ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 11, May 2018

Application of Machine Learning to Immune Disease Prediction

Kuan-Hui Lin, Yuh-Jyh Hu
College of Computer Science, National Chiao Tung University, Hsinchu, Taiwan

Abstract*—The intrusion of viruses, germs or parasites can trigger the immune system to protect our body from the harms done by so-called immunogens. However, these protein antigens can sometimes disable our immune activities and cause immune diseases. Common immune diseases include allergies, autoimmune diseases, and infectious diseases. Recently due to environmental changes, the number of cases of immune diseases has been increasing dramatically. They sometimes take months for the patients to fully recover, or even take lives when the situation gets worse. Therefore, an early accurate prediction of immune diseases can provide valuable information for preventive medicine. Previous studies for the most part focused on the diseases caused by allergens, and thus lacked the analysis of other immune diseases such as autoimmune diseases and infectious diseases. To fill the gap, we applied machine learning techniques to construct accurate classification models for three types of immune diseases, allergy, autoimmune disease and infectious disease, caused by different protein antigens. This study consists of three stages: (a) collected and processed antigen data related to immune diseases, including allergy, autoimmune disease, and infectious disease, (b) analyzed the properties of these protein antigens at the sequence level and the structural level to select and develop new features for classification modeling, and (c) demonstrated the application of machine learning to build classification models for immune disease prediction.

***Index Terms*—immune diseases, immunogen, antigen, B-cell epitopes.**

I. INTRODUCTION

It has been reported that allergic diseases can affect a considerable portion of the general public. For example, asthma and eczema, respectively, affect 10% and 15% of the children in some countries [1]; seafood allergy and general food allergy, respectively, were reported in 2.3% [2] and 4% [3] of the US population. Though up to one third of the human population are affected by one or more of allergic diseases [4], common protein antigens can also cause other immune diseases, such as autoimmune diseases and infectious diseases, in addition to allergies. Despite many efforts into research on immune diseases, most works of antigen prediction were focused on allergies [5-7], and lacked the studies of other immune diseases.

The immune system normally guards against antigens like bacteria and viruses, but in an autoimmune disease, the immune system mistakenly attacks our own body. Though what causes the immune system misfire is still unclear, yet the dramatic change in the climate and living environments in recent years is arguably one primary factor to increase the number and variety of autoimmune diseases. In addition, as a result of globalization and advance of transportation technology, traveling between regions around the world has become more frequent and available; nevertheless, it also increases the risk of spreading infectious diseases from one person to another, and can eventually cause severe pandemics. Unlike most previous works that mainly focused on allergens, our study expanded the research on protein antigens by including other immunogens that may cause other immune diseases such as autoimmune diseases and infectious diseases. We aimed to evaluate the feasibility and potential of machine learning for protein antigen classification according to the immune diseases they cause. Not only can an efficient and accurate computational prediction method save human labor in wet lab tests, it can also serve as a screening tool in medical diagnosis.

II. MATERIALS AND METHODS

In this study, we concentrated on three types of immune diseases: (a) allergies, (b) autoimmune diseases, and (c) infectious diseases. We compared the representative machine learning algorithms respectively from three different categories, namely decision tree learning, instance-based learning, and maximum margin learning.

A. Data Preparation

One of the objectives of this study is to investigate the correlations between the classifications of immunogens and their physicochemical properties. We first collected the information of immunogens from IEDB (Immune Epitope Database http://www.iedb.org/) based on disease states, B-cell responses, locations of discontinuous epitopes, hosts, source organisms, etc. We further divided the immunogens into two classes according to the availability of their 3D structures in PDB (Protein Data Bank https://www.rcsb.org/pdb/home/home.do), and preprocessed the data differently for later analysis. We present the flowchart of data collection in Fig. 1, and show the summary of raw immunogen data in Table I.

