Journal of Universal Computer Science, vol. 19, no. 4 (2013), 563-580
submitted: 3/8/12, accepted: 20/2/13, appeared: 28/2/13 © J.UCS

A Semi-Supervised Ensemble Learning Method for Finding Discriminative Motifs and its Application

Thi Nhan Le[1,2], Tu Bao Ho[1,2], Saori Kawasaki[1]
([1]Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
[2]Vietnam National University, Ho Chi Minh, Vietnam
{lethinhan, bao, skawasa}@jaist.ac.jp)

Tatsuo Kanda, Katsuhiko Takabayashi, Shuang Wu, Osamu Yokosuka
(Graduate School of Medicine, Chiba University, Chiba, Japan
kandat-cib@umin.ac.jp, takaba@ho.chiba-u.ac.jp
wushuang@graduate.chiba-u.jp)

Abstract: Finding discriminative motifs has recently received much attention in biomedicine as such motifs allow us to characterize in distinguishing two different classes of sequences. It is common in biomedical applications that the quantity of labeled sequences is very limited while a large number of unlabeled sequences is usually available. The current methods of discriminative motif finding are powerful and effective with large labeled datasets, but they do not function well on small labeled datasets. In this paper, we present a semi-supervised ensemble method for finding discriminative motifs which is based on the SLUPC algorithm, a separate-and-conquer searching method to discover motifs of type 'discriminative one occurrence per sequence'. The proposed method, named E-SLUPC (Ensemble SLUPC), uses SLUPC to search discriminative motifs from an extended labeled dataset that contains labeled data and unlabeled data with predicted labels. Strong discriminative and frequent motifs characterizing two outcome classes of hepatitis C virus treatment (sustained viral response and non-sustained viral response) were detected and analyzed. Furthermore, the experimental evaluation shows that our method can function considerably well in the common context of medical research when the labeled data is usually difficult to obtain.

Key Words: discriminative motif, separate-and-conquer search, self-training technique, ensemble learning, hepatitis C virus, NS5A region.
Category: I.5

1 Introduction

One of the key interests of biologists is to detect short and highly conserved motifs in a collection of DNA or protein sequences. 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. For example, in the study of hepatitis pathogenesis and therapy by using non structure 5A (NS5A) protein, where we are interested in discriminating two classes of sustained viral response (SVR) sequences and non sustained viral response (non-SVR) sequences, from the biggest resource of LANL[1] database, we can only get 134 NS5A non-SVR sequences and 93 SVR sequences to IFN/RBV therapy, and 13 non-SVR sequences and 12 SVR sequences from Chiba Medical University, but from Genbank[2] and HVDB[3] databases, we obtain about 5000 NS5A unlabeled sequences. In the case of our study on hepatitis C, the combination of interferon and ribavirin (IFN/RBV) is currently the standard therapy for Hepatitis C virus (HCV). However, only fewer than half of the HCV infected individuals achieve sustained viral response by this therapy, and the genetic basis of resistance to antiviral therapy remains unknown, see for example [Gao and Nettles, 2010]. Many studies report that the NS5A in the 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 in NS5A [Enomoto et al., 1996, Sarrazin et al., 1999, Witherell and Beineke, 2001, Pascu et al., 2004, ElHefnawi et al., 2010, El-Shamy et al., 2011].

The research on discriminative motif finding has several newly developed methods using the hidden Markov model (HMM) [Lin et al., 2011], position weight matrix (PWM) [Redhead and Bailey, 2007, Kim and Choi, 2011, Bailey et al., 2010], association mining with domain knowledge [Vens et al., 2011]. However, due to their general purposes, these methods have shown to be ineffective for the situation when only small labeled datasets are available.

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

[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

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.

2 Related work

2.1 Discriminative motif learning

A *sequence motif* is generally understood as a pattern in nucleic or amino acid sequences that is widespread and biologically significant [Sami and Nagatomi, 2008]. For example, in DNA sequences, motifs can be transcription factor binding sites (TFBSs) in the promoter regions; in protein sequences, motifs can be regions corresponding to a specific function/structure or they can be signals playing an important role in controlling the cellular localization [Vens et al., 2011].

A 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 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].

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. 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 Expectation Maximization (EM) 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 maximum mutual information estimate (MMIE) 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.

2.2 Semi-supervised ensemble learning

The combination between semi-supervised learning (SSL) and ensemble learning (EL) 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.

3 The 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 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.

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.

