JOURNAL OF ENGINEERING AND APPLIED SCIENCE, VOL. ??, NO. ?, AUG. 2011, PP. 327-???
FACULTY OF ENGINEERING, CAIRO UNIVERSITY

CONFORMATIONAL B-CELL EPITOPES CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

[1] [2] [3]

ABSTRACT

Vaccination is one of the most important and successful public health interventions to human lives. Computational techniques presented by bioinformatics seem to be believed by many scientists that it can contribute significantly in the vaccine design process. They can overcome the time and money limitations of the traditional used methods. One essential step in the vaccine design process is the B-cell epitope prediction. Synthetic peptides can be developed after reaching an accurate prediction degree. As most B-cell epitopes are known to be conformational (discontinuous), it's vital to develop modules to enhance its prediction results. In this paper, a classification model for classifying conformational b-cell epitopes is proposed. This model depends on the machine learning techniques for classification and clustering. The model achieved high performance on the train data.

KEYWORDS: Vaccine Design, B-Cell Epitopes, Classification, Clustering, Support Vector Machine.

1. INTRODUCTION

Vaccination is the process of artificial induction of immunity in an attempt to protect living organisms from infectious diseases. A Vaccine is a biological preparation from weakened or killed forms of the microbe or its toxins. Vaccines are usually provided using a vaccine design process to improve the immunity to a

[1]

[2]

[3]

particular disease. Vaccine design process can either be achieved by traditional experimental procedures or by computer-aided methods. Traditional vaccine design methods are known to be suffering from three major drawbacks; being expensive, time consuming and sometimes harmful to a researcher or experimental animal. In-silico vaccine design can overcome these problems with the help of the recent availability of the genome, proteome sequencing and the 3D structure of the protein complexes. This is in addition to the wide spread of the data mining and bioinformatics (especially immunoinformatics) computational tools.

One of the important steps in the vaccine design process, which can be improved using computational methods, is the epitope prediction. An epitope (aka antigenic determinant) is that part of an antigen which binds to the antibody molecule produced by the immune system specifically to this antigen. Because experimentally confirming that a specific peptide is an epitope or not is time and money consuming, epitope prediction can be of a great help to introduce new potential epitopes. These epitopes are essential to design a molecule that can replace an antigen in the process of either antibody production or antibody detection [1] [2].

The immune system recognizes the epitope part of an antigen by its antibodies, B-cells or T-cells. The B-cells have the capability to recognize native antigens and then produce antibodies which are effective mediators to the extracellular pathogens. On the other hand, T-cells can't recognize except processed antigens in the antigen presenting cells (APCs) [3]. Epitopes are named after their lymphocytes receptors to have B-cell epitopes and T-cell epitopes. Epitopes take one of two forms; linear (continuous) epitopes and conformational (discontinuous) epitopes. Linear epitope is a short sequence of residues in the primary protein structure. Conformational epitope is a patch of atoms which are far apart in the primary structure but joined on the protein surface in the three-dimensional space (due to protein folding). It's well known and proved that approximately 90% of the B-cell epitopes are conformational [2]. This is the main reason for concentrating in the paper at hand on predicting conformational B-cell epitopes.

This paper objective is to propose solution to some of the drawbacks of the models built to deal with the problem of classification of conformational B-cell epitopes. This classification process undergoes several steps trying to result in a considerable accuracy. As area under the curve (AUC) measurement of classification till now doesn’t exceed 75%, more work is needed to reach higher values and increase the performance.

This paper is organized as follows; section (2) will give an overview of the problem and its importance. Section (3) is to summarize the most important work done handling the conformational b-cell epitopes prediction and classification. The applied techniques and used datasets are discussed in section (4). Results and evaluation criteria are to be found in section (5). Section (6) will conclude the paper.

2. BACKGROUND

To better understand the importance of the B-cell epitope prediction process, the epitope function must be explained first. The epitope function will become obvious after introducing how the immune system responds to a foreign cell. A human immune system is a network of cells, tissues and organs where its main function is to defend him against foreign invasions. Every cell in the body of a human being carries the same set of distinctive surface proteins on its surface which marks this specific person as “self”. This is why the immune system doesn’t attack the body it contains. And, this is how the immune system detects foreign bodies. That’s because their cells are marked as “non self” [4].