Each data item collected from IEDB is indexed by four IDs: B-cell ID, Epitope ID, Source Accession Number and PDB ID (if it has a known 3D structure). Multiple B-cell IDs can map to the same Epitope IDs, and multiple Epitope IDs can map to the same Source Accession Number or the same PDB IDs. The Source Accession Numbers and the PDB IDs are unique and can be mapped to the NCBI or the PDB to obtain the protein sequences or the 3D structures. Consequently we preprocessed the raw data by removing

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 11, May 2018

redundancies such as those epitopes with duplicate Epitope IDs or Source Accession Numbers, and then merging different immunogens when they are mapped to the same Source Accession Numbers or to the same PDB IDs. We show the data preprocess flowcharts for immunogen data with and without 3D structures in Fig. 2, respectively.

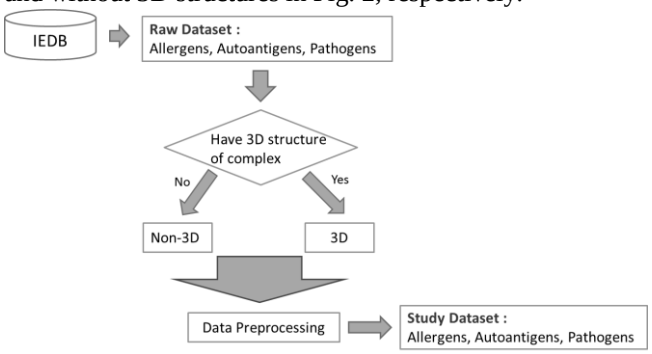


Fig. 1. Flowchart of data collection

TABLE I. SUMMARY OF RAW IMMUNOGEN DATA

Data \ Disease		Allergy	Autoimmune	Infectious
Positive B-cell Response Assays	3D	2	3	64
	Non-3D	141	134	746
Total	3D+Non-3D	143	137	810

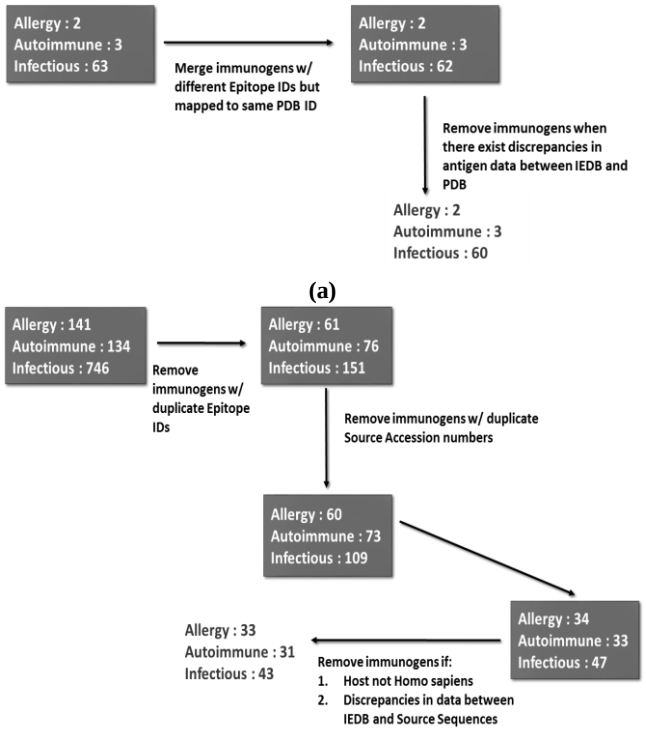


Fig.2. Control flows of data preprocesses. (a) Preprocess of immunogen data with 3D structures, (b) Preprocess of immunogen data without 3D structures

After the data preprocesses, we obtained the immunogens for further studies. We summarized the immunogen data in Table II.

