Applying Stacked and Cascade Generalizations to B-cell Epitope Prediction

Yuh-Jyh Hu, Shun-Chien Lin, Yu-Lung Lin

Institute of Biomedical Engineering, National Chiao Tung University, Hsinchu, Taiwan

Abstract . One of the major challenges in the field of vaccine design is to identify the B-cell epitopes in ever-evolving viruses. Various prediction servers have been developed to predict linear or conformational epitopes, each relying on different physicochemical properties and adopting distinct search strategies. We propose meta learning approaches to epitope prediction based on stacked generalization and cascade generalization. By combining the base prediction servers in a hierarchical architecture, we demonstrated that a meta learner outperformed the best single server in predicting the epitopes of an independent dataset of pathogen proteins.

Keywords: epitope prediction; linear; conformational; meta learning

1 Introduction

The ability of antibodies to respond to an antigen, such as a virus capsid protein fragment, depends on the antibodies' specific recognition of epitopes, which are sites of the antigen to which antibodies bind. Based on structures and interactions with the antibody, epitopes can be divided into two categories: linear and conformational. A linear epitope is formed by a continuous sequence of amino acids, whereas a conformational epitope is composed of discontinuous primary sequences, which are close in the 3D space.

Approaches to predicting linear and conformational epitopes are different. Various physicochemical properties of amino acids have been applied in linear epitope prediction [1-3]. Nevertheless, a study of 484 amino acid scales showed that predictions based on the best scales still produced poor correlation with experimentally confirmed epitopes [4]. This prompted machine learning methods to improve the prediction. BepiPred combines amino acid propensity scales with a Hidden Markov Model (HMM) to yield a slight improvement over methods only based on physicochemical properties [5]. ABCPred uses artificial neural networks for predicting linear B-cell epitopes [6]. Chen et al. proposed a novel scale called the amino acid pair (AAP) antigenicity scale [7]. They trained a support vector machine (SVM) classifier, using the AAP propensity scale, to distinguish epitopes and non-epitopes. BCPREDS uses SVM combined with a variety of kernel methods, including string kernels, radial basis function (RBF) kernel and subsequence kernel (SSK), for linear B-cell epitope prediction [8].

An increase in the availability of protein structures has enabled the identification of conformational epitopes, using various computational methods. For example, DiscoTope is driven by a combination of amino acid composition information, spacial neighborhood information, and a surface measure to make the prediction [9]. Ellipro uses Thornton's propensities and applies residue clustering to identify epitopes [10]. Based on the ideas of unit patch of residue triangle and the clustering coefficient to describe the local spacial context and compactness, SEPPA predicts spacial epitopes [11]. By combining structural and physichochemical features, EPITOPIA adopts a Bayesian classifier to predict epitopes [12]. Liang et al. developed EPSVR, using a support vector regression method to predict conformational epitopes, and proposed a meta learner EPMeta, which incorporates consensus results from multiple prediction servers by a voting mechanism [13].

In this paper, we propose combining multiple servers to improve epitope prediction based on two meta-learning strategies: stacked generalization (stacking) [14,15] and cascade generalization (cascade) [16,17]. They both work in a hierarchical architecture, constituted by meta learners and base learners, in which the input space for meta learners is extended by the predictions of base learners. We selected several linear and conformational epitope prediction servers to be the base learners, and evaluated three inductive learning algorithms as the meta learner. To demonstrate the performance, we tested the combinatorial method on an independent set of pathogen proteins that were not used previously to train the epitope prediction servers. The results showed the potential of the meta-learning method in epitope prediction.

2 Materials and Methods

2.1 Meta Learning: Stacked Generalization

Considering the epitope prediction as an inductive learning problem, given a set of antigens with known epitope and non-epitope regions, the goal is to learn a classifier from the set of antigens, and apply the classifier to novel antigens for epitope detection. Many learning-based epitope prediction servers have been developed [5-13]. Since different learning algorithms employ different knowledge representations and search heuristics, they explore different hypothesis space, and consequently obtain different results. We here propose combining multiple servers to achieve higher prediction performance than a single server.

Stacked generalization is a method of combining the predictions of multiple learning models that have been trained for a classification task [14,15]. It works as a layered process. Each of a set of base learners is trained on a dataset, and the predictions of these base learners become the meta features. A successive layer receives as the input the meta features, and pass its output to the next layer. A single classifier at the top level makes the final prediction. Stacked generalization is considered a form of meta learning because the transformation of the training data for the successive layers contain the information of the predictions of the base learners, which is a form of meta knowledge. Cascade generalization belongs to the family of stacked generaliza-

tion [16]. What distinguishes cascade from stacking is that in cascade generalization intermediate learners have access to the predictions of the lower level learners as well as the base features. The goal of stacking is combining the predictions of the immediately preceding level learners, while the goal of cascade is to obtain a model that can use the features in the representation language for the lower level learners [17].

