TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 17, Số 6 (2020): 950-960
ISSN: 1859-3100
HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 17, No. 6 (2020): 950-960
Website: *http://journal.hcmue.edu.vn*

Research Article

DISCRIMINATIVE MOTIF FINDING TO PREDICT HCV TREATMENT OUTCOMES WITH A SEMI-SUPERVISED FEATURE SELECTION METHOD

Nguyen Thi Tuong Vy [1], Le Thi Nhan[2*]

[1]University of Lyon 1, Lyon, France

[2] University of Science, Vietnam National University Ho Chi Minh City, Vietnam

[*]Corresponding author: Le Thi Nhan – Email: ltnhan@fit.hcmus.edu.vn

Received: January 08, 2020; Revised: February 27, 2020; Accepted: June 02, 2020

ABSTRACT

Hepatitis C treatment is currently facing many challenges, such as high costs of medicines, side effects in patients, and low success rates with Hepatitis C Virus genotype 1b (HCV-1b). In order to identify what characteristics of HCV-1b cause drug resistance, many sequence analysis methods are conducted, and bio-markers helping to predict failure rates are also proposed. However, the results may be imprecise when these methods work with a dataset having a small number of labeled sequences and short length sequences. In this paper, we aim to predict outcomes of the HCV-b treatment and characterize the properties of HCV-b by using the combination of a feature selection and semi supervised learning. Our proposed framework improves the prediction accuracy about 5% to 8% in comparison with previous methods. In addition, we obtain a set of good discriminative subsequences that could be considered as biological signals for predicting a response or resistance to HCV-1b therapy.

Keywords: discriminative motif; hepatitis C virus; sequential forward floating selection; semi-supervised feature selection

1. Introduction

Hepatitis C disease is a kind of transmitted disease primarily caused by Hepatitis C Virus (HCV). This virus affects the liver, and after many years, it could lead to liver cirrhosis, or more serious problems including liver failure or liver cancer. According to World Health Organization (WHO), 71 million people worldwide are chronically infected with HCV and nearly 399,000 people die each year from cirrhosis and liver cancer. Antiviral medicines for chronic HCV are the combinations of pegylated interferon (PegIFN)-alpha and ribavirin (RBV) (Manns et al., 2001). In recent years, this therapy has also associated with the new class of drug, such as sofosbuvir (SOF), simeprevir (SIM), and daclatasvir (DCV) to reduce

Cite this article as: Nguyen Thi Tuong Vy, & Le Thi Nhan (2020). Discriminative motif finding to predict HCV treatment outcomes with a semi-supervised feature selection method. *Ho Chi Minh City University of Education Journal of Science, 17*(6), 950-960.

side effects and shorten the duration of treatment. However, the result of treatments often fails in almost half of cases, especially HCV genotype 1b (HCV-1b) (Gao et al., 2010). Therefore, knowing the sign of response or resistance, also known as sustained viral response (SVR) or non-sustained viral response (non-SVR), to the above drugs before the treatment is very important and necessary to alleviate distressing symptoms and expense for patients.

Several methods for characterizing sequences and discovering motifs were already developed such as position weight matrix (PWM) (Kim, & Choi, 2011; Bailey, Boden, Whitington, & Machanick, 2010), hidden Markov model (HMM) (Lin, Murphy, & Bar-Joseph, 2011), or association mining with domain knowledge (Vens, Rosso, & Danchin, 2011). With a general purpose of pattern searching, these studies showed their ineffectiveness in a case of short input sequences. Consequently, it is very difficult to get the highest probability of patterns when the dataset contains a small number of short and highly similar sequences.

In this paper, we approach the characterization and prediction problems by using a semi-supervised feature selection method. We are developing a framework which uses labeled and unlabeled data to select effective feature subsets. Our proposed framework predicts around 56% of accuracy on average, while 50% is of MEME (Bailey et al., 2010). The results of sequence characterization are promising discriminative motifs which provide physicians hints to understand thoroughly the resistance to IFN/RBV therapy of HCV-1b, as well as to get a better treatment for patients.