These markers are called epitopes. When an antigen is found in a human body, soluble substances known as antibodies secreted by the immune system B-cells attach the epitopes found on that antigen. Epitopes on different antigens can take different shapes. When an antigen-specific antibody on a B-cell matches up with an antigen and binds to it, the B-cell engulfs this antigen. Then, the B-cell changes to a large plasma cell factory after joining of a special helper T-cell. This factory produces up to ten million identical copies of this antibody in less than an hour.

An adaptive immune system is characterized by having memory cells that remember the pattern for every antibody secreted by the B-cells. This is what is called immunological memory. This results in response enhancement to subsequent encounters with that same pathogen. This process of acquired immunity is the basis of vaccination [5]. As conformational B-cell epitopes constitutes the majority of the B-cell epitopes, their prediction is found to be a major step for successful vaccine design process.

3. RELATED WORK

The computational B-cell epitope prediction algorithms take the attention of many researchers during the past few years. That's because they offer a great hope to predict the epitopes in an antigen to mimic them and use synthetic vaccines instead of killed or weakened microbes. This facility will overcome the time and money consumption drawbacks of the traditional methods, in addition to being a safe process. Researchers study these algorithms from different points of view. Some of them concentrate on linear B-cell epitopes prediction while others were interested in the conformational B-cell epitope prediction. Following is a summary of the main contribution in the conformational B-cell epitope prediction field which is the main scope of this paper.

The first attempt to predict conformational epitopes was in 2005 through a web server called CEP (Conformational Epitope Prediction). This server offers the option to predict sequential epitopes as well [6]. In 2006 DiscoTope was presented as a novel method for prediction of residues located in discontinuous B-cell epitopes. The authors of this research project declared that this is the first method to focus explicitly on predicting epitopes. The main contribution of Discotope was being the first reported method combining a propensity scale with 3-Dimensional structural information such as spatial proximity [7]. A year after, a research was carried out to evaluate eight web servers developed for antibody and protein binding sites (antigenic determinants or B-cell epitopes) prediction. This paper focused on structural epitopes inferred from known 3D structures of antibody-protein complexes from Protein Data Bank (PDB). The main contribution here was developing B-cell epitope benchmark datasets inferred from existing 3D structures of antibody protein complexes [8].

In 2008 two important web servers were developed. The first “PEPITO - a state-of-the-art B-cell epitope predictor” overcomes some of the limitations of previous predictors using half sphere exposure values of the amino acid propensity scale, solvent accessibility information along with side chain orientation. PEPITO uses propensity scales and half sphere exposure values at multiple distance thresholds from the target residue. It also derived two additional datasets from the set of protein chains that are common to two famous 3D structure datasets; Epitome and DiscoTope [9]. The second “ELLIPRO” is another web-accessible application where its main contribution is using each residue’s center of mass rather than its Cα atom. The author compared his model against six different structure-based methods [10].

While most of the researchers concentrate on predicting conformational B-cell epitopes using the antigen 3D structure, a research was carried out in 2010 using the primary sequence in prediction. The author declared that his idea is very practical when the tertiary structure of an antigen is not available [11]. More research on this classification problem was introduced by Zhang et al. who proposed a novel concept named “thick surface patch” to take adjacent surface residues into consideration as well as interior residues. They dealt with the imbalanced data problem by bootstrapping and voting procedure [12]. A valued comparison between sequence based and structure based approaches was introduced in [13]. It also presents a novel prediction combining five important features with the powerful Support Vector Machine (SVM) classifier.

3.1 Analysis and Criticism

As previously mentioned, conformational B-cell epitopes prediction constitutes an essential step not only in the vaccine design process but also in diagnostics. From the previous related work, the conformational B-cell classification model general workflow can be summarized in fig. 1. To reach reliable prediction accuracy, all the shown implied process components must be very well designed and tuned. Scientists worked on this important research field (most often) proposed an enhancement to one of the system components. The main drawback is that the recorded accuracy till now doesn't reach a reliable degree. One of the factors that affect the resultant accuracy is the imbalanced characteristic of the 3D structure datasets.

