Discriminative Motif Learning for Hepatitis C Virus Study

by

LE, Thi Nhan

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Ho Tu Bao

School of Knowledge Science
Japan Advanced Institute of Science and Technology

September, 2013

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.

Acknowledgements

This dissertation would not have been possible without the help, support, guidance and effort of a lot of people. It gives me great pleasure to express my sincere thanks to whom I am greatly indebted.

Firstly, I am very grateful to the Ministry of Education and Training (MOET) of Vietnam for its scholarship supports during the period my study. I owe my deepest gratitude to my supervisor, Professor Ho Tu Bao, for his guidance and support from the first day when I am just a fresh PhD student to the present day when I am planing to be researcher. I learned from him much useful experience in the academic life as well as the daily life.

I would like to show my gratitude to Professor Tatsuo Kanda, Professor Katsuhiko Takabayashi, and Professor Osamu Yokosuka (Chiba University) who suggested general questions on the non structure 5A protein for hepatitis C virus treatment, and have provided evaluations on computed results. I have had a great opportunity to work with them based on more than a decade of consecutive research projects with my laboratory.

I would like to express the appreciation to my committee chair, Professor Mitsuru Ikeda (JAIST), Professor Takashi Hashimoto (JAIST), Associate Professor Dam Hieu Chi (JAIST) and Professor Kenji Satou (Kanazawa University) who read and give useful comments to improve this dissertation.

A very special thank to the members of Ho laboratory who shared their research ideas, valuable experiences and useful discussions, and my friends who stood by and helped me to deal with and overcome tough times in my study at JAIST.

Finally, I dedicate this dissertation to my parents and brother who always believed in me, encourage and support throughout my life.

Contents

Abstract **i**

Acknowledgements **iii**

1 Introduction **1**

1.1 Research context 1

1.2 Problem formulation 3

1.3 Major contributions 5

1.4 Dissertation organization 6

2 Background **7**

2.1 Fundamental of motif discovery 7

2.1.1 Motif definition 7

2.1.2 Motif representation 7

2.1.3 Motif finding algorithm 11

2.1.4 Discriminative motif learning 13

2.2 Fundamental of HCV study 15

2.2.1 Hepatitis C Virus 15

2.2.2 NS5A protein 18

2.2.3 ISDR of NS5A protein 22

2.2.4 HCV therapy 22

3 Discriminative motif learning for few labeled data **25**

3.1 Introduction . . . 25
3.1.1 Discriminative motif learning . . . 25
3.1.2 Semi-supervised ensemble learning . . . 27
3.2 Method . . . 28
3.2.1 SLUPC algorithm . . . 29
3.2.2 Self-training technique for semi-supervised learning . . . 31
3.2.3 Majority voting strategy for ensemble learning . . . 32
3.3 Application to HCV study . . . 33
3.3.1 The dataset . . . 33
3.3.2 Finding DMOPS motifs charactering SVR and non-SVR to therapy 34
3.3.3 Evaluating the accuracy of E-SLUPC and SLUPC . . . 35
3.3.4 Comparing the output of E-SLUPC to MEME and DEME . . . 36
3.4 Conclusion . . . 38

4 Discriminative motif learning for short sequences 40
4.1 Introduction . . . 40
4.2 Method . . . 41
4.2.1 Topic model . . . 41
4.2.2 Supervised dimension reduction . . . 42
4.2.3 The two-step framework . . . 42
4.2.4 Methodology . . . 44
4.3 Results . . . 48
4.3.1 Datasets . . . 48
4.3.2 Accuracy of prediction . . . 48
4.3.3 Choosing a good setting for the framework . . . 50
4.3.4 Discriminative subsequences . . . 51
4.4 Conclusion . . . 54

5 Conclusion 56

5.1 Dissertation summary 56
5.2 Future Work 58

Bibliography **60**

Publications **69**

List of Figures

2.1 DNA motifs [D'haeseleer, 2006]. 8

2.2 An example of PWM [Craven, 2011]. 9

2.3 An example of PWM [Wu and Xie, 2010]. 9

2.4 A formula to calculate the likelihood of a sequence motif in PWM [Craven, 2011]. 9

2.5 An example of transition probability matrix in HMM. 11

2.6 An example of length-3 motif representation in MEME [Craven, 2011]. . . 13

2.7 Simplified diagram of the structure of hepatitis C virus [Colm, 2008]. . . . 15

2.8 Milestones in hepatitis C virus research [Moradpour et al., 2007]. 16

2.9 Simplified diagram of the structure of hepatitis C virus [Colm, 2008]. . . . 17

2.10 Hepatitis C Virus life cycle [AIDSInfoNet, 2012]. 19

2.11 The NS5A protein structure [Macdoanldt and Harris, 2004]. 19

2.12 N-terminal amphipathic helix [Macdoanldt and Harris, 2004]. The evidence for functions of NS5A in virus replication. 20

2.13 Interferon signaling pathway [Suhalim, 2007]. 21

2.14 Interferon sensitivity determining region (ISDR) [Macdoanldt and Harris, 2004]. The ability of IFN/RBV resistance can be eliminated by the mutations indicated in bold. 22

3.1 The framework of E-SLUPC. 29

4.1 Sketch of approaches for SDR [Than et al., 2012]. Existing methods for SDR directly find the discriminative space, which is supervised learning (c). The two-step framework consists of two separate steps: (a) first find an initial space in a unsupervised manner; then (b) utilize label information and local structure of data to derive the final space. 43

4.2 The proposed framework for sequence characterization and prediction. . . . 45

4.3 Data representation. A sequence is represented by many subsequences, subsequence frequencies, and class of that sequence. Each subsequence in a vector is represented by its index in the dictionary. 45

4.4 An example of projection of data onto 10-dimensioned topical space. A sequence in the topical space is a mixture of topics. Topics that do not have any contributions will be omitted in representation. 46

4.5 SVR and non-SVR sequences in the original space and in topical space. ◯ and ♢ are respectively non-SVR and SVR sequences. 50

4.6 The impact of the parameters. (a) Changing N_d, while fixing $\lambda = 0$, $R = 1000$, and K=10. (b) Changing R, while fixing $N_d = 5$, $\lambda = 0$, and $K = 10$. (c) Changing K, while fixing $N_d = 5$, $\lambda = 0$, and $R = 1000$. (d) Changing λ, while fixing $N_d = 5$, $R = 1000$, and $K = 10$. 51

List of Tables

2.1 Lengths of the NS5A protein. 20

2.2 The response rate depends on HCV genotype. The Response rate is the percentage of patients whose cancer shrink (a partial response) or disappear (a complete response) after treatment 24

3.1 DMOPS motifs characterizing SVR and non-SVR to IFN/RBV therapy . . 35

3.2 Accuracy of SLUPC adn E-SLUPC 36

3.3 The top twelve motifs found by MEME 37

3.4 The top twelve motifs found by DEME 38

4.1 Accuracies of 7 methods for predicting sequences 49

4.2 Discriminative subsequences characterizing SVR and non-SVR outcomes of HCV treatment 52

Chapter 1

Introduction

This chapter first introduces the research context of discriminative motif learning and hepatitis C virus treatment. Then the research problem and objectives are presented. The next part provides a concise view of our major contributions. And the last one shows the organization of the dissertation.

1.1 Research context

One of the key interests of biologists is to detect short and highly conserved motifs in a collection of DNA (*deoxyribonucleic acid*) or protein sequences. Motif finding is normally the challenging problem in molecular biology and computer science [Das and Dai, 2007]. This problem can be understood simply as follows: given a set of sequences, find an unknown motif that occurs frequently. Found motifs can be used to group data into meaningful classes, to summarize data, or to reveal unusual phenomena in sequences [Conklin et al., 1993]. Traditionally, motif finding has been dominated by generative models using only sequences of one class to produce descriptive motifs of the class. Recently, discriminative motif finding using sequences of two distinct classes to discover selective motifs that can distinguish these two different classes has attracted much attention from the research community. Discriminative motif finding can be seen as the next step of motif finding problem using one more dataset to help motif searching more effectively.

It is well known that labeled data are often difficult and time consuming to obtain, because they require human annotations, knowledge from experts and special devices. In biomedical applications, the number of existing labeled (annotated) sequences in many domains is usually small while a large number of unlabeled sequences are available. In addition, sequence motifs are often embedded in short sequences (known as short sequence intervals or sequence fragments) which contain a few dozen instead of a few hundred

residues [Jr. and Liang, 2010, Mehdi et al., 2013], and frequently have mutations in their pattern [Narang et al., 2010]. These characteristics have brought challenges for motif finding problem.

The research on discriminative motif learning has recently developed pattern discovery methods using HMM (*hidden Markov model*) [Lin et al., 2011], PWM (*position weight matrix*) [Kim and Choi, 2011, Bailey et al., 2010, Redhead and Bailey, 2007], and association mining with domain knowledge [Vens et al., 2011]. Due to their general purposes, these methods have showed to be ineffective when input sequences do not provide enough information to discover discriminative patterns. The main reason of this limitation is that these methods try to obtain PWM and HMM from the training data, therefore it is very difficult for them to learn the best probabilities of patterns if the training data are short in length and small in number.

In the case of our study on hepatitis C virus, a study in 4 consecutive projects between our laboratory and Chiba University with funding from JSPS on computational methods [Ho, 2011, Ho, 2007, Ho, 2004, Motoda, 2001], we focus on hepatitis pathogenesis and therapy by using NS5A (*non structure 5A*) protein, where we are interested in discriminating two classes of SVR (*sustained virologic response*) sequences and non-SVR (*non sustained virologic response*) sequences, from the biggest resource of LANL[1] database, we can only get 134 non-SVR sequences and 93 SVR to IFN/RBV therapy, and 13 non-SVR sequences and 12 SVR sequences from Chiba University, but from Genbank[2] and HVDB[3] databases, we obtain about 5000 NS5A unlabeled sequences.

The combination of IFN/RBV (*interferon and ribavirin*) is currently the standard therapy for HCV (*hepatitis C virus*). However, this therapy is often accompanied by side effects and only fewer than half of the HCV infected individuals achieve sustained viral response by this therapy, especially HCV-1b (*hepatitis C virus genotype 1b*), see for example [Gao et al., 2010, Hoofnagle, 1994]. Many studies have reported that the NS5A in HCV genome is known as the protein implicated in the interferon resistance, and thus much effort has been made to pursue uncovering such resistance mechanisms. Furthermore, several studies have suggested biomarkers for explanations of the resistance to IFN/RBV therapy of HCV to get a deeper understanding how HCV escape from the immune system, such as mutations in a part of the NS5A of HCV [Enomoto et al., 1996], the relation between gene expression and viral level [Murakami et al., 2010, Brodsky et al., 2007], the relation of NS5A protein and TLRs (*toll-like receptors*) [Imran et al., 2012], the variation of IL-28B (*interleukin-28B*) gene [Alestig et al., 2011] and so on. Among these studies, the correlation between NS5A protein and IFN/RBV therapy has

[1]Los Alamos National Laboratory http://hcv.lanl.gov

[2]Genbank http://www.ncbi.nlm.nih.gov/genbank

[3]Hepatitis Virus Database http://s2as02.genes.nig.ac.jp

been reported in numerous paper in biomedicine field [El-Shamy et al., 2011, Guilou-Guillemette et al., 2007, Pascu et al., 2004, Sarrazin et al., 1999, Enomoto et al., 1996], as well as in computational field [ElHefnawi et al., 2010, Aurora et al., 2009, Witherell and Beineke, 2001]. However, the understanding of inhibitions of the HCV NS5A protein with IFN/RBV therapy is still deeply unknown and no clear adaptation patterns to the antiviral treatment were detected [Gao et al., 2010, Cuevas et al., 2009].

One more difficulty in study of NS5A and IFN/BRV therapy is that ISDR (*interferon sensitivity determining region*) sequences tend to resemble one another. ISDR is a small part of NS5A protein consisting of 40 amino acids and has been widely discussed for its correlation with IFN resistance and response. Because two main functions of NS5A protein are supposed to replicate for HCV and resist IFN activity, the polypeptide sequence of NS5A protein should be preserved and has few variants at some positions. Therefore, short and similar characteristics of sequences pose serious drawbacks for traditional methods.

In summary, how to know in advance the signals of response or resistance to IFN/RBV therapy, also known as SVR or non-SVR before the treatment to be able to save the pain and expense for patients is necessary and important to HCV study. We are motivated by the four following points: (1) why is the rate of INF/RBV response low? The SVR rate has achieved less than 50%, especially with HCV genotype 1b [Gao et al., 2010]; (2) what are NS5A biomarkers for SVR and non-SVR outcomes? And what are links among them?; (3) the labeled data in reality are small in number, short in length and similar to each other; (4) the current discriminative motif finding methods are ineffective when input data are not enough information. Therefore, our study objectives are set to develop efficient and effective computational methods for discovering discriminative motifs in two cases: few labeled data and short sequences; and develop a semi-supervised ensemble learning method to exploit a large unlabeled data in order to improve the quality and accuracy of discriminative motifs. To achieve these objectives, we propose computational processes based on well-founded machine learning techniques, address the rational uses of popular methods, as well as apply biomedical background to the proposed methods.

1.2 Problem formulation

Finding motifs having a descriptive ability for a set of sequences is to find common properties and characteristics of that sequence set. And finding discriminative motifs is not only to find descriptive properties of a sequence set but also to find discriminative properties so that these motifs can differentiate this sequence set from other sequence sets. This searching is the main purpose of discriminative motif finding problem. The current discriminative motif learning typically involve building PWM or HMM from sequences

and then using techniques such as EM (*expectation maximization*) or Gibbs sampling to optimize the likelihood of PWM or HMM, and thus do not guarantee to find the global solution. Because PWM and HMM are normally obtained from training data, previous methods require a large number of labeled data to learn the best PWMs and HMMs. Therefore, these methods have shown their limitations when work with the small number of labeled data, short and similar sequences.

In studies of the relation between NS5A protein and IFN/RBV therapy, so far, many biomarkers are proposed to explain IFN/RBV resistance or response of HCV, for example the number of point mutations [Enomoto et al., 1996, Sarrazin et al., 1999, Pascu et al., 2004], type and position of point mutation [Torres-Puente et al., 2008, Witherell and Beineke, 2001] in NS5A region; and no clear patterns to IFN/RBV therapy are detected [Cuevas et al., 2009]. After analyzing the current HCV study, we set our goal to discover discriminative motifs such that these motifs can help to distinguish SVR and non-SVR results of HCV treatment. Once these motifs are found, they will be suggestions or hints for HCV treatment that should be confirmed by experimental works.

To this end, we formulate our research problem to 3 main tasks: (i) develop a discriminative motif learning method for few labeled data, (ii) develop a discriminative motif learning method for short and similar sequences and (iii) apply these methods mentioned above in order to discover new knowledge in biomedical field.

With a small labeled dataset, it is difficult to be able to discover discriminative motifs that have an overall representation for a class and discriminate clearly from other classes. Therefore, in the first task, we develop an appropriate method to find effectively discriminative motifs from two sequence datasets, a limited number of labeled sequences and a large number of unlabeled sequences.

In the second task, in order to deal with the lack of information from short and similar sequences, we first find a way to enrich the representation for sequences, then we develop an advanced machine learning method based on topic model to characterize and predict sequences.

In the third task, HCV treatment results are categorized into two cases: SVR (or IFN/RBV response) and non-SVR (or IFN/RBV resistance), we therefore apply the above two methods to find discriminative motif in a set of two-class sequences. In other words, given a set of two-class sequences, the problem is to find discriminative motifs that help to classify well a sequence into a certain class. These potential and promising motifs present many patterns that were not known previously and need preliminary assessments by physicians.

1.3 Major contributions

Our target is to develop new computational methods for discriminative motif discovery in order to understand comprehensively the relation between NS5A protein and IFN/RBV therapy. Therefore, our contributions are two-fold: *biomedical findings* and *computational methods* summarized concretely in each following situation:

Discriminative motif learning for few labeled data. In this work, we propose a semi-supervised ensemble method for finding discriminative motifs which is based on a separate-and-conquer searching method. Our method firstly searches core motifs from a small labeled dataset, then used these motifs to exploit unlabeled data, and continues searching discriminative motifs with the enlarged labeled dataset. The experimental results show the accuracy of the proposed framework is improved by 8% and discriminative motifs with high accuracies from 80% to 100% found by our new method are able to discriminate better than discriminative motifs of MEME and DEME methods. These motifs, when verified by physicians, lead to a better understanding of the resistance or response to IFN/RBV therapy of HCV.

This work was reported at 4th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2012) and published in the Journal of Universal Computer Science, Special Issue on Hybrid and Ensemble Methods in Machine Learning (HEMML 2012), Vol. 19, Issue 4, pp. 563-580, April 2013.

Discriminative motif learning for short sequences. In this joint work, I contributed the way to represent optimally for input sequences, the comparison of experimental results with other methods, the interpretation for subsequence pattern, the connection between subsequence patterns and HCV genotype 1b, and the experimental performance. Our framework shows its effectiveness through the prediction quality being often higher than the quality of the baseline method, about 30% improvement. Furthermore, characteristics of HCV treatment outcomes we obtained are good discriminative motifs helping to predict signals of response or resistance to HCV therapy. We believe these potential findings provide additional knowledge to studies of HCV treatment as well as viral sequence variation studies. (This work is under consideration to submit).

Application to HCV study. Applying two proposed methods, we believe to contribute strong discriminative motifs, the new additional knowledge, to studies of HCV treatment as well as studies of viral sequence variation. And obviously, we need time to verify these motifs by experimental works.

The scientific significances of our study are in the achievement of understanding of the relation between NS5A protein and IFN/RBV therapy at a molecular biology level and in the novel computational methods that are proposed. In a context of Knowledge

Science, namely Knowledge Media, we address a new computational process for creating new knowledge automatically by using computer algorithms and then using this knowledge to classify the data into different categories in biomedical field. Our biomedical findings (discriminative motifs) are patterns and regularities in data that have not been discovered before. These findings are created in a reliable computational process, judged by extensive experiments and on the way to be verified as new knowledge. Therefore, our study contributes to Knowledge Science are the new knowledge for biomedical field and computational methods for knowledge creation.

1.4 Dissertation organization

The dissertation is organized into five chapters, as follows:

Chapter 1 introduces the research problem and its formulation. This chapter also states our main contributions in term of biomedical findings and computational methods.