TABLE II. SUMMARY OF IMMUNOGEN DATA AFTER PREPROCESSES

Disease	w/ 3D structures	w/o 3D structures	Total
Allergic	2	33	**35**
Autoimmune	3	31	**34**
Infectious	60	43	**103**
Total	65	107	**172**

B. Physicochemical Features of Amino Acids on Immunogens

We analyzed six physicochemical properties as the base features to represent each amino acid on an immunogen in this study. They are: (1) information per position in PSSM (Position Specific Scoring Matrix), (2) side chain polarity, (3) hydropathy index, (d) antigenic propensity, (e) flexibility, and (f) hydrophilic scale. We briefly describe each base feature as follows.

PSSM Information per position

We used PSI-BLAST [8] to search a non-redundant protein database and produced the PSSM profile, from which we obtained the information per position for each amino acid on the immunogen sequences in the study. We show the sample PSSM profile in Fig. 3. The column framed by the red line presents the information per position on an immunogen sequence, and it is used as a base feature.

Fig. 3. A sample PSSM profile

Side chain polarity

According [9], we divided amino acids into four categories: (a) Polar, (b) Basic Polar, (c) Acidic Polar, and (d) Non-polar.

Hydropathy index

Kyte and Doolittle [10] devised the hydopathy index by applying a sliding-window strategy that continuously determined the average hydopathy in a window as it advanced through the sequence. Table III shows the standard amino acids and their side chain polarities and hydropathy indices.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 11, May 2018

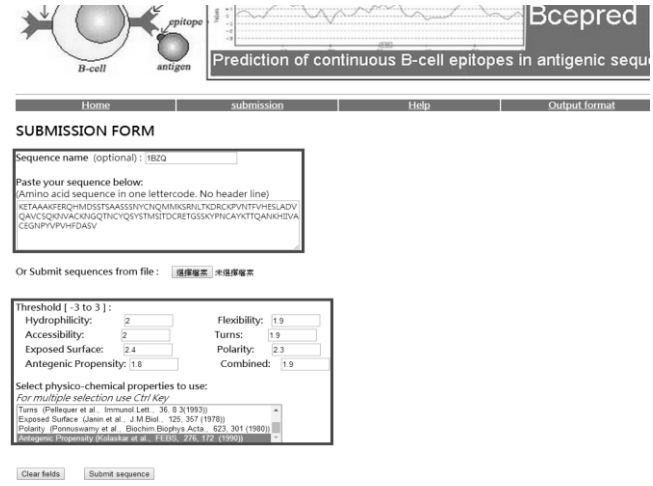
TABLE III. AMINO ACIDS AND PROPERTIES

Amino Acid	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Side Chain Polarity*	NP	BP	P	AP	NP	AP	P	NP	BP	NP	NP	BP	NP	NP	NP	P	P	NP	P	NP
Hydropathy Index	1.8	-4.5	-3.5	-3.5	2.5	-3.5	-3.5	-0.4	-3.2	4.5	3.8	-3.9	1.9	2.8	-1.6	-0.8	-0.7	-0.9	-1.3	4.2

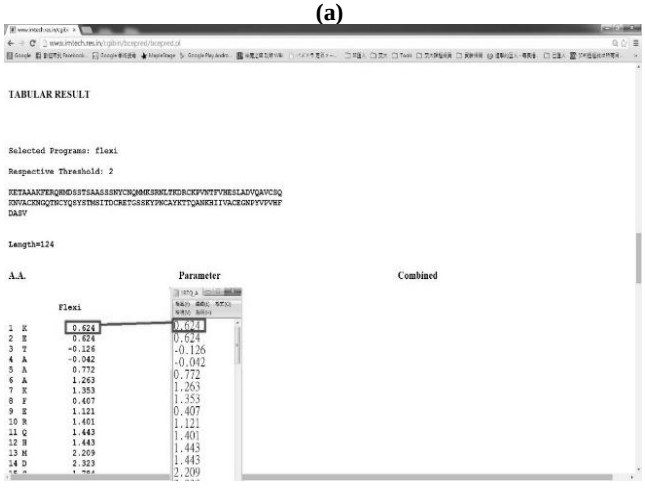
*NP: non-polar; BP: basic polar; P: polar; AP: acidic polar

