Precise Predictions of Linear B Cell Epitopes in Protean 3D

Steven Darnell, Ph.D.* *and Martin Riese, M.Sc.* *DNASTAR, Inc., Madison, WI USA, www.dnastar.com*

October 1, 2012

B cell epitopes–the part of an antigen recognized by an antibody–are conformational in nature; however, accurately predicting these epitopes is difficult for proteins without a known three-dimensional (3D) structure. DNASTAR presents the *NeoClone method*, a machine learning approach that improves the ability to predict linear B cell epitopes (a peptide segment) using only sequence-based information. This method provides better overall predictive accuracy than COBEpro and Epitopia, two leading epitope prediction methods. Free trial software is available at dnastar.com.

Introduction

The state of the art in linear epitope prediction is embodied by the application of supervised machine learning techniques, which train algorithms using empirical data to predict epitopes using probabilistic models or combinations of protein features. Two leading examples are COBEpro and Epitopia. COBEpro relies on propensities of amino acids and dipeptides[1], an over-reduced set of sequence-only features. Conversely, Epitopia considers a rich set of sequence properties, including physical, chemical, structural, and geometric properties[2]. In collaboration with NeoClone® Biotechnology International, LLC (Madison, WI USA), DNASTAR presents a linear B-cell epitope prediction method with a better overall predictive accuracy than COBEpro and Epitopia.

The *NeoClone method*–a predictive model that considers secondary structure, flexibility, hydropathy, and antigenicity–is based on NeoClone's antibody development and bioinformatics approach and is featured in Protean 3D, DNASTAR's application for exploring macromolecular structure, motion, and function. The model benefits from the cross-disciplinary insights and expert knowledge of both NeoClone and DNASTAR. The result is an adaptive, interactive tool enabling researchers to accelerate their immunological research by focusing on a protein's most promising antigenic regions.

Methods

The process of implementing the NeoClone method requires four steps: 1) assemble a data set of peptides with experimentally characterized antigenicity, 2) calculate a set of features describing the localized properties of each peptide, 3) select the most important features and create a predictive machine learning model, and 4) analyze the performance of the model. The resulting model is used by Protean 3D to identify putative antigenic regions in any given protein sequence.

Epitope Data Sets

*Corresponding author. Email: info@dnastar.com

Data Set for Cross-Validation

The main data set is collected from the NeoClone archive. Each peptide is used to inoculate a mouse and each mouse is used to create many unique hybridoma cultures[3]. The relative binding affinity of the antibodies secreted from each culture is characterized by an enzyme-linked immunosorbent assay (ELISA) against the original peptide. The peptides are classified using the following definitions:

- *Highly antigenic*: at least 5% of cultures create tight binding antibodies (20 out of 384 cultures)
- *Antigenic*: at least 2.5% of cultures create tight binding antibodies (10 out of 384 cultures)
- *Not antigenic*: less than 2.5% of cultures create tight binding antibodies

where a "tight binding" antibody has a relative binding affinity fivefold over background.

The data set consists of 44 peptides from 18 proteins where 19 peptides (with 331 total residues) are highly antigenic, 9 (with 172 residues) are antigenic, and 16 (with 245 residues) are not antigenic. The proteins group into 15 clusters where the members between clusters have less than 30% sequence identity.

Independent Test Set

A second data set–independent of the previous set–is collected from Bcipep[4] to further validate our method. Each peptide is experimentally characterized; however, the peptides are not characterized by the frequency of producing tight binding antibodies. Instead, the relative antigenicity is listed in Bcipep as either highly antigenic, antigenic, or not antigenic. The test set consists of 285 peptides from 110 proteins where 129 peptides (with 1591 total residues) are highly antigenic, 128 (with 1799 residues) are antigenic, and 28 (with 406 residues) are not antigenic. Each protein has less than 30% sequence identity to every other protein.