Chapter 2 presents the background of the dissertation. We present studies of HCV, NS5A protein and IFN/RBV therapy. Motif and discriminative motif finding problems of sequence characterization are discussed.

Chapter 3 describes our computational methods for discovering discriminative motifs in situation of few labeled data. First, we present how to find discriminative motifs in a small training dataset. And then we show how to exploit the huge unlabeled dataset in order to enlarge the small training dataset. Next, we present the application to HCV treatment for this problem. Finally, the findings and their significance are discussed.

Chapter 4 presents a new method to discover discriminative motifs for short sequences. The method consists of four main steps. Data representation is the first step. The second step is data transformation into a discriminative space. And the next step is data projection onto this discriminative space. The final step performs prediction and analysis. We also discuss our new findings and their significance in biomedical field.

Chapter 5 concludes the dissertation by summarizing the major contributions, achievements, and limitations of our work. We also talk about open problems for future research on this topic.

Chapter 2

Background

2.1 Fundamental of motif discovery

2.1.1 Motif definition

A *sequence motif* is generally understood as a subsequence of sequences that is widespread and biologically significant [Sami and Nagatomi, 2008]. For examples, with DNA sequences, motifs can be TFBSs (*transcription factor binding sites*) in the promoter regions; with protein sequences, motifs can be regions corresponding to a specific function or structure, or it can be signals playing an important role in controlling the cellular localization [Vens et al., 2011]. The sequence motif usually is short, from 5 to 20 bp (*base-pairse*) long, and is supposed to repeat many times in a sequence [Das and Dai, 2007] (Figure 2.1).

In motif discovery, we often use a popular assumption to find motifs. This assumption is stated that the significant regions are better preserved during the evolution because of their importance in terms of structure and/or function of the molecule, and thus that they appear more frequently than it is expected [Nevill-Manning et al., 1998]. Motifs can be classified into two main types: (1) *simple motif*, no variable gaps are allowed in the motif, and (2) *structured motif or composite motif*, variable gaps are allowed in a motif, in other words, structured motif is a pair of simple motifs separated by a variable but restricted distance [Sinha, 2003].

2.1.2 Motif representation

A sequence motif can be represented by either (i) a string-based model or (ii) a probabilistic model. A string-based model represents a motif as a sequence of letters that may

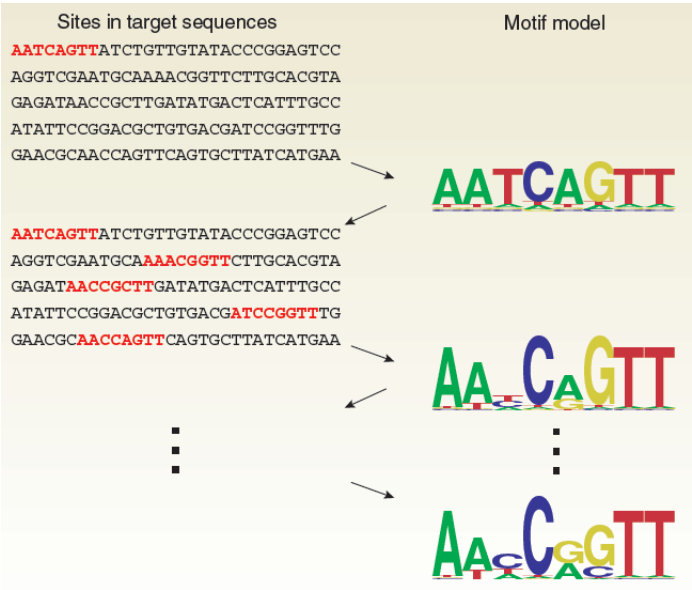

Figure 2.1: DNA motifs [D'haeseleer, 2006].

contain special characters to increase the variability of the motif. Among probabilistic models, PWM (*position weight matrix*) and HMM (*hidden Markov model*) are the most commonly used models to represent motifs. PWM considers a motif as a matrix in which each element has the probability of a given nucleotide or amino 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 character only depends on its previous state with the assumption that these states are not necessarily independent [Wu and Xie, 2010].

1. PWM (*position weight matrix*): PWM specifies a score for each base or amino acid at each position in the motif, assuming independence between positions in the motif. An example of PWM is described as follows: Given a set of aligned sequence, we can construct a profile matrix characterizing a motif having 8 bases in length for example. Each element represents the probability of given character at a specified position (Figure 2.2).

   Another example is described more concretely: Given a sequence 'GCCGCCCTTTC-CTCTTTCTTTCGCGCTCTAGCCACCCGG'. PWM of a motif that has 12 bases in length, 'TTTCGCGCTCTA'. We can see that the first position of motif is more likely to be a 'T' since the probability is 0.759, the highest probability in the first column (Figure 2.3)

   In order to know whether a subsequence is a motif or not, PWM will calculate a likelihood of a sequence given a motif with starting position (Figure 2.4).

2. HMM (*hidden Markov model*): The probability of the current state only depends on

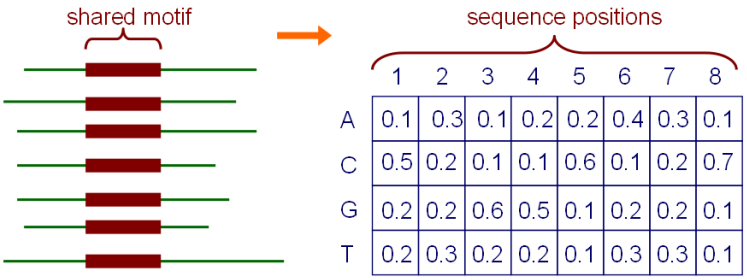

Figure 2.2: An example of PWM [Craven, 2011].

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.069	0.034	0	0.034	0	0	0	0	0.034	0.517	0.448	0.448
C	0.103	0.035	0.034	0.448	0.103	1	0	0.759	0.276	0.31	0.035	0.138
G	0.069	0.069	0.069	0.517	0.897	0	1	0.241	0.586	0.104	0.241	0.31
T	0.759	0.862	0.897	0.001	0	0	0	0	0.104	0.069	0.276	0.104

Figure 2.3: An example of PWM [Wu and Xie, 2010].

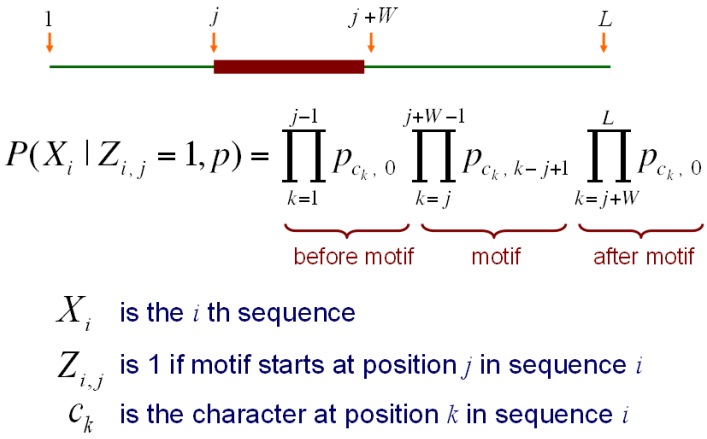

Figure 2.4: A formula to calculate the likelihood of a sequence motif in PWM [Craven, 2011].

its previous state (states are not necessarily independent). These states are hidden since we only observe the letters but not theirs states. Hidden Markov Model includes below components:

(a) The sequence of observations $O = \{o_1, ..., o_L\}, o_l \in \{A, C, G, T\}, l = 1..L, L$ is the length of sequence.

(b) The sequence of the hidden states $X = \{x_1, ..., x_L\}, x_l \in \{0, ..., M\}, M + 1$ is the total number of states.

(c) The transition probability matrix of the hidden state $Q = \{q_{ij}\}, i, j = 0, ..., M$, $q_i j = P(x_{(l+1)} = j | x_l = i)$.

(d) The initial probability of the states $\pi = \pi_0, ..., \pi_M, i = 0, ..., M$.

(e) The probability of observing letters o_l from each state i, $E = P(o_l | x_l = i)$ is known as the emission probability.

A following example is illustrated for HMM: Suppose we know that a sequence contains a motif whose length is 12. Then we have 13 hidden states. Specifically, one state in background is denoted by 0, and one state for each position of motif is denoted from 1 to 12, where E represents the 12 motif states. We view the letters in the top line as observations from the hidden path in the bottom line.

Observation: GCCGCCCTTTCCTCTTTCTTTCGCGCTCTAGCCACCCGG

Hidden state: 000000000000000000EEEEEEEEEEEE000000000

Suppose we know that the motif is likely to follow the background state with probability 0.01 and no two motifs are next to each other. Transition probability matrix will be as Figure 2.5, where $q_{0,1} = 0.01$ (from background state to motif state), $q_{0,0} = 1 - q_{0,1} = 1 - 0.01 = 0.99$ (from background state to background state), $q_{(i,i+1)} = 1, i = 1, .., 11$ (from motif state to motif state), $q_{12,0} = 1$ (from motif state to background state), and $q_{i,j} = 0$ for all other i, j, and we assume no two motifs are next to each other.

The emission probabilities of the observations from the background state is uniform, $P(o_l = A | x_l = 0) = P(o_l = C | x_l = 0) = P(o_l = G | x_l = 0) = P(o_l = T | x_l = 0) = 0.25$.

The emission probabilities for the observation from states 1-12 are given by the PWM (Figure 2.3).

Suppose we know the first position of the given sequence is not a motif then we have the initial probabilities $\pi_0 = 1$ and $\pi_1 = \pi_2 = \pi_1 2 = 0$.

Then the joint probability of an observed sequence O and a state sequence x is $P(O, x) = \pi_{x_1} P(o_1 | x_1) \prod P(o_l | x_l) q_{x_{l-1}} x_l)$.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.99	0.01	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	0	0	1
12	1	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2.5: An example of transition probability matrix in HMM.

2.1.3 Motif finding algorithm

Based on motif representational models, motif finding algorithms are categorized into two major groups [Das and Dai, 2007]: string-based methods that mostly rely on exhaustive enumeration and probabilistic methods that parameters of the motif model are estimated using maximum likelihood principle or Bayesian inference.

1. String-based method

   In this method, motifs are found by enumerative algorithms that cover exhaustively the space of all possible motifs, for a specific motif model description [D'haeseleer, 2006]. The key idea of enumerative methods is that frequencies of each nucleotide or amino acid in sequences will be counted and compared to find the most overrepresented subsequences.

   The work of [Tompa, 1999] is an example of enumerative methods. For each length$-k$ sequence s, the number N_s of sequences containing an occurrence of s is recorded. This occurrence allows for a small and fixed number c of substitution residues in s. Then, a measure of s as a motif would be based on how unlikely it is to have N_s occurrences if the sequences were drawn at random according to the background distribution.

   The statistical significance test for motif occurrences is proposed by [Tompa, 1999] as follows: Let X be a single random sequence of the specified length L, with residues drawn randomly and independently from the background distribution. Supposed

Algorithm 1 EM algorithm, where p is a matrix of probabilities of each character, Z is a matrix of probabilities that a motif starts in a position of a sequence

Given: length parameter W, training set of sequences

1: $t = 0$
2: set initial values for $p^{(0)}$
3: **repeat**
4: $++t$
5: re-estimate Z^t from p^{t-1} (E-step)
6: re-estimate p^t from Z^t (M-step)
7: **until** change in $p^{(t)} < \varepsilon$

return: $p^{(t)}, Z^{(t)}$

that p_s is the probability that X contains at least one occurrence of the length-k sequence s allowing for c substitutions. Under the reasonable assumption that random sequences of X are independent, the expected number of containing at least one occurrence of s among the N random sequences is N_{p_s}, and its standard deviation is $\sqrt{N_{p_s}(1-p_s)}$. Therefore, the associated z-score is $M_s = \frac{N_s - N_{ps}}{\sqrt{N_{ps}(1-p_s)}}$, where M_s is the number of standard deviations by which the observed value N_s exceeds its expectation, and it is called the "z-score".

2. Probabilistic method

   (a) EM (*Expectation Maximization*). EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state* [Lawrence and Reilly, 1990]. In the motif finding problem, the hidden state is where the motif starts in each training sequence. The EM algorithm iteratively computes the expectation of the missing data (E-step) and maximizes the exptected hidden log-likelihood of the data (M-step) illustrated as in Algorithm 1.

   In practice, PWM does not only represent the probabilistic of each residue, but also represent the "background", i.e. sequence outside the motif. An example of motif representation in MEME (*multiple EM for motif elicitation*) tool has shown in Figure 2.6.

   (b) Gibbs sampling. The EM approach can get trapped in local minima because the PWM is generated at the beginning of the EM algorithm. One approach to alleviate this limitation is that we try different initial parameters. Gibbs sampling is an alternative to the EM approach and it can exploit randomized search to a much greater degree. Therefore, we can view Gibbs sampling as a stochastic analog of the EM algorithm. In the EM approach, we maintained a distribution Z_i over the possible motif starting points for each sequence. But

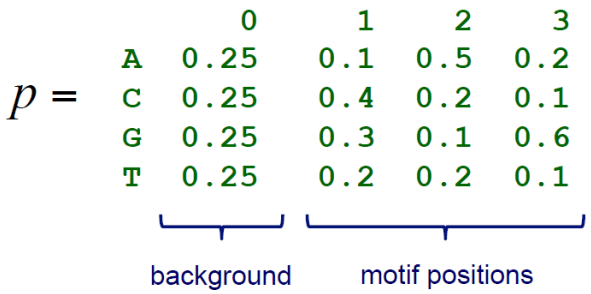

Figure 2.6: An example of length-3 motif representation in MEME [Craven, 2011].

Algorithm 2 Gibbs sampling algorithm, where p represents th probability of being in state u at any given time in a random walk on the chain and the specific stating point a for each sequence

Given: length parameter W, training set of sequences

choose random position for a
repeat
    pick a sequence X_i
    estimate p given current motif positions a
    (using all sequences but X_i) (predictive update step)
    sample a new motif position a_i for X_i (sampling step)
until convergence

return: p, a

in the Gibbs sampling approach, we will maintain a specific starting point for each sequence a_i and we will keep randomly resampling these. The basic Gibbs sampling approach is illustrated as in Algorithm 2.

We can view the motif finding in terms of a Markov Chain Monte Carlo, where "Markov Chain" is the results from every step depends only on the results of the preceding one (like in EM) and "Monte Carlo" is the way to select the next step is not deterministic but rather based on random sampling [Das and Dai, 2007]. Each state represents a configuration of the starting positions and transitions correspond to changing selected starting positions.

2.1.4 Discriminative motif learning

Motif learning is the problem that given a set of sequences thought to contain unknown motifs of interest, then two main tasks for inferring a model of motifs and predicting the locations of motifs in those given sequences are performed. Finding motifs in a class of sequences is to find motifs that share a certain characteristic, such as motifs containing

a large number of wildcard symbols [Hsu et al., 2011], degenerate motifs [Vens et al., 2011], conserved motifs, and so on. However, sequence motifs are usually short and can be highly variable patterns [Redhead and Bailey, 2007], and it is difficult to distinguish them from random patterns that are likely to occur by chance [Bailey et al., 2010]. This has led to a new approach utilizing an additional class of sequences to guide the motif finding process to come near to specialized motifs that we want to seek in one class of sequences, or go far away from other motifs in the other class of sequences. Using the second class of sequences can help to distinguish motifs from randomly occurrences, because it provides additional information to compare and then eliminate early motifs that are overrepresented by chance. Therefore, finding motifs with a set of two-class sequences has opened a new view of discriminative motif finding.

Discriminative motif finding problem is to find motifs occurring more frequently in one set of sequences and not occurring in the other set of sequences. These motifs can help to classify effectively a sequence into a certain class or to describe the discriminative characteristics of a class.

The probabilistic models for motif discovery can be classified as being of one of three types by the assumption on the number of binding sites per sequence [Kim and Choi, 2011]. Then, the models of each class can be further classified into two different versions of discriminative and non-discriminative, resulting six different models.

1. OOPS (*one occurrence per sequence*): Only one subsequence of each sequence is generated from the motif model.

2. ZOOPS (*zero or one occurrence per sequence*): The ZOOPS model is an extension of the OOPS model by allowing each sequence to have at most one binding site.

3. MOPS (*multiple occurrence per sequence*): The MOPS model is further extended from the ZOOPS model by allowing any number of binding sites.

4. DOOPS (*discriminative one occurrence per sequence*): The DOOPS model is derived from the OOPS model and is employed a discriminative learning.

5. DZOOPS (*discriminative zero or one occurrence per sequence*): The DZOOPS model can be derived from the DOOPS model by introducing the true class label determining whether a positive sequence contains a binding site or not.

6. DMOPS (*discriminative multiple occurrence per sequence*): The DMOPS model is derived from the MOPS model and is employed a discriminative learning.

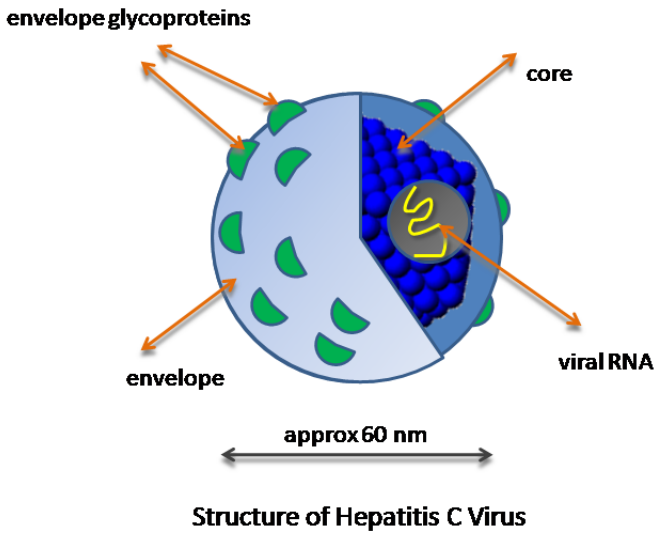

Figure 2.7: Simplified diagram of the structure of hepatitis C virus [Colm, 2008].

2.2 Fundamental of HCV study

2.2.1 Hepatitis C Virus

HCV (*hepatitis C virus*) is an enveloped, approximately 9600 nucleotides, single-stranded RNA (*ribonucleic acid*) virus and is classified in the family *Flaviviridae* [Choo et al., 1989]. HCV is small in size, about 55 - 65 nm (*nanometre*) and its particle consists of a core of genetic material (RNA), surrounded by an icosahedral protective shell of protein, and further encased in a lipid envelope of cellular origin [Beeck and Dubuisson, 2003]. Figure 2.7 illustrates the basic structure of HCV.

