielsen M, Lund O: **Prediction of residues in discontinuous B-cell epitopes using protein 3D structures.** *Protein Sci.* 2006, **15**:2558.
10. EL-Manzalawy Y, Dobbs D, Honavar V: **Predicting flexible length linear B-cell epitopes.** *7th International Conference on Computational Systems Bioinformatics* 2008, 121-131.
11. Larsen J, Lund O, Nielsen M: **Improved method for predicting linear B-cell epitopes.** *Immunome Res.* 2006, **2**:2.
12. Saha S, Raghava G: **Prediction of continuous B-cell epitopes in an antigen using recurrent neural network.** *Proteins* 2006, **65**:40-48.
13. Söllner J, Mayer B: **Machine learning approaches for prediction of linear B-cell epitopes on proteins.** *J. Mol. Recognit.* 2006, **19**:200-208.
14. Söllner J, Grohmann R, Rapberger R, Perco P, Lukas A, Mayer B, Blythe M: **Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins.** *Immunome Res* 2008, **7**:4.
15. Greenbaum J, Andersen P, Blythe M, Bui H, Cachau R, Crowe J, Davies M, Kolaskar A, Lund O, Morrison S, *et al*: **Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools.** *J. Mol. Recognit* 2007, **20**:75-82.
16. Feng Y, Jacobs F, Van Craeyveld E, Lievens J, Snoeys J, Van Linthout S, De Geest B: **The impact of antigen expression in antigen-presenting cells on humoral immune responses against the transgene product.** *Gene therapy* 2009, **17**:288-293.

17. Ponomarenko J, Bourne P: **Antibody-protein interactions: benchmark datasets and prediction tools evaluation.** *BMC Structural Biology* 2007, **7**:64.
18. Blythe M, Flower D: **Benchmarking B cell epitope prediction: underperformance of existing methods.** *Protein Science: A Publication of the Protein Society* 2005, **14**:246.
19. Flower D: **.** *Bioinformatics for vaccinology* Wiley Black-well 2009.
20. Walter G: **Production and use of antibodies against synthetic peptides.** *J. Immunol. Methods* 1986, **88**:149-61.
21. Flower D: **.** *Immunoinformatics: predicting immunogenicity in silico* Quantum distributor, 1 2007.
22. Korber B, LaBute M, Yusim K: **Immunoinformatics comes of age.** *PLoS Comput. Biol.* 2006, **2**:e71.
23. Parker J, Guo HRD: **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* 1986, **25**:5425-5432.
24. Karplus P, Schulz G: **Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen.** *Naturwiss* 1985, **72**:21-213.
25. Pellequer J, Westhof E, Van Regenmortel M: **Correlation between the location of antigenic sites and the prediction of turns in proteins.** *Immunol. Lett.* 1993, **36**:83-99.
26. Emini E, Hughes J, Perlow D, Boger J: **Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide.** *J. Virol.* 1985, **55**:836-839.
27. Hopp T, Woods K: **Prediction of protein antigenic determinants from amino acid sequences.** *Proceedings of the National Academy of Sciences of the United States of America* 1981, **78(6)**:3824.
28. Levitt M: **A simplified representation of protein conformations for rapid simulation of protein folding.** *Journal of molecular biology* 1976, **104**:59.
29. Pellequer J, Westhof E, Van Regenmortel M: **Predicting location of continuous epitopes in proteins from their primary structures.** *Meth. Enzymol.* 1991, **203**:176-201.
30. Alix A: **Predictive estimation of protein linear epitopes by using the program PEOPLE.** *Vaccine* 1999, **18**:311-4.
31. Odorico M, Pellequer J: **BEPITOPE: predicting the location of continuous epitopes and patterns in proteins.** *J. Mol. Recognit.* 2003, **16**:20-22.
32. Saha S, Raghava G: **BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties.** *Artificial Immune Systems, Third International Conference (ICARIS 2004), LNCS* 2004, **3239**:197-204.
33. Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C: **Text classification using string kernels.** *J. Mach. Learn. Res.* 2002, **2**:419-444.
34. Sweredoski M, Baldi P: **COBEpro: a novel system for predicting continuous B-cell epitopes.** *Protein Eng Des Sel* 2009, **22(3)**:113-120.
35. Yan C, Dobbs D, Honavar V: **A two-stage classifier for identification of protein-protein interface residues.** *Bioinformatics* 2004, **20(Suppl 1)**: i371-i378.
36. Chen C, Zhou X, Tian Y, Zou X, Cai P: **Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network.** *Analytical biochemistry* 2006, **357**:116-121.
37. Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, Sun X: **Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature.** *Bioinformatics* 2009, **25**:30.
38. Yan C, Terribilini M, Wu F, Jernigan R, Dobbs D, Honavar V: **Predicting DNA-binding sites of proteins from amino acid sequence.** *BMC bioinformatics* 2006, **7**:262.