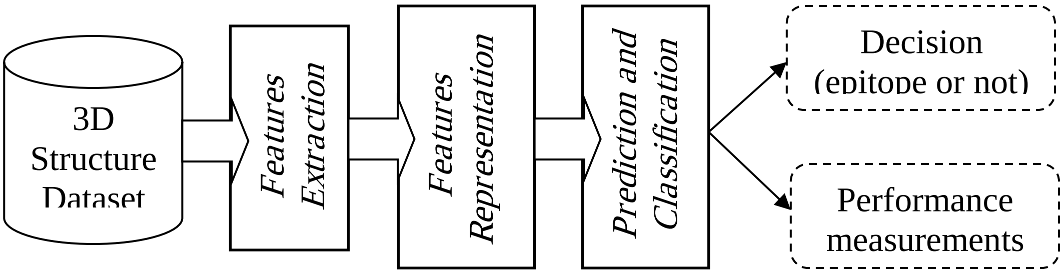

Fig. 1. General Workflow of Conformational B-Cell Epitope classification

3.2 Proposed Solution

Imbalanced data represents a major problem where the number of epitope residues in a dataset is always very small compared to the number of non-epitope residues. Poor models performance and training time high consumption are the consequences of imbalanced data. Some researchers solved this issue by choosing from the non-epitope residues a number equal to the number of epitope residues. This was achieved either by random selection from the original data [14] or by using a balanced classifier [7]. Our proposed solution is to choose this equal number of non-epitopes by clustering them into heterogeneous groups of homologous residues and taking the top residue as a prediction candidate.

In the proposed model, Support vector machine will be used as a classifier. Different kernel functions will accompany the SVM implementation trying to reach the best kernel representing the data.

4. METHODOLOGY

To easily explain the proposed conformational B-cell epitopes classification model, it is divided into two main stages. The first handles the data matrix formation while the second is responsible for the classification process. Next is the explanation of the employed dataset followed by a discussion of the two model stages.

4.1 3D Structure Dataset

A structure dataset contains important spatial information about the antigen-antibody protein complexes. In the proposed model, seventy one antigen antibody complexes are taken from the discotope dataset [7]. Antigen chain is extracted from each complex through a protein data bank identifier. All the information needed for each chain is downloaded from the protein data bank in a ".pdb" file. Features are extracted and calculated using this file. By analyzing the seventy-one antigen chains, it is found that they contain 13,417 residues. 1128 of these residues are detected as epitopes according to the discotope supplementary files. The other 12,289 are detected as non-epitopes. These 71 complexes are divided into five heterogeneous groups. Two complexes are taken from each group for testing while the remaining sixty-one complexes are used for training. To summarize; the whole dataset which contains 71 complexes is further converted into: train dataset (61 complexes) and test dataset (10 complexes).

4.2 Stage 1: Data Matrix Formation

This stage is very important because it is responsible for getting out the final form of the data which will be further introduced to the classifier. The final data must take a matrix shape with size m×n. Let each cell in the matrix contains a value (Xij) where; "i" ranges from 1 to m (rows) and represents the total number of residues in the dataset. Likely, "j" ranges from 1 to n (columns) and represents the total number of features. This stage contains two main steps:

4.2.1 Features extraction

Features are the information which represents the dataset after being extracted from it. For each residue there must be a feature vector for its representation. The chosen features influence the classification process to a great extent. Then, the process of choosing which features are able to discriminate between different classes is very critical. A parameter which helps deciding which features to use is the location of the B-cell epitope residues in the antigen 3D structure. According to research articles, most of the antigenic epitope residues are found on the surface of the antigen protein structure [7][13]. This important finding is illustrated after a proposed analysis of the antibody interacting sites. Then, chosen features by many researchers reflect the characteristics of the surface exposed residues. Taking this fact into consideration, three features are chosen according to the recommendations of the previous work.

1. Relative Solvent Accessibility: RSA is the surface area of a biomolecule that is accessible to a solvent divided by the maximal Accessible Surface Area (ASA) of all residues in an antigen chain. RSA is calculated by the NACCESS which is a program running under linux. The input to this program is a ".pdb" file and its output is a ".rsa" file which contains the RSA value of each of the antigen chain residues as one of its contents [15].
2. Temperature Factor (B-Factor): B-factor indicates the true static or dynamic mobility of an atom. Thermal motion and crystal imperfections causes deviation of the position of the atom from that present in the atomic coordinates. B-factor is the measure of that deviation [16]. The temperature factor in this study is extracted directly from the downloaded ".pdb" file. Values of the extracted RSA and B-Factor features for all the residues constituting the discotope dataset are visualized in fig. 3 and fig. 4 respectively.
3. Amino acid physio-chemical properties: The 20 amino acids encoded directly by the genetic code can be divided into several groups based on their properties. These properties are important for protein structure and protein–protein interactions [17]. A record of these properties and their different values recorded by different scientists is present in the AAindex database