A great progress has been made in the study of HCV over the past 18 years using heterologous expression systems that enable the study of viral entry under reproducible and conveniently measurable conditions and complete cell-culture systems [Moradpour et al., 2007]. Selected milestones in HCV research are shown in Figure 2.8.

1. *Genome*: The genome consists of 5'-NTR (*non-translated region*), which includes an IRES (*internal ribosome entry site*), a single open reading frame that is translated to produce 10 active proteins, and a 3'-NTR. As illustrated in Figure 2.9, three structural proteins are termed core, E1 (*envelope 1*) and E2 (*envelope 2*), and a protein maned p7; and six NS (*nonstructural*) proteins are termed NS2, NS3, NS4A, NS4B, NS5A and NS5B. The 5' and 3' NTR are not translated into proteins, but are important to translation and replication of the viral RNA.

   (a) Core protein: The core protein has 191 amino acids and can be divided into

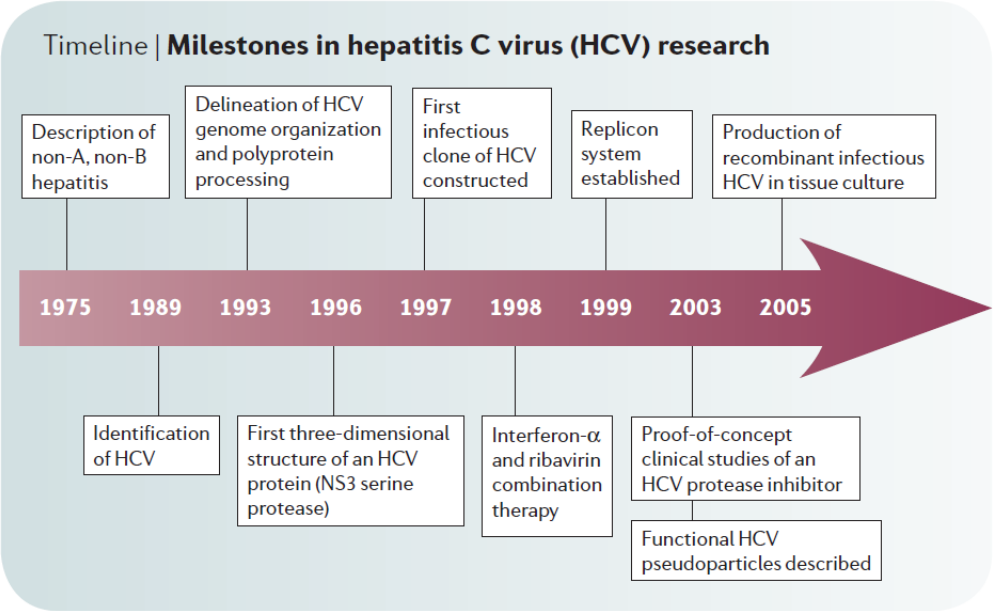

Figure 2.8: Milestones in hepatitis C virus research [Moradpour et al., 2007].

three domains on the basis of hydrophobicity,

i. Domain 1 contains mainly basic residues with two short hydrophobic regions (residues 1 - 117)

ii. Domain 2 is less basic and more hydrophobic (residues 118 - 174)

iii. Domain 3 is highly hydrophobic and acts as a signal sequence for E1 envelope protein (residues 175 - 191)

(b) E1 and E2 proteins: Both envelope proteins are highly glycosylated and important in cell entry. E1 serves as the fusogenic subunit and E2 acts the receptor binding protein.

(c) p7 protein: The p7 protein consists of 63 amino acids and is a spanning membrane that locates in the ER (*endoplasmic reticulum*). This protein is dispensable fro viral genome replication but plays a critical role in virus morphogenesis.

(d) NS2 protein: The NS2 protein is a 21 - 23 kDa (*kiloDalton*) transmembrane protein related to protease activity.

(e) NS3 protein: The NS3 protein is a 67 kDa protein whose N-terminal has serine protease activity and whose C-terminal has helicase activity. It is located within the ER and forms a heterodimeric complex with NS4A.

(f) NS4A protein: The NS4A protein is a 54 amino acid membrane protein that acts as a cofactor of the proteinase.

(g) NS4B protein: The NS4B protein is a small (27 kDa) hydrophobic integral membrane protein that contains 4 transmembrane domains. This protein is

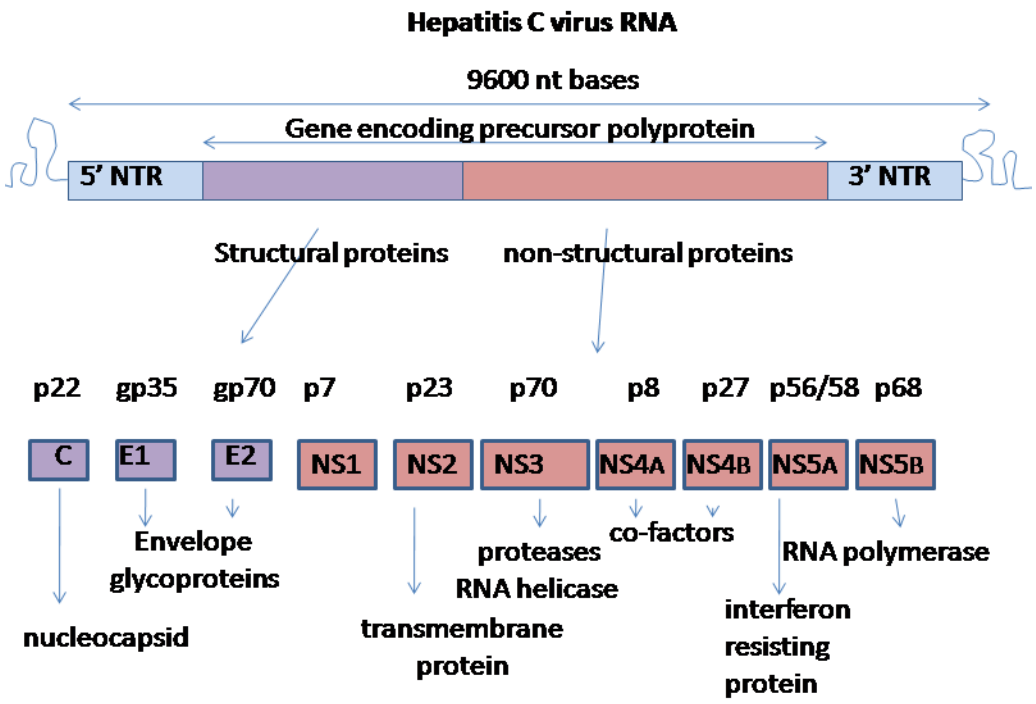

Figure 2.9: Simplified diagram of the structure of hepatitis C virus [Colm, 2008].

located within the ER and plays an important role for recruitment of other viral proteins. It induces morphological changes to the ER forming a structure termed the membranous web.

(h) NS5A protein: The NS5A protein is a hydrophilic phosphoprotein which plays an important role in viral replication, modulation of cell signaling pathways and the interferon response.

(i) NS5B protein: The NS5B protein is the viral RNA-dependent RNA polymerase (RdRp). This protein has the key function of replicating the HCV by using the viral positive RNA strand as its template and catalyzes the polymerization of rNTP (*ribonucleoside triphosphatess*) during RNA replication.

2. *Genotype*: A genotype is a classification of a virus based on the genetic material in the RNA strands of the virus. HCV is divided into 6 distinct genotypes with multiple subtypes in each genotypes class based on the genomic sequence heterogeneity. Following is a list of the different genotypes of HCV summarized by [Simmonds et al., 2005]:

   (a) Genotype 1a, 1b and 1c

   (b) Genotype 2a, 2b, 2c and 2k

   (c) Genotype 3a, 3b, 3c, 3d, 3e and 3f

(d) Genotype 4a

(e) Genotype 5a

(f) Genotype 6a, 6b, 6d, 6g, 6h and 6k

Genotype 1 - 3 have a worldwide distribution. Subtypes 1a and 1b are the most common, accounting for about 60% of global infections [WHO, 2012]. They predominate in Northern Europe and North America, and in Southern and Eastern Europe and Japan. respectively. Genotype 2 is less frequently represented than genotype 1. Genotype 3 is endemic in South-East Asia and is variably distributed in different countries. Genotype 4 is principally found in the Middle East, Egypt and central Africa. Genotype 5 is almost exclusively found in South Africa, and genotype 6 is distributed in Asia.

3. *Replication*: HCV replicates mainly in the hepatocytes of the liver, where it is estimated that daily each infected cell produces approximately 50 virions (virus particles) with a calculated total of one trillion virions generated. HCV probably follows the replication strategy of other positive-strand RNA viruses. Its replication process consists of the following steps:

   (a) Virus binding and internalization

   (b) Cytoplasmic release and uncoating

   (c) IRES (*internal ribosome entry site*) mediated translation and polyprotein processing

   (d) RNA replication

   (e) Packing and assembly

   (f) Virion maturation and release

   The life cycle of HCV is illustrated in Figure 2.10.

2.2.2 NS5A protein

NS5A is a nonstructural protein of HCV which is the protein most reported to be implicated in the interferon resistance [Guilou-Guillemette et al., 2007]. NS5A protein is chemically bonded to a substance containing phosphoric acid. This protein is predicted to be predominantly hydrophilic and to contain no transmembrane helices [Macdoanldt and Harris, 2004].

1. *Structure*: The structural features of NS5A protein have been derived experimentally and four regions of interest are shown in Figure 2.11.

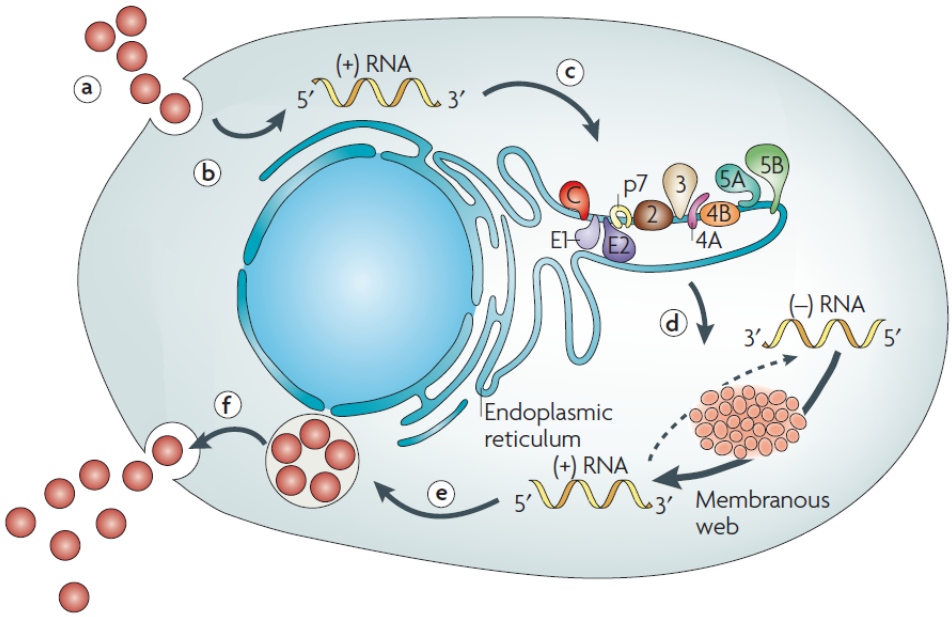

Figure 2.10: Hepatitis C Virus life cycle [AIDSInfoNet, 2012].

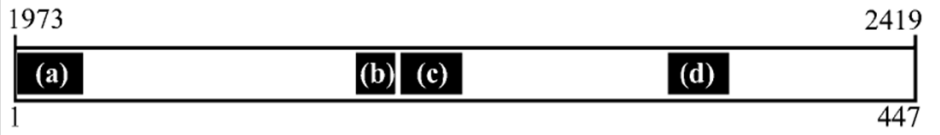

Figure 2.11: The NS5A protein structure [Macdoanldt and Harris, 2004].

(a) N-terminal amphipathic helix: This sequence contains 30 amino acid that were predicted to form a highly conserved amphipathic alpha helix, and were shown to be both necessary and sufficient to mediate association of NS5A with the ER (*endoplasmic reticulum*).

(b) Hyperphosphorylation cluster: In this region, 4 sites of serine were identified as sites of hyperphosphorylation, they are Ser2194, Ser2197, Ser2201, and Ser2204.

(c) Interferon sensitivity determining region: This region was originally associated with resistance or sensitivity of viral isolates to IFN treatment.

(d) Polyproline cluster: This region contains two closely spaced proline-rich motifs. Two classes of these motifs have been defined: class 1 (consensus sequence KxxPxxP) and class 2 (consensus sequence PxxPxR).

Depending on the genotype, NS5A protein varies in length [Guilou-Guillemette et al., 2007] as in the Table 2.1 below.

2. *Function*: NS5A is supposed to have two main functions, virus replication and interferon resistance.

Table 2.1: Lengths of the NS5A protein.

Genotype	Length (amino acid)
1a and 1c	448
1b	447
2a and 2b	452
3	452
4a	445
5a	540
6a	451

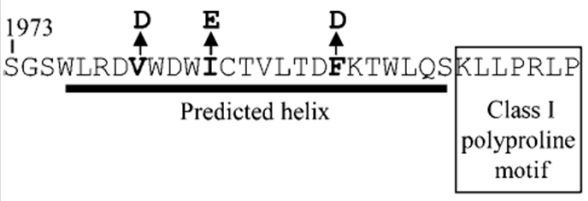

Figure 2.12: N-terminal amphipathic helix [Macdoanldt and Harris, 2004]. The evidence for functions of NS5A in virus replication.

(a) Functions of NS5A in virus replication: NS5A is a part of a multi-protein, membrane-bound replication complex. Together with other nonstructural proteins, NS5A co-localized with viral RNA in a cytoplasmic membrane structure termed the membranous web, the generation of which required the NS4B protein. Replicon cell membrane fractions isolated by differential centrifugation contained both p56 (*56-kDa*) and p58 (*58-kDa*) forms of NS5A and were competent for synthesis of HCV RNA in vitro. A critical role of NS5A came from experiments in which the amphipathic membrane-targeting helix was mutated in the context of he replicon. Introduction of three helix-disrupting mutations (Figure 2.12) implies that NS5A membrane association is an indispensable event during HCV RNA replication [Macdoanldt and Harris, 2004].

(b) Interferon resistance: The interferon resistance of NS5A is supposed in two evidences: (i) the interaction of NS5A with cellular interferon pathway and (ii) the variation with mutations in NS5A.

Interaction of NS5A with cellular interferon pathway: HCV resistance to IFN could be explained by the ability of NS5A protein to bind and inhibit PKR (*protein kinase R*) [Gale et al., 1997] and IL-8 (*interleukin-8*) [Polyak et al., 2001] protein. The cellular interferon pathway (Figure 2.13) is one of the cell signaling pathways in human in which Jak/STAT (*janus kinase - signal trans-*

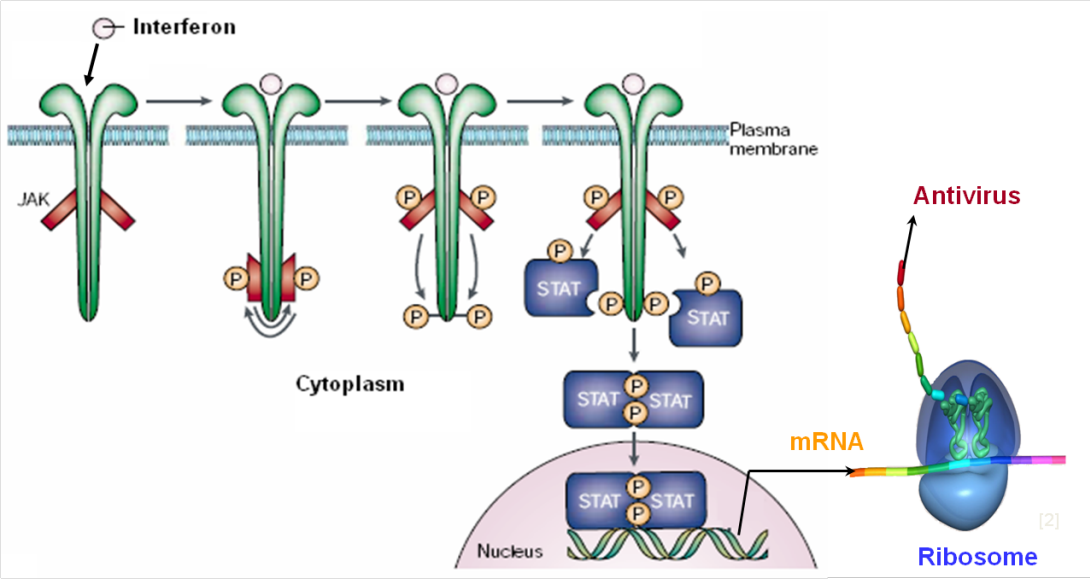

Figure 2.13: Interferon signaling pathway [Suhalim, 2007].

ducer and activator of transcription) signaling pathway of interferon system is an example. In this pathway, the double-stranded RNA-activated protein kinase R, an enzyme that are induced by IFN and has antiviral properties in the cell, is a key regulator of the innate immune response. And the NS5A protein has been suggested to balance the interferon cellular antiviral pathway and to be involved in the resistance to the interferon based-therapy [Guilou-Guillemette et al., 2007]. It has been demonstrated in vitro that NS5A induces the expression of the pro-inflammatory chemokine interleukin 8 at both the mRNA (*messenger RNA*) and protein levels [Guilou-Guillemette et al., 2007]. Note that signaling pathway is defined as a group of molecules in a cell works together to control one or more cell functions. After the first molecule in a pathway receives a signal, it activates another molecule. And this process is repeated until the last molecule is activated and the cell function involved is carried out.

