2014 The 2nd International Conference on Information Technology and Science (ICITS 2014)
IPCSIT vol. 59 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2014.V59.33

Prediction of SARS Coronavirus Main Protease by Support Vector Machine

Shinyoung Lee[+], Yoonjoo Kim, Jisue Kang, Jiwoo Oh, Jungwon Baek and Taeseon Yoon

Natural Science, Hankook Academy of Foreign Studies, Republic of Korea

Abstract. Replicase gene of SARS(Severe Acute Respiratory Syndrome) corona virus is surrounding the polyprotein*la* and *lab*. The polyprotein sequence includes functional proteins essential for the virus replication. Therefore, the inactivation of protease is one of the effective means to prevent the virus from replicating itself. We focused our research on corona virus main protease (CoVMpro).The experiment based on the "distorted key" theory was performed in order to inactivate it. The "distorted key" theory refers to the explanation stated below: substrate of the protein originally combines wit[1]h the active site of enzyme. However, when the scissile bond of peptide becomes modified with the strong hybrid peptide bond, it loses its cleavability. With the substrate strongly combined with the active site of the enzyme, these peptides become competitive inhibitors. The result of the experiment about the polypeptides based on this theory was reanalyzed by Support Vector Machine. (SVM)

Keywords: Support Vector Machine, SARS virus, CoVMpro, kernel function

1. Introduction

1.1. Background

Severe acute respiratory syndrome (SARS) is fatal to both humans and animals. SARS corona virus is a pathogen of the disease, and is classified as a single-strand RNA virus of a zoonotic origin. From November 2011 to July 2003, total 8,273 cases and 775 deaths(9.6% fatality rate)[1] was reported from a variety of countries with the majority of cases in Hong Kong. After the exposure to this virus, initial symptom reported was fever above 38℃ at first, and shortness of breath came up after. SARS corona virus main proteinase(CoVMpro) is an enzyme that participate in the replication process of the virus through the replicasepolyprotein. Considering its catastrophic harm to human being, we decided to do a research for CoVMpro. This research would be a key step for producing drugs against SARS, according to its position as a culprit of SARS presence. We performed an experiment using neural network (NN) on the basis of "distorted key theory", to analyze cleavage sites and improve the prediction accuracy of NN.

1.2. Distorted Key Theory

In developing inhibitors against SARS Distorted Key Theory is widely employed. Protein cleavage sites of CoV protease is very helpful for creating effective inhibitors against work of enzyme[2]. According to Koshland's induced fit theory and Fisher's lock-and-key model, the essential condition for CoVMpro cleaving a peptide with high possibility is binding the active site of enzyme and the substrate. Whereas, although peptide performs an adaption of its scissile bond to strong hybrid chemical bond, it can maintain the binding to the active site and reduce the whole cleavability extremely. This effect is explained by the Distorted Key Theory, summarized in Figure 1. Comparing "the Distorted Key" to the molecule, it can enter the lock but cannot unlock or come out from it. This results in strong binding due to altered peptide to

[+] Corresponding author. Tel.: + 82 10-9319-7048; fax: +82 31-332-0042
E-mail address: leeshinyoung309@gmail.com.

become a competitive inhibitor against the CoVMpro. The altered peptide can perform an efficient inhibitor against SARS. The theory also provides key information to predict the cleavage sites of CoVMpro and the binding groups.

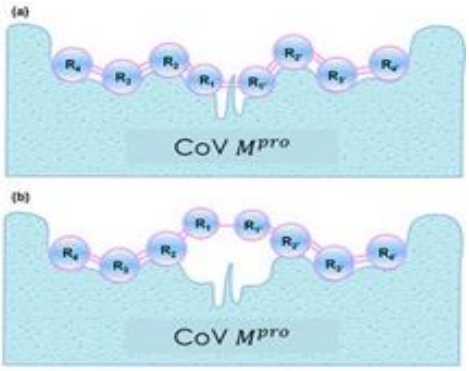

Fig. 1: illustration of distorted key theory. Plate (a) shows CoVMpro cleaves the peptide and binds the peptide with the active site of the protease, while the peptide in plate (b) is not cleavable through modification but still bound to the active site.

