

# Finding HCV NS5A Discriminative Motifs for Assessment of IFN/Ribavirin Therapy Effect

Tu Bao Ho<sup>†</sup>, Saori Kawasaki<sup>†</sup>, Ngoc Tu Le<sup>†</sup>, Tatsuo Kanda<sup>‡</sup>,  
Nhan Le<sup>†</sup>, Katsuhiko Takabayashi<sup>\*</sup>, Osamu Yokosuka<sup>‡</sup>

<sup>†</sup>Japan Advanced Institute of Science and Technology

<sup>‡</sup>Graduate School of Medicine Chiba University, <sup>\*</sup>Chiba University Hospital, Japan

**Abstract.** The objective of this paper is twofold. On one hand, it aims to develop two algorithms of discriminative motif discovery from a small labeled dataset and a large unlabeled dataset. One uses exhaustive search for motifs of type ‘discriminative one occurrence per sequence’ and the other uses separate-and-conquer search for motifs of type ‘discriminative multiple occurrences per sequence’. On the other hand, the two algorithms are applied to detect sequence motifs in NS5A protein that characterize sustained viral response and non-sustained viral response when treating hepatitis C virus (HCV) with standard interferon/ribavirin. Strong discriminative and frequent motifs characterizing the therapy effect in two sub-genotypes *1ac* and *1b* of HCV were detected and analysed.

## 1 Introduction

Using interferon combined with ribavirin (IFN/RBV) is currently the standard therapy of hepatitis C virus (HCV). However, only near a half of the HCV infected individuals achieves sustained viral response (SVR) to this therapy, and the genetic basis of resistance to antiviral therapy remains unknown [1], [2].

The NS5A protein in HCV genome is known as the protein most reported to be implicated in the interferon resistance [3]. Indeed, understanding new functions and mysteries of NS5A has attracted tremendous attention in the HCV field and presents a promising target for anti-HCV therapy [1], [4]. In particular, there are numerous studies about the relation between NS5A and IFN/RBV therapy, e.g., [2], [5], [1], [6], [7]. However, deep understanding of inhibitions of the therapy effect remains unknown [8]. One critical difficulty in study of NS5A and IFN/RBV therapy is the very limited number of labeled data.

The research on motif discovery recently focuses on finding for discriminative motifs. There are several newly developed methods using HMM [9], seeding DNA [10], probabilistic models [11] or association mining with domain knowledge [12], however, those methods may not function well when the data is small.

This work aims to develop two algorithms of discriminative motif discovery from a small labeled dataset and a large unlabeled dataset, and apply them to detect sequence motifs in NS5A protein that characterize sustained viral response and non-sustained viral response when treating hepatitis C virus (HCV) with standard interferon/ribavirin.

## 2 Background

### 2.1 NS5A protein

The typical functions of NS5A include: (i) NS5A interacts with IFN cellular antiviral pathways and thus inhibits the IFN $\alpha$  activity, especially in PKR (Protein Kinase R) [3], [8]; (ii) The mutations in the NS5A resist the IFN therapy, especially in Interferon Sensitivity Determining Region (ISDR) [3].

The HCV NS5A protein is also viewed in three domains I, II and III, which are separated by low complexity sequence blocks of repetitive amino acids. Domains II and III are more variable among HCV genotypes. Recent studies showed that domain II contains the PKR and the HCV NS5B binding domains and domain III could influence RNA replication [13].

- Concerning the role of NS5A in RNA replication, most replication phenotypes mapping to the domain I of NS5A. What is the remained enigmatic role of the domains II and III in the HCV lifecycle [4]?
- Can V3 region be a more accurate biomarker than the ISDR region [7]?
- Larger size of NS5A samples may be needed for investigating its role in the therapy outcome of HCV 1a/1b infection [2]. How to deal with it when NS5A data related to IFN/RBV therapy SVR and non-SVR are limited?

### 2.2 Interferon and ribavirin therapy

Interferons (IFNs) are a group of small proteins made by the human body in response to viral infections. IFN therapy is thought to work by stimulating processes within cells which help to slow down the reproduction and growth of the virus.