Mutations in NS5A: The relationship of mutations in the NS5A protein and the IFN resistance were reported to happen in 2 small subregions of the NS5A protein. The first 40 amino acids of PRK binding domain (residues 2209 - 2248) present a high level of variability. This subregion is called ISDR (*interferon sensitivity determining region*). Mutations in ISDR are involved in response and resistance to IFN [Enomoto et al., 1996, Pascu et al., 2004]. An increasing number of mutations in ISDR is an increasing probability of SVR to IFN therapy. The next 27 amino acids of C-terminal (*carboxy-terminal*) region (residues 2353-2379) of NS5A has a high variability level. This subregion is called V3 domain. A accumulation of muatations around V3 within the C-

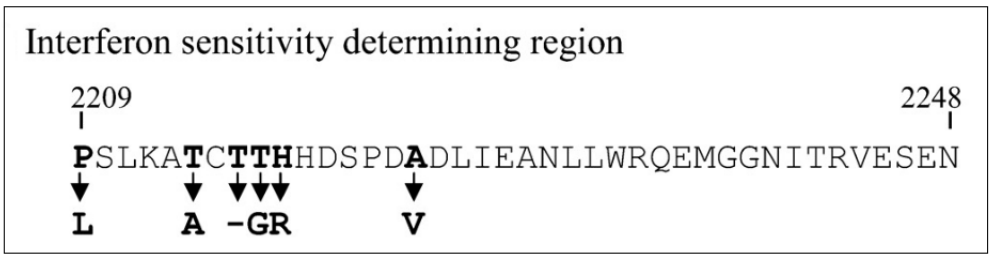

Figure 2.14: Interferon sensitivity determining region (ISDR) [Macdoanldt and Harris, 2004]. The ability of IFN/RBV resistance can be eliminated by the mutations indicated in bold.

terminal part of the NS5A protein correlates with treatment response in HCV. Note that a *mutation* is a change in a genomic sequence, for example DNA sequence of the genome of a cell, or DNA (*deoxyribonucleic acid*) or RNA (*ribonucleic acid*) sequence of a virus. Mutations cause effects on the structure and function of protein sequences.

2.2.3 ISDR of NS5A protein

ISDR (*Interferon sensitivity determining region*) was first identified a stretch of 40 amino acids in the center of NS5A protein related to the total number of substitutions by [Enomoto et al., 1996]. This sequence region is associated with either resistance or sensitivity to IFN treatment or IFN plus RBV combination therapy which is collected and reported in [Chayama and Hayes, 2011]. An example of ISDR sequence is shown in Figure 2.14.

In addition, ISDR is able to bind and inhibit PKR (*protein kinase R*), an IFN-induced protein produced by human immune system during a virus genome replication [Macdoanldt and Harris, 2004]. Therefore, the ISDR binding site has been suggested to involve in the resistance to IFN/BRV therapy [Guilou-Guillemette et al., 2007].

2.2.4 HCV therapy

HCV infection is a major factor leading to the progression of liver cirrhosis and the development of hepatocellular carcinoma [Saito et al., 1990]. It is estimated that 150 million people worldwide are chronically infected with HCV and more than 350,000 people die from HCV-related liver disease each year [WHO, 2012]. The combination therapy with IFN/RBV (*interferon and ribavirin*) is currently the standard and effective treatment for chronic HCV infection [Manns et al., 2001].

1. *Interferon*: IFN is a member of the cytokine family produced by cells in response to viral infections or other stimuli by binding to their specific receptors on the surface of target cells. IFN comes in (i) *the standard form*, is administered 3 times in a week and (ii) *the pegylated form*, is administered once in a week. Types of IFN conclude:

   (a) Type I: IFN- α, IFN- β, and IFN- ω are secreted in response to viral infection by various cell types.

   (b) Type II: FN- γ is induced by mitogenic or antigenic stimulation of the immune system in activated T cells and crophages.

2. *Ribavirin*: RBV is a guanosine analogue that was synthesized more than 35 year and that possesses broad-spectrum antiviral activity against several RNA and DNA viruses in vitro.

3. *IFN/RBV therapy*: The current standard treatment for patients with chronic hepatitis C consists of pegylated alpha interferon and nucleoside analogue ribavirin for 24 to 48 weeks.

4. *Results of HCV treatment*: The result of IFN-α/RBV treatment includes:

   (a) SVR (*sustained virological response*): is defined as undetectable HCV RNA by a sensitive assay at the end of a 24-week follow up period after the end of treatment.

   (b) Non-SVR (*non sustained virological response*): (a) may be found to be HCV RNA negative during therapy but may relapse thereafter, (b) may be virological non-responders showing detectable HCV RNA levels throughout the complete treatment period.

   The response rate of IFN-α/RBV therapy is influenced mainly by the HCV genotype [Pascu et al., 2004], are shown in Table 2.2.

Table 2.2: The response rate depends on HCV genotype. The Response rate is the percentage of patients whose cancer shrink (a partial response) or disappear (a complete response) after treatment

Genotype	Response rate
1	42% - 52%
2	78% - 86%
3	78% - 86%
4	55% - 69%
5	50% - 77%
6	50% - 77%

Chapter 3

Discriminative motif learning for few labeled data

This chapter presents a computational process of discriminative motif discovery for few labeled data. An application to hepatitis C virus treatment brings additional knowledge of the resistance and response to hepatitis C virus therapy. Experimental results and biomedicine significances are discussed as well.

3.1 Introduction

3.1.1 Discriminative motif learning

Discriminative motif finding problem is to find motifs occurring more frequently in one set of sequences and not occurring in the other set of sequences. These motifs can help to classify effectively a sequence into a certain class or to describe the discriminative characteristics of a class. Many methods have been developed to search discriminative motifs so far.

MERCI (*Motif EmeRging and with Classes Identification*) [Vens et al., 2011] uses a string-based model to represent motifs and adapts an Apriori algorithm, a well-known sequential pattern finding technique, to find discriminative motifs. MERCI introduces two parameters which are the minimal frequency threshold for one sequence set and the maximal frequency threshold for the other sequence set to prune early motifs which are not chosen as candidates during the search process.

MEME (*Multiple EM for Motif Elicitation*) [Bailey et al., 2010] represents a motif as a PWM and assumes that each sequence has zero or one motif. Given a PWM, MEME

calculates the likelihood of PWM by the EM (*Expectation Maximization*) algorithm. 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 are 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.

DEME (*Discriminatively Enhanced Motif Elicitation*) [Redhead and Bailey, 2007] is an adaptation of the discriminative framework in [Segal et al., 2002]. DEME also represents a motif as a PWM and uses conjugate gradient to find the best PWMs with the assumption that each sequence may contain no or one motif occurrence. The difference between DEME and Segal's work is that DEME uses the combination of two algorithms called "substring search" and "pattern branching" to learn the parameters of the motif model that is used to maximize the discriminative objective function.

In the work of [Kim and Choi, 2011], a hybrid generative and discriminative model is developed to learn discriminative motifs. The generative model plays the role to maximize the likelihood of PWM, and the discriminative model is responsible for selecting the most discriminative feature. These models are combined by a joint prior distribution over two parameter sets of two models.

Discriminative HMM [Lin et al., 2011] uses profile HMM to represent a motif and this representation is more flexible for insertion or deletion than PWM's representation. Under the HMM, finding motifs in sequences is equivalent to finding hidden states of sequences. The parameters of HMM are estimated by using the MMIE (*maximum mutual information estimate*) technique applied to speech recognition to train the model and get the optimum of discriminative criterion.

In summary, the methods typically involve building PWM or HMM from sequences and then using techniques such as EM or Gibbs sampling to optimize the likelihood of PWM or HMM, and thus do not guarantee to find the global solution, whereas string-based methods can yield the global solution but have to deal with drawbacks such as a large number of input data or discovering lengthy motifs because they can lead to the high complexity of computation. In addition, because PWM and HMM are normally obtained from the input data, all the above mentioned methods require a large number of labeled data to learn good PWMs and HMMs. If these methods work with small labeled datasets, PWMs and HMMs may not return good results as expected.

3.1.2 Semi-supervised ensemble learning

The combination between SSL (*semi-supervised learning*) and EL (*ensemble learning*) is discussed in [Zhou, 2009] for improving generalization, where the combination of learners can be helpful to SSL and unlabeled data can be helpful to EL. So far, many studies have proposed hybrid methods working both in SSL and EL. It could say that semi-supervised ensemble methods are gradually interested in and have been applied to many tasks, for example natural language processing, image processing, document retrieval, and so on.

To improve the task of word alignment, [Huang et al., 2010] uses a semi-supervised learning method, namely Tri-training [Zhou and Li, 2005], to iteratively train three classifiers and assign labels to the unlabeled data. Then it uses some data among the unlabeled one to expand the labeled training set of each individual classifier.

In the work of [Vajda et al., 2011], a semi-automatic labeling procedure is proposed to recognize handwritten characters. This procedure considers a data representation as a component of EL. A voting strategy is used to label for unlabeled data. However, the main distinction between other SSL strategy and this method lies in the fact that the label assignment does not based on the votes. The final classifier is built on top of the inferred labels.

[Dong and Schafer, 2011] applies three classifiers in order to select the new labeled data in the process of self-training for the problem of citation classification. To make the final prediction for a given instance, an adopt majority voting is used.

The combination of label propagation and ensemble learning are applied in semi-supervised learning [Woo and Park, 2012]. A subset of unlabeled data is randomly selected, and it composes a training set together with original labeled data. For the label prediction of the selected unlabeled data, a graph-based label propagation method is used. Then, a classifier is trained on the composed training set.

In stream mining, as data streams are infinite, arrive continuously and there should be online classification, labeling all of the arrived data is impossible. [Admadi and Beigy, 2012] proposed a semi-supervised ensemble learning method to label data in a window. For each learner, a set of labeled instances is determined from unlabeled data by using the majority vote.

Taken together, this work aims to develop a semi-supervised ensemble method for discriminative motif finding from a limited number of labeled sequences and then apply it to detect sequence motifs in NS5A protein that characterize SVR and non-SVR treatment result when using IFN/RBV therapy. Our method is based on the SLUPC algorithm [Ho et al., 2011] which is a separate-and-conquer searching method to discover motifs of type 'discriminative one occurrence per sequence' (DMOPS). Concretely, the proposed

method, named E-SLUPC (*Ensemble SLUPC*), firstly searches core motifs from a small labeled dataset, then uses these motifs to exploit unlabeled data, and continues searching discriminative motifs with the enlarged labeled dataset.

Experiments have been performed to investigate the accuracy of E-SLUPC compared with SLUPC, and the quality of discriminative motifs found by E-SLUPC, MEME and DEME. The experimental results show the accuracy of the proposed framework is improved about by 8% and DMOPS motifs with high accuracies from 80% to 100% found by our new method are able to discriminate better than discriminative motifs of MEME and DEME.

3.2 Method

Because the number of labeled sequences is small, the predictive power of learned motifs is often low. This motivated us to develop a semi-supervised learning method using unlabeled dataset to seek DMOPS with higher predictive power. In order to obtain a higher degree of accuracy of label assignment, we also have develop an ensemble learning method by combining appropriately multiple label assignment approaches. These semi-supervised and ensemble learning methods work together to boost the ability to learn discriminative motifs when labels are assigned more precisely.

In general, our semi-supervised ensemble learning method works under the cluster assumption: if sequences are in the same cluster, they are likely to be of the same class [Chapelle et al., 2006]. Concretely we use two assumptions for clusters in our label assignment approaches, one is based on motif matching and the other is based on the gene distance of sequences. The former uses discriminative motifs to assign unlabeled sequences to different clusters, while the later uses the gene distance between sequences to make clusters.

The framework of E-SLUPC in Figure 3.1 is described below with input sequences from a small labeled dataset and a large unlabeled dataset.

1. Applying SLUPC to labeled sequences to find a set of DMOPS motifs considered as core motifs.

2. Using the core motifs to enlarge the labeled dataset by adding to it unlabeled sequences that well match with the core motifs determining by the following ensemble procedure: each unlabeled sequence that matches well the core motifs by three ensemble components (described in Subsection 3.3) will be finally assigned a label by the majority voting. Then, the pseudo labeled dataset is determined.

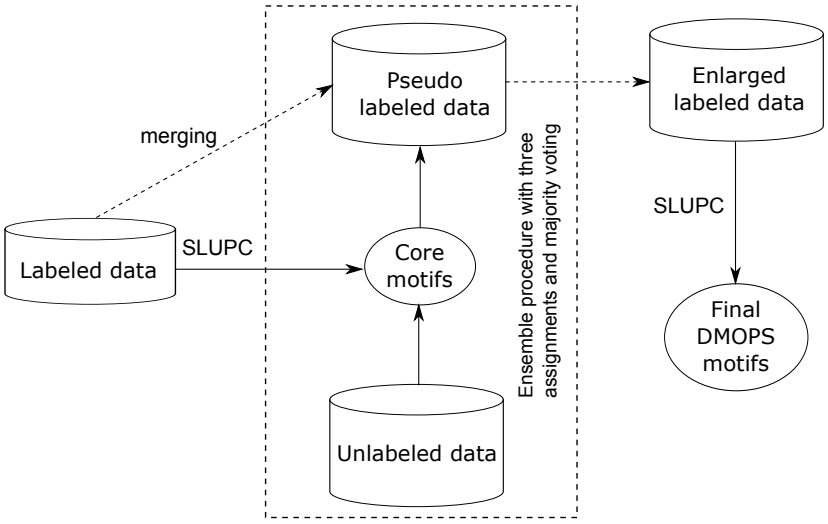

Figure 3.1: The framework of E-SLUPC.

3. Applying SLUPC to the enlarged labeled dataset, which consists of the labeled and pseudo labeled data, to learn the final set of DMOPS motifs.

4. The steps 1-3 are repeated until either (i) the core motif set is stable, or (ii) the maximum number of iterations is achieved.

5. To recognize a new unlabeled sequence, applying the ensemble procedure to the unlabeled dataset.

3.2.1 SLUPC algorithm

Discriminative multiple occurrence per sequence (DMOPS) is one of the motif types categorized by [Kim and Choi, 2011] based on counting the number of total occurrences of motifs in sequences. It shows a structural assumption that is used to generate motifs from the motif model. In this section, we describe the DMOPS motif discovery method that uses the SLUPC algorithm in [Ho et al., 2011] to learn a set of descriptive subsequences for the two-class problem. The algorithm SLUPC is an extended version of LUPC (Learning the Unbalanced Positive Class) [Ho and Nguyen, 2002] for sequential data.

Denote $S = \{(S_1, C_1), (S_2, C_2) \ldots, (S_n, C_n)\}$, where S_i is a sequence of length $|S_i|$ over the alphabet $\Sigma = \{A, U, G, T\}$ or $\Sigma = \{amino\ \ acid\}$ and $C_i \in \{C_1, C_2, ..., C_c\}$ of the class labels. When there are only two classes we call one as positive denoted by Pos and the other as negative denoted by Neg, and thus the labeled set $S = Pos \cup Neg$. The problem is to find a minimal set of DMOPS motifs satisfying two conditions: (1) $Complete$: each

sequence contains at least one found motif, (2) *Consistent*: motifs found for Pos do not match any negative sequences in Neg and vice versa.

Given parameters α ($0 < \alpha < 1$) and β ($0 < \beta < 1$), a subsequence P is an α-*coverage* for Pos if

$$\frac{|cover_{Pos}(P)|}{|Pos|} \geq \alpha,$$

and is a β-*discriminant* for Pos if

$$\frac{|cover_{Pos}(P)|}{|cover_S(P)|} \geq \beta,$$

where $cover_{Pos}(P)$ is the set of sequences in Pos that contains P and $cover_S(P) = cover_{Pos}(P) \cup cover_{Neg}(P)$. If P is both α-coverage and β-discriminant for Pos, we will say P is $\alpha\beta$-strong for Pos. Similar concepts can be defined for Neg. A subsequence will be a DMOPS motif when it satisfies both α-*coverage* and β-*discriminant* thresholds.

Note that if sequence P_1 is a subsequence of a sequence P_2, then we have $cover(P_2) \subseteq cover(P_1)$, i.e., the coverage of P_1 is larger and the discrimination ability of P_1 is smaller than that of P_2. Given an α-coverage pattern P, the most informative pattern related to P in terms of coverage is the longest α-coverage pattern containing P. Alternatively, given a β-discriminant pattern P, the most informative pattern related to P in terms of discrimination is the shortest β-discriminant pattern contained in P.

The DMOPS motif finding of SLUPC algorithm is described in Algorithm 3. Given two sets of positive sequences Pos and negative sequences Neg, Algorithm 3 will find a minimal set of DMOPS motifs satisfying Complete and Consistent requirements. In this algorithm, $Motif(Pos, Neg, \alpha, \beta)$ is an exhaustive search procedure that expands a subsequence one position to the left or to the right, starting with the length is 1.

In the procedure finding an $\alpha\beta$-strong motif, the subroutine *Adjacentaa* searches for letters that can be added to $S(i)$ if making $S(i+1)$ satisfies α and β. The subroutine *StopCond* checks if *Adjacentaa* is successful. If 'not', it returns an empty new motif. If 'yes', the subroutine $CandMotifs$ ranks $S(i+1)$ by the number of occurrences in Pos if there is more than one amino acid that make $S(i+1)$ satisfy both α and β.

The subroutine $CandMotifs$ may require a lot of checks on Neg to see if a generated motif candidate is $\alpha\beta$-strong. However, thanks to the property "*given a threshold α, a pattern P is not $\alpha\beta$-strong for any arbitrary β if* $cover_{Neg}(P) \geq ((1-\alpha)/\alpha) \times cover_{Pos}(P)$" [Ho et al., 2011], many motif candidates are quickly rejected if they are found to match the condition $cover_{Neg}(P) \geq ((1-\alpha)/\alpha) \times cover_{Pos}(P)$ during the scan of Neg. It is easy to count $cover_{Pos}(P)$ for each motif candidate P as Pos is small, and we need only to accumulate the count of $cover_{Neg}(R)$ when scanning Neg until either we can reject the

Algorithm 3 SLUPC algorithm

Given: Labeled sequences in Pos and Neg, and parameters $minalpha$, $minbeta$.

Find: $\alpha\beta$-strong DMOPS motifs for Pos

DMOPS Motif $(Pos, Neg, minalpha, minbeta)$

```
1: MotifSet = φ
2: α, β ← Initialize(Pos, minalpha, minbeta)
3: while Pos ≠ φ & (α, β) ≠ (minalpha, minbeta) do
4:     NewMotif ← Motif(Pos, Neg, α, β)
5:     if NewMotif ≠ φ then
6:         Pos ← Pos \ Cover+(NewMotif)
7:         MotifSet ← MotifSet ∪ NewMotif
8:     else
9:         Reduce(α, β)
10:    MotifSet ← PostProcess(MotifSet)
11: return(MotifSet)
```

motif candidate as the constraint holds or we completely go throughout Neg and find the motif has satisfied accuracy.