2. Related Research

2.1. Support Vector Machine

Support Vector Machine (SVM)is a supervised learning model[3], based on learning algorithms which are supplementary to recognizing patterns, analyzing data, and classifying it. The algorithm builds a model that allocates training examples into one of two categories, as training examples are given in order to train the model in a repetitive way. Specifically, a number of data sets are randomly divided into training data and test data (Fig. 2), but the ratio of the numbers in the two data groups should remain steady. Training data, also the group with larger portion of data is used to teach the analyzing ability to the SVM. After the process, SVM analyzes the test data with the previously obtained ability.

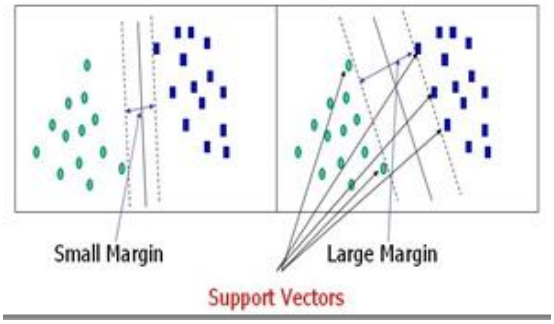

Fig. 2: It is illustration of SVM margin dividing the data. Rather than other computational programs SVM has larger margin which helps to increase the accuracy.

A systematically mapped SVM model represents examples as points in space, which are divided by a distinct gap with maximum margin(Figure 3). Then the new examples are inserted into that same space, and are predicted based on the model. The model efficiently maps the inputs into hyper planes, using the method known as kernel trick[4][5].

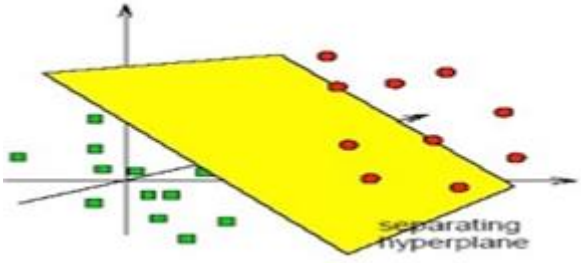

Fig. 3: Shows separating hyperplane of SVM. It sets the complicated boundaries to divide vectors.

2.2. Kernel Function

SVM implements prediction of new data by learning how to classify two types of data. It could not perform efficiently in non-linear classification though it applies soft margin. But it could utilize mapping method using kernel function in SVM, SVM starts to its effectiveness in non-linear problems[6][7]. Mapping using kernels translocates non-linear problem which is hard to divide in the input space where data is actually located to the feature space, the higher dimensional space. It carries out the linear discrimination analysis in this new space and achieves effect as it solves the complicated non-linear discrimination analysis problem in the first input place.

Normal kernel, sigmoid kernel, radical basis function kernel and polynomial kernel are commonly used in the SVM. There are parameters assisting optimization in each kernel[8]. Generally, there is no method to notify best parameter, so we have to find out the condition that shows optimized prediction rate by changing every condition and iterating SVM' learning and prediction[9]. We used four models of SVM in the process to analyze the shape of SARS corona virus: Normal, Sigmoid, RBF, and Polynomial.

2.2.1. Normal Kernel

Normal kernel is a linear classifier, which is the simplest one of all kernels.

$$k(x, y) = x^T y + c$$

Normal kernel is comprised of inner product <x,y> and an optional constant c.

2.2.2. Sigmoid Kernel

Sigmoid kernel, also known as hyperbolic tangent kernel and multilayer perceptron (MLP) kernel has its origin in the Neural Networks field.

$$k(x, y) = \tanh(ax^t y + c)$$

Sigmoid kernel has two adjustable parameters, the slope alpha and the intercept constant c. N means the data dimension, and a common value for alpha is 1/N.

2.2.3. RBF kernel

Radius basis function (RBF) is also a kernel function used in SVMs. The Gaussian kernel is an example of radial basis function kernel.

$$k(x, y) = \exp(-\frac{||x - y||^2}{2\sigma^2})$$

As an alternative, it could also be carried out using

$$k(x, y) = \exp(-\gamma\, ||\mathrm{x} - \mathrm{y}||^2)$$

The adjustable parameter sigma plays a significant role in the performance of this kernel, and should be prudently managed to the problem.