Antigenic propensity

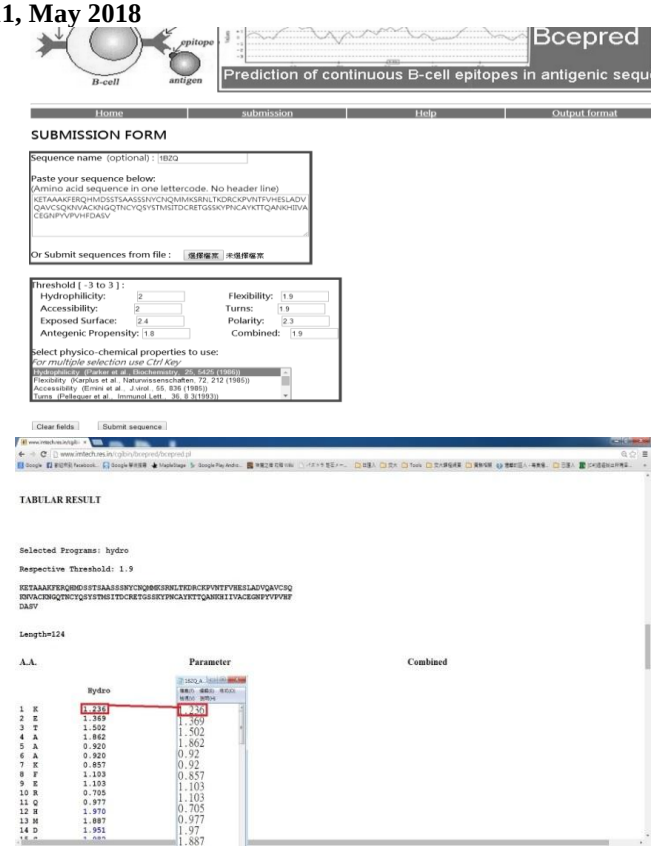
Kolaskar and Tongaonkar [11] analyzed 156 antigenic determinants (<20 residues per determinant) in 34 different proteins to compute the antigenic propensities of amino acid residues.



(a)



(b)



(c)

Fig. 4. Sample output of BcePred. (a) Antigenic propensity, (b) flexibility, and (c) hydrophilic scale.

Flexibility

Karplus and Schulz [12] developed the flexibility scale based on the mobility of the protein segments on 31 proteins with known structures.

Hydrophilic scale

Parker et al. [13] developed the hydrophilic scale based on the high-performance liquid chromatography (HPLC) peptide retention data.

We obtained the scores of antigenic propensity, flexibility and hydrophilic scale, respectively, using BcePred [14]. We show the sample output of BcePred in Fig. 4.

C. Immunogen Data Representation and Learning Algorithms

The goal of this study is to develop the classification models to predict the type of immune diseases caused by immunogens. The learning algorithms evaluated in the experiments were C4.5 [15], k-nearest-neighbor (k-NN)[16], and support vector machines (SVM) [17], each of which is considered as a representative *per se* in its own paradigm, namely decision tree learning, instance-based learning, and maximum margin learning. The predictive performance of a learning algorithm strongly depends on the choice of data representations because different representations convey various amount of information [18]. The expressiveness of an oversimplified representation is limited, and it consequently constrains the learning ability of a learning algorithm. By contrast, an overcomplicated representation can cause a

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 11, May 2018

learning algorithm to over fit the data and produce poor predictions.