3.2.2 Self-training technique for semi-supervised learning

We develop the semi-supervised method based on the idea of self-training technique to enlarge the labeled dataset. Self-training is one of the most common techniques used in semi-supervised learning [Zhu, 2008]. In this technique, a learner is first trained with the small amount of available labeled data. The learner is then used to learn the unlabeled data. Only unlabeled data with their predicted labels having the most confident score are added to the training dataset. After that, the learner is re-trained and this procedure is repeated until convergence is reached.

Self-training is a wrapper method that requires a predetermined learning method and uses its results to teach itself. In our practical point of view, self-training technique is appropriate in a case that the existing learning method is complicated and difficult to modify for doing semi-supervised learning. Our SLUPC algorithm is an example of this case. In addition, evidence shows that doing semi-supervised learning with the cluster assumption, self-training is an effective approach [Rosenberg et al., 2005].

Procedure Finding an $\alpha\beta$-strong motif

Motif $(Pos, Neg, \alpha, \beta)$

1: $CandMotifSet = \phi$
2: **Adjacentaa**$(Pos, Neg, \alpha, \beta)$
3: **while StopCond**$(Pos, Neg, \alpha, \beta)$ **do**
4: **CandMotifs**$(Pos, Neg, \alpha, \beta)$
5: **end while**
6: $Motif \leftarrow FirstCandMotifinCandMotifSet$
7: $return(Motif)$

3.2.3 Majority voting strategy for ensemble learning

In ensemble learning, strategies that combine outputs of learning methods are categorized in three groups, linear combination, product combination and voting combination [Brown, 2010]. The linear and product combinations are used when learning methods output real-valued numbers, while voting combination is applicable to results of class labels. The idea of majority voting strategy is that each learning method votes for a certain class, and the class with the most votes will be chosen as the ensemble output.

Based on majority voting strategy, we develop three ensemble components to explore the unlabeled dataset. Each ensemble component is an approach to assign labels for unlabeled sequences under the cluster assumption. After these three components assign labels for an unlabeled sequence, the plurality label can be the final label for that unlabeled sequence.

Label assignment 1. In this ensemble component, the more an unlabeled sequence contains core motifs of a class, the more it belongs to this class. To apply this rule, each unlabeled sequence will be matched to core motifs by counting how many times this sequence contains core motifs of a class, and then these number of times are used to assess how much an unlabeled sequence can be considered as a sequence of a class. In order to decide which unlabeled sequence will belong to which class, we choose unlabeled sequences that contain the most core motifs and just contain motifs in one class.

Label assignment 2. We use the same label assignment rule of the first component (the more an unlabeled sequence contains core motifs of a class, the more it belongs to this class), however we make a different decision of choosing labels for unlabeled sequences. We choose unlabeled sequences that contain more motifs of a class than those of the remaining class, with the ratio between two classes being larger than a threshold γ (for

example 80%), to assign labels.

Label assignment 3. The gene distance between two sequences is used to assign labels for unlabeled sequences. The gene distance shows the similarity or dissimilarity among sequences and is represented by the optimal local gapped alignment score between two sequences [Altschul et al., 1990, Smith and Waterman, 1981]. According to BLAST[1], the higher the score is, the more similar two sequences are. Therefore, the assignment is that if two sequences have a high score, they are likely to be of the same class. To apply this assignment rule, the score of an unlabeled sequence and a representative of each class is calculated and we choose the larger score to decide to label for that unlabeled sequence. We obtain the representative of a class by choosing a sequence having the minimum deviation between scores of sequences and the average of these scores.

3.3 Application to HCV study

We are given a set of sequences of the NS5A region that are hypothesized to contain several instances of SVR and non-SVR signals. The problem is to find SVR and non-SVR motifs in the NS5A regions. Solving this problem provides a biomarker or additional knowledge to the relation between NS5A region and IFN/RBV therapy. This hypothesis, when verified, leads to a better understanding of the resistance or response to IFN/RBV therapy of HCV.

3.3.1 The dataset

In this study, all sequences, each containing 447 amini acids, are in NS5A region of HCV genotype 1b. We used two kinds of datasets as follows:

- *Labeled dataset*: including 28 sequences SVR, 49 sequences non-SVR from LANL database, and 13 sequences SVR, 12 sequences non-SVR from Chiba University database.
- *Unlabeled dataset*: including 1424 sequences from HVDB and 168 sequences from GenBank.

[1]Basic Local Alignment Search Tool http://blast.ncbi.nlm.nih.gov

3.3.2 Finding DMOPS motifs charactering SVR and non-SVR to therapy

The experiments aim to evaluate the performance of discovered motifs in terms of discrimination. A 3-fold cross validation on labeled data was done with the algorithms parameters as follows: $minalpha = 0.1, minbeta = 0.5$. We obtained these values by performing the SLUPC algorithm many times to pick out the best parameters that are suitable to the training dataset. In this experiment, the initial value of α and β are with high values of 0.7 and 0.95, respectively and alternatively reduced them, $\alpha = \alpha - \Delta\alpha, \beta = \beta - \Delta\beta$ with $\Delta\alpha = 0.05$ and $\Delta\beta = 0.02$, in order to firstly find the strongest $\alpha\beta$-motifs, then step by step reduce α and β to find as strong as possible $\alpha\beta$-motifs that each training sequence contains at least one motifs found.

Because of the small labeled dataset, the widespread of DMOPS motifs is not ensured in the whole dataset and the accuracy of prediction is not stable. To get the good quality motifs as well as the stable prediction accuracy, we perform 3-fold cross validation 5 times. Following the idea of ensemble learning, we add up DMOPS motifs of each run time to create a set of integrated motifs, assess the widespread and effect on the prediction accuracy of each motifs in 3 testing sets, and then eliminate motifs which are infrequent and make prediction accuracy low. The average accuracy of the SLUPC algorithm is represented in Table 3.2. Though the average accuracy on testing data is low (about 66%), it is very encouraging in the biomedical field.

Table 3.1 presents DMOPS motifs that are found in 15 times of experimental running (5 times of 3-fold cross validation). Each four columns stands for DMOPS motifs found in SVR and non-SVR sequences, together with the number of SVR sequences and non-SVR sequences containing a motif and the number of occurrences of that motif in 15 times, respectively. The number of SVR sequences and non-SVR sequences containing a motif are calculated on the whole dataset. These motifs are selected from the set of integrated motifs after filtering motifs that have the low accuracy and coverage. However, some DMOPS motifs that have the low number of occurrences still exist in this table. That is because if they are removed out of the integrated motif set, the prediction accuracy will be decreased.

It can be observed from Table 3.1: the SVR motif "LAI" occurs in 7 SVR sequences and does not occur in non-SVR sequences. Its coverage is 17% ($7/41 = 0,17$) and its accuracy is 100% ($7/(7+0) = 1$). In addition, this motif occurs 13 times in the cross validation experiment. Another SVR motif "AI" also has the high coverage (24%) and accuracy (100%). Similar observations can be done for non-SVR motifs "NM", "DK", or "NR". It could say that these DMOPS motifs can be viewed as the good biological signals for charactering SVR and non-SVR to IFN/RBV therapy. The group of SVR motifs, such

Table 3.1: DMOPS motifs characterizing SVR and non-SVR to IFN/RBV therapy

SVR motifs	SVR sequences	Non-SVR sequences	No. of occurrences	Non-SVR motifs	SVR sequences	Non-SVR sequences	No. of occurrences
LAI	7	0	13	NM	0	7	14
AF	4	1	12	ND	1	10	13
AI	10	0	12	DK	0	3	9
VEA	5	2	10	RS	1	5	8
FN	2	0	8	VDLIEA	1	4	7
TAA	2	0	6	AKA	1	6	6
HN	1	0	5	NR	0	4	6
VN	1	0	4	MA	0	3	3
KAA	2	0	3	PAS	0	4	3
AAC	2	0	3	WC	0	4	2

as "AF", "VEA", "FN", "TAA", "HN", "KAA", and "ACC", and non-SVR motifs, such as "ND", "VDLIEA", and "AKA", have high accuracies from 80% to 100% that show the high ability of discrimination. However, their occurrences in 15 times of conducting the experiments are insufficiently large to conclude that they are good DMOPS motifs.

3.3.3 Evaluating the accuracy of E-SLUPC and SLUPC

In this part, we present the experiment that focuses on validating and comparing the accuracy assessment of SLUPC algorithm before and after enlarging labeled dataset. Therefore we perform 3-fold cross validation 5 times with parameters *minalpha*, $\Delta\alpha$, *minbeta*, $\Delta\beta$ are set to 0.05, 0.05, 0.4, and 0.05 respectively which are different from values of parameters in SLUPC algorithm. This adjustment is essential because the number of sequences in the training dataset will be increased, the old values of parameters are not the most appropriate values in the case of the new training dataset. However, these parameters are fixed during the iteration process of semi-supervised ensemble learning because the number of sequences added to training dataset after one iteration is not significant.

In this experiment, 1424 unlabeled sequences are used and repeated for each iteration to pick out sequence candidates. The maximum number of iterations is set to 5 and the highest rank of a sequence is 1. Because the number of sequences in the training set is small, we consider one match between a DMOPS motif and an unlabeled sequence is enough for the first and second ensemble components to assign a label for that unlabeled sequence.

Table 3.2 shows the experiment results of comparing the accuracy of SLUPC and E-SLUPC (about 8% increase in accuracy). Accuracies in Table 3.2 are average accuracies of folds in each time of doing 3-fold cross validation. These accuracies are computed on our testing dataset. In 5 times of 3-fold cross validation, accuracies of E-SLUPC are

Table 3.2: Accuracy of SLUPC adn E-SLUPC

No. of 3-fold cross validation	SLUPC	E-SLUPC
The 1^{st} 3-fold	0.83	0.85
The 2^{nd} 3-fold	0.65	0.76
The 3^{rd} 3-fold	0.63	0.73
The 4^{th} 3-fold	0.58	0.68
The 5^{th} 3-fold	0.63	0.70
The average accuracy	0.66	0.74

increased from 2% to 10%. This can be explained by the quality of DMOPS motifs found during semi-supervised ensemble learning process. When the label assignment is more effective and precise, DMOPS motifs are better and more qualified.

3.3.4 Comparing the output of E-SLUPC to MEME and DEME

1. MEME

   We choose MEME to compare the output of E-SLUPC because MEME is currently one of the most well-known and powerful types of software for motif finding. Using the web version of the MEME[2], we perform a 5 times 3-fold cross validation experiment with the following parameters: the occurrence of a single motif among the sequences is set to the multiple occurrence per sequence, the length of each motif is between 2 and 6, and the maximum number of motifs is 30. The first two parameters, the multiple occurrence per sequence and the length of a motif, are chosen in a similar way to our previous experiment for E-SLUPC. It allows us to do a comparative assessment of results between E-SLUPC and MEME when setting the same values for two sets of parameters. Because MEME yields only a motif at each runtime, and we also want to get as many motifs as possible, we let MEME repeat 30 times. After 15 times of MEME running, we collect about 163 SVR motifs and 170 non-SVR motifs. In this result, we compare between the set of SVR motifs and non-SVR motifs, and we find about 57 motifs appeared in both SVR and non-SVR motif sets. Table 3.3 shows the top 12 motifs found by MEME which have the highest frequency in a total of 15 times of MEME running.

   Observing Table 3.3, we see that although MEME allows us to search discriminative motifs with two sets of positive and negative sequences, the discriminative ability

[2]MEME http://meme.sdsc.edu/meme/cgi-bin/meme.cgi

Table 3.3: The top twelve motifs found by MEME

SVR motifs	SVR sequences	Non-SVR sequences	No. of occurrences	Non-SVR motifs	SVR sequences	Non-SVR sequences	No. of occurrences
WRQEMG	39	60	15	TFQVGL	18	49	15
RKSRKF	21	32	14	GDFHYV	21	49	14
WKDPDY	27	47	12	WKDPDY	27	47	14
EEDERE	30	54	11	QITGHV	17	40	12
CTTHHD	11	22	10	DLIEAN	35	60	11
GDFHYV	21	49	9	RLHRYA	27	47	10
SHITAE	41	54	8	KNGSMR	25	47	8
DPSHIT	41	56	7	LLREEV	11	37	7
EPDV	40	59	6	SQLASAP	34	61	5
PVVHGC	37	57	5	TSMLTD	39	61	4
LKAT	35	59	3	PEFF	28	49	3
SPDA	32	55	2	EEYV	27	48	2

of these motifs is not high. The motifs such as "WKDPDY", or "GDFHYV" have the high frequency in 15 times of MEME running, but their appearances in both SVR and non-SVR motif sets make them difficult to be reliable discriminators when distinguishing two classes. In addition, MEME does not return motifs that have the high accuracy such as "LAI", "VEA", or "VDLIEA" found by E-SLUPC. Therefore, MEME has just worked effectively in the case of finding motifs that describe characteristics of a sequence dataset.

2. DEME

   DEME is one of the efficient discriminative motif finding methods. DEME combines two times of search, global and local search, to learn the parameters of the PWM motif model that maximize the discriminative objective function. Moreover, DEME uses an informative Bayesian prior to incorporate the prior knowledge of reside characteristics of protein sequences. Using the free program DEME[3], we also perform a 5 times 3-fold cross validation experiment in order to compare discriminative motifs of the proposed method and DEME. Parameters are chosen as follows, the length of each motif is from 2 to 6 amino acids, the occurrence of a single motif is set to one occurrence per sequence and the input sequences are protein sequences. Other parameters use default values of DEME. After 15 times of DEME running, we obtain 248 SVR motifs and 387 non-SVR motifs, where 11 motifs appear in both SVR and non-SVR motif sets. Table 3.4 shows the top 12 motifs found by DEME which have the highest frequency in a total of 15 times of DEME running.

   In Table 3.4, the SVR motifs "LAIKT", "LAIK", "LVGLNW", "LSALSL" and "VSLK" occur in SVR sequences and do not occur in non-SVR sequences. The similar observation is concluded for non-SVR motifs, such as "DQPSND", "NMWH",

[3]DEME http://bioinformatics.org.au/deme/

Table 3.4: The top twelve motifs found by DEME

SVR motifs	SVR sequences	Non-SVR sequences	No. of occurrences	Non-SVR motifs	SVR sequences	Non-SVR sequences	No. of occurrences
KK	15	9	13	FQ	23	56	11
RK	41	61	12	TFQ	19	50	10
LAIKT	7	0	12	DQASD	1	7	8
KSKK	6	5	11	DQDSD	12	20	7
LAIK	7	0	10	EMGGN	36	61	6
LVGLNW	10	0	9	DQPSND	0	4	6
LATKT	19	45	9	SFD	25	48	5
LSALSL	3	0	8	ASQ	41	61	4
PSLK	32	59	8	YN	40	60	4
VSLK	1	0	6	NMWH	0	5	4
RKT	10	11	6	MWHGT	0	5	3
KSRK	22	34	5	ATCTT	32	54	3

and "MWHGT", that occur in non-SVR sequences and do not occur in SVR sequences. The frequency of these motifs in a total of 15 times cross validation experiment are high. Two SVR motifs, "LAIKT" and "LAIK", and the non-SVR motif "NMWH" contain the SVR motif "LAI" and non-SVR motif "NM" respectively that are found by E-SLUPC. This shows that the ability of searching longer length motifs of DEME is better than the one of E-SLUPC. However, DEME cannot limit the search to the discriminative motifs only. Besides finding discriminative motifs, DEME finds motifs in both SVR and non-SVR sequences. For example, "KK", "LATKT", "PSLK", "RKT" and "KSRK" are SVR motifs, but they appear in several non-SVR sequences. The same remark is also made for the group of non-SVR motifs, the motifs "FQ", "TFQ", "EMGGN", "ASQ", "YN", and "ATCTT" are found in many SVR sequences. The searching results of DEME do not completely discriminate SVR and non-SVR properties of sequences. Therefore, a step of the comparative assessment is necessary to pick out discriminative motifs after using DEME.

3.4 Conclusion

We have presented the algorithm for discovering discriminative motifs which can function well when the labeled dataset is small, but the unlabeled dataset is large. Our algorithm is applied to detect the relationship between HCV NS5A protein and IFN/RBV therapy effect. The results are promising as they present many patterns that were not known previously. However, the SLUPC algorithm quickly eliminates the cases that do not satisfy two thresholds *coverage* and *discriminant* during recursively expand a subsequence. This can lead to ignoring some potential motifs neglected one or more positions if we want to

find gap motifs.

We have also explored the use of self-training-based semi-supervised ensemble learning to enlarge the training set of the discriminative motif finding problem in case the number of labeled data is small. This method works in an iterative procedure to choose the best match sequences among the unlabeled sequences. The experiment results show that with more data for the training dataset, the SLUPC algorithm can obtain higher accuracy.

Chapter 4

Discriminative motif learning for short sequences

To solve the problem of discriminative motif learning for short sequences, this chapter presents a new method for sequence characterization and prediction. Experimental results and biomedicine significances are discussed later on.

4.1 Introduction

Sequence characterization and prediction with a small number of short and similar sequences usually has low performance. It is not effective to apply directly classification methods but preprocessing, especially some appropriate data transformation, is always needed. The key approach to this problem is dimensionality reduction with feature extraction that is essentially based on transforming original data into a new feature space where the original data will be represented by new latent components. In addition to traditional linear dimensionality reduction methods such as PCA (*principal component analysis*), ICA (*independent component analysis*), projection pursuit, nonlinear dimensionality reduction methods have been actively developed such as polynomial or kernel PCA [Bernhard et al., 1997], locally-linear embedding [Roweis and Saul, 2000], Isomap [Tenenbaum et al., 2000].

A topic model is a statistical model that analyze the words of the original texts to discover the themes or hidden topics consisting of a set of words that frequently occur together [Blei, 2012]. Topic models can be seen as a tool for dimensionality reduction for different data types. The typical topic models include LSA (*latent semantic analysis*) [Deerwester et al., 1990], PLSA (*probabilistic latent semantic analysis*) [Hofmann, 2001], and LDA (*latent Dirichlet allocation*) [Blei et al., 2003], the most currently used.