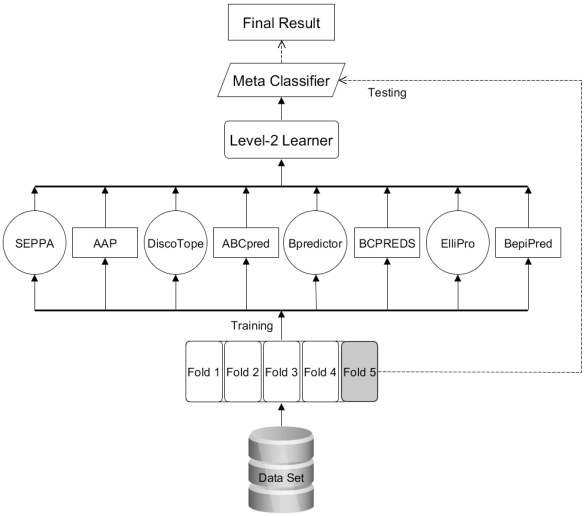

Figure 1. Two-Layer Architecture for Stacked Generalization

In this study, we adopted a 2-layer architecture for stacked generalization, as shown in Figure 1. We used C4.5 [18], 1-nearest-neighbor [19] or SVM [20] as a level-2 learner because C4.5 learns comprehensible decision trees, the nearest-neighbor rule is capable of constructing local approximation to the target, and SVM has demonstrated promising performance in various applications. We selected several state-of-the-art linear and conformational epitope prediction tools as the candidate base learners, including BepiPred [5], ABCpred [6], AAP [7], BCPREDS [8], DiscoTope [9], Ellipro [10], SEPPA [11], and Bpredictor [21]. In the training stage, their outputs will be passed to the level-2 learner to learn a meta classifier. In the test stage, for a given protein, the learned classifier predicts the epitopes based on the predictions of the base learners.

For cascade generalization, we developed a 2-level and a 3-level architectures, as presented in Figure 2, with the epitope prediction servers at the bottom level, and C4.5, 1-nearest-neighbor or SVM at the higher levels. We analyzed and compared the base features exploited by previous prediction methods, and selected those that characterize physicochemical propensities and structural properties. We adopted 18 base features derived from the studies: epitope propensity [22], residue accessibility [23], secondary structure [24], B factor [25], solvent excluded surfaces, solvent accessible surfaces [26], protein chain flexibility [27], hydrophilicity [28], PSSM [29], and atom volume [30]. The predictions of the servers are combined with the base feature values,

and passed to the higher-level learners to train a meta classifier. The learned classifier predicts the epitopes of a novel protein based on both the predictions of the base learners, and the base features about that antigen protein.

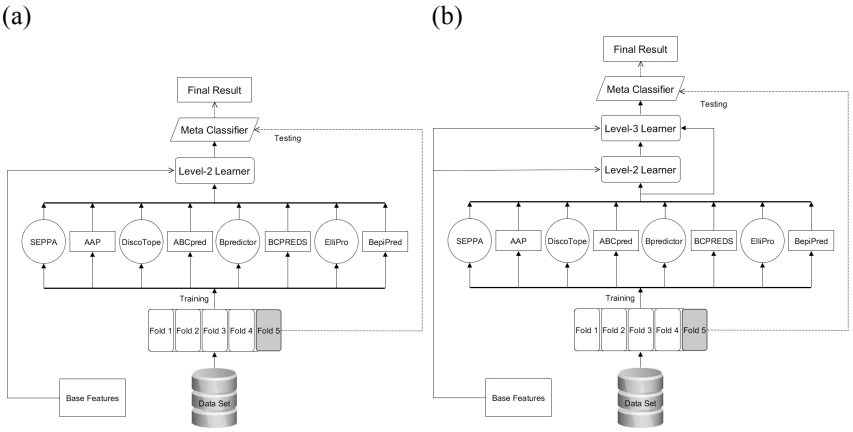

Figure 2. Cascade Generalization. (a) 2-level Cascade Architecture, (b) 3-level Cascade Architecture