Ribavirin (RBV) is a drug that has activity against some viruses. When ribavirin is used in combination with the drug interferon, researchers have found that about twice as many people as those using interferon alone show a long term clearance of detectable HCV from the blood. The general mechanisms of resistance to IFN/ribavirin, in particular NS5A and IFN $\alpha$  resistance and many questions are open to uncover are described in [14]. Some questions have not yet received the answers from the research community:

- No significantly different virologic response rates between HCV subtypes were reported.
- It must be assumed that additional mutations outside the NS5A but within HCV ORF are important for determinations of IFN sensitivity.
- Nothing is known about the frequency of UA and UU dinucleotides before and during IFN based therapy in correlation with HCV RNA clearance.

## 3 Discriminative motif discovery

A *sequence motif* is generally understood as a nucleotide or amino acid pattern that is widespread and has a biological significance, commonly as transcription factor binding sites (TFBSs). Traditionally, the motif finding problem has

been dominated by generative models using only one sequence class to produce descriptive motifs of the class. Recently, research focuses on discovering of discriminative motifs those can be used to distinguish sequences belonging to two different classes [4], [11], [10], [12].

Motif discovery methods can broadly be divided into two groups, inductive (or ‘sequence-driven’) methods and enumerative (or ‘pattern-driven’) methods. The former methods involve in building a search space of potential motifs from sequence data then using search strategies to look for solutions. Though routinely fast these methods do not guarantee to yield the global solution. The latter, on the other hand, can yield the global solution but is computationally expensive and cannot be applied to large datasets and limited to short motifs.

Denote  $S = \{(S_1, C_1), (S_2, C_2) \dots, (S_n, C_n)\}$ , where  $S_i$  is a sequence of length  $|S_i|$  over the alphabet  $\Sigma = \{A, U, G, T\}$  or  $\Sigma = \{amino\ acid\}$  and  $C_i \in \{C_1, C_2, \dots, C_c\}$  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  $S = Pos \cup Neg$ .

Denote by  $cover_{Pos}(P)$  and  $cover_{Neg}(P)$  the set of sequences in  $Pos$  and  $Neg$  each contains a given sequence  $P$  as a subsequence, respectively, and  $cover_S(P) = cover_{Pos}(P) \cup cover_{Neg}(P)$ . A large number of total occurrences of  $P$  in sequences of  $S$ , either counted by DOOPS motif model (discriminative one occurrence per sequence) or DMOPS motif model (discriminative multiple occurrences per sequence, see [11]), usually indicates that the sequences in  $cover_S(P) \subseteq S$  are biologically related. A subsequence  $P$  is an  $\alpha$ -coverage for class  $Pos$  if  $\frac{|cover_{Pos}(P)|}{|Pos|} \geq \alpha$ , given parameter  $\alpha$  ( $0 < \alpha < 1$ ), and is a  $\beta$ -discriminant for class  $Pos$  if  $\frac{|cover_{Pos}(P)|}{|cover_S(P)|} \geq \beta$ , given parameter  $\beta$  ( $0 < \beta < 1$ ). Given  $\alpha$  and  $\beta$ , if  $P$  is both  $\alpha$ -coverage and  $\beta$ -discriminant for  $Pos$ , we say that  $P$  is  $\alpha, \beta$ -strong for  $Pos$ . Similar concepts can be defined for  $Neg$ .

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 those of  $P_2$ , respectively. Given an  $\alpha$ -coverage pattern  $P$ , the most informative pattern related to  $P$  in terms of coverage is the longest  $\alpha$ -coverage pattern containing  $P$ . Alternatively, given a  $\beta$ -discriminant pattern  $P$ , the most informative pattern related to  $P$  in terms of discrimination is the shortest  $\beta$ -discriminant pattern contained in  $P$ .

In the next section we present our investigation of NS5A motifs, in particular enumerative finding of DOOPS motifs in subsection 3.1 and inductive finding of DMOPS motifs in subsection 3.2.

### 3.1 Finding discriminative one occurrence per sequence (DOOPS)

Given two sets of positive sequences  $Pos$  and negative sequences  $Neg$ , and a set of unlabeled sequences  $UL$  (assuming that  $|S = Pos \cup Neg|$  is relative small and  $|UL| \gg |S|$ ). The problem is to find all potential  $\alpha, \beta$ -strong motifs for  $Pos$  given  $\alpha$  and  $\beta$ .