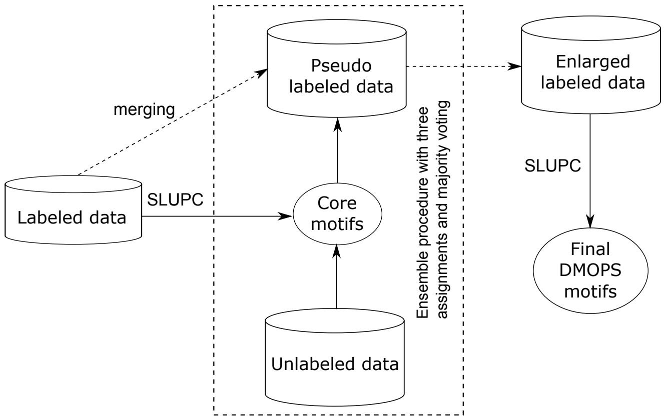

Figure 1: The framework of E-SLUPC.

3.1 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 $\Sigma = \{amino \quad acid\}$ 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.

Algorithm 1 SLUPC algorithm

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

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

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

$MotifSet = \phi$
$\alpha, \beta \leftarrow$ **Initialize**$(Pos, minalpha, minbeta)$
while $Pos \neq \phi$ & $(\alpha, \beta) \neq (minalpha, minbeta)$ **do**
  $NewMotif \leftarrow$ **Motif**$(Pos, Neg, \alpha, \beta)$
  if $NewMotif \neq \phi$ **then**
    $Pos \leftarrow Pos \setminus Cover^+(NewMotif)$
    $MotifSet \leftarrow MotifSet \cup NewMotif$
  else
    Reduce(α, β)
  end if
  $MotifSet \leftarrow$ **PostProcess**$(MotifSet)$
end while
return$(MotifSet)$

The DMOPS motif finding of SLUPC algorithm is described in Algorithm 1. Given two sets of positive sequences Pos and negative sequences Neg, Algorithm

1 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 length's subsequence is 1.

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: $return(Motif)$

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

3.2 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].

3.3 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[4], 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.

4 Application to study HCV

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.

4.1 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.

4.2 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

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

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 2. Though the average accuracy on testing data is low (about 66%), it is very encouraging in the biomedical field.

Table 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 DM-OPS 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.

Table 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

It can be observed from Table 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 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.

4.3 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 2 shows the experiment results of comparing the accuracy of SLUPC and E-SLUPC (about 8% increase in accuracy). Accuracies in Table 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 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.

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

4.4.1 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

Table 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

finding. Using the web version of the MEME[5], 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 shows the top 12 motifs found by MEME which have the highest frequency in a total of 15 times of MEME running.

Table 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

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

Observing Table 3, we see that although MEME allows us to search discriminative motifs with two sets of positive and negative sequences, the discriminative ability 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.

4.4.2 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[6], 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 4 shows the top 12 motifs found by DEME which have the highest frequency in a total of 15 times of DEME running.

Table 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

In Table 4, the SVR motifs "LAIKT", "LAIK", "LVGLNW", "LSALSL"

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

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”, 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.

5 Conclusions

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.

Acknowledgments

This work is supported by JSPS’s project “Computational Methods for Discovering Molecular Mechanisms of Hepatitis Pathology and Therapy” and Vietnam National Foundation for Science and Technology Development (NAFOSTED). We would like to thank Ngoc Tu Le for providing the source code of DMOPS motif finding algorithm. Thi Nhan Le would like to express her sincere thanks to the 322 project of the Ministry of Education and Training (MOET) of Vietnam for their scholarship support.

References

[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.

[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.

[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).

[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.

[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. M., Zada, S., and El-Azab, I. E. (2010). Prediction of prognostic biomarkers for interferon-based therapy to hepatitis C virus patients: a metaanalysis of the ns5a protein in subtypes 1a, 1b, and 3a. *Virology Journal*, 7.

[Enomoto et al., 1996] Enomoto, N., Sakuma, I., Asahina, Y., Kurosaki, M., Murakami, T., Yamamoto, C., 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(2).

[Gao and Nettles, 2010] Gao, M. and Nettles, R. (2010). Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. *Nature*, 465:953–960.

[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 and Nguyen, 2002] Ho, T. B. and Nguyen, D. D. (2002). Chance discovery and learning minority classes. *New Generation Computing*, 21(2).

[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.

[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).

[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).

[Pascu et al., 2004] Pascu, M., Martus, P., Hohne, M., Wiednmann, B., Hopf, U., Schreir, E., and Berg, T. (2004). Sustained virological response in hepaptitis 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.

[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.

[Sami and Nagatomi, 2008] Sami, A. and Nagatomi, R. (2008). *Data mining in medical and biological research.* 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. *Proceeding of the Sixth Annual International Conference on Computational Biology, ACM New York*, pages 263–272.

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

[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 classifcation-based motifs in biological sequences. *Bioinformatics*, 27(9):1231–1238.

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

[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.

[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.