(http://www.genome.jp/aaindex/). One of these properties or a combination of them was mainly used for predicting linear epitopes. Their usage is extended to conformational B-cell epitopes prediction. According to previous recommendations three important properties (hydrophilicity, flexibility, antigenicity) are used in this study. The propensity score of these three properties results in three different residue features. An algorithm to calculate the propensity score of physio-chemical properties is shown in fig. 2. This flowchart is repeated for each residue in each complex.

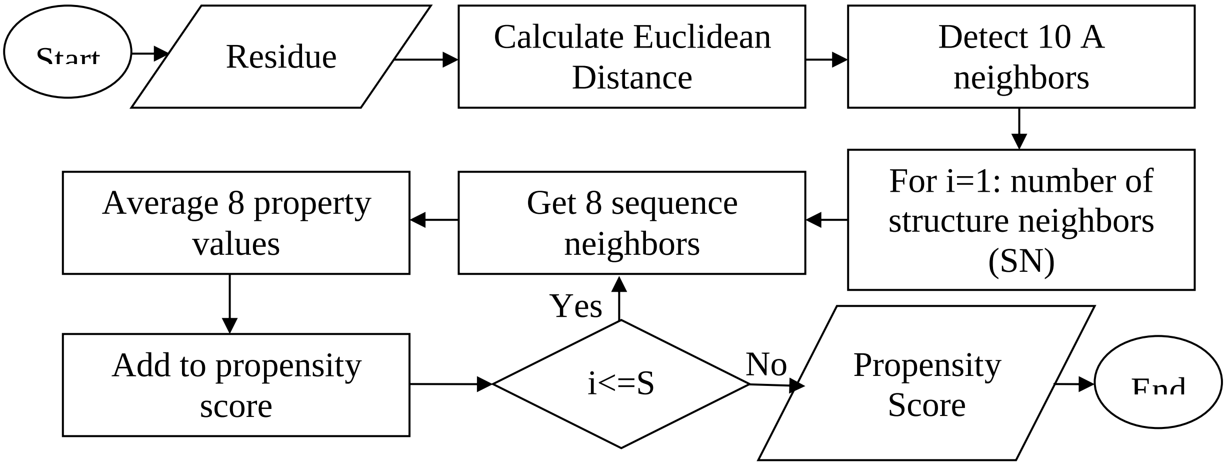

Fig. 2. Flow Chart for Calculation of Propensity Scores

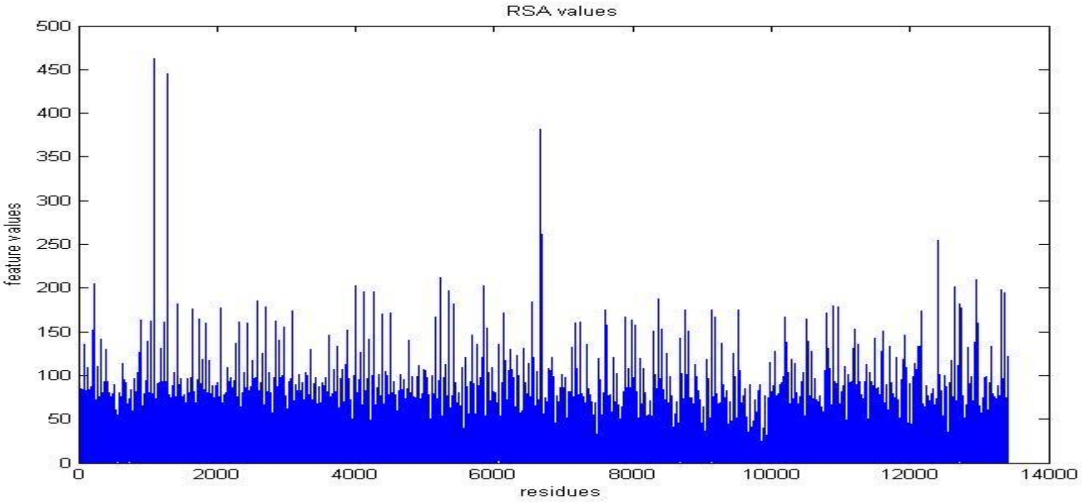

Fig. 3. RSA Values of the 13417 residues in the 71 ag-ab complexes

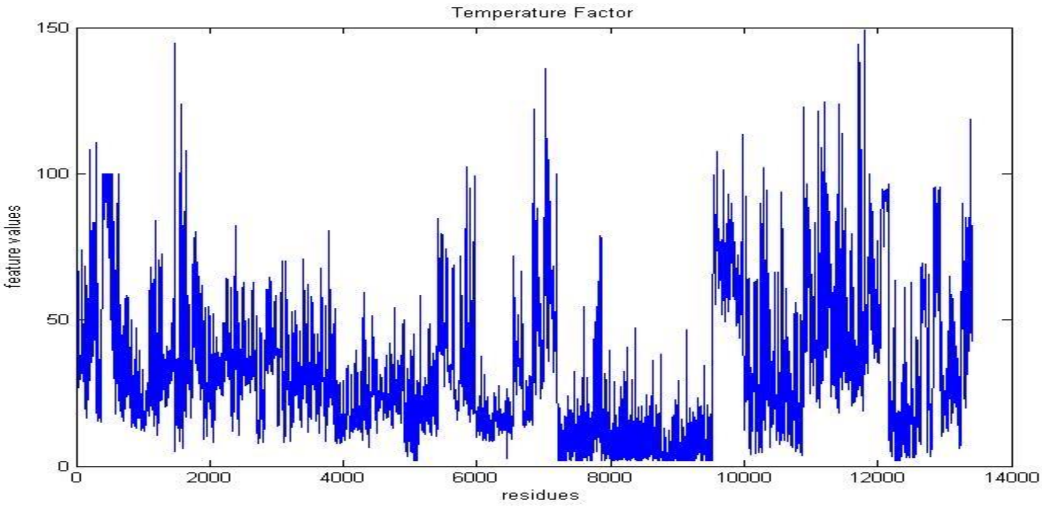

Fig. 4. Temperature Factor Values of the 13417 residues in the 71 ag-ab complexes

4.2.2 Features representation

After detecting the features and deciding a number of them which will contribute in the classification process, these chosen features must be correctly represented. Then for configuring the data, a (m×n) data matrix is composed. Where, its rows (m) represent the total number of residues in all the antigen chains in the training dataset and its columns (n) represent the features. This configuration is true only if residues are independent. But, it is proved that residues are dependent. This means that a function of a residue is not solely affected by itself but also by its neighboring residues [18]. Then, the data matrix columns (n) will be equal to the number of features multiplied by number of neighbors plus one.

Target residue and its neighbors are represented by a sliding window with a pre-specified size. There are two types of windows; either sequence based or structure based window. Sequence based window considers the target residue and its preceding and following residues in the primary structure. Structure based window is the window constructed with the target residue in the first position and its neighboring residues are determined according to the 3D structure. For predicting conformational epitopes, structure based window is the appropriate choice. Two main factors affect the window construction;

