氏名	LE THI NHAN		
学位の種類	博士(知識科学)		
学位記番号	博知第 147 号		
学位授与年月日	平成 25 年 9 月 24 日		
論文題目	Discriminative Motif Learning for Hepatitis C Virus Study(識別モチーフ学習による C 型肝炎の研究)		
論文審査委員	主査 Ho Tu Bao	北陸先端科学技術大学院大学	教授
	橋本　敬	同	教授
	池田　満	同	教授
	Dam Hieu Chi	同	准教授
	佐藤　賢二	金沢大学	教授

論文の内容の要旨

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.

論文審査の結果の要旨

Discriminative motif learning is to find motifs occurring more frequently in one sequence set and not occurring in the other sequence sets by using a set of two-class sequences. Recently, discriminative motif learning has received much attention from the research community and could be the next step of motif learning. So far, many methods have been developed to discover discriminative motifs with a probabilistic model and with a string-based model. However, in the case of our study on HCV treatment, previous methods have shown to be ineffective because of the input sequences are similar, short in length and small in number. The main reason of these limitations is that traditional methods require a large number of sequences to learn the optimal motif models. Therefore, our research aimed to develop new methods to discover discriminative motifs in two situations: few labeled data and short sequences, and then we applied these new methods to HCV study to discover the new knowledge from the relationship between NS5A protein and IFN/RBV therapy. Obtaining this new knowledge, we believe that our potential findings can provide additional

knowledge to answer two main research questions: what are NS5A biomarkers of IFN/RBV resistance and response? And what are links among these biomarkers? Our contributions were made in following,

Discriminative motif learning for few labeled data: In many research fields, labeled data are difficult to have, because they need a lot of factors such as human annotations, expert knowledge, special devices, and so on. For example, in our HCV study, we can obtain a very small number of existing labeled protein sequences (147 NS5A sequences for non-SVR and 105 NS5A sequence for SVR) while a large number of unlabeled sequences (more than 5,000 NS5A sequences) are available at public databases. The objective of this situation is to develop a semi-supervised ensemble method that has ability to discovery discriminative motifs from an extended labeled sequence that contains labeled sequences and unlabeled sequences with predicted labels. We proposed a semi-supervised ensemble method for discriminative motif learning based on the SLUPC algorithms, a separate-and-conquer searching method. The proposed method, named E-SLUPC (Ensemble SLUPC), firstly search a core motif set from a small number of labeled data, and then use these core motifs to extend the training dataset by exploiting a large unlabeled data with the majority voting strategy in ensemble learning. Strong discriminative and frequent motifs characterizing two outcome classes of HCV treatment (SVR and non-SVR) were detected and analyzed. These motifs are promising as they represent many patterns that have not been known before. E-SLUPC can improve the quality of discriminative motifs when compare to discriminative motifs found by MEME and DEME, and the accuracy when compare to the SLUPC algorithm. The proposed method showed the ability to find strong discriminative motifs and obtain higher accuracy when provide more data for the training dataset. However, using two thresholds coverage and discriminant, the SLUPC algorithm could eliminate quickly some potential candidates during recursively expand a subsequence because the two thresholds must be satisfied simultaneously. Working with a small labeled dataset, even though we used a self-training technique of semi-supervised learning to enlarge the training dataset, the over-fitting problem is inevitable.

Discriminative motif learning for short sequences: The input sequences are short in length and similar to each other that are serious obstacles for current methods, because these sequences do not provide enough information for the motif searching process. In our HCV study, the ISDR of NS5A protein contains 40 amino acids only and has few variants at some positions. The objective of this situation is to develop an effective computational method for discriminative motif discovery. We approached to discriminative motif learning in a new way by using topic modeling. The short sequences were first enriched by a representation in a high dimensional space. Then we constructed a discriminative space (topical space) by using unsupervised learning with a topic model. We used labels and neighborhood information to infer a new representation of data as well as new latent

components of the discriminative space. Next, we project the data onto the discriminative space by using the inference procedure of the topic model. Finally, we performed the prediction and analysis of discriminative patterns from the projected data. This method was applied to get insight into SVR and non-SVR properties of IFN/RBV therapy. We found a large number of discriminative subsequences for non-SVR, even though the number of non-SVR sequence is small. This suggests that non-SVR sequences are very diverse and complicated. And this is in coincidence with experimental researches of HCV genotype 1b. The proposed method has shown its effectiveness through the prediction quality being often higher than the quality of the baseline method, about 30% improvement. However, in topic model, the data or documents are often represented sparsely. If the representation of short sequences is dense, we cannot get a high accuracy for prediction.

The study has shown the candidate's strong ability of independently conducting the scientific research, which should be sufficient to be a considered for the doctoral degree of knowledge science.