2. Background

2.1. Sequence characterization

Data characterization is a summarization of common features of objects in a target class of data in order to know properties of that class (Han, & Kamber, 2006). In the case of sequence characterization, we usually summarize to find certain patterns or motifs which can represent a class of sequence data. However, a discriminative motif is defined that it occurs frequently in one class of sequences and hardly occurs in the other classes of sequences. Therefore, these discriminative motifs help us to describe characteristics of a class and then classify a sequence into a certain class.

Motif discovery methods often use a string-based model or probabilistic-based model to represent discriminative motifs. In a string-based model, a motif is a short sequence of letters which are nucleic acids in DNA/RNA sequences or amino acids in protein sequences. Moreover, letters are special characters to increase the variability of the motif. A method represents this model is MERCI (Motif EmeRging and with Classes Identification) (Vens et al., 2011). It uses an idea of Apriori algorithm to generate candidate motifs during the sequential pattern mining. This method accepts or eliminates motifs based on two parameters which are the minimal frequency in a dataset and the maximal frequency in another dataset.

In a probabilistic-based model, a motif is represented by Position Weight Matrix (PWM) or Hidden Markov Model (HMM). PWM considers a motif as a matrix that each element is the probability of a given acid at a specified position with an independence assumption among positions. HMM describes a motif as a Markov process of hidden states where the probability of the current state of a letter only depends on its previous state with the assumption that these states are not necessarily independent (Wu, & Xie, 2010). A very popular tool to find disciminative motifs nowadays is MEME (Multiple EM for Motif Elicitation) (Bailey et al., 2010). MEME represents a motif as a PWM and assumes that each sequence has zero or one motif. To discriminate motifs, MEME calculates a "position-specific prior" (PSP) of each position in a sequence in order to measure the likelihood that a motif starts at each position of a sequence. PSP plays the role of additional information to assist the search by increasing the probability of start positions containing subsequences that is commonly found in sequences of interest, as well as decreasing the probability of start positions characterizing for sequences that do not contain features of interest. In a work by Lin et al. (2011), a motif is represented by using the profile HMM (Hidden Markov Model). Because this model allows to insert or delete a position in a sequence, and finding motifs is as to find hidden states of sequences. It is more flexible than MEME. The parameters of HMM were estimated by the maximum mutual information estimate technique to obtain the optimum discriminative motifs.

In brief, these methods definitely do not converge into the global optimal solution because they used expectation-maximization (EM) or Gibbs sampling algorithm to optimize the likelihood of PWM or HMM. Furthermore, these methods need to learn from a large enough training sequences in order to have precise motifs. If the learning process works with a small number of short sequences, PWM or HMM will not present good discriminative motifs due to a lack of information. As regards string-based methods, the exhaustive search was applied so that the global solution can be achieved. However, some disadvantages may occur such as a large amount of data will make the searching process time-consuming, or finding long length motifs can lead to a high computational complexity.

2.2. Semi-supervised feature selection

Feature selection is a significant step of the machine learning area with an aim to improve the learning performance by removing irrelevant features from the training dataset. In the supervised learning, feature selection methods work on labeled data to find the most useful feature subsets that help to increase the prediction accuracy or shorten the training time of classifiers on high dimensional datasets. However, as we all know, the size of labeled data is very limited because they need many human annotation efforts including time and expense, as well as expert-level knowledge. The use of a small labeled dataset together with a large unlabeled dataset to identify relevant features was first introduced by Zhao and Liu (2007). Therefore, conducting a feature selection from mixed labeled and unlabeled data is

a definition for a semi-supervised feature selection.

The survey presented by Sheikhpour, Sarram, Gharaghani, and Chahooki (2017) provides two taxonomies of a semi-supervised feature selection. They are the combination of the basic taxonomy of a feature selection and a semi-supervised learning. In the first taxonomy, methods are classified into groups based on the feature selection such as filter, wrapper, and embedded methods. Each group is then divided into smaller groups based on how to use the unlabeled data to learn the feature subsets. In contrast to the first taxonomy, methods in the second taxonomy are divided into five groups based on semi-supervised learning, such as graph-based, self-training, co-training, support vector machine based (SVM-based), and others. These groups are also divided into smaller groups based on the procedure of a feature selection. Overall, the first taxonomy is the most mentioned one in many studies (Chin, Mirzal, Haron, & Hamed, 2016; Xu, King, Lyu, & Jin, 2010; Chen, Nie, Yuan, & Huang, 2017).