i. Window size: it presents the number of target residue neighbors. A window size equal to 9 is used in the study at hand which means the nearest 8 neighbors are chosen. Then total number of features equals 45 (9×5).

ii. Metric used for neighborhood detection: a measure used to decide whether a specific residue is a neighbor to the target residue or not, the Euclidean distance is the one used here.

4.3 Stage 2: Classification Using Machine Learning Techniques

A machine learning classifier is known to have two phases; a train phase and a test phase. As mentioned before the whole 71 antigens are splitted into two groups. The previous stage (stage 1) is repeated two times to result in; a train data matrix of size (11445×45) and a test data matrix of size (1972×45). These are the inputs to stage 2 as shown in fig. 3. First, the train dataset is divided into two datasets; class (0) dataset which contains all the residues previously classified as non-epitopes and class (1) dataset which contains epitope residues. This stage is composed of three main phases.

4.3.1 Clustering

After the division it is noticed that class (0) have 10461 residues whereas class (1) contains only 984 residues. To solve this imbalanced data problem, a K-means clustering technique is used. K-means clustering is an unsupervised machine learning technique where its main objective is to divide its input data into pre-specified K clusters using the extracted features. These clusters are characterized by being mutually exclusive. K-means clustering depends on the idea of decreasing within-class distances and increases between-classes distances [19]. The distance used in this study is the Euclidean distance.