In this work, we approach to the characterization and prediction problems in a new

way. Instead of characterizing and predicting directly from the original data as traditional methods often do, we propose a framework to characterize and predict outcomes of HCV treatment by using topic modeling. This framework is based on the two-step framework [Than et al., 2012] which is developed for SDR (emphsupervised dimension reduction) with topic modeling. Concretely, before prediction, our framework performs a transformation of data into a topical space by using label information and nearest neighbor information of the data. Because the input sequences are short in length and very similar to each other, the information of nearest neighbors plays an important role for characterizing properties of SVR and non-SVR. We believe that our framework represents an effective approach more than previous methods, especially when input sequences do not provide enough information and the number of labeled data is small.

The proposed framework is adapted to three topic models: PLSA (*probabilistic latent semantic analysis*) [Hofmann, 2001], LDA (*latent Dirichlet allocation*) [Blei et al., 2003], and FSTM (*fully sparse topic models*) [Than and Ho, 2012], resulting in PLSA^c, LDA^c, and FSTM^c for prediction of SVR and non-SVR. Extensive experiments have been performed to test the quality of these methods, with more than 2000 experiments. Compared with the baseline method SVM (*support vector machine*) for prediction, the three methods (PLSA^c, LDA^c, and FSTM^c) often perform significantly better. As an example, with 5 folds cross validation, FSTM^c often predict 91.94% correctly, as compared with 69.5% by SVM.

The characterization results are reliable subsequences with high contributions to a set of SVR or non-SVR sequences. These discriminative subsequences can be considered as potential patterns for predicting SVR or non-SVR sequences, concurrently suggest evidences to the better comprehension of resistance to IFN/RBV therapy of HCV.

4.2 Method

4.2.1 Topic model

In natural language processing, a topic model is a statistical model that analyze the words of the original texts to discover the themes or hidden topics consisting of a set of words that frequently occur together [Blei, 2012]. The early topic model, LSA (*latent semantic analysis*), was developed by [Deerwester et al., 1990]. The next one, PLSA, was introduced by [Hofmann, 2001]. The most common topic model currently in use, LDA, was described by [Blei et al., 2003]. And a recent topic model, FSTM, was proposed by [Than and Ho, 2012].

The goal of topic modeling is to automatically discovery topics from a collection of

unlabeled documents. The documents themselves are observed while topics are hidden. The central computational problem for topic modeling is to use the observed documents to infer the hidden topics [Blei, 2012].

In a topic model, a document is often assumed to contain multiple topics. Hence it has a latent representation in the topical space, and such a representation can be inferred once the topic model had been learned. Note that the topical space is often lower dimensions than the original space of documents. For this reason, topic modeling provides a potential approach to dimension reduction.

4.2.2 Supervised dimension reduction

SDR (*supervised dimension reduction*) is the problem that we are asked to find a low-dimensional space which preserves the predictive information of the response variable. Projection on that space should keep the discrimination properties of the data in the original space. Once the new space is determined, we can work with projections in that low-dimensional space instead of the high-dimensional one. In text applications, a data point or an instance of data is a document, the formal definition of SDR is stated as follows.

Given a corpus $D = \{d_1, ..., d_M\}$ consisting of M documents which are composed from a vocabulary of V terms. Each document d is represented as a vector of term frequencies, i.e. $d = (d_1, ...d_V) \in \mathbb{R}^V$, where d_j is the number of occurrences of term j in d. Let $\{y_1, ..., y_M\}$ be the class labels assigned to those documents, respectively. The task of SDR is to find a new space of K dimensions which preserves the predictiveness of the response/label variable Y.

4.2.3 The two-step framework

In SDR problem, many studies often find directly a low-dimensional space that preserves the discriminative properties of the data classes in the original space. With the two-step framework, Than et al. [Than et al., 2012] proposed a novel approach to SDR, in which the first step tries to find an initial topical space and the second step tries to utilize label information and local structure of the data, to find the discriminative space (Figure 4.1). The first step learns an unsupervised topic model to obtain topics. And the second step consists of four tasks performing the projection of documents onto the initial space so that inner-class local structure is preserved and inter-class margin is widen. Therefore, the discriminative property is not only preserved, but also better in the final step. The details of the two-step framework are shown in Algorithm 4.

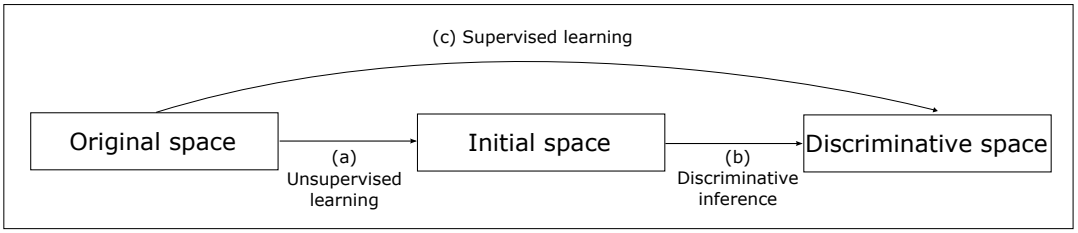

Figure 4.1: Sketch of approaches for SDR [Than et al., 2012]. Existing methods for SDR directly find the discriminative space, which is supervised learning (c). The two-step framework consists of two separate steps: (a) first find an initial space in a unsupervised manner; then (b) utilize label information and local structure of data to derive the final space.

Algorithm 4 Two-step framework for SDR

Step 1: Learn a unsupervised model to get K topics $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$.

$\mathfrak{A} = span\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K\}$ is the initial space.

Step 2: Build the discriminative space

(2.1) for each class c, select a set S_c of topics which are potentially discriminative for c.

(2.2) for each document d, select a set N_d of its nearest neighbors which are in the same class as d.

(2.3) infer new representation $\boldsymbol{\theta}_d^*$ for each document d in class c by the Frank-Wolfe framework with the objective function

$f(\boldsymbol{\theta}) = \lambda.L(\widehat{d}) + (1-\lambda).\frac{1}{|N_d|}\sum_{\mathbf{d}' \in N_d} L(\widehat{d'}) + R.\sum_{j \in S_c} \sin(\theta_j)$,

where $L(\widehat{d})$ is the log likelihood of document $\widehat{d} = d/||d||_1$; $\lambda \in [0, 1]$ and R are nonnegative constants.

(2.4) compute new topics $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ from all d and $\boldsymbol{\theta}_d^*$.

$\mathfrak{B} = span\{\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*\}$ is the discriminative space.

4.2.4 Methodology

From the perspective of sequence characterization and prediction, why do we think topic modeling approaches are appropriate to solve these problems? In our observations, the answer comes from the two key properties: interpretability and practical effectiveness. More concretely, the following attractive characteristics provide us an affirmative answer:

- *Discovery of latent factors:* probabilistic topic models are models of latent factors. Each factor is called *a topic*, and is unobservable. They can be learned efficiently from data by the EM algorithm or Gibbs sampling. Hence topic modeling provides a reasonable way to explore factors/topics hidden in our data.

- *Uncovering contributions of latent factors:* many topic models are admixture models such as LDA, PLSA, and FSTM. They often assume that a document is a mixture of hidden topics, and that each topic is a distribution over terms. Therefore, once a model is learned, we can easily interpret which topics drive the theme of a specific document and which terms are important for a hidden topic. In addition, we can inspect easily the contribution/effect of a hidden topic to the whole data.

- *Uncovering effects of latent factors to a class:* when dealing with supervised data, e.g., ISDR sequences, topic models provide us a principled way to investigate the effect of a hidden factor to a class of data. Indeed, by estimating contributions of hidden topics to a class, we can see explicitly which topics play an important role in a class. More importantly, by comparison from different classes, we can uncover which topics are discriminative for a class.

- *Effectiveness in classification for discrete data:* recent work [Than et al., 2012, Yhu et al., 2012] demonstrate that topic models can exploit well label information when learning form data. Excellent performance on document classification was observed by various researches [Than et al., 2012, Yhu et al., 2012]. This observation will be further supported from our problem of predicting SVR or non-SVR, as investigated later.

1. **Framework for prediction**

   To perform the characterization and prediction for sequences by using topic modeling, our framework consists of 3 main steps: data representation, characterization, and prediction. The graphical framework is shown in below Figure 4.2.

   (a) Data representation

   To represent a sequence in the context of topic modeling, we firstly perform the subsequence extraction from a sequence dataset by using sliding windows.

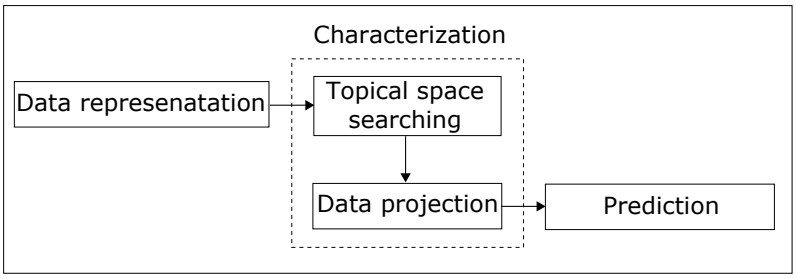

Figure 4.2: The proposed framework for sequence characterization and prediction.

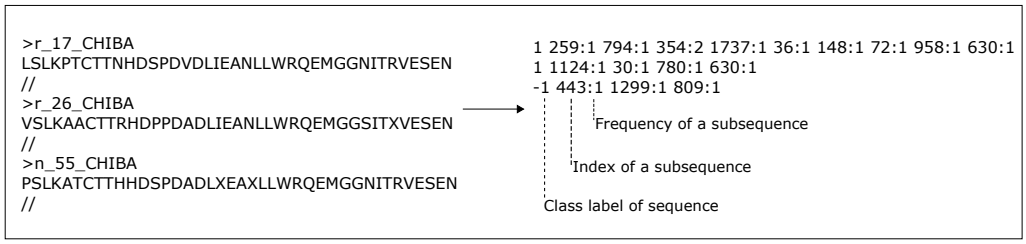

Figure 4.3: Data representation. A sequence is represented by many subsequences, subsequence frequencies, and class of that sequence. Each subsequence in a vector is represented by its index in the dictionary.

During the extraction process, the occurrence frequency of each subsequence is calculated. After this process ends, we obtain a dictionary of subsequences. Secondly, basing on the assumption that the significant regions appear more frequently than they are expected to preserve their structure and function during the evolution [Nevill-Manning et al., 1998], we remove subsequences having few occurrences in the whole dataset from the dictionary. Finally, a sequence is represented by a vector of frequencies of subsequences occurring in that sequence (Figure 4.3).

In this representation, we do not limit the length of a subsequence and thus a sequence is represented by subsequences with different lengths. Our approach differs from common representation methods, in which a sequence is represented by subsequences with identical lengths, however with this approach we want to keep information of a short sequence as much as possible. And in order to avoid dense representations, we do not choose subsequences contained in other subsequences to represent a sequence. We bias to choose subsequences whose length is longer.

(b) Characterization

The characterization step consists of 2 tasks, topical space searching and data projection, that work as follows.

Topical space searching: By using the two-step framework [Than et al., 2012],

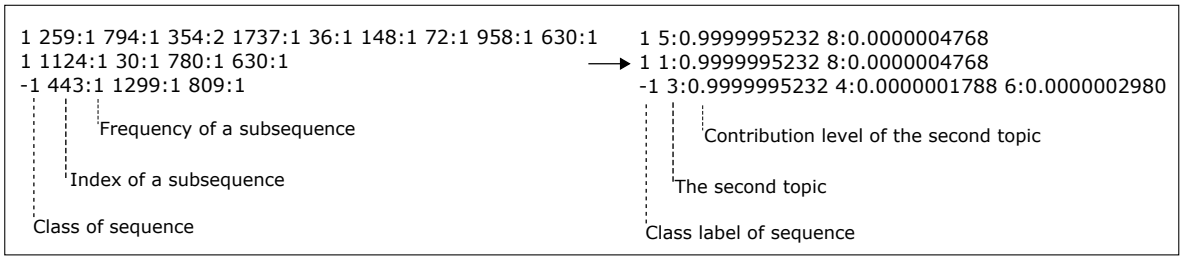

Figure 4.4: An example of projection of data onto 10-dimensioned topical space. A sequence in the topical space is a mixture of topics. Topics that do not have any contributions will be omitted in representation.

we find a discriminative space, or topical space, on which data are well separated. The KL (*Kullback-Leibler*) divergence is used in this step to find nearest neighbors of the data. KL divergence is a measure of the difference between two probability distributions, and often applied to measure the similarity among data points. Our data points are data vectors in which frequencies of subsequences are discrete values (Figure 4.4). Therefore, the use of KL divergence to choose k-nearest neighbors is reasonable for our problem. Despite of the reasons stated in [Than et al., 2012], many researches show the excellent performance of KL divergence for document classification, such as [McCallum and Nigam, 1998] and [Madsen et al., 2005]. In addition, finding k-nearest neighbors directly with sequence data instead of discrete data is also a potential approach that we will investigate in the future.

Data projection: We finally project data onto the new space, the topical space resulting from the previous step. In this step, data are projected without labels by using inference methods, such as variational methods for LDA [Blei et al., 2003], folding-ing in PLSA [Hofmann, 2001], and sparse inference in FSTM [Than and Ho, 2012].

(c) Prediction

To learn a classifier from data, we use support vector machine (SVM) [Fan et al., 2008]. In our framework, SVM learns a linear function working in the topical space and finds an optimal hyperplane such that the margin from two classes to this hyperplane is maximized. Note that, other methods such as decision tree or boosting can be used for prediction.

2. **Analysis of discriminative subsequences and topics**

Next, we discuss how to find subsequences and hidden topics (factors) that characterize each class (SVR/non-SVR) of sequences. In other words, we want to see which subsequences and which topics are potentially discriminative for a class. To

this end, some inherent properties of admixture topic models will be exploited. In the following, we will use concepts document and term in the argument. However, one can readily map documents to sequences, terms to subsequences, and topics to hidden factors.

It is often assumed in admixture topic models that a document is a distribution over hidden topics, and each topic is a distribution over words. These assumptions are employed by various models including LDA, PLSA, and FSTM. Let $\theta_{dk} = P(z_k|\boldsymbol{d})$ be the probability that topic k appears in document $\boldsymbol{d}$, and $\beta_{kj} = P(w_j|z_k)$ be the probability that term j contributes to topic k. These definitions basically imply that $\sum_{k=1}^{K} \theta_{dk} = 1$ for each document $\boldsymbol{d}$, and $\sum_{j=1}^{V} \beta_{kj} = 1$ for each topic k. Once a topic model is learned, we can assess $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K)$. The topic proportion $\boldsymbol{\theta}_d = (\theta_{d1}, ..., \theta_{dK})$ of document $\boldsymbol{d}$ can be found by doing inference for $\boldsymbol{d}$. Note that we can obtain $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_d$'s just after doing the first two steps of the proposed framework in Fig 4.2.

Discriminative topics: We first consider which topics are potentially discriminative for a class by assessing $\boldsymbol{\theta}_d$'s. Our key idea is to estimate the contributions of a topic to classes, and then contrast those contributions to find discriminative topics. Let $\mathcal{D}_1$ (and $\mathcal{D}_2$) be the set of sequences in class SVR (and non-SVR, resp.). The contribution of topic k to class c is approximated by

$$T_{ck} = \frac{\sum_{\boldsymbol{d} \in \mathcal{D}_c} P(z_k|\boldsymbol{d})}{\sum_{i=1}^{K} \sum_{\boldsymbol{d} \in \mathcal{D}_c} P(z_i|\boldsymbol{d})} = \frac{\sum_{\boldsymbol{d} \in \mathcal{D}_c} \theta_{dk}}{\sum_{i=1}^{K} \sum_{\boldsymbol{d} \in \mathcal{D}_c} \theta_{di}}. \tag{4.1}$$

If a topic k characterizes class c, then it is expected to contribute to class c significantly greater than to the other class. Hence, by contrasting T_{1k} and T_{2k}, one can decide which class topic k is discriminative for.

Discriminative terms: We find discriminative terms for a class by first estimating contributions of terms to classes and then contrasting those contributions. Note that the probability of term j appearing in document $\boldsymbol{d}$ is $P(w_j|\boldsymbol{d}) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|\boldsymbol{d}) = \sum_{k=1}^{K} \theta_{dk}\beta_{kj}$. Hence the contribution of term j to class c can be approximated by

$$W_{cj} = \frac{\sum_{\boldsymbol{d} \in \mathcal{D}_c} P(w_j|\boldsymbol{d})}{\sum_{r=1}^{V} \sum_{\boldsymbol{d} \in \mathcal{D}_c} P(w_r|\boldsymbol{d})} = \frac{\sum_{\boldsymbol{d} \in \mathcal{D}_c} \sum_{k=1}^{K} \theta_{dk}\beta_{kj}}{\sum_{r=1}^{V} \sum_{\boldsymbol{d} \in \mathcal{D}_c} \sum_{k=1}^{K} \theta_{dk}\beta_{kr}}. \tag{4.2}$$

Similar to the above argument, one can contrast W_{1j} and W_{2j} to see which class term j has more significant contribution. In practice, we are mostly interested in terms that have high contributions.

Discriminative terms by assembling models: It is possible to find discriminative terms by combining results from different models. Our framework for prediction of SVR/non-SVR allows employment of various topic models. Each employment

will result in a pretty different topics (or topical space). Such a randomness can yield different results when finding discriminative terms as described above. This behavior would potentially give us unstable sets of discriminative terms.

To obtain a stable set of discriminative terms, our idea is to assemble different models. More concretely, we first find a set of potentially discriminative terms for each topic model, as described before. We then remove all, but keep only terms that appear at least τ times. In order to guarantee statistically significant terms, the number of topic models should be larger than 30.

4.3 Results

4.3.1 Datasets

In this study, all sequence data are ISDR sequences before treatment of HCV genotype 1b. We use datasets as follow:

- 14 sequences (9 SVR and 5 non-SVR) from Chiba University
- 20 sequences (11 SVR and 9 non-SVR) from LANL[1] database
- 90 sequences (59 SVR and 31 non-SVR) from 4 published studies that analyzed the relationship between IFN/RBV therapy and ISDR sequences [Yoon et al., 2007, Rueda et al., 2008, Chayama et al., 1997, Enomoto et al., 1996].