3. Methodology

To characterize and predict motifs from sequences by using a semi-supervised feature selection, we develop a framework (Figure 1) consisting of four main steps: data vectorization, a feature selection, semi-supervised learning, and comparative analysis.

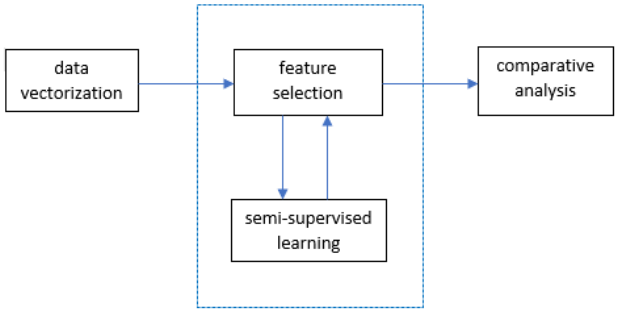

Figure 1. *The proposed framework for sequence characterization and prediction*

3.1. Data vectorization

In this step, we extract subsequences or motifs from a sequence dataset by using a sliding window technique and consider a subsequence as a feature in the feature selection problem. This means a sequence will be represented by many features, and the value of each feature is the occurrence frequency of that feature in a sequence. Then, we continue to eliminate subsequences which have a low frequency of occurrence in the dataset. Figure 2 demonstrates a sequence vectorized by the occurrence frequency of subsequences in that sequence.

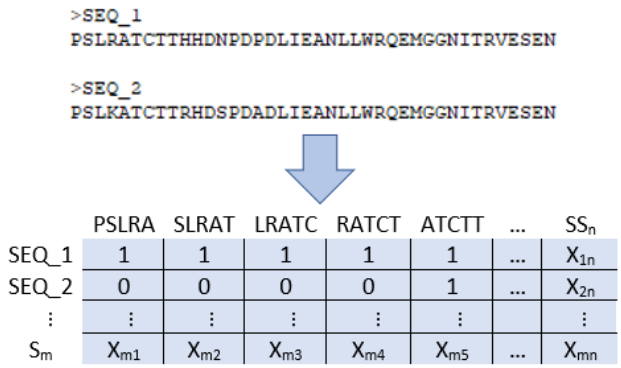

Figure 2. *Data vectorization. A sequence S(i) is represented by many subsequences SS(j). The value Xij is the frequency of occurrence of a subsequence SS(j) in the sequence S(i).*

With this vectorization, extracted subsequences are in different lengths because we want to keep information of short sequences as much as possible. We consider subsequences as motifs in biology and call them features in machine learning. Therefore, the problem of discriminative motif finding leads to the feature selection so that a classification algorithm running on a set of selected features obtains the highest possible accuracy.

3.2. Integration of feature selection and semi-supervised learning

The characterization task consists of two integrated steps, feature selection and semi-supervised learning, based on the work by Ren et al. (2008). Concretely, we use a wrapper-based method and the SFFS (Sequential Forward Feature Selection) algorithm for a feature selection and a self-training technique for semi-supervised learning. The idea of this task is shown in Algorithm 1, where *nSelect* denotes a predefined number of features, *selectedFeatures* denotes the output features, *mat* and *labels* denote the training dataset and their labels respectively. The algorithm begins with the empty set of selected features. While the number of selected features has not yet reached the predefined number, unselected features will be pseudo added into the set of selected features one by one. If the added features can help to increase the accuracy of the learner, we will officially add them to the set of output features. On the other hand, we do nothing with features not improving the accuracy.

