A Data Mining Approach for the Prediction of Hepatitis C Virus protease Cleavage Sites

Ahmed mohamed samir ali gamal eldin
Bio-inforamtics
Helwan University
Cairo, Egypt

Abstract— **Summary: Several papers have been published about the prediction of hepatitis C virus (HCV) polyprotein cleavage sites, using symbolic and non-symbolic machine learning techniques. The published papers achieved different Levels of prediction accuracy. the achieved results depends on the used technique and the availability of adequate and accurate HCV polyprotein sequences with known cleavage sites. We tried here to achieve more accurate prediction results, and more Informative knowledge about the HCV protein cleavage sites using Decision tree algorithm. There are several factors that can affect the overall prediction accuracy. One of the most important factors is the availably of acceptable and accurate HCV polyproteins sequences with known cleavage sites. We collected latest accurate data sets to build the prediction model. Also we collected another dataset for the model testing.**

Motivation: Hepatitis C virus is a global health problem affecting a significant portion of the world's population. The World Health Organization estimated that in1999; 170 million hepatitis C virus (HCV) carriers were present worldwide, with 3 to 4 million new cases per year. Several approaches have been performed to analyze HCV life cycle to find out the important factors of the viral replication process. HCV polyprotein processing by the viral protease has a vital role in the virus replication. The prediction of HCV protease cleavage sites can help the biologists in the design of suitable viral inhibitors.

Results: The ease to use and to understand of the decision tree enabled us to create simple prediction model. We used here the latest accurate viral datasets. Decision tree achieved here acceptable prediction accuracy results. Also it generated informative knowledge about the cleavage process itself. These results can help the researchers in the development of effective viral inhibitors. Using decision tree to predict HCV protein cleavage sites achieved high prediction accuracy.

Keywords-component; HCV polyprotein; decision tree; protease; decamers

I. Introduction

Hepatitis C virus (HCV) is a virus that infects liver cells and causes liver inflammation. It is a global disease with a worldwide expanding incidence and prevalence base. Hepatitis C virus presents supremely challenging problems in view of its adaptability and its pathogenic capacity. The strategies that HCV utilizes to parasitize its hosts make it formidable enemy. Therapeutic interventions need considerable sophistication to counter its progress. It is estimated that 3–4 million people are infected with HCV each year. Some 130–170 million people are chronically infected with HCV and at risk of developing liver cirrhosis

and/or liver cancer. More than 350 000 people die from HCV related liver diseases each year.

HCV infection is found worldwide. Countries with high rates of chronic infection are Egypt (22%), Pakistan (4.8%) and China (3.2%). these countries are attributed to unsafe injections using contaminated equipment. [1].

HCV protease cleavage sites are considered one of the most important inhibitor targets, cause of the cleavage of polyprotein Sequences plays an important role in the viral replication [2].

The prediction of the viral proteases cleavage sites will help in the development of suitable protease inhibitor. Several data mining techniques have been used in solving and analyzing several biological problems. One of the interesting problems is the analyzing of HCV life cycle, using Data mining techniques to find useful knowledge which may help the biologist to develop suitable HCV vaccine. Many data mining techniques have been used to analyze different viral proteases cleave sites. For example artificial neutral network has been used to predict both Human immunodeficiency virus (HIV) and HCV proteases cleavage sites and achieved high prediction accuracy [3-5]. Finding more accurate and simpler prediction model is considered a challenging point.

Decision tree is one of the most common data mining techniques. It has been used in analyzing and solving several classification problems. Decision tree has a great advantage which its ability to provide us with informative rules about the classification problem itself. The biologists and the researchers can use these rules to understand the cleavage

Process characteristic. In spite of that decision tree does not have prediction accuracies than the other classification techniques, but its ease to understand and also its informative rules make it an interesting method. Decision tree prediction results depends on the availability of the datasets which it will train the classification model. Decision tree has been used in the prediction of HCV protease cleavage sites, but it did not achieve an acceptable prediction accuracies cause of the lake of accurate cleaved sequences [6]. We tried

Here to collect and find more accurate HCV cleaved sequences to build a decision tree to predict the proteases cleavage sites.

II. System And Methods

A. Viral protease cleavage process

The cleavage process of the protein is look like the ‘Lock and key’ model where a sequence of amino acids fits as A key to the active site in the protease, which in the HCV protease Case is estimated to be ten residues long. The protease active site pockets are denoted by S (Schechter and Berger, 1967) [7].

S = S5, S4, S3, S2, S1, S1’, S2’, S3’, S4’, S5’

Corresponding to residues P in the peptide

P=P5, P4, P3, P2, P1, P1’, P2’, P3’, P4’, P5’

The scissile bond is located between positions P1 and P1’, and Pi can take on any one of the following 20 amino acid values {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}. There are 2010 possible values for string P. If the amino acids in P (the ‘key’) fit the positions in S (the ‘lock’), then the protease will cleave the decamer (ten amino acids) between positions P1 and P1’. The goal of the decision tree model to learn the ‘lock and key’ rules, from the available datasets.