2.2 Analysis of Prediction Performance

An epitope prediction server must be trained to obtain its prediction model before it can make a prediction. Because the servers used in our study are web-based servers, they cannot be retrained by new training data. To make an unbiased comparison of the prediction performance of these servers, we compiled an independent dataset of antigens with known epitopes. We first collected the test datasets used in DiscoTope [9], SEPPA [11], and Bpredictor [21]. Combining them with the data of the Epitome [31] and the IEDB [32] databases, we obtained 254 antigen structures in total. To build an independent dataset for prediction performance evaluation, we first removed the antigens previously used to train the prediction servers, and then combined the remaining with Liang et al.'s independent dataset [13]. Among these antigens, there are 94 antigens with the real epitopes annotated in the IEDB [32]. We used the independent dataset of these 94 antigens for a consistent and fair comparison of different prediction methods. The antigen structures were used as input to the structure-based servers; the corresponding antigen sequences, the input to the sequence-based servers.

The performance measures we used are true positive rate (i.e., sensitivity), false positive rate, precision (i.e. positive predictive value), Accuracy, F-score, and Matthew's Correlation Coefficient (MCC). We consider a predicted antigenic residue a true positive if it is within a known epitope part, otherwise, a false positive. In contrast, a predicted non-antigenic residue is a true negative if it lies outside the known epitopes, or a false negative when it is actually part of a known epitope. We tested the prediction servers on an independent antigen data, and for each amino acid we ob-

tained: (1) the epitope prediction score, or (2) the classification (epitope or non-epitope), according to a score threshold.

Table 1. Definitions of Performance Measures

Performance Measure	Definition
TPR[a]	TP/(TP+FN)
FPR	FP/(FP+TN)
Precision[b]	TP/(TP+FP)
Accuracy	(TP+TN)/(TP+TN+FP+FN)
F-score	2*TPR*Precision/(TPR+Precision)
MCC	$\frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

[a]True Positive Rate is also known as Sensitivity or Recall.
[b]Precision is also known as Positive Predictive Value.

2.3 Study of Correlation

A meta learner can consist of an arbitrary number of base learners, and its overall performance depends on these learning components. If the learning components have complementary predictive strengths, a meta learner can search a wider variety of hypotheses in the hypothesis space, and consequently generalize better for novel test data than a single component learner [14,17]. We used the statistical techniques to analyze the prediction servers. By a scatterplot of the prediction scores, we visualized the strength of the relationship between a pair of prediction servers. We ranked the prediction scores, and then employed the Spearman's rank correlation coefficient to investigate the correlation of the predictions made by different epitope prediction servers. A sparser scattering and a lower correlation of the prediction scores indicates a higher probability that the predictors have complementary strengths. The results of the scatterplots and the correlation analysis provide the basis for selecting the appropriate base learners in meta learning.

3 Results

3.1 Performance Correlation

For a meta learning-based method to work well, the base learners need to have complementary predictive capabilities, which can be reflected by relatively low correlations. To investigate the complementary strengths of the base learners, for each pair of conformational and linear epitope prediction servers, we can visualize the correlations in their predictions by scatterplots. We showed four scatterplots in Figure 3. In each scatterplot, the x-axis and y-axis represent the prediction scores given by the two servers in comparison. A sparse plot indicates a low correlation between the two servers. The scatterplots, as shown in Figure 3, for either conformational or linear predictions servers are relatively sparse, which suggests that the correlation between these

servers is relatively low. We also observed similar prediction distributions in the other scatterplots of different servers (data not shown to save space). To further analyze the performance correlation, we did the Spearman's rank correlation analysis of the predictions of the base servers. We present the pairwise prediction correlations in Table 2, which indicates that the performance correlations between the servers are relatively weak, and the correlation between linear epitope prediction servers (0.157~0.377) is generally smaller than that between conformational epitope prediction servers (0.413~0.639). These results suggest that the base learners have complementary strengths, and a meta learner built upon these learners can demonstrate the synergy of their predictive capabilities.