Algorithm 1 The sequential feature selection algorithm

Input: mat, labels, nSelect
Output: selectedFeatures
selectedFeatures ← NULL
accuracy ← 0
selected ← 0
while selected < nSelect **do**
    for i ← 0 **to** length(features) − 1 **do**
        if features[i] ∉ selectedFeatures **then**
            Pseudo add features[i] to selectedFeatures

```
            mat_i ←mat + selectedFeatures
            accuracy_i ← Learning from mat_i, labels
        end if
    end for
    if max(accuracy_i) ≤ accuracy then
        break
    else
        accuracy ← max(accuracy_i)
        idx = arg max(accuracy_i)
        Officially add features[idx] to selectedFeatures
        mat ← mat + selectedFeatures
        selected++
    end if
end while
return selectedFeatures
```

In the prediction task, we use the SVM (Support Vector Machine) method to learn a linear function which separates the data into two classes, SVR and non-SVR. We investigate the effectiveness of SVM because none of the previous studies to predict SVR/non-SVR in the case the dataset with few sequences and they are very short in length. Furthermore, SVM is the most widely used model in machine learning.

3.3. Comparative analysis

In this section, we discuss how to find subsequences that characterize the SVR/non-SVR class of sequences. That means we want to know which subsequences are potentially discriminative for a class. We firstly obtain a set of potentially discriminative subsequences from the results of the cross-validation experiment and then find discriminative subsequences by estimating and contrasting their contributions to classes. In order to get the final set of subsequences, our idea is to combine the results from different folds of the experiment. Concretely, we keep subsequences that appears at least τ folds. The contribution of subsequences to classes can be approximated by counting the frequency of subsequences in the SVR and non-SVR sequences. After that, we contrast these frequencies to see the class to which a subsequence contributes significantly. In practice, subsequences in which we are mostly interested must have a high contribution.

4. Experiment

4.1. Dataset

In this work, the data are sequences before treatment and belong to HCV-1b. The dataset consists of (a) 43 sequences including 21 SVR and 22 non-SVR sequences downloaded from Chiba University; (b) 254 sequences including 141 SVR and 113 non-SVR sequences taken from five published studies (Enomoto et al., 1996; Chayama et al., 1997; Yoon et al., 2007; Rueda et al., 2008; El-Shamy et al., 2011), and (c) 1,444 unlabeled sequences downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/).

4.2. Experimental setting

In the semi-supervised feature selection (called SemiFS for short) experiment, a set of parameters consists of *sizeFS*, *startfn*, *maxIteration*, *samplingTimes*, *samplingRate*, and *fnstep*, where *sizeFS* is the number of output features; *startfn* is the number of initial features in labeled dataset that are used to train a classifier; *maxIteration* is the number of times for learning and predicting labels for the unlabeled dataset; *samplingTimes* is the number of times for adding unlabeled data to labeled data; *samplingRate* is the number of unlabeled data with predicted labels in order to find more useful features; *fnstep* is the number of selected features after adding unlabeled data. They are set to 30, 5, 30, 10, 50%, and 6, respectively. We performed the SemiFS many times to obtain these best parameters.

In the Algorithm 1, the parameters *selectedFeatures* is first set to *startfn* when starting to learn with labeled data, and then set to *fnstep* when adding unlabeled data to labeled data. We also initialize the parameter *accuracy* of a learner with 0, a default value.

4.3. Accuracy of prediction

To evaluate the effectiveness of our framework, we conduct predictions of SVR and non-SVR with three methods: Perceptron, SVM-linear, and k-NN which are experimented by scikit-learn (Pedregosa et al., 2011), a tool for data mining and analysis. The perceptron learning algorithm is a classification algorithm for the simplest case that has only two classes. In our experiments with SemiFS, we find the setting *randomstate=0* to be a reasonable choice, and use default values for other parameters. The SVM-linear classifier is a parametric method and is used with a hypothesis space where the dataset is linearly separable. We choose the regularization parameter *C=1.0* and use a default setting for the parameter *gamma*. And the k-NN, a non-parametric method, is a *k* nearest neighbor classifier whose the parameter *k* is a small positive integer. In practice, we chose *k=5*. We also compare the SemiFS framework with MEME (http://meme-suite.org), a previous study of a discriminative motif finding. The tool MEME is a popular and powerful web-based software to discover motifs in biology. Therefore, we easily conduct the experiment with the following parameters: the length of a motif is between 3 and 32 residues, the occurrence frequency of a single motif per sequence is set to zero or one, and the maximum number of motifs is 10. For all prediction methods, we perform a 5-fold cross validation experiment and the prediction accuracy is averaged from these five folds.