B. Date representation

HCV protein sequences are represented as a long chain of letters. Each letter represents one amino acid. We interested in the10 amino acids where the HCV protease can cleave it. There is a poplar technique used by the previous researches to generate non-cleaved sequences [6]. It depends on considering the regions between known cleaved sequences as a non-cleaved.

C. Building the classification model

We used here one of the most common used classification algorithms which is the decision tree algorithm. We will summarize basic concepts of the decision trees and its advantages over the other classification methods. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of Decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to Decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome [8]. The following is a summary of the important characteristics of decision

- Decision tree induction is a nonparametric approach for building classification models. In other words, it does not require any prior assumptions regarding the type of probability distributions satisfied by the class and other attributes.
- Techniques developed for constructing decision trees are computationally inexpensive, making it possible to quickly construct models even when the training set size is very large. Furthermore, once a decision tree has been built, classifying a test record is extremely fast.
- Decision trees, especially smaller-sized trees, are relatively easy to interpret. The accuracies of the trees are also comparable to other Classification techniques for many simple data sets.
- Decision trees provide an expressive representation for learning discrete valued functions.
- Decision tree algorithms are quite robust to the presence of noise, especially when methods for avoiding overfitting.

For the HCV protein cleave sites prediction problem. We used the decision tree model with Gini index splitting rule [9]. Each sample in the training dataset was consisting of 11 items 10 items represent the amino acids where the protease can cleave it. The last item represents the class label of the amino acids sample. In our problem we have two classes cleavage ‘positive’ or non-cleavage ‘negative’.

D. Data collection

The process of collecting enough and accurate HCV cleaved decamers, is the core of our research. We searched a lot of the published papers that have discussed HCV polyprotein analysis. Also we contacted a lot of researchers interested in this area. The availability of the online protein databases provided us with some accurate and valid HCV polyprotein sequences for the training and testing our model. To generate more non- cleaved ‘negative’ sequences we used the technique which has been used by the previous researchers as we mentioned in the previous section.

There are several conflicts and uncertainties in the data which have been used in the previous published papers. We tried to found the most recent and accurate samples to build the prediction model. We used the last accurate datasets used [10-18] in previous work. The collected datasets are divided into two parts:

- Training dataset
- out of sample or testing dataset

We collected 939 decamers as training dataset 199 as cleaved ‘positive’ samples and 706 as non-cleaved ‘negative’ samples. We collected three out of samples dataset to the proposed model [19]:

- Four proteins from the TLR3 pathway were used for another test data set: IκB kinase ε (IKKε) [GenBank: AAC51216]; TRAF family member-associated NF-κB activator-binding kinase 1 (TBK1) [GenBank: NP_037386]; Toll-like receptor 3 (TLR3) [GenBank: NP_003256]; and Toll-IL-1 receptor domain containing adaptor inducing IFN-β (TRIF or TICAM-1) [GenBank: BAC55579].the four proteins created dataset contains 2806 samples of which two are reported as cleaved samples by HCV protease enzyme[20].
- There are 69 samples reported in vivo as cleaved samples[19].

We used the same datasets for training and testing which have been used by Thorsteinn Rögnvaldsson et al [19]. They collected a new datasets rather than the previous datasets used by the other researchers, which contains a lot of conflict and uncertainties as we mentioned before.

III. Results and discussion

We implemented the decision tree using classification and regression tree (CART) Mat lab toolbox with GINI index as spitting criteria. The training dataset was consisting of 939 samples.199 as cleaved sample and 740 as non-cleaved samples. Each sample was consisting of 10 amino acids where the HCV protease can cleave.

We used ten-fold cross validation to be able to evaluate the overall performance of the prediction model. For the training dataset we got Prediction accuracy 99 %. Also we got 98% as Sensitivity and Specificity as 99%. Table I show the confusion matrix for the training data.

After apply the ten-fold cross validation got overall accuracy 96% and we got Sensitivity is 95.5% and the model Specificity 98.6%. Table II shows the average achieved confusion matrix for the tenfold cross validation.

We applied our model on the out of samples dataset. For the first test set which is consist of 2806 (2 cleaved and the remaining are non-cleaved samples) sample. Our model successfully predicted one of the cleaved samples. But it got 89 as false positive or false cleaved samples.

For the 69 in vivo cleaved samples our model successfully predicted 59 of the 69 as cleaved samples.

Using the decision tree as a classification model has achieved an overall prediction accuracy 96% which can be considered as an acceptable results, if we compared the presented model with the other techniques that achieved the

Highest prediction accuracy like support vector machine (SVM) [5]. We can find that our results are comparable with SVM which achieved 97% as overall prediction accuracy.

The presented work is a try to achieve more accurate prediction accuracy using easy and simple classification technique like the decision.

IV. Conclusions and future work

The prediction of HCV polyproetin cleavage sites, using Decision tree, has achieved acceptable prediction accuracies. The achieved results are not the best, but the created rules by the decision tree prediction model make the achieved results more informative. In the future work we can add more factors like the amino acids secondary structure as training attribute to find out its effect in the overall prediction accuracy. Also we can enhance the decision tree prediction results by using the ensembles of decision tree technique which can enhance the prediction results of the proposed model.