Clustering is carried in this study on the class (0) data with number of clusters (k) equals 984 which is equal to the number of residues in the class (1) data. Then 984 clusters are formed. By random choice of one residue from each cluster, 984 are present in class (0) data instead of 10461 which guaranteed balanced data.

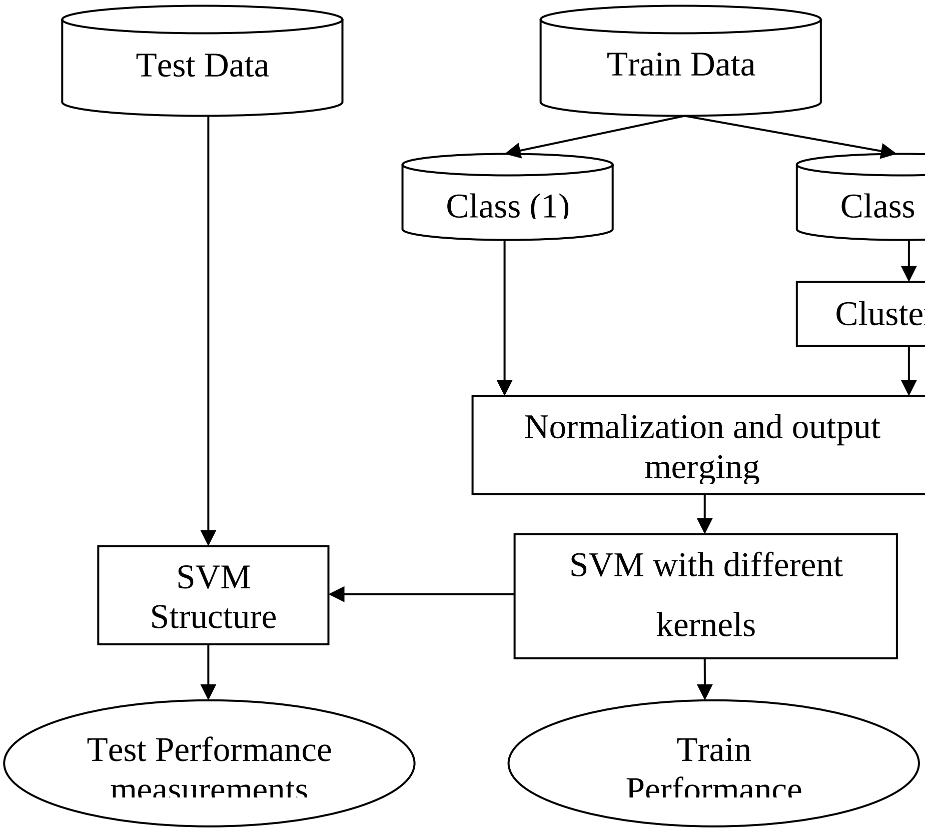
Test Data
Train Data
Class (1)
Class (0)
Clustering
Normalization and output merging
SVM Structure
SVM with different kernels
Test Performance measurements
Train Performance

Fig. 5. Workflow of the Classification Stage

4.3.2 Normalization

Now, there exist two equal datasets created from the train dataset class (0) and class (1) datasets with size 984×45 each. A normalization procedure is carried on each dataset alone where the mean and standard deviation is calculated along each column (feature). Then, each data entry in each column is normalized by subtracting the mean and dividing by the standard deviation of that column. The two normalized datasets (normalized class (0) and normalized class (1)) are then merged together to give the output of this phase. This output is a normalized train dataset containing equal number of positive and negative instances.

4.3.3 Classification

Support Vector Machine (SVM) is known to be a very powerful machine learning classifier. So, it is chosen to be trained using the train data and then classify the test data. SVM idea of operation is to construct a hyperplane between the two classes. Then, it uses an optimization solution to find the maximum margin hyperplane which is characterized by having the longest distance to the nearest points from the two classes. SVM is superb with linear as well as non-linear separable data. It solves the non-linearity problem by using different kernel functions which map the non-linear separable samples into the feature space. Different kernel functions include; Gaussian, polynomial, and Radial Basis Function (RBF) [20]. In this study SVM is first trained using the output of the normalization phase. A structure is formed which contains all the SVM parameters needed to classify the test samples. SVM is trained in this study using a k-fold cross validation technique with k=10.