Table 1. *Accuracies of three methods for prediction*

	Perceptron			SVM-linear			k-NN		
	Full	SemiFS	MEME	Full	SemiFS	MEME	Full	SemiFS	MEME
Fold 1	0.64	0.57	0.59	0.52	0.50	0.54	0.49	0.44	0.40
Fold 2	0.47	0.56	0.52	0.52	0.57	0.56	0.47	0.47	0.57
Fold 3	0.64	0.62	0.50	0.52	0.54	0.52	0.47	0.62	0.40
Fold 4	0.44	0.59	0.44	0.52	0.57	0.55	0.47	0.62	0.44
Fold 5	0.51	0.50	0.55	0.51	0.55	0.51	0.43	0.53	0.48
Avg. Acc.	0.54	**0.57**	0.52	0.52	**0.55**	0.54	0.46	**0.54**	0.46

From Table 1, the average accuracies of each classifier working on our framework are better than the average accuracies of MEME. For example, with the k-NN classifier, the accuracy of SemiFS is 54%, while the accuracy of MEME is 46% (about 8% improvement). In a similar manner, SemiSF has a 57% accuracy with the perceptron classifier, while MEME has a 52% accuracy (about 5% improvement). However, with the SVM-linear classifier, the accuracy of SemiFS is just 1% higher than the accuracy of MEME. Table 1 also shows the effectiveness of SemiSF compared to the classification without doing feature selection (called Full for short). With three classifiers, SemiFS is more accurate from 3% to 5% higher than that of Full. This gives a strong significance for our work, because features or subsequences found by SemiSF are the potential discriminative ones. They increase the classification accuracy, which means that they can contribute to characterizing classes in a dataset. In our case, classes are SVR and non-SVR.

4.4. Discriminative subsequences

To find reliable subsequences characterizing SVR and non-SVR sequences, we conduct a subsequences analysis. We collect selected subsequences discovered by SemiFS in five experiments, and then choose a collection of subsequences appearing at least in three folds. Table 2 presents 10 subsequences along with the number of SVR sequences and non-SVR sequences containing these subsequences. These numbers are calculated on the whole labeled dataset.

Table 2. *Discriminative subsequences characterizing SVR and non-SVR by semifs*

Subsequence	No. of SVR sequences	No. of non-SVR sequences
ACTT	6	0
AHH	4	0
DANLLWRQEM	7	0
GGS	6	0
HRDSPDA	2	0
MGGS	6	0
QHDSPDADLI	1	0
RDSPDA	2	0
VDLV	4	0
WQQ	6	0

It can be observed from Table 2 that 10 subsequences appear many times in SVR sequences, and two or three times in non-SVR sequences. Typically, subsequences such as “WQQ”, “ACTT”, or “DANLLWRQEM” do not appear in non-SVR sequences. Their coverage are from 3.2% to 3.8% (6/184 = 0.043 or 7/184 = 0.038), and their accuracy are 100% (6/(6+0) = 1 or 7/(7+0) = 1). These subsequences are likely to discriminate the SVR property. In addition, subsequences such as “ACC,” “KAA,” “VSL,” “LSLKA,” or

"GGDITR" have the higher coverage from 3.2% to 7.6% and the lower accuracy from 76% to 89%. However, they also help to discriminate the SVR property thanks to a majority rule. These subsequences can be discriminative motifs in order to predict the SVR property in a HCV study because they are considered to be significant to the SVR class and not significant to the non-SVR class.

A comparison with the results of MEME is presented in Table 3. After five times of experiments of MEME, we collect around 10 motifs, and most of them appear in both SVR and non-SVR sequences. MEME found only three motifs, "RGK," "TAC," and "SLKATCTFHHDSPDADLIEANLLWRQEMGGNI," that appear in the SVR class and do not appear in the non-SVR class. The coverage of "TAC" is 3.8% while the coverage of the other two motifs is 0.5%, a very low coverage. The rest of motifs in Table 3 are found many times in both classes, for example "SLK" appears 121 times in SVR and 147 times in non-SVR, or "THHDSPDADLIEANLLWRQEMGGNITRVESEN" appears 29 times in SVR and 37 times in non-SVR. In our opinion, MEME discovered motifs which are not good enough to differentiate SVR and non-SVR characteristics. Therefore, it just works effectively in the case of finding common motifs describing certain characteristics of a large sequence dataset.

