# HCV 단백질과 상호작용하는 인간 단백질을 예측하기 위한 SVM 모델

방초°, 최광우, 한경숙*
인하대학교 정보공학과
fczqx@inha.edu, cuigy119@inhaian.net, khan@inha.ac.kr

# An SVM model for predicting human proteins interacting with HCV proteins

Chao Fang°, Guangyu Cui, and Kyungsook Han*
Department of Information Engineering, Inha University
° presenting author, *corresponding author

## 1. Introduction

Several computational methods have been developed for predicting protein-protein interactions, but most of these methods are intended for finding the protein-protein interactions within a species rather than for the interactions across different species. Methods for predicting the interactions between homogeneous proteins are not appropriate for predicting the interactions between heterogeneous proteins since they do not distinguish the interactions between proteins of the same species from those of different species.

In this study we developed a computational method to predict the interactions between hepatitis C virus (HCV) and human proteins. Although much progress has been made in clinical and basic research on HCV, interactions between HCV proteins and human proteins is not fully understood. Identifying more interactions between HCV and human proteins should help elucidate the interaction mechanism of HCV with host cells and can be helpful in the design of molecules that target the new interacting proteins.

## 2. Representation and Method

We obtained the interaction data between HCV proteins (core, E1, E2, F, NS2, NS3, NS4A, NS4B, NS5A, NS5B and P7) and human proteins from the infection mapping project (I-MAP) [1]. The data contains 481 interactions between 11 HCV proteins and 421 human proteins. By searching Gene IDs of the 421 human proteins in HPRD (http://www.hprd.org), we identified a total of 695 interactions between HCV and human proteins.

We implemented a support vector machine (SVM) model using LIBSVM (http://www.csie.ntu.edu.tw/~cjlim/libsvm/) with the radial basis function (RBF) as a kernel function. A positive data set for the SVM model consists of the 695 protein-protein interactions. To construct a negative data set, we randomly selected human proteins from HPRD, which are not included in the 695 interactions of the positive data set. We built a negative data set of 695 interactions for the balance with the positive data set.

We constructed a training set with 500 positive data and 500 negative data, which were selected from the 695 positive data and 695 negative data. The remaining 195 positive data and 195 negative data were used to construct a test set.

One of the main computational challenges in using an SVM model to predict PPIs is to find a suitable way to describe the important information of protein. We represent a protein sequence using three consecutive amino acids called *amino acid triplet*. For example, in the amino acid sequence TVAVTVA, there are four overlapping amino acid triplets: TVA, VAV, AVT and VTV, which appear 2, 1, 1, and 1 times in the sequence, respectively.

To reduce further the dimension of the vector space, we represent an amino acid sequence using the class of amino acids. Based on the biochemical similarity of amino acids, twenty amino acids were classified into six categories: {IVLM}, {FYW}, {HKR}, {DE}, {QNTP}, and {ACGS}. According to this classification, there are 6×6×6 = 216 possible triplets. We use a binary space (V, F) to represent a protein sequence, in which V is a vector space of feature vectors with a fixed number of features and F is a vector space of frequency vectors. A

protein sequence of variable length is first mapped to a feature vector v of fixed length. A feature vector $v$ is then mapped to a frequency vector $f_i$ ($i$=1, 2, …, 216), which represents the frequency of each triplet type.

The difference between the frequency values of triplets can be too small to discriminate interacting sequences from others, so we modified the equation used in the study of Shen *et al.* [2] as follows:

$$d_i = \{e^{\frac{fi - \min\{f_1, f_2, ..., f_{216}\}}{\max\{f_1, f_2, ..., f_{216}\} - \min\{f_1, f_2, ..., f_{216}\}}}\} - 1$$

where $d_i$ is the relative frequency of the $i$-th triplet type in one sequence. The value of $d_i$ ranges from 0 to e-1. the modified $d_i$ has a value in a wider range than the frequency value used in the study of Shen *et al.* Two features were added at the end of a feature vector: index of an HCV protein (1 to 11 for 11 HCV proteins) and classification of the feature vector (1 for interaction and -1 for non-interaction). By encoding the index of an HCV protein, the SVM model can find a human protein interacting with the HCV protein.

## 3. Results and Summary

We evaluated the performance of the SVM model using three measures: sensitivity, specificity and accuracy. Our method achieved a sensitivity of 77%, a specificity of 86.3% and an accuracy of 81.5% on average. Our method outperformed the method of Shen *et al.* [2], which showed a sensitivity of 75.5%, a specificity of 79.2% and an accuracy of 77.3% on average.

To find potentially new human proteins that interact with an HCV protein (called $H_{HCV}$ hereafter), we searched human proteins in NCBI that are similar to the 695 human proteins known to interact with an HCV protein. After running BLASTP (http://www.ncbi.nlm.nih.gov/BLAST/) with the E-value ≤ $10^{-20}$ and removing the redundant sequences with the 695 human proteins, we obtained a total of 4,209 human proteins. To predict reliable interactions, we selected the human proteins that have the same cellular component gene ontology (GO) IDs with $H_{HCV}$ for each HCV protein. For instance, there are 29 $H_{HCV}$ proteins that are known to interact with the HCV E2 protein, and the 29 $H_{HCV}$ proteins have a total of 15 cellular component GO IDs. The SVM model predicted 33 $H_{HCV}$ proteins as interacting partners of HCV E2 protein, but 10 out of the 33 candidates that have the same cellular component GO IDs were left as reliable candidates.

In summary, we represent a protein sequence using three consecutive amino acids called amino acid triplet. We map a protein sequence of variable length to a feature vector of fixed length, and then map the feature vector to a frequency vector that represents the relative frequency of each triplet within the protein sequence. The SVM model showed an average accuracy of 81.5% in predicting the interactions between HCV proteins and human proteins, which is higher than the previous method by others. Using the SVM model and the Gene Ontology (GO) annotations of proteins, we also predicted a total of 456 new human proteins that potentially interact with HCV proteins, which are being validated by biochemical experiments. As a future work, we plan to use additional biochemical properties of amino acids to improve the performance of the SVM model.

## Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0017213).

## References

[1] B. de Chassey, *et al.*, Hepatitis C Virus Infection Protein Network, International Journal of Infectious Diseases, 12, E175-E175, 2008.

[2] J. W. Shen, *et al.,* Predicting protein-protein interactions based only on sequences information, Proceedings of the National Academy of Sciences of the United States of America, 104, 4337-4341, 2007.