---

**Given:** Labeled sequences in  $Pos$  and  $Neg$  and a set  $UL$  of unlabeled sequences.

**Find:**  $\beta$ -DOOPS motifs for  $Pos$   $\alpha$  and  $\beta$  ( $0 < \alpha, \beta < 1$ ).

**Exhaustively Finding DOOPS Motif**( $Pos, Neg, UL, \alpha, \beta, \gamma$ )

1.  $k = 1, MotifSet = \phi$
  2.  $MotifSet_k = \phi$
  3.  $U \leftarrow$  set of all possible subsequences  $S(k)$  of length  $k$  from  $Pos$ .
  4. **for** each  $S(k) \in U$  **do**
  5.     **if**  $S(k)$  is  $\alpha, \beta$ -strong for  $Pos$  and  $|cover_{UL}(S(k))| > \gamma$  **do**
  6.          $MotifSet_k = MotifSet_k \cup S(k)$ .
  7. **if**  $MotifSet_k \neq \phi$  or  $\{S(k) \in U : S(k) \text{ is } \alpha\text{-coverage}\} \neq \phi$  **do**
  8.      $MotifSet = MotifSet \cup MotifSet_k$
  9.      $k = k + 1$
  10.    **goto** step 2.
  11. **else** return( $MotifSet$ )
- 

**Fig. 1.** Algorithm for exhaustively finding  $\beta$ -DOOPS motifs with unlabeled data.

The coverage for  $Pos$  of a pattern  $P$  measures its frequent occurrence in a region, or in other words, its conservation in evolution. As the labeled dataset for learning SVR and non-SVR to INF/RBV therapy are small, we can think of an exhaustive learning strategy to find all potential discriminative motifs.

Figure 1 describes the key idea of our proposed algorithm to exhaustively find DOOPS motifs from  $Pos$  while exploiting unlabeled sequences given  $\alpha$  and  $\beta$ . The exhaustive search is rational when  $Pos$  and  $Neg$  are of small sizes and short length. The algorithm starts with an empty set  $MotifSet$  and step by step adds to it the set of motifs which are  $\alpha, \beta$ -strong for  $Pos$  and of length  $k$ . These motifs need to occur in a large enough set of sequences in the unlabeled set, where  $k$  is increased from 1 by one unit in each loop in lines 2–10. The returned  $MotifSet$  is the union of all  $\alpha, \beta$ -strong motifs which are widespread in all available sequences.

### 3.2 Finding discriminative multiple occurrences per sequence (DMOPS)

Given two sets of positive sequences  $Pos$  and negative sequences  $Neg$ . Find a minimal set of DMOPS motifs satisfying two conditions: (1) *Complete*: Each sequence in  $Pos$  contains at least one found motif, (2) *Consistent*: Found motifs occur in sequences of  $Pos$  but do not occur in sequences of  $Neg$  (in fact, this condition is relaxed by allowing some error).

We introduce the algorithm SPLUPC to learn a “best” set of DMOPS motifs. The algorithm SLUPC is an extended version of LUPC that learns a set of descriptive rules for the two-class problem where the data objects are vectors [15]. Algorithm SLUPC learns discriminative motifs where the data objects are

---

**DMOPS Motif**( $Pos, Neg, UL, mina, minb, \gamma$ )    **Motif**( $Pos, Neg, UL, \alpha, \beta, \gamma$ )

|   |  |
|---|--|
| 1. $MotifSet = \phi$<br>2. $\alpha, \beta \leftarrow \mathbf{Initialize}(Pos, mina, minb)$<br>3. <b>while</b> ( $Pos \neq \phi \ \& \ (\alpha, \beta) \neq (mina, minb)$ )<br>4. $NewMotif \leftarrow \mathbf{Motif}(Pos, Neg, UL, \alpha, \beta, \gamma)$<br>5. <b>if</b> ( $NewMotif \neq \phi$ )<br>6. $Pos \leftarrow Pos \setminus Cover^+(NewMotif)$<br>7. $MotifSet \leftarrow MotifSet \cup NewMotif$<br>8. <b>else Reduce</b> ( $\alpha, \beta$ )<br>9. $MotifSet \leftarrow \mathbf{PostProcess}(MotifSet)$<br>10. <b>return</b> ( $MotifSet$ ) | 11. $CandMotifset = \phi$<br>12. <b>Adjacentaa</b> ( $Pos, Neg, \alpha, \beta$ )<br>13. <b>while StopCond</b> ( $Pos, Neg, UL, \alpha, \beta, \gamma$ )<br>14. <b>CandMotifs</b> ( $Pos, Neg, UL, \alpha, \beta, \gamma$ )<br>15. $Motif \leftarrow$<br>First $CandMotif$ in $CandMotifset$<br>16. <b>return</b> ( $Motif$ ) |
|---|--|

---

**Fig. 2.** SLUPC: Learning discriminant and minimal descriptors

sequences. For the simplicity, we work with a two-class problem where the two classes can be two genotypes or two groups of SVR and non-SVR sequences.