Table 3. *Discriminative subsequences characterizing SVR and non-SVR by SemiFS*

Subsequence	No. of SVR sequences	No. of non-SVR sequences
DLI	149	126
ESE	140	154
RGK	1	0
SLK	121	147
TAC	7	1
TRVESEN	136	151
THHDSPDADLIEANLLWRQEMGGNITRVESEN	29	37
SLKATCTFHHDSPDADLIEANLLWRQEMGGNI	1	0
ATCTTHHDSPDADLIEANLLWRQEMGGNITRV	27	37
CTTHHDSPDADLIEANLLWRQEMGGNITRVES	27	37

5. Conclusion

We developed a framework for characterization and prediction of HCV treatment outcomes by using a semi-supervised feature selection. Our approach was demonstrated to represent well sequence data into numeric vectors, analyze and interpret clearly results of the computational process. This approach works effectively with the data containing short sequences and being similar to another while the traditional methods could not overcome this case of data. Furthermore, it has shown to be a general and flexible method that can be applied to other kinds of sequence data. Potentially discriminative motifs that we found can be good patterns for predicting SVR/non-SVR sequences after being verified by physicians.

❖ ***Conflict of Interest:*** *Authors have no conflict of interest to declare.*

❖ ***Acknowledgements:*** *The authors would like to thank Dr. Tatsuo Kanda from the Graduate School of Medicine, Chiba University for his data support.*

This work was supported by the university-level research project, T2017-02, at University of Science, Vietnam National University Ho Chi Minh City.

REFERENCES

Bailey, T. L., Boden, M. B., Whitington, T., & Machanick, P. (2010). The value of position-specific priors in motif discovery using meme. *BMC Bioinformatics, 11*(1).

Chayama, K., Tsubota, A., Kobayashi, M., Okamoto, K., Hashimoto, M., Miyano, Y.,… & Kumada, H. (1997). Pretreatment virus load and multiple amino acid substitutions in the interferon sensitivity - determining region predict the outcome of interferon treatment in patients with chronic genotypes 1h hepatitis C virus infection. *Journal of Hepatology*, *25*(3), 745-749.

Chen, X., Nie, F., Yuan, G., & Huang, J. Z. (2017). *Semi-supervised feature selection via rescaled linear regression*. Proceedings of the 26th International Joint Conference on Artificial Intelligence.

Chin, A., Mirzal, A., Haron, H., & Hamed, H. (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene election. IEEE/ACM *Transactions on Computational Biology and Bioinformatics*, *13*.

El-Shamy, A., Shoji, I., Saito, T., Watanabe, H., Ide, Y., Deng, L.,… & Hotta, H. (2011). Sequence heterogeneity of NS5A and core proteins of hepatitis C virus and virological responses to pegylated-interferon/ribavirin combination therapy. *Microbiology and Immunology*, *55*, 418-426.

Enomoto, N., Sakuma, N., Asahina, I., Kurosaki, Y., Murakami, M., Yamamoto, T.,… & Chifumi Sato, M. D. (1996). Mutations in nonstructural protein 5A gene and response to interferon in patients with chronic hepatitis C virus 1b infection. *The New England Journal of Medicine*, *334*, 77-81.

Gao, M., Nettles, R. E., Belema, M., Snyder, L. B., Nguyen, V. N., Fridell, R. A.,… & Hamann, L. G. (2010). Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. *Nature Letters*, *465*, 96-100.

Han, J., & Kamber, M. (2006). *Data mining concepts and techniques*. Diane Cerra.

Kim, J. K., & Choi, S. (2011). Probabilistic models for semi-supervised discriminative motif discovery in DNA sequences. IEEE/ACM *Transactions on Computational Biology and Bioinformatics*, *8*(5).

Lin, T., Murphy, R. F., & Bar-Joseph, Z. (2011). Discriminative motif finding for predicting protein subcellular localization. IEEE/ACM *Transactions on Computational Biology and Bioinformatics*, *8*(2).