5. EVALUATION AND RESULTS

For evaluating the proposed model, three measurements are recorded with each kernel function used accompanied by SVM. These measurements are; a) accuracy to present the total number of correctly classified residues, b) sensitivity (true positive rate) to represent the total number of correctly classified epitopes and c) specificity (true negative rate) to represent the total number of correctly classified non-epitopes. These are applied once for the train dataset with results recorded in Table 1. Then, they are applied again for the test dataset with results written in Table 2.

A step more is carried when recording the test dataset measurements. As the class of the test dataset is needed to be detected and assumed not to be previously known, its normalization mean and standard deviation are not defined and can’t be calculated. To solve this issue three assumptions are proposed;

1. Use the mean and standard deviation of class (1) assuming all residues are epitopes. If the output class is “1” then the assumption is true. Otherwise, the residue belongs to class “0”.

2. Use the mean and standard deviation of class (0) assuming all residues are non-epitopes. If the output class is “0” then the assumption is true. Otherwise, the residue belongs to class “1”.
3. If the output of assumption 2 is “0”, then the residue is of class “0”. If not, take the output of assumption 1 whatever it is “1” or “0”.

Note that train dataset is balanced after clustering where number of epitopes equals to number of non-epitopes (984 residues each). However; the test dataset can’t be balanced because number of epitopes and non-epitopes are not pre-determined. This is the main reason of the approximately equal measurements resulted from the classification of the train data. This is more illustrated in tables 3 and 4 for the train and test datasets respectively. They record the number of mis-classified epitopes (False Positives (FP)) in addition to the number of mis-classified non-epitopes (False Negatives (FN)).

Table 1. Train Dataset Performance Measurements

Kernel fn Performance	Linear	RBF with sigma equals		
		1	2	3
Accuracy	0.506	0.9685	0.9776	0.9695
Sensitivity	0.506	0.937	0.9756	0.9654
Specificity	0.506	1	0.9798	0.9736

Table 2. Test Dataset Performance Measurements

		Assumption 1			Assumption 2			Assumption 3		
		Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
Linear		0.3834	0.6042	0.366	0.4746	0.507	0.4721	0.476	0.507	0.473
RBF	1	0.935	0.125	0.999	0.929	0.0417	0.9989	0.929	0.0417	0.9989
	2	0.5375	0.5486	0.5366	0.8199	0.3263	0.8589	0.8311	0.236	0.878
	3	0.455	0.5556	0.4469	0.7794	0.396	0.8096	0.7854	0.326	0.8217

Table 3. Number of mis-Classified Residues in the Train Dataset

Kernel fn Total no of	Linear	RBF with sigma equals		
		1	2	3
FP+FN(from 1968)	972	62	44	60
FN (from 984)	486	62	24	34
FP (from 984)	486	0	20	26

Table 4. Number of mis-Classified Residues in the Test Dataset

		Assumption 1			Assumption 2			Assumption 3		
		FP+FN (From 1972)	FN (From 144)	FP (From 1828)	FP+FN (From 1972)	FN (From 144)	FP (From 1828)	FP+FN (From 1972)	FN (From 144)	FP (From 1828)
Linear		1216	57	1159	1036	71	965	1033	71	962
RBF	1	128	126	2	140	138	2	140	138	2
	2	912	65	847	355	97	258	333	110	223
	3	1075	64	1011	435	87	348	423	97	326

6. CONCLUSION

The paper at hand proposed a classification model with an objective to distinguish between epitope and non-epitope residues. The model used a benchmark structure dataset “DiscoTope” which is further divided into train and test datasets. The train dataset is the only used during the design procedure. Five different features are extracted from the 61 train antigen-antibody complexes using their files which are downloaded from the Protein Data Bank. We proposed to use a k-means clustering technique to choose a proper informative number of residues. The total chosen residues along with their features values are represented using the structure window neighborhood criteria. A powerful classifier “SVM” was employed with three different kernel functions; linear, RBF and polynomial. While using the polynomial kernel with

SVM, it was found that it consumed too much time and its performance wasn't comparable with the other kernels. Also, the linear kernel function couldn't achieve good results as well. SVM with RBF was the best amongst the three. On the train dataset, the recorded measurements were very satisfactory leading to a minimum error rate equals 2.24% when using RBF sigma value equals 2. The sensitivity and specificity recorded at these parameters was also very high; 0.9756 and 0.9798 respectively. On the test dataset, assumption 2 accompanied by using RBF sigma value equals 1 results in a very high accuracy and specificity; 92.9% and 99.89% respectively.