The procedure **DMOPS Motif** starts with an empty  $MotifSet$  (line 1) and two adaptive parameters  $\alpha$  and  $\beta$ , initialized by **Initialize** (lines 2). If the bias is on one accuracy, then  $\beta$  will be set to  $minb$ , and vice-versa. If the bias is on both accuracy and cover rate or there is no bias specified by the user, then both  $\alpha$  or  $\beta$  are set as the biggest value. Lines 3-8 describe a recursive procedure to learn one the best rule among  $\alpha\beta$ -strong rules, to add it to the  $MotifSet$ , to remove positive instances covered by this rule under some conditions, and to change adaptively thresholds  $\alpha$  and  $\beta$ .

If there are any instances remain in  $Pos$ , and  $\alpha$  and  $\beta$  are still equal or greater than  $mina$  and  $minb$  (line 3), **DMOPS Motif** calls the subroutine **Motif** to learn a new motif that is “best” with respect to the user-specified search bias (line 4). If **Motif** yields a motif (line 5), some positive instances covered by that motif will be removed from  $Pos$  (line 6) and the learned motif is added to the  $MotifSet$  (line 7). A sequence is removed from  $Pos$  when learning a new motif if it is covered by the new motif, and previously covered by  $\delta - 1$  rules in  $MotifSet$ . If **Motif** cannot find a motif,  $\alpha$  and/or  $\beta$  will be adaptively reduced by the subroutine **Reduce** (line 8). The loop between lines 4-8 is repeated until the stopping condition (line 3) holds. The obtained  $MotifSet$  can be optionally post-processed by **PostProcess**( $MotifSet$ ) (line 9) before the procedure returns the final  $MotifSet$  (line 10). The removal of only positive sequences covered by new motif (line 6) is an *one-sided selection*.

The procedure **Motif** searches for a  $\alpha\beta$ -strong motifs (motifs satisfy both  $\alpha$  and  $\beta$ ) in the remained  $Pos$  (sequences which are not removed), with the biggest occurrences in remained  $Pos$ . The key idea is to recursively expand a

subsequence  $S(i)$  of length  $i$  which is not satisfy both  $\alpha$  and  $\beta$  to  $S(i+1)$  with one position to the left or to the right, starting with  $i = 1$ .

The subroutine **Adjacentaa** searches for amino acids that can be added to  $S(i)$  if making  $S(i+1)$  satisfies  $\alpha$  and  $\beta$ . The subroutine **StopCond** checks if **Adjacentaa** successful. If ‘no’, it return an empty new motif ( $NewMotif = \phi$ , line 5) motif. If ‘yes’, the subroutine **CandMotifs** ranks  $S(i+1)$  by their number of occurrences in  $Pos$  if there are more than one amino acid that make  $S(i+1)$  satisfying both  $\alpha$  and  $\beta$ , and  $cover_{UL}(S(i+1)) > \gamma$ .

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 in Proposition 1, 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 motif candidate as the constraint holds or we completely go throughout  $Neg$  and find the motif has a satisfied accuracy. Two parameters  $\eta$  and  $\gamma$  can influence the findings of SLUPC.

Generally, the higher value of  $\eta$  and  $\gamma$  the higher chance to discover better motifs. When searching for  $\alpha\beta$ -strong motifs for  $Pos$ , a candidate motif can be eliminated without scanning throughout  $Neg$  if the following property holds during scanning.

**Proposition 1.** *Given a threshold  $\alpha$ , a pattern  $P$  is not  $\alpha\beta$ -strong for any arbitrary  $\beta$  if  $cover_{Neg}(P) \geq ((1 - \alpha)/\alpha) \times cover_{Pos}(P)$ .*

SLUPC distinguishes three alternatives that occur in practice and that lead to the three corresponding types of search heuristics:

1. *Bias on discrimination:* Sequentially find motifs whose cover ratio equal and greater than  $mina$  but discriminant power is as large as possible.
2. *Bias on cover ratio:* Sequentially find motifs whose discriminant power equal and greater than  $minb$  but the occurrences in  $Pos$  is as large as possible.
3. *Alternative bias on cover ratio and discrimination.* SLUPC starts with two highest estimated values of  $\alpha$  and  $\beta$ , and alternatively learns motifs with bias on either discrimination or cover ratio, then reduces one of the corresponding  $\alpha$  or  $\beta$  while keeping the other. The search is done until reaching the stopping condition.