Manns, M., McHutchison, J. G., Gordon, S. C., Rustgi, V. K., Shiffman, M., Reindollar, R.,… & Albrecht, J. K. (2001). Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: A randomised trial. *The Lancet*, *358*, 985-965.

Pedregosa., F., Varoquaux, G., Gramfort., A., Michel., V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Ren, J., Qiu, Z., Fan, W., Cheng, H., & Yu, P. S. (2008). *Forward semi-supervised feature selection.* Proceedings of the 12th Pacific-Asia Conference in Knowledge Discovery and Data Mining.

Rueda, P. M., Casado, J., Paton, R., Quintero, D., Palacios, A., Gila, A.,… & Salmeron J. (2008). Mutations in E2-PePHD, NS5A-PKRBD, NS5A-ISDR, and NS5A-V3 of hepatitis C virus genotype 1 and their relationship to pegylated interferon-ribavirin treatment responses. *Journal of Virology, 82*, 6644-6653.

Sami, A., & Nagatomi, R. (2008). *A new definition and look at DNA motif*. Intech.

Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. Z. (2017). A survey on semi-supervised feature selection methods. *Pattern recognition, 64*.

Vens, C., Rosso, M. N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics, 27*(9), 1231-1238.

Wu, J., & Xie, J. (2010). Hidden Markov model and its application in motif findings. Statistical *Methods in Molecular Biology, 620*, 405-416.

Xu, Z., King, I., Lyu, M. R. T., & Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization. IEEE *Transactions on Neural Networks, 21*.

Yoon, J., Lee, J. I., Baik, S. K., Lee, K. H., Sohn, J. Y., Lee, H. W., … & Yeh, B. I. (2007). Predictive factors for interferon and ribavirin combination therapy in patients with chronic hepatitis C. *World Journal of Gastroenterology, 13*(46), 6236-6242.

Zhao, Z., & Liu, H. (2007). *Semi-supervised feature selection via spectral analysis*. Proceeding of the 7th SIAM International Conference on Data Mining.

TÌM MOTIF PHÂN BIỆT ĐỂ DỰ ĐOÁN KẾT QUẢ ĐIỀU TRỊ HCV VỚI PHƯƠNG PHÁP CHỌN LỌC ĐẶC TRƯNG BÁN GIÁM SÁT

Nguyễn Thị Tường Vy[1], Lê Thị Nhàn[2*]

[1]Trường Đại học Lyon, Pháp

[2]Trường Đại học Khoa học Tự nhiên, ĐHQG TPHCM, Việt Nam

[*]Tác giả liên hệ: Lê Thị Nhàn – Email: ltnhan@fit.hcmus.edu.vn

Ngày nhận bài: 08-01-2020; ngày nhận bài sửa: 27-02-2020; ngày duyệt đăng: 02-6-2020

TÓM TẮT

Điều trị viêm gan C hiện đang phải đối mặt với nhiều thách thức, ví dụ như chi phí chữa trị cao, thuốc có tác dụng phụ và tỉ lệ thành công thấp với kiểu gen viêm gan C 1b (HCV-1b). Để xác định đặc tính nào của HCV-1b gây ra kháng thuốc, nhiều phương pháp phân tích chuỗi đã được tiến hành để tìm ra các dấu hiệu sinh học giúp dự đoán tỉ lệ thất bại. Tuy nhiên, kết quả vẫn có thể không chính xác khi các phương pháp này thực hiện trên một tập dữ liệu nhỏ gồm các chuỗi được gán nhãn và có độ dài ngắn. Trong bài báo này, chúng tôi hướng đến dự đoán kết quả điều trị HCV-1b và mô tả đặc trưng của HCV-1b bằng cách kết hợp hai phương pháp lựa chọn đặc trưng và học có giám sát bán. Phương pháp đề xuất của chúng tôi cải thiện độ chính xác dự đoán khoảng từ 5% đến 8% so với các phương pháp trước đó. Ngoài ra, chúng tôi tìm được một tập các motif phân biệt tốt có thể được xem là tín hiệu sinh học để dự đoán đáp ứng hoặc kháng thuốc của điều trị HCV-1b.

Từ khóa: motif phân biệt; virus viêm gan C; phương pháp lựa chọn thay đổi liên tiếp; chọn lọc đặc trưng bán giám sát