Sequence Features

The features used to train our method are inspired from the NeoClone bioinformatics approach[3]. We characterize the biochemical properties within a protein using ten sequence-based bioinformatics analyses describing beta turns, antigenicity, hydrophobicity, flexibility, and transmembrane propensity. Each analysis produces a score describing the local sequence environment for each residue in each protein.

Machine Learning Approach

We use a supervised machine learning approach to search for subsets of features that generate robust and predictive models. The approach is set up to predict whether individual residues are part of an antigenic peptide. The considered algorithms include decision trees[5–7], random forests[5], and support vector machines[8]. In practice, all combinations of the ten bioinformatics analyses are screened using all machine learning algorithms.

The NeoClone method is comprised of two models: an antigenic predictor and a highly antigenic predictor. The antigenic model is trained using features describing beta turns, antigenicity, hydrophobicity, and flexibility. The highly antigenic model is trained using features describing beta turns and hydrophobicity. For each model, an exhaustive search guarantees we identify the optimal features and algorithm for describing an antigenic signature.

F1 Score, Precision, and Recall

Our data sets contain an uneven distribution of antigenic peptides; as such, a naïve prediction that all residues are part of an epitope leads to an arbitrarily high statistical accuracy, or the fraction of correctly classified residues. To avoid this situation, we evaluate model performance in terms of the widely used *F1* score ($F1$)–a robust metric of overall accuracy defined in terms of precision and recall:

$$F1 = \frac{2PR}{P+R} \tag{1}$$

where *precision* (P) is the fraction of predicted antigenic residues that belong to true epitopes and *recall* (R) is the fraction of true epitope residues that are predicted as antigenic. Each measure's worst value is 0 and its best value is 1.

Cross-validation

To guarantee that our models are not overtrained on a single data set, the F1 score for the NeoClone method is estimated by cross-validation–a resampling

technique that averages the analysis over a series of training and testing events on different subsets of the same data set[9;10]. This approach guards against calculating overly optimistic F1 scores for models suggested from the data.

A 15-fold "leave one protein cluster out" cross-validation is performed to estimate the accuracy of the NeoClone method when presented with a new sequence. This method is more logical than a typical tenfold cross-validation (where residues are randomly divided into 10 equal partitions) given the inherent dependencies between residues in the same protein.

Results and Discussion

In this section, we compare the accuracy of our new knowledge-based B cell epitope prediction model–the *NeoClone method*–against two leading sequence-based methods, COBEpro and Epitopia. It is important to emphasize that the F1 score for a random model represents the practical baseline when analyzing performance. The following results illustrate that the NeoClone method performs considerably better than randomly guessing and consistently outperforms both COBEpro and Epitopia. We conclude by discussing how to use the NeoClone method in Protean 3D.

DNASTAR vs. COBEpro and Epitopia

The NeoClone method from DNASTAR more accurately predicts linear B cell epitopes than COBEpro and Epitopia using two different data sets (Figure 1). This analysis defines an epitope as the set of residues belonging to a peptide classified as antigenic or highly antigenic (see Epitope Data Sets for definitions). Table I summarizes the precision, recall, and F1 Score for each method as it performs on the NeoClone and Bcipep data sets. We used the NeoClone antigenic model for this analysis. In both cases, our model's precision (predicted epitope residues are correctly classified) remains high and its recall (true epitope residues are correctly classified) leads across all data sets. This indicates that the NeoClone method has the greatest likelihood of discovering new linear epitopes and its high precision will keep false positive predictions low.