39. Terribilini M, Lee J, Yan C, Jernigan R, Honavar V, Dobbs D: **Prediction of RNA binding sites in proteins from amino acid sequence.** *Rna* 2006, **12(8)**:1450-1462.
40. Kumar M, Gromiha M, Raghava G: **Prediction of RNA binding sites in a protein using SVM and PSSM profile.** *Proteins* 2008, **71**:189-194.
41. Fleury D, Daniels R, Skehel J, Knossow M, Bizebard T: **Structural evidence for recognition of a single epitope by two distinct antibodies.** *Proteins* 2000, **40(4)**:572.
42. Mirza O, Henriksen A, Ipsen H, Larsen J, Wissenbach M, Spangfort M, Gajhede M: **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.** *Journal of immunology (Baltimore, Md.: 1950)* 2000, **165**:331.
43. Rapberger R, Lukas A, Mayer B: **Identification of discontinuous antigenic determinants on proteins based on shape complementarities.** *Journal of Molecular Recognition* 2007, **20(2)**:113-121.
44. Mitchell J, Kerr R, Ten Eyck L: **Rapid atomic density methods for molecular shape characterization.** *J Mol Graph Model* 2001, **19(3-4)**:325-330.
45. Camacho C, Zhang C: **FastContact: rapid estimate of contact and binding free energies.** *Bioinformatics* 2005, **21(10)**:2534-2536.
46. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The Protein Data Bank.** *Nucleic Acids Research* 2000, **28**:235-242.
47. Comeau S, Gatchell D, Vajda S, Camacho C: **ClusPro: an automated docking and discrimination method for the prediction of protein complexes.** *Bioinformatics* 2004, **20**:45-50.
48. Mandell J, Roberts V, Pique M, Kotlovyi V, Mitchell J, Nelson E, Tsigelny I, Ten Eyck L: **Protein docking using continuum electrostatics and geometric fit.** *Protein Eng* 2001, **14(2)**:105-113.
49. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson H: **PatchDock and SymmDock: servers for rigid and symmetric docking.** *Nucleic acids research* 2005, **33(Web-Server-Issue)**:W363.
50. Bradford J, Westhead D: **Improved prediction of protein-protein binding sites using a support vector machines approach.** *Bioinformatics* 2005, **21(8)**:1487-1494.
51. Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R: **PIER: protein interface recognition for structural proteomics.** *Proteins* 2007, **67(2)**:400-417.
52. Neuvirth H, Raz R, Schreiber G: **ProMate: a structure based prediction program to identify the location of protein—protein binding sites.** *Journal of molecular biology* 2004, **338**:181-199.
53. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N: **ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.** *Nucleic acids research* 2005, **33(Web Server Issue)**:W299-W302.
54. Ponomarenko J, Bui H, Li W, Fusseder N, Bourne P, Sette A, Peters B: **ElliPro: a new structure-based tool for the prediction of antibody epitopes.** *BMC bioinformatics* 2008, **9**:514.
55. Thornton J, Edwards M, Taylor W, Barlow D: **Location of 'continuous' antigenic determinants in the protruding regions of proteins.** *The EMBO Journal* 1986, **5(2)**:409-413.
56. Sweredoski M, Baldi P: **PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure.** *Bioinformatics* 2008, **24(12)**:1459-1460.
57. Hamelryck T: **An amino acid has two sides: a new 2D measure provides a different view of solvent exposure.** *Proteins* 2005, **59**:38-48.
58. Rubinstein N, Mayrose I, Halperin D, Yekutieli D, Gershoni J, Pupko T: **Computational characterization of B-cell epitopes.** *Molecular Immunology* 2008, **45(12)**:3477-3489.
59. Pizzi E, Cortese R, Tramontano A: **Mapping epitopes on protein surfaces.** *Biopolymers* 1995, **36(5)**:675-680.
60. Halperin I, Wolfson H, Nussinov R: **SiteLight: binding-site prediction using phage display libraries.** *Protein Science: A Publication of the Protein Society* 2003, **12(7)**:1344-1359.
61. Moreau V, Granier C, Villard S, Laune D, Molina F: **Discontinuous epitope prediction based on mimotope analysis.** *Bioinformatics* 2006, **22(9)**:1088-1095.
62. Bublil E, Freund N, Mayrose I, Penn O, Roitburd-Berman A, Rubinstein N, Pupko T, Gershoni J: **Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm.** *Proteins* 2007, **68**:294-304.
63. Mayrose I, Penn O, Erez E, Rubinstein N, Shlomi T, Freund N, Bublil E, Ruppin E, Sharan R, Gershoni J, *et al*: **Pepitope: epitope mapping from affinity-selected peptides.** *Bioinformatics* 2007, **23(23)**:3244-3246.
64. Mayrose I, Shlomi T, Rubinstein N, Gershoni J, Ruppin E, Sharan R, Pupko T: **Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm.** *Nucleic Acids Research* 2007, **35**:69-78.