REFERENCES

1. Sobolev, B.N., Olenina, L.V., Kolesanova, E.F., Poroikov, V.V., and Archakov, A.I., “Computer Design of Vaccines: Approaches, Software Tools and Informational Resources”, Current Computer-Aided Drug Design, Vol. 1, pp. 207-222, 2005.
2. Ponomarenko, J.V., and Regenmortel, M.H.V., “B-Cell Epitope Prediction”, in Structural Bioinformatics, John Wiley & Sons, 2009.
3. Wang, J.O., and Watanabe, T., “Antigen Presentation to Lymphocytes”, Encyclopedia of Life Sciences- Nature Publishing Group, pp. 1-5, 2001.
4. Schindler, L., Kerrigan, D., Kelly, J., and Hollen, B., “Understanding Cancer and Related Topics:Understanding the Immune System”, National Cancer Institute, 2005.
5. Beck, G., and Habicht, G. S., “Immunity and the Invertebrates”, Scientific American, Vol. 275, no. 5, pp. 60-66, 1996.
6. Kulkarni-Kale, U., Bhosle, S., and Kolaskar, A.S., “CEP: A Conformational Epitope Prediction Server”, Nucleic Acids Research, Vol. 33, pp. W168–W171, 2005.
7. Andersen, P.H., Nielsen, M., and Lund, O., “Prediction Of Residues In Discontinuous B-Cell Epitopes Using Protein 3D Structures”, Protein Science, Vol. 15, pp. 2558–2567, 2006.
8. Ponomarenko, J.V., and Bourne, P.E., “Antibody-Protein Interactions: Benchmark Datasets and Prediction Tools Evaluation”, BMC Structural Biology, Vol. 7:64, 2007.
9. Sweredoski, M.J., and Baldi, P., “PEPITO: Improved Discontinuous B-Cell Epitope Prediction Using Multiple Distance Thresholds and Half Sphere Exposure”, Bioinformatics, Vol. 24, no. 12, pp. 1459–1460, 2008.
10. Ponomarenko, J., Bui, H.H., Li, W., Fusseder, N., Bourne, P.E., Sette, A., and Peters, B., “ElliPro: A New Structure-Based Tool for the Prediction of Antibody Epitopes”, BMC Bioinformatics, Vol. 9, no. 514, 2008.
11. Ansari, H.R., and Raghava, G.P., “Identification of Conformational B-Cell Epitopes in an antigen from its primary sequence”, Immunome Research, Vol. 6, 2010.
12. Zhang, W., Xiong, Y., Zhao, M., Zou, H., Ye X., and Liu, J., “Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature”, *BMC* Bioinformatics, Vol. 12, no. 341, 2011.

13. Hassan, K.A., Badr, A., and Hegazy, A.-F., “On Predicting Conformational B-cell Epitopes: a Comparative Study and a New Model”, *American* Journal of Bioinformatics Research, Vol. 1, no. 1, pp. 6-17, 2011.
14. Rubinstein, N.D., Mayrose, I., Martz, E., and Pupko, T., “Epitopia: a Web-Server for Predicting B-Cell Epitopes”, BMC Bioinformatics, Vol. 10, no. 287, 2009.
15. Hubbard, S.J., and Thornton, J.M., “NACCESS computer program”, Department of Biochemistry and Molecular Biology, University College of London, UK, 1993.
16. Stroud, R.M., and Fauman, E.B., “Significance of Structural Changes in Proteins: Expected Errors in Refined Protein Structures”, Protein Sci., Vol. 4, pp. 2392-2404, 1995.
17. Creighton, and Thomas, H., “Proteins: Structures and Molecular Properties”, W. H. Freeman, San Francisco, 1993.
18. Garnier, J., Gibrat, J.-F, and Robson, B., “GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence”, Methods Enzymol, Vol. 266, no. 32, pp. 540-553, 1996.
19. MacQueen, J., “Some methods for classification and analysis of multivariate observations”, Proc. of the Fifth Berkeley Symp. On Math. Stat. and Prob., Vol. 1, pp. 281-296, 1967.
20. Noble, W.S, “What is a support vector machine?”, Nature Biotechnology, Vol. 24, pp. 1565-1567, 2006.