## 4 Finding discriminative motifs characterizing IFN/RBV therapy effect.

### 4.1 The datasets

Two datasets are available for this study:

- *Labeled datasets:* This is NS5A sequences of non-response (134) and response patients (93) to IFN/RBV therapy from Los Amalos HCV database [16].

- *Unlabeled dataset* (in terms of two SVR and non-SVR groups): About 5000 NS5A sequences belonging to 6 genotypes of which the genotype 1 divided into subtypes 1a, 1b and 1c, and the genotype 2 divided into subtypes 2a and 2b. These data were taken from the database of Nagoya City University and GenBank.

In order to analyze and compare the labeled and unlabeled data as well as SVR and non-SVR to IFN/RBV therapy, according to the analysis purpose we extracted different datasets for the labeled and unlabeled data, and for several sub-genotypes, accordingly: (i) The ISDR region of 40 amino acids in length, (ii) The V3 region of 27 amino acids in length, (iii) The domains I, II and III have 213, 93 and 111 amino acids in length, and (iv) The full NS5A sequences. This work focuses on exhaustively finding descriptive motifs of SVR and non-SVR patients and their comparative study between two typical subgenotypes 1b and 1ac, as well as finding discriminative motifs of SVR and non-SVR patients of those subtypes.

#### 4.2 Finding NS5A DOOPS motifs characterizing SVR and non-SVR to IFN/RBV therapy.

The algorithm is applied to the V3 and ISDR regions, domains I, II, and III of NS5A of subtypes 1ac and 1b where the number of available sequences are

- 14 SVR, 73 non-SVR and 1559 unlabeled sequences of subtype 1ac,
- 30 SVR, 49 non-SVR and 1159 unlabeled sequences of subtype 1b.

Table 1 shows some typical  $\beta$ -DOOPS motifs in subtypes 1ac and 1b found from the V3 region (the left part) and ISDR region (the right part), respectively. The motifs selected in this table not only have high discrimination power but also distinguish well two the most popular sub-genotypes 1b and 1ac.

In Table 1, the first row shows results of V3 region in the left and ISDR region in the right. In the left part, the second row indicates two groups of columns standing for sub-genotypes 1ac or 1b, the third row shows groups of three columns corresponding to SVR, non-SVR and unlabeled sequences. The fourth row shows the number of training sequences in SVR, non-SVR classes and the number of unlabeled sequences, respectively, and for each of sub-genotypes 1ac and 1b. Each row from the fifth one presents a motif: the first column shows the motifs and the next six cells show the number of SVR, non-SVR and unlabeled sequences in 1ac and 1b data, respectively, containing the motif.

Consider some motifs, for example, the pattern “SGC” in Table 1 is a typical discriminative non-SVR motif of sub-genotype 1ac. This motif matches 66 out of 73 non-SVR sequences and only 2 SVR sequences of 1ac. It also occurs in 678 out of 1559 unlabeled sequences of 1ac. Moreover, this motif “SGC” only occurs in 1 unlabeled sequence and none of labeled sequences of the sub-genotype 1b.

The subsequences “SGC”, “GCP”, “PSG”, “APSG”, “NTTA” and “EPAS” are typical for non-SVR sequences in 1ac, while “PDC”, “DDT”, “GDD”, “GDD”,