65. Castrignanò T, De Meo P, Carrabino D, Orsini M, Floris M, Tramontano A: **The MEPS server for identifying protein conformational epitopes.** *BMC bioinformatics* 2007, **8(Suppl 1)**:S6.
66. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3390-3402.
67. Caragea C, Sinapov J, Silvescu A, Dobbs D, Honavar V: **Glycosylation site prediction using ensembles of Support Vector Machine classifiers.** *BMC bioinformatics* 2007, **8**:438.

68. Ouali M, King R: **Cascaded multiple classifiers for secondary structure prediction.** *Protein Science* 2000, **9(6)**:1162-1176.
69. Kedarisetti K, Kurgan L, Dick S: **Classifier ensembles for protein structural class prediction with varying homology.** *Biochem Biophys Res Commun* 2006, **348(3)**:981-988.
70. Han P, Zhang X, Norton R, Feng Z: **Large-scale prediction of long disordered regions in proteins using random forests.** *BMC bioinformatics* 2009, **10**:8.
71. Yan X, Chao T, Tu K, Zhang Y, Xie L, Gong Y, Yuan J, Qiang B, Peng X: **Improving the prediction of human microRNA target genes by using ensemble algorithm.** *FEBS letters* 2007, **581(8)**:1587-1593.
72. Söllner J: **Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins.** *Journal of Molecular Recognition* 2006, **19(3)**:209-214.
73. Jones S, Thornton J: **Prediction of protein-protein interaction sites using patch analysis.** *Journal of molecular biology* 1997, **272**:133-143.
74. Bradford J, Westhead D: **Improved prediction of protein—protein binding sites using a support vector machines approach.** *Bioinformatics* 2005, **21(8)**:1487-1494.
75. Martin W, Ewgenij P, Schneider G: **PocketPicker: analysis of ligand binding-sites with shape descriptors.** *Chemistry Central Journal* 2007, **1**:7.
76. Weisel M, Kriegl J, Schneider G: **PocketGraph: graph representation of binding site volumes.** *Chemistry Central Journal* 2009, **3(Suppl 1)**:P66.
77. Grosdidier S, Fernández-Recio J: **Identification of hot-spot residues in protein-protein interactions by computational docking.** *BMC bioinformatics* 2008, **9**:447.
78. Peters B, Sidney J, Bourne P, Huynh-Hoa B, Buus S, *et al*: **The immune epitope database and analysis resource: From vision to blueprint.** *PLoS Biol* 2005, **3(3)**:e91.
79. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne P, Bui H, Buus S, Frankild S, Greenbaum J, *et al*: **Immune epitope database analysis resource (IEDB-AR).** *Nucleic Acids Research* 2008, **36(Web Server issue)**: W513.
80. Vaughan K, Blythe M, Greenbaum J, Zhang Q, Peters B, Doolan D, Sette A: **Meta-analysis of immune epitope data for all Plasmodia: overview and applications for malarial immunobiology and vaccine-related issues.** *Parasite Immunology* **31(2)**:78-97.
81. Zarebski L, Vaughan K, Sidney J, Peters B, Grey H, Janda K, Casadevall A, Sette A: **Analysis of epitope information related to Bacillus anthracis and Clostridium botulinum.** *Expert Review of Vaccines* 2008, **7**:55-74.
82. Blythe M, Zhang Q, Vaughan K, de Castro R, Salimi N, Bui H, Lewinsohn D, Ernst J, Peters B, Sette A: **An analysis of the epitope knowledge related to Mycobacteria.** *Immunome Research* 2007, **3**:10.
83. Moult J: **A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.** *Current opinion in structural biology* 2005, **15(3)**:285-289.
84. Henrick K, Moult J, Sternberg M, Vajda S, Vakser I, Wodak S: **CAPRI: a critical assessment of predicted interactions.** *Proteins* 2003, **52**:2-9.
85. EL-Manzalawy Y, Honavar V: **Epitopes Toolkit.** 2009 [http://ailab.cs.iastate.edu/epit/index.html], Software available at.
86. Witten I, Frank E: **.** *Data mining: Practical machine learning tools and techniques* Morgan Kaufmann, 2 2005.
87. Saha S, Raghava G: **ABCPred benchmarking datasets.** 2006 [http://www.imtech.res.in/raghava/abcpred/dataset.html], Available at.
88. Blythe M, Flower D: **Benchmarking B cell epitope prediction: underperformance of existing methods.** *Protein Sci.* 2005, **14**:246-248.
89. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res.* 2000, **28**:45-48.
90. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16**:412-424.
91. Hand D: **Measuring classifier performance: a coherent alternative to the area under the ROC curve.** *Machine Learning* 2009, **77**:103-123.

doi:10.1186/1745-7580-6-S2-S2
Cite this article as: EL-Manzalawy and Honavar: **Recent advances in B-cell epitope prediction methods.** *Immunome Research* 2010 **6**(Suppl 2):S2.