TABLE I. THE CONFUSION MATRIX FOR THE TRAINING DATA

	Non cleavage	cleavage	Total
None cleavage	735	5	740
Cleavage	4	195	199
Total	739	200	939

TABLE II. THE CONFUSION MATRIX FOR 10-FOLD CROSS VALIDATION

	Non cleavage	cleavage	Total
None cleavage	730	10	740
Cleavage	9	190	199
Total	739	200	939

References

[1] World healt oragnization Media centre. “Hepatitis C ." http://www.who.int. 2011. 5 October 2011 http://www.who.int/mediacentre/factsheets/fs164/en/

[2] Sarah Welbourn and Arnim Pause,” The Hepatitis C Virus NS2/3 Protease,” Molecular Biology (2007), in press

[3] T. Rognvaldsson, Liwen You , “No Algorithm Beats the Simple Perceptron on HIV Protease Function Prediction,” unpublsihed .

[4] Thompson, T., Chou, K, and Zheng, C. , “Neural network prediction of the HIV-1protease cleavage sites". Journal of Theoretical Biology (1995)177, 369-379,” inpress .

[5] T Cai, Y.-D. and Chou, K.-C., “Artificial neural network model for predicting HIV protease cleavage sites in protein,” Advances in Engineering Software (1998) 29, 119-128 .

[6] Ajit Narayanan, Xikun Wu and Z. Rong Yang,” Mining viral protease data to extract cleavage knowledge,” Bioinformatics (2002) 18 (suppl1): S5-S13,In press

[7] T. Rognvaldsson, Liwen You , “Why neural networks should not be used for HIV-1 protease cleavage site prediction,” Bioinformatics (2004), in press .

[8] W. Peng, J. Chen and Haiping Zhou,” An Implementation of ID3 Decision Tree Learning Algorithm,” unpublished

[9] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984),”Classification and regression trees.” Wadsworth, Belmont. Refrence

[10] Jarman IH, Etchells TA, Martin JD, Lisboa PJ (2008) an integrated framework for risk profiling of breast cancer patients following surgery.
Artificial Intelligence in Medicine, 42:165-188

[11] Grakoui A, McCourt DW, Wychowski C, Feinstone SM, Rice CM: Characterization of the hepatitis C virus-encoded serine proteinase: determination of proteinase-dependent polyprotein cleavage sites. Journal of Virology 1993, 67:2832-2843.

[12] Leinbach SS, Bhat RA, Xia SM, Hum WT, Stauffer B, Davis AR, Hung PP, Mizutani S: Substrate specificity of the NS3 serine proteinase of hepatitis C virus as determined by mutagenesis at the S3/NS4A junction. Virology 1994, 204:163-169.

[13] Kolykhalov AA, Agapov EV, Rice CM: Specificity of the hepatitis C virus NS3 serine protease: effects of substitutions at the 3/ 4A, 4A/4B, 4B/5A, and 5A/5B cleavage sites on polyprotein processing. Journal of Virology 1994, 68:7525-7533.

[14] Bartenschlager R, Ahlborn-Laake L, Yasargil K, Mous J, Jacobsen H: Substrate determinants for cleavage in cis and in trans by the hepatitis C virus NS3 proteinase. Journal of Virology 1995, 69:198-205.

[15] Urbani A, Bianchi E, Narjes F, Tramontano A, Francesco RD, Steinkühler C, Pessi A: Substrate specificity of the hepatitis C virus serine protease (NS3). The Journal of Biological Chemistry 1997, 272:9204-9209.

[16] Zhang R, Durkin J, Windsor WT, McNemar C, Ramanathan L, Le HV: Probing the substrate specificity of hepatitis C virus NS3 serine protease by using synthetic peptides. Journal of Virology 1997, 71:6208-6213.

[17] Kwong AD, Kim JL, Rao G, Lipovsek D, Raybuck SA: Hepatitis C virus NS3/4A protease. Antiviral Research 1998, 40:1-18.

[18] Attwood MR, Bennett JM, Campbell AD, Canning GGM, Carr MG, Conway E, Dunsdon RM, Greening JR, Jones PS, Kay PB, Handa BK,

[19] Hurst DN, Jennings NS, Jordan S, Keech E, O'Brien MA, Overton HA, Wilkinson TCI, Wilson FX: The design and synthesis of potent inhibitors of hepatitis C virus NS3-4A proteinase. Antiviral Chemistry & Chemotherapy 1999, 10:259-273.

[20] T. Rögnvaldsson, T. A Etchells, L. You,"How to find simple and accurate rules for viral protease cleavage specificities," BMC Bioinformatics 2009, in press

[21] Li K, Foy E, Ferreon JC, Nakamura M, Ferreon ACM, Ikeda M, Ray SC," Immune evasion by hepatitis C virus NS3/4A protease-mediated cleavage of the Toll-like receptor 3 adaptor protein TRIF", .

[22] Proceedings of the National Academy of Sciences of the United States Of America 2005, in press.