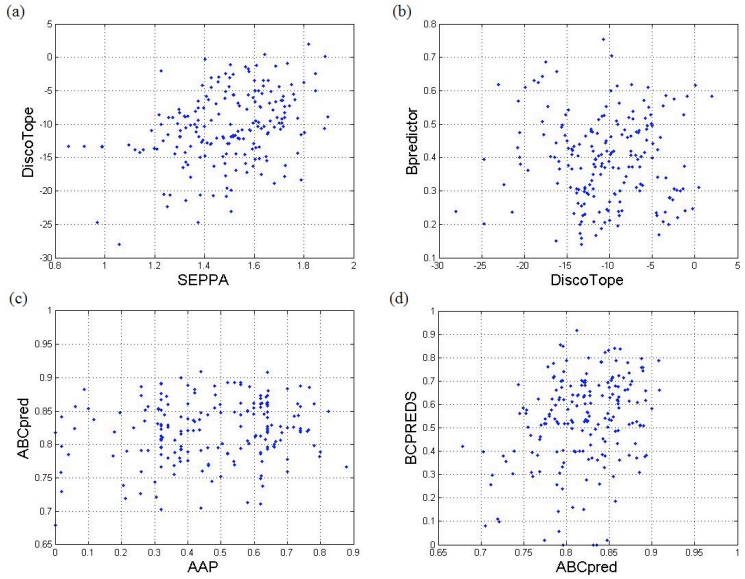

Figure 3. Scatterplots of base learner predictions

3.2 Performance Comparison

We conducted a stratified 5-fold cross validation experiment, as shown in Figures 1 and 2, to evaluate the performance of the base servers and the proposed meta learners. The independent data were randomly divided into 5 disjoint folds (i.e. subsets), each of approximately equal size. The folds were also stratified to maintain the same distribution of the epitopes and the non-epitopes as that in the original independent dataset. One fold of data was used for testing the prediction performance, and the remaining four were all used for training. We repeated the same training-testing process on each fold iteratively. Each run produced a performance result based on the fold selected for testing. The overall performance was taken as the average of the results obtained from all iterations.

Table 2. Results of Spearman's Rank Correlation Analysis

Conformational			
	SEPPA	DiscoTope	Bpredictor
SEPPA	1	-	-
DiscoTope	0.497	1	-
Bpredictor	0.561	0.413	1
ElliPro	0.585	0.499	0.639
Linear			
	AAP	ABCpred	BCPREDS
AAP	1	-	-
ABCpred	0.157	1	-
BCPREDS	0.290	0.195	1
BepiPred	0.253	0.205	0.377

We tested C4.5, 1-NN, and SVM separately in the 2-level stacking and cascade architectures. For the 3-level cascade generalization, we used SVM as the top-level meta learner, combined with either C4.5 or 1-NN as the level-2 learner. We summarized the results of the 5-fold cross validation on the meta classifiers in Table 3, and presented the performance of each base prediction server, based on the same cross validation, in Table 4 for reference. In addition, we included two recent epitope prediction methods, CEKEG [33] and LBtope [34], for comparison. Tables 3 and 4 showed that both stacking and cascade outperformed all the prediction tools by a marked amount in accuracy, F-score, and MCC. The results demonstrate the advantages of exploiting the complementary capabilities of the prediction servers. Comparing the performance results of the 2-level architecture between stacking and cascade, we noted that the 18 base features played a crucial role in meta learning. With the 18 base features, we increased the performance of stacking in F-score and MCC by 35%~149% and 29%~102% respectively. In addition, the comparison between 2-level and 3-level cascade generalization indicated that the 3-level cascade obtained markedly better F-score and MCC than the 2-level cascade architectures except the one based on SVM.

Table 3. Performance Results of Meta Classifiers Based on Stacking and Cascade

	TPR	**FPR**	**Precision**	**Accuracy**	**F-score**	**MCC**
2-level Stacking						
C4.5	0.218	0.015	0.516	0.930	0.305	0.304
1-NN	0.375	0.048	0.377	0.911	0.376	0.328
SVM	0.144	0.005	0.695	0.934	0.234	0.291
2-level Cascade						
C4.5	0.337	0.022	0.540	0.932	0.414	0.393
1-NN	0.502	0.034	0.530	0.933	0.515	0.479
SVM	0.452	0.008	0.814	0.954	0.582	0.587
3-level Cascade						
C4.5→SVM	0.452	0.008	0.809	0.954	0.580	0.584
1-NN→SVM	0.452	0.008	0.821	0.954	0.583	0.589

Table 4. Performance Results of Prediction Servers

	TPR	FPR	Precision	Accuracy	F-score	MCC
SEPPA	0.880	0.643	0.095	0.394	0.171	0.129
DiscoTope	0.082	0.025	0.202	0.912	0.116	0.088
Bpredictor	0.695	0.448	0.106	0.562	0.184	0.127
ElliPro	0.689	0.514	0.093	0.500	0.164	0.090
AAP	0.594	0.459	0.090	0.545	0.157	0.070
ABCpred	0.684	0.603	0.080	0.418	0.143	0.043
BCPREDS	0.367	0.220	0.114	0.751	0.174	0.090
BepiPred	0.624	0.462	0.094	0.544	0.163	0.083
CEKEG	0.513	0.287	0.075	0.704	0.131	0.101
LBtope	0.187	0.101	0.125	0.848	0.149	0.072

4 Conclusion

A profound understanding of the interaction between antibodies and epitopes provides basis for the rational design of preventive vaccines. Our analysis results showed that the epitope prediction servers revealed complementary strengths, which suggested the synergy of these servers. We proposed the use of meta learning-based approaches to B-cell epitope prediction. Combining these servers in a hierarchical structure, we demonstrated a better prediction performance of meta learners than that of each single server.