**Table 1.** Typical  $\beta$ -DOOPS motifs with high value of  $\beta$  in the V3 and ISDR regions.

| V3 region |     |      |     |     |      |     | ISDR region |     |      |     |     |      |     |
|-----------|-----|------|-----|-----|------|-----|-------------|-----|------|-----|-----|------|-----|
| Type      | 1ac |      |     | 1b  |      |     | Type        | 1ac |      |     | 1b  |      |     |
| # Seq.    | SVR | nSVR | UL  | SVR | nSVR | UL  | # Seq.      | SVR | nSVR | UL  | SVR | nSVR | UL  |
| SGC       | 2   | 66   | 678 | 0   | 0    | 1   | ANH         | 1   | 61   | 649 | 0   | 0    | 3   |
| GCP       | 2   | 63   | 592 | 0   | 0    | 1   | CTA         | 1   | 61   | 673 | 4   | 3    | 47  |
| PSG       | 2   | 57   | 455 | 0   | 0    | 5   | TAN         | 1   | 61   | 653 | 0   | 0    | 3   |
| APSG      | 1   | 57   | 392 | 2   | 0    | 2   | AAN         | 0   | 25   | 87  | 0   | 0    | 0   |
| NTTA      | 0   | 29   | 69  | 0   | 0    | 0   | IAA         | 0   | 25   | 84  | 0   | 0    | 0   |
| EPAS      | 0   | 9    | 300 | 0   | 0    | 0   | LIA         | 0   | 25   | 85  | 0   | 0    | 0   |
|           |     |      |     |     |      |     | WNQ         | 0   | 25   | 0   | 0   | 0    | 0   |
| PDC       | 7   | 0    | 28  | 0   | 0    | 0   | ESES        | 0   | 24   | 6   | 0   | 0    | 0   |
| DDT       | 5   | 0    | 204 | 0   | 0    | 2   |             |     |      |     |     |      |     |
| GDD       | 5   | 0    | 249 | 0   | 0    | 7   | DAN         | 0   | 0    | 1   | 3   | 0    | 3   |
| PPDC      | 7   | 0    | 9   | 0   | 0    | 0   | LVD         | 0   | 0    | 0   | 3   | 0    | 1   |
| SEPT      | 4   | 0    | 67  | 0   | 0    | 0   | VDA         | 0   | 0    | 0   | 3   | 0    | 1   |
| EPTP      | 4   | 0    | 35  | 0   | 0    | 0   | CTTH        | 0   | 0    | 10  | 2   | 15   | 317 |
| DQA       | 0   | 0    | 0   | 2   | 12   | 264 |             |     |      |     |     |      |     |
| QSA       | 0   | 0    | 0   | 1   | 11   | 277 |             |     |      |     |     |      |     |

PPDC”, “SEPT” and “EPTP” are typical for SVR sequences in sub-genotype 1ac in the V3 region. Those are all 0.95-DOOPS motifs for SVR or non-SVR of sub-genotype 1ac. However, it is much more difficult to find such motifs for sub-genotype 1b. In fact, for this sub-genotype, two best motifs “DQA” and “QSA” found for non-SVR sequences are only 0.85-DOOPS motifs and no DOOPS motifs with threshold  $\beta > 0.6$  could be found in 1b. The last row in Table 1 (in the right part) shows the shortest motif containing subsequence “TTH” that can distinguish non-SVR from SVR sequences with accuracy of 0.88. Note that the ISDR regions of 1ac and 1b are different at only 4 positions, “TAN” in 1ac versus “TTH” in 1b and “E” versus “D” at the position of 17th. The subsequence “TTH” is a 0.98-discriminant pattern for 1ac and also with large coverage (0.71-coverage pattern for non-SVR 1ac sequences) while subsequence “TTH” is the core of a 0.88-discriminant and 0.21-coverage pattern for SVR 1b sequences). It can be observed from Table 1:

- There are more discriminant patterns with larger coverage for non-SVR sequences in 1ac than those for SVR sequences.
- Some discriminant patterns can be found for SVR sequences in the V3 region but with smaller coverage.
- Less DMOPS motifs were found in V3 and ISDR for SVR or non-SVR sequences in the subtype 1b, and they also appear less frequent.

When applying the method to datasets extracted from the domains I, II and III we could find a number of 0.95-DOOPS motifs for sub-genotype 1ac, and