2.2.4. Polynomial kernel

Polynomial kernel is commonly used in data mining with SVMs and other kernelizing models. It is known as a non-stationary kernel. It allows learning for non-linear models and determining similarities of vectors (training samples) in feature. These kernels are commensurate with problems where all the training data are normalized.

$$k(x, y) = (\alpha x^T y + c)^d$$

Adjustable parameters are the slope alpha, the constant term c and the degree d polynomial.

3. Method and Experiment

3.1. Experiment Design

SARS virus amino acid data is encoded in binary form, +1 is positive one and -1 is negative one[10]. There are 154 positive data, 146 negative data and total 300 data. To take 10-fold cross validation, we broke data into 10 sets and trained on 9 datasets and test on 1. Each datasets are randomly chosen. Then we repeated 10 times on the dataset and took a mean accuracy, precision and recall. This task is practiced on normal kernel, sigmoid kernel, RBF kernel and polynomial kernel. SARS virus amino acid data is encoded in binary form, +1 is positive one and -1 is negative one[10]. There are 154 positive data, 146 negative data

and total 300 data. To take 10-fold cross validation, we broke data into 10 sets and trained on 9 datasets and test on 1. Each datasets are randomly chosen. Then we repeated 10 times on the dataset and took a mean accuracy, precision and recall. This task is practiced on normal kernel, sigmoid kernel, RBF kernel and polynomial kernel.

3.2. Experiment Result

3.2.1. Normal

	1	2	3	4	5	6	7	8	9	10	Average
Accuracy(%)	58.06	48.39	51.61	50.00	51.61	50.00	54.84	50.00	48.39	51.61	51.45
Precision(%)	55.17	50.00	51.72	50.00	51.72	50.00	53.57	50.00	50.00	51.62	51.38
Recall(%)	100.00	93.75	93.75	100.00	93.75	93.75	93.75	100.00	93.75	100.00	96.25

The table above shows the result of data analysis based on Normal kernel. According to the table, the average is 51.451% in accuracy, 51.379% in precision, 96.25% in recall. While all three values are measured higher than sigmoid analysis, they are still lower than polynomial analysis.

3.2.2. Sigmoid

	1	2	3	4	5	6	7	8	9	10	Average
Accuracy(%)	48.39	48.39	48.39	50.00	48.39	50.00	48.39	50.00	48.39	48.39	48.87
Precision(%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.00	0.00	0.00	5.00
Recall(%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	10.00

The table shows the result of data analysis based on Sigmoid kernel. From the table, the average is 48.81% in accuracy, 5.00% in precision, 10.00% in recall. All three values are measured lowest among all types of analyses.

3.2.3. RBF

	1	2	3	4	5	6	7	8	9	10	Average
Accuracy(%)	87.10	83.87	90.32	100.00	77.42	87.50	80.65	50.00	80.65	90.32	82.78
Precision(%)	100	100.00	100.00	100.00	100.00	100.00	100.00	50.00	100.00	84.21	93.00
Recall(%)	75.00	68.75	81.25	100.00	56.25	75.00	62.50	100.00	62.50	100.00	78.13

The table shows the result of data analysis based on RBF kernel. It can be noted that the average rate is 82.78% in accuracy, 93% in precision, and 78.13% in recall. Each rate has evidently increased from the previous analysis, yet it is still quite lower than the rate based on Polynomial.

3.2.4. Polynomial

	1	2	3	4	5	6	7	8	9	10	Average
Accuracy(%)	100.00	96.77	96.77	100.00	96.77	96.88	96.77	100.00	96.77	93.55	97.43
Precision(%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	88.89	98.89
Recall(%)	100.00	93.75	93.75	100.00	93.75	93.75	93.75	100.00	93.75	100.00	96.25

The table shows the result of data analysis based on Polynomial kernel: average rate is 97.43% in accuracy, 98.89% in precision, and 96.25% in recall. Each average rate is higher than in any other types of kernel models. The result that these rates attain the highest number suggests that the 3-dimentional structure of SARS CoV corresponds to the polynomial model. That is, the non-linear features rule the overall shape of SARS CoV.