Acknowledgment

This work was partially supported by National Science Council of Taiwan (NSC 102-2221-E-009-125).

References

1. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinant from amino acid sequences. Proc. Nat. Acad. Sci USA 78: 3824-3828.
2. Pellequer J, Westhof E, Van Regenmortel M (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunol. Lett. 36: 83-99.
3. Pellequer J, Westhof E (1991) Predicting location of continuous epitopes in proteins from their primary structures. Meth. Enzymol. 203: 176-201.
4. Blythe MJ, Doytchinova IA, Flower DR (2002) JenPep: A database of quantitative functional peptide data for immunology. Bioinformatics. 18: 434-439.
5. Larsen J, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. Immu. Res. 2: 2. doi:10.1186/1745-7580-2-2.
6. Saha S, Raghava G (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 65: 40-48.
7. Chen J, Liu H, Yang J, Chou K (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423-428.

8. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. J. Mol. Recognit. 21: 243-255.
9. Andersen PH, Nielsen M, Lund O (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci 15: 2558–2567.
10. Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, et al. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics 9: 514.
11. Sun J, Wu D, Xu T, Wang X, Xu X, et al. (2009) SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Res 37(suppl_2): W612–W616.
12. Rubinstein ND, Mayrose I, Martz E, Pupko T (2009) Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinformatics 10: 287.
13. Liang S, Zheng D, Standley DM, Yao B, Zacharias M, Zhang C (2010) EPSVR and EP-Meta: prediction of antigenic epitopes using support vector regression and multiple server results. BMC Bioinformatics 11: 381.
14. Wolpert D (1992) Stacked Generalization. Neural Networks, 5: 241-259.
15. Ting KM, Witten IH (1997) Stacked generalization: When does it work?. In Proc. 15th IJCAI: 866-873.
16. Gama J (1998) Combining classifiers by constructive induction. In Proc. ECML-98: 178-189.
17. Gama J, Brazdil P (2000) Cascade Generalization, Machine Learning 41(3): 315-343.
18. Quinlan JR (1993) C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers.
19. Duda RO, Hart PE, Stork DG (2001) Pattern Classification, 2nd edn. New York: Wiley.
20. C.-C. Chang C.-C and C.-J. Lin C.-J (2011) LIBSVM: a library for support vector machines. ACM Trans. Intelligent Systems and Technology, vol. 2(27): 1-27.
21. Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J (2011) Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature. BMC Bioinformatics 12: 341.
22. Kringelum JV, Lundegaard C, Lund O, Nielsen M (2012) Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. PLOS Computational Biology 8(12).
23. Hubbard SJ , Thornton JM (1993) NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College London.
24. Nagano K (1973) Logical analysis of the mechanism of protein folding: I. predictions of helices, loops and beta-structures from primary structure. Journal of Molecular Biology 75(2): 401-420.
25. Kossiakoff AA, Chambers JL, Kay LM,Stroud RM (1977) Structure of Bovine Trypsinogen at 1.9 Å Resolution. Biochemistry 16(4): 654-664.
26. Michel Sanner, Arthur J. Olson, Jean Claude Spehner (1996). Reduced Surface: an Efficient Way to Compute Molecular Surfaces. Biopolymers 38(3): 305-320.
27. Karplus P, Schulz G (1985) Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen. Naturwissenschaften 72: 212-213.
28. Parker J, Guo D, Hodges R (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.
29. Zhang Z, Schäffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26(17): 3986-90.

30. Gerstein M, Tsai J, Levitt M (1995) The Volume of Atoms on the Protein Surface: Calculated from Simulation, using Voronoi Polyhedra. Journal of Molecular Biology 249(5): 955-966.
31. Schlessinger A, Ofran Y, Yachdav G, Rost B (2006) Epitome: database of structure-inferred antigenic epitopes. Nucleic Acids Res 34(Database issue): D777–D780.
32. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2006) The Immune Epitope Database 2.0. Nucleic Acids Res 38 (suppl 1): D854-D862.
33. Pai TW, Lo YT, Wu WK, Chang HT (2013) Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics. BMC Bioinformatics 14(Suppl 4): S3.
34. Singh H, Ansari HR, Raghava GPS (2013) Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence. PLOS ONE 8(5) : e62216.