Abstract

A motif is an abstraction over a set of repeated patterns observed in a dataset. It captures the essential features shared by a set of related data. Motif finding can be understood simply that given a set of sequences, one will find an unknown motif that occurs frequently in that sequence dataset. Finding discriminative motifs has recently received much attention in biomedicine as such motifs allow us to characterize in distinguishing two different classes of sequences. The obvious difference between discriminative motif finding and motif finding is that the former uses sequences of two different classes to discover motifs while the latter searches motifs in one class of sequences only. Discriminative motif finding can be seen as the next step of motif finding problem using one more dataset to help motif searching more effectively.

In many biomedical domains, the quantity of labeled sequences is very limited while a large number of unlabeled sequences are usually available. Discovering discriminative motifs in a small number of labeled data is a challenge for sequence motif finding methods at present. These methods usually require a large amount of labeled data to search optimal parameters for models representing motifs. Furthermore, because motifs are often embedded in conserved sequence fragments, the labeled sequences are short in length and tend to resemble one another. Therefore, these characters also pose serious drawbacks for traditional motif finding methods.

In our study of hepatitis therapy by using NS5A (non structure 5A) protein, where we are interested in discriminating two classes of SVR (sustained virologic response) and non-SVR (non sustained virologic response) sequences, few labeled sequences are collected from public sequence database, but thousands of unlabeled sequences are obtained. Working with ISDR (interferon sensitivity determining region), a small part of NS5A protein consisting of 40 amino acids, we are dealing with one more difficult case of data, short and similar sequences. Because the function of ISDR is supposed to do the replication for HCV, the polypeptide sequence of ISDR should be preserved and has a few variants at some positions.

It is well known that the current treatment, a combination of interferon and ribavirin (IFN/RBV), for HCV (hepatitis C virus) is expensive, often causes side effects, and its success rate is only a half of cases. Sequence analyzing to find characteristics of response or resistance to HCV treatment is necessary to be able to predict failures before the treatment. Several studies were conducted for explanations of the resistance to IFN/RBV therapy of HCV to get a deeper understanding how HCV escape from the immune system. In addition, the correlation between NS5A protein and IFN/RBV therapy has been reported in numerous papers in biomedicine, as well as in computational field. However, the understanding of inhibitions of the HCV NS5A protein with IFN/RBV therapy is still unknown deeply and no clear adaptation patterns to the antiviral treatment were detected. And existing methods for sequence characterization work ineffectively when input sequences do not provide enough information for searching because they are short in length and very similar to one another and the number of labeled sequences is small.

Therefore, our research focuses on developing computational methods to discover the new knowledge from NS5A protein in two situations, few labeled data and short sequences. From this knowledge, we aim at a comprehensive understanding the relation between NS5A protein and IFN/RBV therapy in order to answer two main questions: what NS5A biomarkers for IFN/RBV resistance and response are and what links among these biomarkers are. Our contributions consist of new biomedical findings that can help to predict signals of response or resistance to IFN/RBV therapy and new computational methods for knowledge creation.