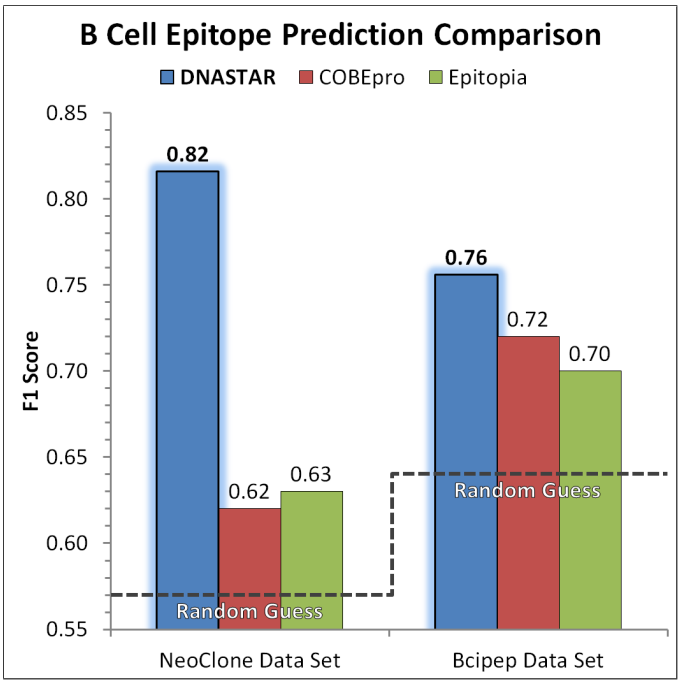

Figure 1: *DNASTAR outperforms COBEpro and Epitopia.* The dotted line represents the expected F1 score when guessing randomly (NeoClone set: 0.57, Bcipep set: 0.64).

	NeoClone			Bcipep		
	P	*R*	*F1*	*P*	*R*	*F1*
DNASTAR	0.75	0.90	0.82	0.90	0.65	0.76
COBEpro	0.64	0.59	0.62	0.90	0.59	0.72
Epitopia	0.68	0.58	0.63	0.92	0.56	0.70

Table I: *Performance of DNASTAR, COBEpro, and Epitopia on two different data sets*

Using the NeoClone Method in Protean 3D

It is easy to predict linear B cell epitopes for any given protein sequence or structure in Protean 3D. After the NeoClone method is activated from the Methods Panel, the results from the antigenic and highly antigenic models are instantaneously displayed in the Analysis View (Figure 2). A confidence score is assigned to each residue, where a value of 1 indicates high confidence the residue is part of an epitope and a value of 0 indicates high confidence the residue is not part of an epitope. By default, residues with $confidence \geq 0.5$ are predicted to be antigenic; however, that threshold can be changed using the Parameters Bar.

