Predicting Continuous Epitopes in Proteins

Reeti Tandon[1], Sudeshna Adak[1], Brion Sarachan[2], William FitzHugh[3], Jeremy Heil[3],
Vaibhav A. Narayan[3]
Computational Biology & Biostatistics Lab, GE Global Research, Bangalore, India[1]
Computational Biology & Biostatistics Lab, GE Global Research, Niskayuna, NY, USA[2]
Celera Genomics, 45 West Gude Drive, Rockwille, MD, USA[3]
E-mail: reeti.tandon@geind.ge.com

Abstract

The ability to predict antigenic sites on proteins is crucial for the production of synthetic peptide vaccines and synthetic peptide probes of antibody structure. Large number of amino acid propensity scales based on various properties of the antigenic sites like hydrophilicity, flexibility/mobility, turns and bends have been proposed and tested previously. However these methods are not very accurate in predicting epitopes and non-epitope regions. We propose algorithms that combine 14 best performing individual propensity scales and give better prediction accuracy as compared to individual scales.

1. Introduction

The antigenic regions of protein that are recognized by the binding sites of immunoglobin molecules are called B-cell epitopes. Identification of epitopes on proteins is of vital importance for developing synthetic peptide vaccines, immunodiagnostic tests and antibody production. Previous algorithms predict the position of epitopes based separately on certain protein properties like hydrophilicity, mobility/flexibility, surface accessibility, structure etc. Algorithms using a logical combination of scales have been developed but report poor accuracy ranging from 40%-60% [4]. Kolaskar and Tangaonkar developed a combination scale using hydrophilicity, flexibility and surface accessibility, known as the Antigenic Propensity (AP) scale which is considered a gold standard in epitope prediction with an accuracy of 75% [3]. We propose using learning algorithms to combine selected scales to improve the accuracy of epitope prediction.

2. Methods

2.1. Antigenic Propensity Scales

Blythe and Flower [2] report 14 best performing scales (Table 1) from the AAIndex database (http://www.genome.jp/aaindex/), which is a collection of 484 scales based on different physicochemical and biological properties of amino acids. We have normalized these 14 scales between 0 and 1 and combined them using learning algorithms [3].

Table 1. Scales used for combination algorithms

Scale Tag	Description
A098	Alpha-helix indices for alpha-proteins (Geisow-Roberts, 1980)
A335	Relative preference value at C1 (Richardson-Richardson, 1988)
C137	Sequence frequency (Jungck, 1978)
H215	Long range non-bonded energy per atom (Oobatake-Ooi, 1977)
H364	Zimm-Bragg parameter sigma x 1.0E4 (Sueki et al., 1984)
P063	Size (Dawson, 1972)
P214	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)
P219	Optimized propensity to form reverse turn (Oobatake et al., 1985)
P280	Weights for beta-sheet at the window position of 3 (Qian-Sejnowski, 1988)
P353	Mean area buried on transfer (Rose et al., 1985)
Z019	Normalized positional residue frequency at helix termini C''' (Aurora-Rose,1998)
Z021	Delta G values for the peptides extrapolated to 0 M urea (O'Neil-DeGrado,1990)
Z022	Helix formation parameters (delta delta G) (O'Neil-DeGrado, 1990)
Z031	Free energy in beta-strand region (Munoz-Serrano, 1994)
Antigenicity	Antigenic propensity scale by Kolaskar and Tangaonkar(1990)

2.3. Data

The performance of the scales and the combination algorithms were evaluated on the BCIPep database, which is a collection of approximately 3000 B-cell epitopes.

The ~3000 epitopes in the BCIPep database come from 971 proteins. The 971 proteins were each represented as sub-sequences of length 10 with each sub-sequence being labeled as 1 (any amino acid in the subsequence is a part of epitope) or 0 (if none of the amino acids in the sub-sequence are a part of the epitope). The sub-sequences were chosen to be of length 10 because (1) literature reports commonly occurring epitopes of length 5-15 and (2) the frequency of epitopes of length ten was highest in BCIPep database (Fig 1). The method resulted in ~38000 subsequences from 971 proteins. The generated sub-sequences were divided randomly into a training dataset (90% of the observations) and a validation dataset (10% of the observations).