**Table 2.** Typical  $\beta$ -DOOPS motifs with high  $\beta$  in the domain I and domains II-III.

| Domain I |     |      |      |     |      |      | Domain II, III |     |      |      |     |      |      |
|----------|-----|------|------|-----|------|------|----------------|-----|------|------|-----|------|------|
| Type     | 1ac |      |      | 1b  |      |      | Type           | 1ac |      |      | 1b  |      |      |
| # Seq.   | SVR | nSVR | UL   | SVR | nSVR | UL   | # Seq.         | SVR | nSVR | UL   | SVR | nSVR | UL   |
|          | 14  | 73   | 1559 | 30  | 49   | 1159 |                | 14  | 73   | 1559 | 30  | 49   | 1159 |
| EYP      | 1   | 58   | 846  | 0   | 0    | 1    | AAN            | 0   | 25   | 86   | 0   | 0    | 0    |
| LHE      | 1   | 58   | 848  | 0   | 0    | 39   | LWN            | 0   | 25   | 0    | 0   | 0    | 0    |
| RVD      | 1   | 18   | 20   | 0   | 0    | 3    | NQE            | 0   | 25   | 0    | 0   | 0    | 0    |
| CDF      | 0   | 20   | 4    | 0   | 0    | 0    | WNQ            | 0   | 25   | 0    | 0   | 0    | 0    |
| MWD      | 0   | 14   | 1    | 0   | 0    | 0    | SES            | 0   | 24   | 6    | 0   | 0    | 7    |
| CDF      | 0   | 20   | 4    | 0   | 0    | 0    | SKV            | 0   | 24   | 6    | 0   | 0    | 15   |
| APP      | 0   | 0    | 2    | 0   | 7    | 9    | KNP            | 0   | 0    | 0    | 8   | 0    | 31   |
| DYA      | 0   | 0    | 0    | 0   | 7    | 7    | NPD            | 0   | 0    | 0    | 8   | 0    | 40   |
| KNP      | 0   | 0    | 0    | 8   | 0    | 31   | WKN            | 0   | 0    | 0    | 8   | 0    | 31   |
| NPD      | 0   | 0    | 0    | 8   | 0    | 40   | GEI            | 0   | 0    | 1    | 7   | 0    | 11   |
| WKN      | 0   | 0    | 0    | 8   | 0    | 31   |                |     |      |      |     |      |      |
| SGT      | 11  | 59   | 1241 | 6   | 0    | 8    |                |     |      |      |     |      |      |
| WSG      | 11  | 59   | 1247 | 6   | 0    | 3    |                |     |      |      |     |      |      |
| DSH      | 0   | 0    | 42   | 6   | 0    | 71   |                |     |      |      |     |      |      |
| GDS      | 0   | 0    | 42   | 6   | 0    | 72   |                |     |      |      |     |      |      |
| SNM      | 0   | 0    | 2    | 0   | 7    | 213  |                |     |      |      |     |      |      |

more important a number of 0.95-DOOPS motifs for sub-genotype 1b as shown in Table 2.

#### 4.3 Finding NS5A DMOPS motifs characterizing SVR and non-SVR to IFN/RBV therapy.

As a preliminary study, algorithm SLUPC is applied to labeled NS5A sequences of sub-genotype 1b to learn DMOPS motifs. The experiments aim to evaluate the performance of discovered motifs in terms of discrimination.

The numbers of SVR, non-SVR and unlabeled sequences of subtype 1b used in the experiments are 30, 49 and 1159, respectively. A 3-fold cross validation on labeled data was done with the algorithms parameters as follows:  $mina = 0.05$ ,  $minb = 0.5$ ,  $\gamma = 0.05$ . In this experiment, SLUPC initialized  $\alpha$  and  $\beta$  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  $\alpha$  and  $\beta$  to find as strong as possible  $\alpha\beta$ -motifs that each training sequence contains at least one motifs found. The given  $\gamma$  ensures that each motif found is contained in at least  $0.05 \times |UL| = 58$  unlabeled sequences, i.e., it is expectedly a widespread pattern. Table 3 presents DMOPS motifs found in each run of the 3-fold cross validation where each two

**Table 3.** NS5A MDOPS motifs characterizing SVR and non-SVR to IFN/Ribavirin therapy.

| Fold 1 |      | fold 2 |       | fold 3 |      |
|--------|------|--------|-------|--------|------|
| SVR    | nSVR | SVR    | nSVR  | SVR    | nSVR |
| DAE    | QA   | AI     | NM    | AI     | ND   |
| AI     | AEA  | DI     | AKA   | DAN    | GR   |
| AAG    | NM   | AAC    | RRRLA | TAA    | NM   |
| TRAL   | GR   | TRA    | DT    | HA     | MA   |
| AF     | DK   | RAC    | IKA   | FR     | RS   |
| CR     | RF   |        |       |        | DK   |
|        | DI   |        |       |        | PNA  |
|        |      |        |       |        | EAT  |
|        |      |        |       |        | LGA  |
|        |      |        |       |        | AEA  |

columns stand for the run when each fold was taken as testing data and union of the other two as training data. Two columns under each fold show DMOPS motifs found in SVR and non-SVR sequences, respectively.