4.3.2 Accuracy of prediction

We now investigate the effectiveness of our framework on prediction of SVR and non-SVR. Our framework is very general and flexible, and hence can result in various methods for prediction, by adapting to existing topic models. In our investigation however, only three topic models were taken into consideration, FSTM [Than and Ho, 2012], PLSA [Hofmann, 2001], and LDA [Blei et al., 2003]. The resulting methods for prediction are respectively FSTM^c, PLSA^c, and LDA^c. Because of no existing work for prediction of SVR/non-SVR in the case of very short and few sequences, we also investigate the effectiveness of support vector machines (SVM). We would like to remark that SVM is one of the state-of-the-art methods for doing classification in Machine Learning and Data Mining. We took SVM into comparison as a baseline.

[1]Los Alamos National Laboratory (http://hcv.lanl.gov).

Table 4.1: Accuracies of 7 methods for predicting sequences

No. of topics	FSTM^c	LDA^c	PLSA^c
5	91.9355	95.1613	79.8387
10	91.9355	91.9355	94.3548
20	91.9355	86.2903	89.5161
30	91.9355	90.3226	90.3226
40	91.9355	89.5161	91.1290

SVM-NormalizedPoly	SVM-Puk	SVM-RBF	SVM-Linear
68.14	69.50	63.70	59.75

In summary, we conducted prediction with 7 methods: FSTM^c, PLSA^c, LDA^c, SVM-linear, SVM-NormalizedPoly, SVM-Puk, and SVM-RBF[2]. Different kernels for SVM are investigated to ensure that the investigation is extensive for our problem, and also to show the advantages of our framework. For SVM-linear, SVM-NormalizedPoly, SVM-Puk, and SVM-RBF, we chose the best regularization parameter C from 1, 10, 100, 1000 by 5-folds cross-validation. High values of C essentially mean high penalties on the method when making error in the trainning data. When learning unsupervised topic models, we used default settings for their parameters and varied the number of topics. In our experiments with the two-steps framework, we found the following setting to be reasonable for our data: $K = 30, N_d = 1, R = 0, \lambda = 0$. This setting basically says that only the nearest neighbor plays a crucial role when doing projection for a document. $K = 30$ says that the topical space is rich enough to characterize our data. For all prediction methods, 5-folds cross-validation was used and prediction accuracy is averaged from 5 folds.

It can be observed from Table 4.1 that accuracies of each topic model working on our framework are better than the accuracies of SVM with 4 types of kernel functions. For example, with the number of topics is 10, the accuracies of FSTM^c, LDA^c, and PLSA^c are 91.9355%, 91.9355% and 94.3548% respectively, while the best accuracy of SVM is 69.50%. Among three topic models, with different settings, the accuracies of FSTM^c do not change much when compared with those of LDA^c, and PLSA^c. Therefore, it can be said that FSTM^c is more stable than LDA^c, and PLSA^c. The experimental results show that the setting of $\{K = 30, N_d = 1, R = 0, \lambda = 0\}$ will be good for all of three topic models. The reason making this setting reasonable is explained in Figure 4.6 of

[2]We use Weka to experiment SVM with different kernels (http://www.cs.waikato.ac.nz/ml/weka/).

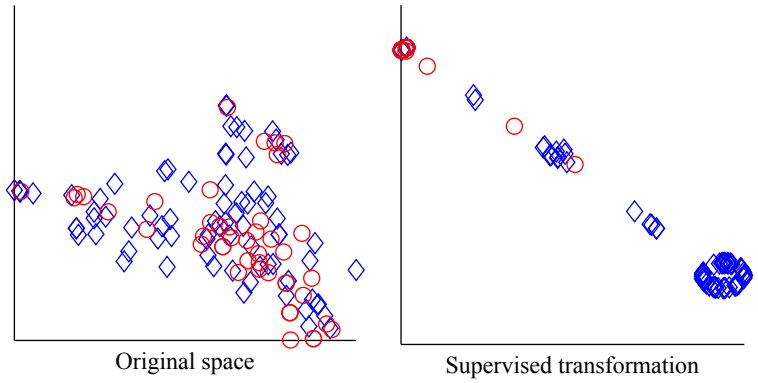

Figure 4.5: SVR and non-SVR sequences in the original space and in topical space. ◯ and ♢ are respectively non-SVR and SVR sequences.

the section C below where we will discuss the sensitivity of choosing a good setting for FSTM^c, LDA^c, and PLSA^c.

In the SVM experiments, the SVM-linear is the least effective method, because its accuracy is the lowest (59.75%). Other kernel functions such as SVM-RBF, SVM-NomalizedPoly, and SVM-Puk, work better than the SVM-linear, however their accuracy differences are negligible, about 10% comparing to the best accuracy (69.50%) of SVM-Puk.

Figure 4.5 compares the discrimination of data projected onto the original space and topical space. It is obviously that data on the topical space are discriminated better than data on the original space. Therefore, prediction on topical space is done more exactly and effectively. The bad separability of data in the original space may be the main drawback that worsen performance of SVM (even with different kernels).

4.3.3 Choosing a good setting for the framework

In this section, we will analyze the effect of the parameters to see which parameters play an important role in the proposed framework. A set of parameters consists of N_d, R, K, and λ, where N_d is a number of neighbors for a document to be used and is chosen from $\{1, 5, 10\}$; λ is a constant that combines a document with its neighbors, and is chosen from $\{0, 0.1, 0.5, 1\}$; R is a regularization constant and is chosen from $\{0, 1000, 10000\}$; and K is a number of topics that can receive values in $\{5, 10, 20, 30, 40\}$.

Parameters are examined one after another by making a parameter change its values while keeping the values of three remaining parameters. For example, when the parameter K is considered, parameters λ, N_d, and R are fixed and set to 0, 5, and 1000 respectively. Observing Figure 4.6, the parameter N_d plays an important role when $\lambda = 0$, in order

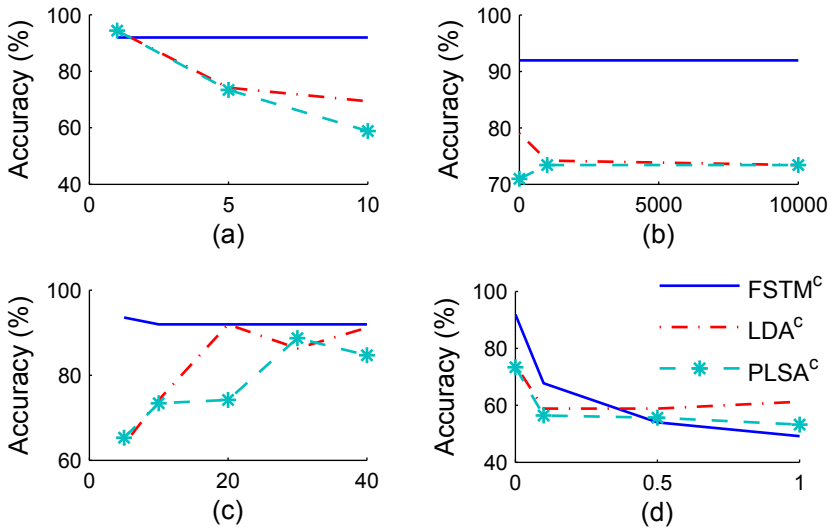

Figure 4.6: The impact of the parameters. (a) Changing N_d, while fixing $\lambda = 0$, $R = 1000$, and K=10. (b) Changing R, while fixing $N_d = 5$, $\lambda = 0$, and $K = 10$. (c) Changing K, while fixing $N_d = 5$, $\lambda = 0$, and $R = 1000$. (d) Changing λ, while fixing $N_d = 5$, $R = 1000$, and $K = 10$.

words when data are projected onto the new space, the nearest neighbor information is the most important. When the value of λ increases, the role of N_d decreases in searching a topical space. In our observation, parameter R does not affect much the results of experiments.

4.3.4 Discriminative subsequences

To find reliable subsequences that characterize SVR and non-SVR sequences, we opt to analyze qualified topic models. In our analysis, we chose a collection of topic models for which the associated prediction accuracy is at least 90%. The resulting collection consists of 75 models [3]. We believe that a model with a good quality of prediction (or a high accuracy) can provide good discriminative subsequences. Furthermore, in order to guarantee the statistical significance of a subsequence, selected subsequences have to be highly evaluated by at least in 30 models. Table 4.2 presents 39 potential discriminative subsequences that are found from 75 models for each class of SVR and non-SVR.

Observing Table 4.2, all subsequences are considered important in a class label of sequences, and not important in the remaining class. In other words, subsequences are rated for one class by many models while being rated for the remaining class by very few models. This suggests that these subsequences have the ability to discriminate classes very well. Let us take some examples of subsequences that have very high contributions

[3]Note that each choice of $\{K, N_d, \lambda, R\}$ and $\{PLSA, LDA, FSTM\}$ will result in a specific model.

Table 4.2: Discriminative subsequences characterizing SVR and non-SVR outcomes of HCV treatment

Subsequence	No. of models rating for SVR	No. of models rating for non-SVR
AA	0	33
ANLLW	0	33
ATCTTRHDSPD	2	43
ATY	15	35
CTTNHDSPD	3	51
DSPDADLIEANLLW	8	40
DSPDVDLIEANLLWRQEMGGNITRVESEN	0	64
EANLLWRQEMGGSITRVESEN	5	31
EK	11	35
EV	18	42
GGDITRVESEN	6	32
HDSPDADLIEANLLWRQEMGGNITRV	38	11
HDSPDV	3	32
HDSPDVDLIEANLLWRQEMGG	4	61
HHDSPD	6	38
KATCTTHHDSPD	3	32
LIEANLLW	3	32
LSLK	2	58
LSLKAACT	2	48
LSLKATC	6	39
PDL	33	11
PSLKATC	19	31
PSLKATCTA	44	6
PSLKATCTTH	6	38
PSLKATCTTHHDSPDADLI	0	33
PSLRATCTT	14	39
PSSK	9	38
QE	6	35
QEMGGNITRVESEN	0	45
RH	11	38
RHDSPD	2	39
SE	34	6
SKAT	9	38
SLKAACTT	5	38
TCTTNHDS	9	47
THHDSPDADLIEANLLWRQEMGGNITRVESE	5	32
THRDSPD	6	48
TQ	17	41
TYIT	9	35

to only one class, "HDSPDVDLIEANLLWRQEMGG" is rated by 61 models for non-SVR while only 4 models vote it for SVR (its contributions to SVR/non-SVR are respectively $W_{SVR} = 0.0005$ and $W_{non-SVR} = 0.0602$) or "ANLLW" has no models rating for SVR, but 33 models rate it for non-SVR (its contributions are $W_{SVR} = 0,0005$ and $W_{non-SVR} = 0.0054$) or "PSLKATCTA" is rated by 6 models for non-SVR while 44 models vote it for SVR (its contributions are $W_{SVR} = 0.0463$ and $W_{non-SVR} = 0.0086$). One can say that subsequences such as "ANLLW", "DSPDVDLIEANLLWRQEMGGNITRVESEN", or "QEMGGNITRVESEN" can be good discriminative subsequences for predicting non-SVR sequences, and subsequences such as "PSLKATCTA" or "SE" can be potential candidates to predict SVR sequences.

Comparison with existing method: By using the web version of MEME[4] to find discriminative subsequences (or motifs) with ISDR data, we want to know the differences between the results of MEME and ours. We choose MEME because it is currently one of the most well known and widely used software for searching discriminative motifs. For SVR, MEME found 3 motifs, "HHDSPDADLIEANLLWRQEMGGNITRVES", "PSLKATCT", and "EN"; for non-SVR, MEME found only 1 motif with a length of 40 amino acids, "PSLKATCTTHHDSPDADLIEANLLWRQEMGGNITRVESEN".

To evaluate the discriminative ability of SVR motifs found by MEME, we count the number of occurrences of them in both SVR and non-SVR sequences. "HHDSPDADLIEANLLWRQEMGGNITRVES" is found 6 times in non-SVR and 5 times in SVR; "PSLKATCT" is found 31 times in non-SVR and 26 times in SVR. It is clearly that these motifs are not good enough to discriminate SVR and non-SVR sequences. In the experimental results of DPA, not many models voted for these motifs: with "HHDSPDADLIEANLLWRQEMGGNITRVES", we have 11 models voting for SVR and 17 models voting for non-SVR; and with "PSLKATCT", we just have 8 models voting for SVR and 9 models voting for non-SVR. In our observation, "EN" is a basic characteristic of data and cannot be a discriminative motif for SVR or non-SVR, because it occurs in nearly almost sequences of the data.

Complication of non-SVR sequences and consequences: Table 4.2 shows that subsequences voted for non-SVR dominate ones voted for SVR. About 35 subsequences were found to potentially characterize for non-SVR in total of 39 subsequences. Further, each subsequence which is assumed to characterize for non-SVR often receives a high number of rates from 75 models. These observations suggest the diversity and complication of non-SVR sequences. And because of these diversity and complication, the prediction problem can meet some difficulties to give exact results. That is also the reasons that SVM methods work ineffectively.

[4]Multiple Em for Motif Elicitation (http://meme.nbcr.net/meme/cgi-bin/meme.cgi).

Connection to experimental research: In our observation, although the number of SVR sequences (79 sequences) is larger than the number of non-SVR sequences (45 sequences) in our dataset, most subsequences in Table 4.2 are rated highly for non-SVR. This helps us confirm that our findings are appropriate to the current state of the art of HCV study in which HCV genotype 1b (HCV-1b) is the least response to IFN/RBV therapy. The SVR rate of HCV-1b (42 to 52%) is lower than SVR rates of other HCV genotypes (50 to 80%) [Wohnsland et al., 2007, El-Shamy et al., 2011].

Significance in practice of HCV treatment: With characteristics that are found for non-SVR in our experiments, we believe that it will lead to better understandings of the resistance to IFN/RBV therapy of HCV. The diversity and complication of HCV-1b through our findings contribute to explain the reason why it is very difficult to treat HCV-1b. Addition to, these promising findings provide physicians hints or biomarkers in order to get a better treatment that avoids side effects and saves expense for patients.

4.4 Conclusion

We have proposed a novel computational approach to characterize and predict SVR/non-SVR outcomes of HCV treatment by using topic modeling. Our framework was demonstrated to search effectively a topical space (or discriminative space), represent well sequence data into a document space as well as discriminative space, and interpret results of computational process. The proposed framework also works effectively in a special case of data (sequences are short in length and resemble each other) that traditional methods could not overcome. Further, it has shown to be general and flexible and can be applied many kind of data.

The quality of the prediction method in this framework often outperforms the baseline method and can reach more than 30% improvement. The subsequences we obtained are promising and when verified by physicians, they can be good discriminative patterns for predicting SVR/non-SVR sequences.

Regarding my own contributions in this work, I have contributed the following points:

1. The optimal representation for data. Concretely, from the properties of sequence data, I found the new way to represent them into the properties of documents in order to get the effectiveness of the proposed method. The main property of the document representation is very sparse in a multiple dimensional space. Therefore, finding the way to represent sequence data so that this representation is a sparse high-dimensional representation is the important step for the whole framework. In other words, with short sequence data, 40 amino acids in length, it is very hard

for methods to discover the optimal discriminative motifs, because short input data are lacking in information for searching process. I decided to approach the problem in two stages: (i) enrich short sequences by a representation in a high dimensional space containing subsequences and then (ii) discriminate these subsequences.

2. Computational elucidation for experimental results. In details, I first carried out the experimental performance, next I performed a comparison of the results of the proposed method with those of other methods that are popular in biological field, interpreted the meaning of subsequence patterns and finally found connections between subsequence patterns and HCV genotype 1b.

Chapter 5

Conclusion

This final chapter summarizes the research problem and solutions that the dissertation has done and achieved. Limitations and future research directions of the dissertation are also addressed.

5.1 Dissertation summary

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, such as [Bailey et al., 2010, Redhead and Bailey, 2007, Segal et al., 2002, Lin et al., 2011, Huggins et al., 2011, Kim and Choi, 2011, Smith et al., 2005, Leung and Chin, 2006, Sinha, 2006], and with a string-based model, such as [Vens et al., 2011, Fauteux et al., 2008, Mason et al., 2010, Mehdi et al., 2013, Narang et al., 2010, Jr. and Liang, 2010]. 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?

We have presented our study on discriminative motif learning for HCV treatment with considerations of computational and biomedical aspects. Results of the dissertation are the novel computational methods for discover discriminative motifs and potential discriminative motifs that are able to help to predict signals of response or resistance to IFN/RBV therapy, together with the additional insight of the relation between NS5A protein and IFN/RBV therapy. 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, 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.

5.2 Future Work

In order to go further in the computational perspective from this dissertation, we aim to extend our study at the following points:

- During topical space searching of the second work, the KL (*Kullback-Leibler*) divergence is used to find nearest neighbors of the data. Many researches have shown the excellent performance of KL divergence for document classification, but finding k-nearest neighbors for sequence data should be performed directly with sequences instead of discrete data.

- The process of searching neighbors is currently using labeled data. In order to exploit the available unlabeled data, we suggest to use semi-supervised learning to search neighbors.

- In the discriminative motif analysis, we need to connect more experimental evidences to findings of the dissertation, for example discriminative motifs contain mutation or not, which types of mutation do discriminative motifs contain?, and which position do mutations happen? Or we should analyze the simultaneous occurrences of two or more discriminative motifs.

Regarding future works for HCV study, we have the following remarks:

- With the appearance of a new drug, TVR (*telaprevir*), for HCV treatment, patients who had failed with IFN/RBV therapy are highly likely to achieve a SVR when their treatment is with the addition of telaprevir [McHutchison et al., 2010, Zeuzem et al., 2011, Jacobson et al., 2011]. Telaprevir offers an additional to the HCV treatment because the standard therapy of IFN/RVB is less effective, especially on genotype 1. Although telaprevir has a side effect, the most common side effect is rash, telaprevir is currently indicated for use against HCV genotype 1 in USA and Japan. Therefore, the study of IFN/RBV therapy for HCV genotype 1 is not necessary in the future, but the study of triple combination therapy may be considered.

- The IL-28B polymorphism can help to explain individual and racial differences in response to HCV treatment. Therefore, the study on variations of IL-28B gene may be needed.

Bibliography

[Admadi and Beigy, 2012] Admadi, Z. and Beigy, H. (2012). Semi-supervised Ensemble Learning of Data Streams in the Presence of Concept Drift. *Hybrid Artificial Intelligent Systems*, pages 526–537.

[AIDSInfoNet, 2012] AIDSInfoNet (2012). Hepatitis C Virus Life Cycle, Fact Sheet, No. 670.