A profile for each of the proteins and each scale was computed using a moving average method where the average value for 9 residues is assigned to the center residue. Inputs to the learning algorithm were the maximum and the minimum of each of 14 scale profiles on the sub-sequences of length 10.

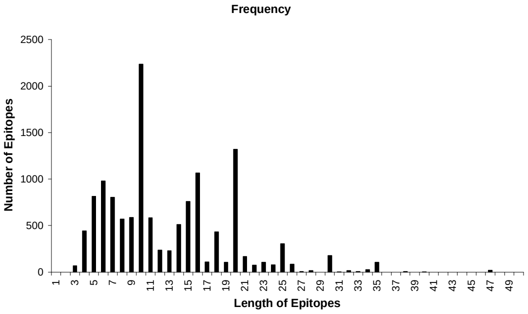

Figure 1. Frequency of occurrence of epitope lengths in BCIPep database

2.4. Evaluation of individual scales and combination algorithms

Instead of comparing sensitivity and specificity at a specific threshold, we have compared the sensitivity and specificity across all the possible thresholds to generate a ROC plot [2].

The 14 scales have been combined using Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Classification Trees. The models give a posterior probability for each sub-sequence being an epitope, which is used to generate a ROC plot.

3. Results

The individual scales have sensitivity ranging from 2%-32% (Fig. 2) while the combination algorithms give approximately 46% sensitivity at 10% false positive rate (Fig 3). The combination algorithms give approximately 80% sensitivity at 40% false positive rate whereas the individual scales have a poorer false positive rate (50%) at the same sensitivity. The QDA gives the best performance among the combination methods and has better sensitivity as compared to any individual scale and antigenic propensity scale though other combination algorithms were comparable.

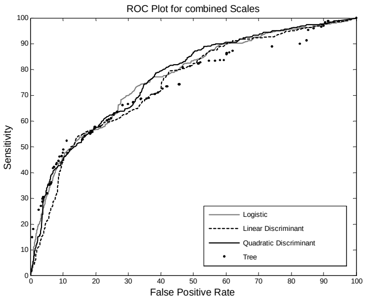

Figure 2. ROC plot for combination algorithms on validation data

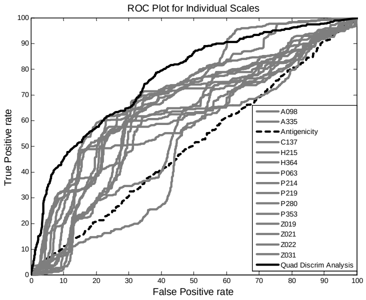

Figure 3. ROC plot for individual scales on validation data

4. Discussions and Conclusions

The performance of individual scales was improved by combining them. However, the AP scale does not give good performance probably because the scale was developed on a rather small dataset. Efforts are ongoing to recalculate the AP scale for the BCIPep database. We are also working on to incorporate structural information into the combination algorithms to further improve their performance.

5. References

[1] M.J. Blythe and D.R. Flower (2005). “Benchmarking B cell epitope prediction: Underperformance of existing methods”, *Protein Science*, Cold Spring Harbor Laboratory Press,pp 246-248.

[2] H. Delacour, A. Servonnet, Perrot A, J.F. Vigezzi, J.M. Ramirez (2005). “ROC (receiver operating characteristics) curve: principles and application in biology”, *Ann Biol Clin*, John Libbey Eurotext, pp 145-154.

[3] A.S. Kolaskar and P.C. Tangaonkar (1990). “A semi-emperical method for prediction of antigenic determinants on protein antigens”, *FEBS*, Elsevier, pp 172-174.

[4] M.H.V. Regenmortel and G. Marcillac (1988). “An assessment of prediction methods for locating continuous epitopes in proteins”, *Immunology Letters*, Elsevier, pp 95-107.