The accuracy of assessment on testing in each fold is 54.8%, 64.4%, and 85%, and thus the average is 68.8%. Though this accuracy is very encouraging, note that the prediction accuracies in three run are largely different caused by the small number of labeled sequences. One the other hand, the medical researchers are being evaluated this computational results.

## 5 Conclusion

We have presented two algorithms for discovering discriminative motifs which can function when labeled data is small but a large set of unlabeled sequences is available. The algorithms were applied to detect the relationship between HCV NS5A protein and the interferon/ribavirin therapy effect. These results are promising as they present many patterns that were not known previously.

This research can be improved in several directions. One is to improve the algorithms and their implementation, to link the discovered motifs to IFN pathways and mutations to interpret those NS5A mechanisms of IFN/ribavirin resistance. The other is to develop semi-supervised methods to find discriminative motifs in imbalanced data.

## Acknowledgments

This work is partially supported by MEXT project “Advanced computation methods for analyzing scientific data”, JSPS bilateral Japan-Vietnam project on “Computational methods in Biomedicine”, JSPS’s kakenhi project No. 19300045, and NAFOSTED (National Foundation for Science and Technology Development).

## References

1. Gao M., Nettles R.E., and et al. Chemical genetics strategy identifies an hcv ns5a inhibitor with a potent clinical effect. *Nature*, 465:953–960, 2010.
2. Jardim A.C.G., Yamasaki L.H.T., Queiro A.T.L., Bittar C., Rahal R.P.R., and de Carvalho Mello I.M. V. Quasispecies of hepatitis c virus genotype 1 and treatment outcome with peginterferon and ribavirin. *Infection, Genetics and Evolution*, 9:689–698, 2009.
3. Guillou-Guillemette H.L., Vallet S., Gaudy-Graffin C., Payan C., Pivert A., Goudeau A., and Lunel-Fabiani F. Genetic diversity of the hepatitis c virus: Impact and issues in the antiviral therapy. *World Journal of Gastroenterology*, 13(17):2416–2426, 2007.
4. T.H. Lin, R.F. Murphy, and Z. Bar-Joseph. Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 8(2):441–451, 2011.
5. Feld J.J. and Hoofnagle J.H. Mechanism of action of interferon and ribavirin in treatment of hepatitis c. *Nature*, 436:967–972, 2005.
6. Aurora R., Donlin M.J., and Cannon N.A. Genome-wide hepatitis c virus amino acid covariance networks can predict response to antiviral therapy in humans. *The Journal of Clinical Investigation*, 119:225–236, 2009.
7. AlHefnawi M.M., Zada S., and El-Azab I.A. Prediction of prognostic biomarker for interferon-based therapy to hepatitis c virus patients: a metaanalysis of the ns5a protein in subtypes 1a, 1b, and 3a. *Virology Journal*, 7:1–9, 2010.
8. Macdonaldt A. and Harris M. Hepatitis c virus ns5a: Tales of a promiscuous protein. *Journal of General Virology*, 85:2485–2502, 2004.
9. Lemon S.M., Keating J.A., Pietschmann T., Frickd D.N., Glenne J.S., Tellinghuisen T.L., Symonsg J., and Furman P.A. Development of novel therapies for hepatitis c. *Antiviral Research*, 86:79–92, 2010.
10. F. Fauteux, M. Blanchette, and M. Stromvik. Seeder: Discriminative seeding dna motif discovery. *Bioinformatics*, 24(20):2303–2307, 2008.
11. J.K. Kim and S. Choi. Probabilistic models for semi-supervised discriminative motif discovery in dna sequences. *IEEE/ACM Trans. Computational Biology and Bioinformatics*, (DOI: 10.1109/TCBB.2010.84), 2011.
12. C. Vens, M.N. Rosso, and E.G.J. Danchin. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, 27(9):1–12, 2011.
13. Sharma S.D. Hepatitis c virus: Molecular biology & current therapeutic options. *Indian Journal of Medical Research*, 131:17–34, 2010.
14. Wohnsland A., Hofmann W.P., and Sarrazin C. Viral determinants of resistance to treatment in patients with hepatitis c. *Clinical Microbiology Reviews*, 20, 2007.
15. Ho T.B. and Nguyen D.D. Chance discovery and learning minority classes. *Journal of New Generation Computing*, 21(2):1477–160, 2002.
16. Kuiken C., Yusim K., Boykin L., and Richardson R. The los alamos hepatitis c sequence database. *Bioinformatics*, 21(3):379–384, 2005.