To train a classification model for immune disease prediction from a set of immunogens, using different learning algorithms, we defined an immunogen representation based on the aforementioned six base features. Note that the immunogen protein sequences in a training data set may not be of the same length while it is necessary to ensure each immunogen to be represented in the same vector form, so the data can be a legal input to the learning algorithms for training. To resolve the issue of different immunogen lengths, we first took the average of each base feature values for each amino acid on an immunogen, and then combined all the averages into a 6-element vector to represent this immunogen. Take an immunogen of L amino acids for example, of which the base feature values are listed below,

Site	F1	F2	F3	F4	F5	F6
1	$F1_1$	$F2_1$	$F3_1$	$F4_1$	$F5_1$	$F6_1$
2	$F1_2$	$F2_2$	$F3_2$	$F4_2$	$F5_2$	$F6_2$
3	$F1_3$	$F2_3$	$F3_3$	$F4_3$	$F5_3$	$F6_3$
…	…	…	…	…	…	…
L-1	$F1_{L-1}$	$F2_{L-1}$	$F3_{L-1}$	$F4_{L-1}$	$F5_{L-1}$	$F6_{L-1}$
L	$F1_L$	$F2_L$	$F3_L$	$F4_L$	$F5_L$	$F6_L$

L is the length of the immunogen. F1~F6 are the base features. *F* in italic denotes a feature value. Take the averages of F1~F6, denoted by $F1_{avg}$~$F6_{avg}$. We represent this immunogen by the vector < $F1_{avg}$, $F2_{avg}$, $F3_{avg}$, $F4_{avg}$, $F5_{avg}$, $F6_{avg}$>.

As a result, we represented each immunogen in the data set by a vector of six average feature values. In addition, to mitigate the effects of variance among these feature averages, we standardized each element in the vectors by taking its z-score, respectively. Each standardized vector was used as a training example to train a classification model, using C4.5, k-NN or SVM.

III. RESULTS

We adopted the LOOCV (Leave-One-Out Cross-Validation) to evaluate the predictive performance because the total of immunogens in the experiments was relatively small (172 immunogens). We used F-score and percentage accuracy as the performance measures. We show the results in Table 4. We marked the highest score among the three algorithms in boldface. From Table IV we noticed that k-NN (k=1) outperformed C4.5 and SVM markedly for autoimmune disease classification. By contrast, the difference in the predictive performances for the other two immune diseases, allergic and infectious, was modest among the three learning algorithms.

We also conducted an ablation test on the effects of the base features on the predictive performances of C4.5, k-NN and SVM. After the removal of one base feature at a time, we re-ran the LOOCV to evaluate the F-score and the percentage accuracy. We present the results in Tables V and VI. For each learning algorithm, we highlighted the maximum increase in performance in boldface, and underlined the maximum decrease after the removal of one base feature. For example, C4.5 had the maximum increase in F-score after PSSM was removed, and had the maximum decrease in F-score if hydrophilic scale was not considered.

TABLE IV. RESULTS OF IMMUNE DISEASE CLASSIFICATION

Immune Disease	C4.5	k-NN	SVM
Allergy F-score	0.568	0.559	**0.576**
Autoimmune F-score	0.481	**0.657**	0.400
Infectious F-score	0.815	**0.833**	0.798
Average F-score	*0.621*	***0.683***	*0.591*
Accuracy	*0.709*	***0.744***	*0.698*

TABLE V. RESULTS OF ABLATION TESTS FOR F-SCORE

Ablation Test	C4.5	k-NN	SVM
all 6 base features	*0.621*	*0.683*	*0.591*
w/o PSSM	**0.650**	0.690	0.613
w/o side chain polarity	0.621	0.692	0.566
w/o hydropathy index	0.586	**0.696**	0.586
w/o antigenic propensity	0.641	0.690	0.491
w/o flexibility	0.520	0.671	0.456
w/o hydrophilic scale	0.515	0.657	**0.636**

TABLE VI. RESULTS OF ABLATION TESTS FOR ACCURACY

Ablation Test	C4.5	k-NN	SVM
all 6 base features	*0.709*	*0.744*	*0.698*
w/o PSSM	**0.727**	0.750	0.715
w/o side chain polarity	0.709	0.680	0.686
w/o hydropathy index	0.680	0.744	0.703
w/o antigenic propensity	0.703	**0.756**	0.657
w/o flexibility	0.651	0.738	0.622
w/o hydrophilic scale	0.640	0.733	**0.727**

These findings indicate that different base features had different effects on predictive performances. What is worth notice from Tables 5 and 6 is that k-NN was the least sensitive to the removal of base features, compared with C4.5 and SVM.