[Alestig et al., 2011] Alestig, E., Arnholm, B., Eilard, A., Lagging, M., Nilsson, S., and ans et al, G. N. (2011). Core mutations, IL28B polymorphisms and response to peginterferon/ribavirin treatment in Swedish patients with hepatitis C virus genotype 1 infection. *BMC Infectious Diseases*, 11.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). A basic local alignment search tool. *Journal of Melecular Biology*, 215:403–410.

[Aurora et al., 2009] Aurora, R., Donlin, M., A.Cannon, N., and Tavis, J. E. (2009). Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans. *The Journal of Clinical Investigation*, 119.

[Bailey et al., 2010] Bailey, T. L., Boden, M. B., Whitington, T., and Machanick, P. (2010). The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics*, 11(1).

[Beeck and Dubuisson, 2003] Beeck, A. O. D. and Dubuisson, J. (2003). Topology of hepatitis C virus envelope glycoproteins. *Medical Virology*, 13(4):233–241.

[Bernhard et al., 1997] Bernhard, S., Alexander, S., and Klaus-Robert, M. (1997). Kernel principal component analysis. In Wulfram, G., Alain, G., Martin, H., and Jean-Daniel, N., editors, *Artificial Neural Networks (ICANN'97)*, volume 1327 of *Lecture Notes in Computer Science*, pages 583–588. Springer Berlin Heidelberg.

[Blei, 2012] Blei, D. M. (2012). Probabilistic Topic Models. *Communicate of the ACM*, 55(4):77–84.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[Brodsky et al., 2007] Brodsky, L. I., Wahed, A. S., Tavis, J. E., Tsukahara, T., and et al (2007). A novel unsupervised method to identify genes important in the anti-viral response: Application to interferon/ribavirin in hepatitis C patients. *PLoS ONE*, 2.

[Brown, 2010] Brown, G. (2010). *Ensemble Learning - Encyclopedia of Machine Learning*. Springer Press, Berlin Heidelberg.

[Chapelle et al., 2006] Chapelle, O., Shoolkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press, Cambridge, Massachusetts.

[Chayama and Hayes, 2011] Chayama, K. and Hayes, C. N. (2011). Hepatitis C virus: how genetic variability affects pathobiology of disease. *Journal of Gastroenterology and Hepatology*, 26:83–95.

[Chayama et al., 1997] Chayama, K., Tsubota, A., Kobayashi, M., Okamoto, K., Hashimoto, M., Miyano, Y., and et al (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.

[Choo et al., 1989] Choo, Q.-L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W., and Houghton, M. (1989). Isolation of a cDNA Clone Derived from a Blood-Gorne Non-A, Non-B Viral Hepatitis Genome. *Science*, 224:359–362.

[Colm, 2008] Colm, G. (2008). Structure of Hepatitis C Virus.

[Conklin et al., 1993] Conklin, D., Fortier, S., and Glasgow, J. (1993). Representation for Discovery of Protein Motifs. *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, pages 101–108.

[Craven, 2011] Craven, M. (2011). Lecture Notes - Learning Sequence Motif Models using EM.

[Cuevas et al., 2009] Cuevas, J. M., Torres-Puente, M., Jimenez-Hernandez, N., Bracho, M. A., Garcia-Robles, I., Carnicer, F., and et al (2009). Combined therapy of interferon plus ribavirin promotes multiple adaptive solutions in hepatitis C virus. *Journal of Medical Virology*, 81:650–656.

[Das and Dai, 2007] Das, M. K. and Dai, H. K. (2007). A survey of DNA motif finding algorithms. *MBC Bioinformatics*, 8(7).

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[D'haeseleer, 2006] D'haeseleer, P. (2006). How does DNA sequence motif discovery work? *Nature Biotechnology*, 24(8):959–961.

[Dong and Schafer, 2011] Dong, C. and Schafer, U. (2011). Ensemble-style Self-training on Citation Classification. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 623–631.

[El-Shamy et al., 2011] El-Shamy, A., Shoji, I., Saito, T., Watanabe, H., Ide, Y., Deng, L., and et al. (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.

[ElHefnawi et al., 2010] ElHefnawi, M., Zada, S., and El-Azab, I. A. (2010). Prediction of prognostic biomarkers for interferon-based therapy to hepatitis C virus patients: A meta-analysis of the NS5A protein in subtypes 1a, 1b and 3a. *Virology Journal*, 7.

[Enomoto et al., 1996] Enomoto, N., Sakuma, N., Asahina, I., Kurosaki, Y., Murakami, M., Yamamoto, T., and et al (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.

[Fan et al., 2008] Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

[Fauteux et al., 2008] Fauteux, F., Blanchette, M., and Stromvik, M. V. (2008). Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, 24(20):2303–2307.

[Gale et al., 1997] Gale, M. J. J., Korth, M. J., Tang, N. M., Tan, S. L., Hopkins, D. A., Dever, T. E., and et al (1997). Evidence that hepatitis C virus resistance to interferon is mediated through repression of the PKR protein by the nonstructural 5A protein. *Virology*, 230:217–227.

[Gao et al., 2010] Gao, M., Nettles, R. E., Belema, M., Snyder, L. B., Nguyen, V. N., Fridell, R. A., and et al (2010). Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. *Nature Letters*, 465:96–100.

[Guilou-Guillemette et al., 2007] Guilou-Guillemette, H. L., Vallet, S., Gaudy-Graffin, C., Payan, C., Pivert, A., Goudeau, A., and Lunel-Fabiani, F. (2007). Genetic diversity of the hepatitis C virus: Impact and issues in the antiviral therapy. *World Journal of Gastroenterology*, 13:2416–426.

[Ho et al., 2011] Ho, T., Kawasaki, S., Le, N., Kanda, T., Le, N., Takabayashi, K., and Yokosuka, O. (2011). Finding HCV NS5A discriminative motifs for assessment of INF/RBV therapy Effect. *Workshop Data Mining in Genomics and Proteomics, International Conference ECML/PKDD*.

[Ho, 2004] Ho, T. B. (2004). Grant-in-Aid for Scientific Research (B) (2004-2006) Discovery of Hepatitis Knowledge by Data Mining Methods from Various Sources.

[Ho, 2007] Ho, T. B. (2007). Grant-in-Aid for Scientific Research (B) (2007-2009) Advanced Computing Techniques for Scientific Data.

[Ho, 2011] Ho, T. B. (2011). Grant-in-Aid for Scientific Research (B) (2011-2014) Elucidation of the molecular mechanism related to pathogenesis and treatment of hepatitis by computational approach.

[Ho and Nguyen, 2002] Ho, T. B. and Nguyen, D. D. (2002). Chance Discovery and Learning Minority Classes. *New Generation Computing*, 21(2).

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

[Hoofnagle, 1994] Hoofnagle, J. H. (1994). Therapy of acute and chronic viral hepatitis. *Advances in Internal Medicine*, 39:241–275.

[Hsu et al., 2011] Hsu, C. M., Chen, C. Y., and Liu, B. J. (2011). WildSpan: mining structured motifs from protein sequences. *Algorithms for Molecular Biology*, 6(6).

[Huang et al., 2010] Huang, S., Li, K., Dai, X., and Chen, J. (2010). Improving Word Alignment by Semi-supervised Ensemble. *Proceedings of the Fourth Conference on Computational Natural Language Learning*, pages 135–143.

[Huggins et al., 2011] Huggins, P., Zhong, S., Shiff, I., Beckerman, R., Laptenko, O., Prives, C., Schulz, M. H., Simon, I., and Bar-Joseph, Z. (2011). DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics*, 27(17):2361–2367.

[Imran et al., 2012] Imran, M., Waheed, Y., Manzoor, S., Bilal, M., Ashraf, W., Ali, M., and Ashraf, M. (2012). Interaction of hepatitis C virus with pattern recognition receptors. *Virology Journal*, 9.

[Jacobson et al., 2011] Jacobson, I. M., McHutchison, J., Dusheiko, G., and et al (2011). Telaprevir for previously untreated chronic hepatitis C virus infection. *The New England Journal of Medicine*, 364(25):2405–2416.

[Jr. and Liang, 2010] Jr., R. J. and Liang, J. (2010). Combinatorial Analysis for Sequence and Spatial Motif Discovery in Short Sequence Fragments. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):524–536.

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

[Lawrence and Reilly, 1990] Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7:41–51.

[Leung and Chin, 2006] Leung, H. C. M. and Chin, F. Y. L. (2006). Finding motifs from all sequences with and without binding sites. *Bioinformatics*, 22(18):2217–2223.

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

[Macdoanldt and Harris, 2004] Macdoanldt, A. and Harris, M. (2004). Hepatitis C virus NS5A: tales of a promiscuous protein. *Journal of General Virology*, 85:2485–2502.

[Madsen et al., 2005] Madsen, R. E., Kauchak, D., and Elkan, C. (2005). Modeling word burstiness using the Dirichlet distribution. In *Proceeding of the 22nd International Conference on Machine Learning*, ICML' 05, pages 545–552, New York, NY, USA. ACM.

[Manns et al., 2001] Manns, M., McHutchison, J. G., Gordon, S. C., Rustgi, V. K., Shiffman, M., Reindollar, R., and et al (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.

[Mason et al., 2010] Mason, M. J., Plath, K., and Zhou, Q. (2010). Identification of Context-Dependent Motifs by Contrasting ChIP Binding Data. *Bioinformatics*, 26(22):2826–2832.

[McCallum and Nigam, 1998] McCallum, A. and Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48.

[McHutchison et al., 2010] McHutchison, J. G., Manns, M. P., Muir, A. J., Terrault, N. A., Jacobson, I. M., Afdhal, N. H., and et al (2010). Telaprevir for Previously Treated Chronic HCV Infection. *The New England Journal of Medicine*, 362(14):1292–1303.

[Mehdi et al., 2013] Mehdi, A. M., Sehgal, M. S. B., Kobe, B., Bailey, T. L., and Boden, M. (2013). DLocalMotif: a discriminative approach for discovering local motifs in protein sequences. *Bioinformatics*, 29(1):39–46.

[Moradpour et al., 2007] Moradpour, D., Penin, F., and Rice, C. M. (2007). Replication of hepatitis C virus. *Nature Reviews*, 5:453–463.

[Motoda, 2001] Motoda, H. (2001). Grant-in-Aid for Scientific Research on Priority Areas (B) (2001-2004) Realization of Active Mining in the Information Flood Age.

[Murakami et al., 2010] Murakami, Y., Tanaka, M., Toyoda, H., Hayashi, K., Kuroda, M., Tajima, A., and Shimotohno, K. (2010). Hepatic microRNA expression is associated with the response to interferon treatment of chronic hepatitis C. *BMC Medical Genomics*, 3.

[Narang et al., 2010] Narang, V., Mittal, A., and Sung, W.-K. (2010). Localized motif discovery in gene regulatory sequences. *Bioinformatics*, 26(9):1152–1159.

[Nevill-Manning et al., 1998] Nevill-Manning, C. G., Wu, T. D., and Brutlag, D. L. (1998). Highly specific protein sequence motifs for genome analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11):5865–5871.

[Pascu et al., 2004] Pascu, M., Martus, P., Hohne, M., Wiednmann, B., Hopf, U., E.Schreir, and Berg, T. (2004). Sustained virological response in hepatitis C virus type 1b infected patients is predicted by the number of mutations within the NS5A-ISDR: A meta-analysis focused on geographical differences. *Gut*, 53:1345–1351.

[Polyak et al., 2001] Polyak, S. J., Khabar, K. S., Paschal, D. M., Ezelle, H. J., Duverlie, G., Barber, G. N., and et al (2001). Hepatitis C virus nonstructural 5A protein induces interleukin-8, leading to partial inhibition of the interferon-induced antiviral response. *Journal of Virology*, 75:6095–6106.

[Redhead and Bailey, 2007] Redhead, E. and Bailey, T. L. (2007). Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, 8.

[Rosenberg et al., 2005] Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. *Proceeding of the Seventh Workshop on Applications of Computer Vision*, 1:29–36.

[Roweis and Saul, 2000] Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

[Rueda et al., 2008] Rueda, P. M., Casado, J., Patn, R., Quintero, D., Palacios, A., Gila, A., and et al (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.

[Saito et al., 1990] Saito, I., Miyamura, T., Ohbayashi, A., Harada, H., Katayama, T., Kikishi, S., and et al (1990). Hepatitis C virus infection is associated with the development of hepatocellular carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, 87:6547–6549.

[Sami and Nagatomi, 2008] Sami, A. and Nagatomi, R. (2008). *A new definition and look at DNA motif*, chapter 13. Intech.

[Sarrazin et al., 1999] Sarrazin, C., Berg, T., Lee, J., Teuber, G., Dietrich, C., Roth, W., and Zeuzem, S. (1999). Improved correlation between multiple mutations within NS5A region and virological response in European patients chronical infected with hepatitis C virus type 1b undergoing combination therapy. *Journal of Hepatology*, 30:1004–1013.

[Segal et al., 2002] Segal, E., Barash, Y., Simon, I., Friedman, N., and Koller, D. (2002). From promoter sequence to expression: a probabilistic framework. pages 263–272.

[Simmonds et al., 2005] Simmonds, P., Bukh, J., Combet, C., Delage, G., Enomoto, N., Feinstone, S., and et al (2005). Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology*, 42(4):962–973.

[Sinha, 2003] Sinha, A. (2003). Discriminative Motifs. *Journal of Computation Biology*, 10.

[Sinha, 2006] Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, 22(14):e454–e463.

[Smith et al., 2005] Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proceeding of the National Academy of Sciences of the United States of America*, 102(5):1560–1565.

[Smith and Waterman, 1981] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197.

[Suhalim, 2007] Suhalim, J. (2007). Modeling the Interferon Signaling Process of the Immune Response.

[Tenenbaum et al., 2000] Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323.

[Than and Ho, 2012] Than, K. and Ho, T. B. (2012). Fully sparse topic models. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 7523:490–505.

[Than et al., 2012] Than, K., Ho, T. B., Pham, N. K., and Nguyen, D. K. (2012). Supervised dimension reduction with topic model. *Asian Conference on Machine Learning (ACML)*, 25.

[Tompa, 1999] Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proceedings of the Seventh International Conference on Intelligent Systems on Molecular Biology*, pages 262–271.

[Torres-Puente et al., 2008] Torres-Puente, M., Cuevas, J. M., Jimenez-Hernandez, N., Bracho, M. A., Garcia-Robles, I., Carnicer, F., and et al (2008). Hepatitis C virus and the controversial role of the interferon sensitivity determining region in the response to interferon treatment. *Journal of Medical Virology*, 80:247–253.

[Vajda et al., 2011] Vajda, S., Junaidi, A., and Fink, G. A. (2011). A Semi-Supervised Ensemble Learning Approach for Character Labeling with Minimal Human Effort. *IEEE International Conference on Document Analysis and Recognition*, pages 259–263.

[Vens et al., 2011] Vens, C., Rosso, M. N., and Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, 27(9):1231–1238.

[WHO, 2012] WHO (2012). Hepatitis C, Fact Sheet, No. 164.

[Witherell and Beineke, 2001] Witherell, G. and Beineke, P. (2001). Statistical analysis of combined substitutions in nonstructural 5A region of hepatitis C virus and interferon response. *Journal of Medical Virology*, 63:8–16.

[Wohnsland et al., 2007] Wohnsland, A., Hofmann, W. P., and Sarrazin, C. (2007). Viral determinants of resistance to treatment in patients with Hepatitis C. *Clinical Microbiology Reviews*, 20(1):23–38.

[Woo and Park, 2012] Woo, H. and Park, C. H. (2012). Semi-supervised Ensemble Learning Using Label Propagation. *Proceedings of the 12th International Conference on Computer and Information Technology*, pages 421–426.

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

[Yhu et al., 2012] Yhu, J., Ahmed, A., and Xing, E. P. (2012). Medlda: maximum margin supervised topic models. *The Journal of Machine Learning Research*, 13:2237–2278.

[Yoon et al., 2007] Yoon, J., Lee, J. I., Baik, S. K., Lee, K. H., Sohn, J. Y., Lee, H. W., and et al (2007). Predictive factors for interferon and ribavirin combination therapy in patients with chronic hepatitis C. *World Journal of Gastroenterology*, 13(46):6236–6242.

[Zeuzem et al., 2011] Zeuzem, S., Andreone, P., Pol, S., and et al (2011). Telaprevir for retreatment of HCV infection. *The New England Journal of Medicine*, 364(25):2417–2428.

[Zhou, 2009] Zhou, Z.-H. (2009). When Semi-Supervised Learning Meets Ensemble Learning. *Multiple Classifier Systems*, pages 529–538.

[Zhou and Li, 2005] Zhou, Z.-H. and Li, M. (2005). Tri-Training: Exploiting Unlabeled Data Using Three Classifiers. *IEEE Transaction on Knowledge and Data Engineering*, 17(11):1529–1541.

[Zhu, 2008] Zhu, X. (2008). Tutorial on Semi-Supervised Learning - ICML.

Publications

[1] T.N. Le, T.B. Ho, T. Kanda, S. Kawasaki, K. Takabayashi, S. Wu and O. Yokosuka: "A Semi-Supervised Ensemble Learning Method for Finding Discriminative Motifs and Its Application," Journal of Universal Computer Science, Special Issue on Hybrid and Ensemble Methods in Machine Learning, Vol. 19, Issue 4, pp 563-580, April 2013.

[2] T.N. Le and T.B. Ho: "A Semi-Supervised Method for Discriminative Motif Finding and Its Application to Hepatitis C Virus Study," Proceedings of The 4th Asian Conference on Intelligent Information and Database Systems, Kaohsiung, Taiwan, pp.377-384, March 2012.

[3] T.B. Ho, S. Kawasaki, N.T. Le, T. Kanda, T.N. Le, K. Takabayashi and O. Yokosuka: "Finding HCV NS5A Discriminative Motifs for Assessment of IFN/Ribavarin Therapy Effect," Workshop Data Mining in Genomics and Proteomics, International Conference ECML/PK, Athens, September 2011.

[4] S. Kawasaki, T.B. Ho, T. Kanda, O. Yokosuka, K. Takabayashi and T.N. Le: "Discovering Relationship between Hepatitis C Virus NS5A Protein and Interferon/Ribavirin Therapy," The 5th International Conference on Knowledge, Information and Creativity Support Systems, Chiangmai, Thailand, November 2010.