4. Conclusion

From the 4 tables above, the following conclusion is drawn.

	Normal	Sigmoid	RBF	Polynomial
Accuracy Average(%)	51.45	48.87	82.78	97.43
Ranking	3	4	2	1

The accuracy average is highest in polynomial kernel. Thus, the shape of SARS Corona virus is similar with the shape of polynomial kernel, non-linear form.

	1	2	3	4	5	6	7	8	9	10
highest prediction value order	9	16	4	11	14	15	1	1	5	7
amino acid sequence	V-V-L-Q-S-K-G-H	S-Q-F-Q-S-K-L-T	V-K-L-Q-N-N-E-L	S-T-L-Q-A-G-L-R	T-V-L-Q-A-A-G-L	V-K-L-Q-N-N-E-I	A-V-L-Q-S-G-F-R	V-V-L-Q-S-K-G-H	V-R-L-Q-A-G-N-A	S-Q-F-Q-S-K-L-T

Furthermore, the probable amino acid sequence of virus can be perceived by prediction data. According to the polynomial prediction data, below fact is come out. Thus, the most probable SARS virus amino acid sequence is Valine-Valine-Leucine-Glutamine-Serine-Glycine-Leucine-Arginine.Find out which kernel function has the highest accuracy, and consider the kernel's prediction value. Then, the encoded amino acid number can be easily known. Consequently, we can realize the SARS virus's amino acid sequence. Given the fact that the SARS corona virus had brought severe consequences to the world in the past, researchers performed a number of experiments based on "distorted key theory" in order to find the effective means to prevent the virus from activating. We, as the authors of the work on the paper, reanalyzed the results of the experiment with Support Vector Machine(SVM). The models of SVM used to execute the analysis were 4 types of kernel function; Normal, Sigmoid, RBF, and Polynomial. We found out the shape and the amino acid sequence of the virus after the performance of the analysis and percentage measurement of accuracy, precision, and recall, with the existing data of SARS virus amino acid. And unfortunately, the number of datasets is not enough. On the following study, we should get sufficient data and refine them more prudently.

5. References

[1]. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. World Health Organization (WHO).

[2]. Chou K.C. (1996). "Review: Prediction of human immunodeficiency virus protease cleavage sites in proteins". Analytical Biochemistry 233: 1–14.doi:10.1006/abio.1996.0001. PMID 8789141.

[3]. Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, 1995. HTTP://WWW.SPRINGERLINK.COM/CONTENT/K238JX04HM88J80G/

[4]. Aizerman, Mark A.; Braverman, Emmanuel M.; and Rozonoer, Lev I. (1964). "Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control 25: 821–837.

[5]. Boser, Bernhard E.; Guyon, Isabelle M.; and Vapnik, Vladimir N.; A training algorithm for optimal margin classifiers. In Haussler, David (editor); 5th Annual ACM Workshop on COLT, pages 144–152, Pittsburgh, PA, 1992. ACM Press

[6]. Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, B. P. (2007). "Section 16.5.Support Vector Machines". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8.

[7]. Crammer, Koby; and Singer, Yoram (2001). "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines". J. of Machine Learning Research 2: 265–292.

[8]. Joachims, Thorsten; "Transductive Inference for Text Classification using Support Vector Machines", Proceedings of the 1999 International Conference on Machine Learning (ICML 1999), pp. 200-209.

[9]. Hsu, Chih-Wei; Chang, Chih-Chung; and Lin, Chih-Jen (2003). A Practical Guide to Support Vector Classification (Technical report). Department of Computer Science and Information Engineering, National Taiwan University.

[10]. Duan, Kai-Bo; and Keerthi, S. Sathiya (2005). "Which Is the Best Multiclass SVM Method? An Empirical Study".Proceedings of the Sixth International Workshop on Multiple Classifier Systems. Lecture Notes in Computer Science 3541: 278. doi:10.1007/11494683_28. ISBN 978-3-540-26306-7.