IV. CONCLUSION

To our best knowledge, this is the first study on machine learning for the prediction of immune disease caused by particular protein antigens (i.e. immunogens). We have demonstrated the feasibility and potential of machine learning for immunogen classification though there is still room for improvement. We intend to investigate alternative immunogen representations and evaluate other learning algorithms in the future work.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 11, May 2018

REFERENCES

[1] W. Cookson, "The immunogenetics of asthma and eczema: a new focus on the epithelium," Nat. Rev. Immunol. vol. 4, pp. 978-988, 2004.

[2] S. Sicherer, A. Munoz-Furlong, H. Sampson, "Prevalence of seafood allergy in the United States determined by a random telephone survey," J. Allergy Clin. Immunol. vol. 114, pp. 159-165, 2004.

[3] H. Sampson, "Update on food allergy," J. Allergy Clin. Immunol. vol. 113, pp. 805-819, 2004.

[4] S. Ono, "Molecular genetics of allergic diseases," Ann. Rev. Immunol. vol. 18, pp. 347-366, 2000.

[5] L. Zhang, Y. Huang, Z. Zou, Y. He, X. Chen, A. Tao, "SORTALLER: predicting allergens using substantially optimized algorithm on allergen family featured peptides," Bioinformatics, vol. 18(16), pp. 2178-2179, 2012.

[6] H. X. Dang, C. B. Lawrence, "Allerdictor: fast allergen prediction using text classification techniques," Bioinformatics, vol. 30(8), pp. 1120-1128, 2014.

[7] A. M. Barrio, D. Soeria-Atmadja, A. Nister, M. G. Gustafsson, U. Hammerling, E. Bongcam-Rudloff, "EVALLER: a web server for in silico assessment of potential protein allergen city," Nucleic Acid Research, vol. 35, pp. 694-700, 2007.

[8] Z. Zhang, A. A. Schäffer, W. Miller, T. L. Madden, D. J. Lipman, E. V. Koonin, S. F. Altschul, "Protein sequence similarity searches using patterns as seeds," Nucleic Acids Res 1998, vol. 26(17), pp. 3986-3990, 1998.

[9] R. E. Hausman, G. M. Cooper, The cell: a molecular approach, Washington, D.C: ASM Press, 2003.

[10] J. Kyte, R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," J Mol Biol, vol. 157(1), pp. 105-132, 1982.

[11] A. S. Kolaskar, P. C. Tongaonkar, "A semi-empirical method for prediction of antigenic determinants on protein antigens," FEBS Lett, vol. 276(1-2), pp. 172-174, 1990.

[12] P. A. Karplus, G. E. Schulz, "Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen," Naturwissenschaften, vol. 72, pp. 212-213, 1985.

[13] J. M. Parker, D. Guo, R. S. Hodges, "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, vol. 25(19), pp. 5425-5432, 1986.

[14] S. Saha, G. P. S. Raghava, "BcePred: Prediction of Continuous B-Cell Epitopes in Antigenic Sequences Using Physico-chemical Properties," ICARIS, LNCS 3239 Springer, pp. 197-204, 2004.

[15] J. R. Quinlan, C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1993.

[16] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd edn. New York: Wiley, 2003.

[17] C. C. Chang, C. J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intelligent Systems and Technology, vol. 2(3), pp. 1-27, 2011.

[18] T. Mitchell, "Generalization as search," Artif Intell, vol. 18, pp. 203-26, 1982.

AUTHOR BIOGRAPHY

Kuan-Hui Lin received her MS in Computer Science at National Chiao Tung University, Hsinchu, Taiwan. Her research interests include machine learning, data mining, and bioinformatics.

Yuh-Jyh Hu is a professor in the College of Computer Science, National Chiao Tung University, Hsinchu, Taiwan His research interests include machine learning, data mining, and biomedical informatics.