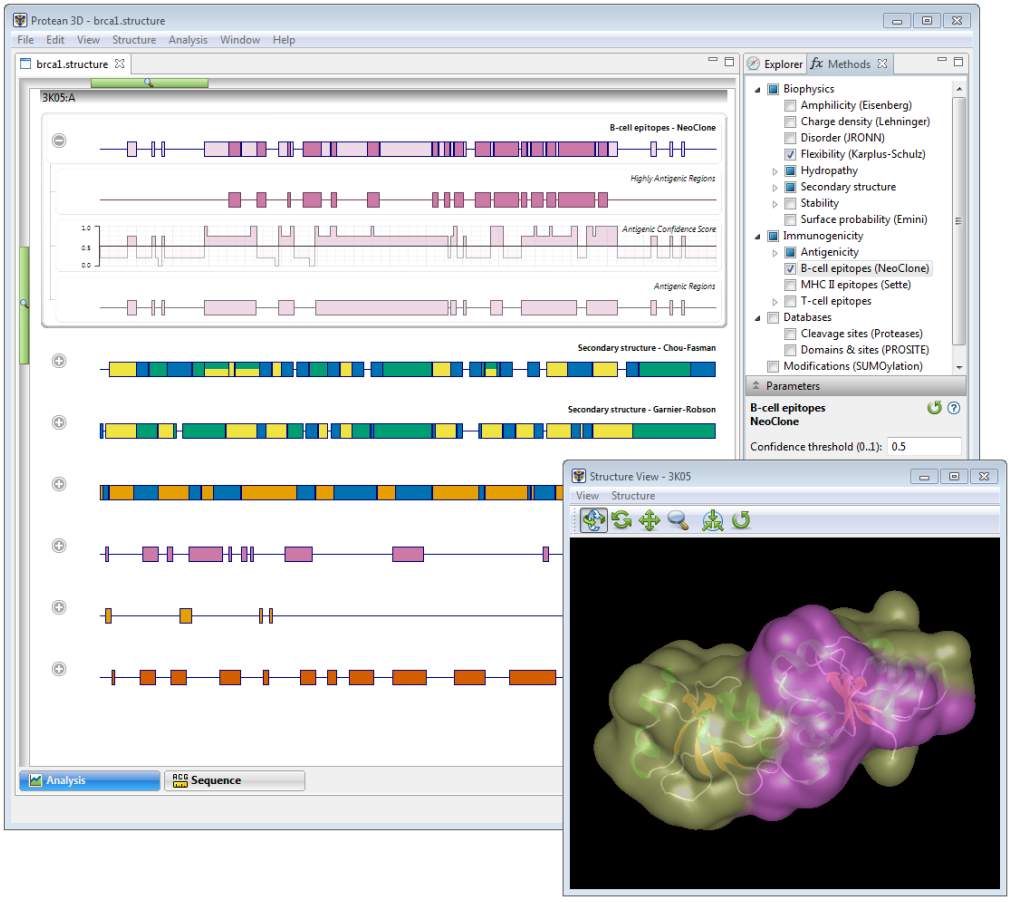

Figure 2: *BRCA1 analyzed by the NeoClone method in Protean 3D.* The largest predicted linear epitope is highlighted in pink on the solvent accessible surface of the breast cancer type 1 susceptibility protein. In the Analysis View, antigenic predictions are highlighted in light pink and highly antigenic predictions are highlighted in dark pink.

Conclusions

This study demonstrates our success in using machine learning to select features that best predict the identity of linear B cell epitopes, thus limiting the complexity of the problem. Benefiting from the strong cross-disciplinary knowledge of NeoClone and DNASTAR, the NeoClone method clearly outperforms COBEpro and Epitopia, two leading epitope prediction methods. In addition, our method demonstrates the ability to keep false positive predictions low as well as the greatest aptitude for discovering new linear epitopes compared to other surveyed methods. The NeoClone method is automated, precise, and instantaneously fast–making it a valuable tool for the interactive exploration of protein immunogenicity. Requests for a fully-functional free trial of Protean 3D featuring the *NeoClone method* can be submitted at www.dnastar.com.

References

[1] Sweredoski, M. & Baldi, P. COBEpro: a novel system for predicting continuous B-cell epitopes. *Protein Eng. Des. Sel.* **22**, 113–120 (2009).

[2] Rubinstein, N., Mayrose, I. & Pupko, T. A machine-learning approach for predicting B-cell epitopes. *Mol. Immunol.* **46**, 840–847 (2009).

[3] NeoClone. Custom antibodies: The NeoAb® process (2012). URL http://www.neoclone.com/index.php?option=com_content&view=article&id=2&Itemid=3.

[4] Saha, S., Bhasin, M. & Raghava, G. Bcipep: a database of B-cell epitopes. *BMC Genomics* **6**, 79 (2005).

[5] Hall, M. *et al.* The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10–18 (2009).

[6] Stiglic, G., Kocbek, S., Pernek, I. & Kokol, P. Comprehensive decision tree models in bioinformatics. *PLoS ONE* **7**, e33812 (2012).

[7] RuleQuest. Data mining tools See5 and C5.0 (2012). URL http://www.rulequest.com/see5-info.html.

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

[9] Stone, M. Cross-validatory choice and assessment of statistical predictions. *J Roy Stat Soc B Met* **36**, 111–147 (1974).

[10] Geisser, S. The predictive sample reuse method with applications. *J Am Stat Assoc* **70**, 320–327 (1975).