Novel metrics for feature extraction stability in protein sequence classication

Rabie Saidi, Aridhi Saber, Maddouri Mondher, Mephu Nguifo Engelbert

Novel metrics for feature extraction stability in protein sequence classication

Rabie Saidi *†‡ Sabeur Aridhi *†‡ Mondher Maddouri ‡ Engelbert Mephu Nguifo *†

Abstract

Feature extraction is an unavoidable task, especially in the critical step of preprocessing biological sequences. This step consists for example in transforming the biological sequences into vectors of motifs where each motif is a subsequence that can be seen as a property (or attribute) characterizing the sequence. Hence, we obtain an object-property table where objects are sequences and properties are motif extracted from sequences. This output can be used to apply standard machine learning tools to perform data mining tasks such as classification. Several previous works have described feature extraction methods for bio-sequence classification, but none of them discussed the robustness of these methods when perturbing the input data. In this work, we introduce the notion of stability of the generated motifs in order to study the robustness of motif extraction methods. We express this robustness in terms of the ability of the method to reveal any change occurring in the input data and also its ability to target the interesting motifs. We use these criteria to evaluate and experimentally compare four existing extraction methods for biological sequences.
keywords : Motif / feature extraction, classification of protein sequences, motif stability, sensibility, stable motif interest

1 Introduction.

Classification of biological sequences is a fundamental task in bioinformatics [7]. In fact, biologists are continuously interested in identifying the family to which a novel sequenced protein belongs [6]. This makes it possible to study the evolution of this protein and to discover its biological functions. Generally, biologists use alignment to classify new biological sequences into already known families/classes by searching similarity and homology among sequences. However, this approach is often inefficient. For instance, in metagenomics, one of the major problems encountered during the application of such approach is that between 25% and 65% of the sequences have no homologous (orphan sequences) in the databases, making these sequences unusable [9].

Machine learning techniques are one way to deal with such a problem. But, with the format of biological sequences where characteristics are encoded into the sequence itself, it is not possible to use the well-known classification algorithms which have proved to be very efficient in real data mining tasks [19] with data described in a relational data format. Consequently, a preprocessing step is necessary in order to parse the biological data into a new format suitable for different data mining tools [1].

For protein sequences, motif extraction represents a widely used solution to perform this preprocessing phase. The protein sequences are chains of amino acids residues where each residue is represented by a character within an alphabet of size 20. Discovering motifs from these sequences is a delicate task aiming to find substrings or words that can serve as descriptors. These descriptors form the feature space that allows transforming the biological sequences into vectors of values, thus facilitating the processing of such data by machine learning and data mining tools. This preprocessing phase is the key step towards a reliable process of knowledge discovery because it directly affects the quality of obtained results [23]. Each protein sequence is described by a set of motifs to achieve a vector of values where the values depend on the nature of the description function that binds the sequence and the motif such as:

- Presence / absence of the motif in the sequence,
- Number of occurrences of the motif in the sequence,
- First position of the motif in the sequence...

These motifs, which will serve as descriptors, are extracted from biological sequences according to predefined parameters such as frequency, length, composition. However, designing and developing a suitable method for finding motifs that may reduce any loss of information due to the format change and help solve problems of data mining, remains a nontrivial task [33] [36].

Several motif extraction methods have been proposed [22]. Meanwhile, various studies have made assessments and comparisons between these methods and have tried to study the impact of one method or another

*LIMOS - Blaise Pascal University - Clermont University, BP 10448, Clermont-Ferrand 63000, France. name.lastname@isima.fr
†LIMOS - CNRS UMR 6158, Aubire 63173, France.
‡URPAH - FST - University of Tunis El Manar, Academic Campus, Tunis 2092, Tunisia. name.lastname@fst.rnu.tn

on the quality of the learning task to be performed (classification, prediction, shape recognition ...). So, the best methods are those that allow having the best values of quality metrics such as accuracy rate in the case of supervised classification.

In this paper we introduce the concept of stability to compare motif extraction methods. We call stability of a motif extraction method from a dataset, the non-variability in its set of motifs, when applying a technique of variation on the input data. The robustness of a method is the coupling of the non-stability and the ability to retain or improve the quality of the associated data mining task. In our case we will use the supervised classification accuracy as quality measure.

Concrete motivations behind the above-mentioned stability can be found within distributed systems and grid computing environments that are mining huge amount of data. In such environments, the variation of data is a common fact. This variation can be due to various events such as failed transfer of data portions, loss of communication between nodes of the distributed system, data updating... Another motivating application is information retrieval from biological databases (such as GenBank[5], EMBL[29], UniProt[4]). The problem here lies in the fact that every database has its own terminology and procedures which sometimes yield related but not identical data [38]. Since the retrieved data are the main materials of several delicate processes in both industry and research like disease management and drug development, it is crucial for database researcher, bioscience user and bioinformatics practitioner to be aware of any change in the preparation data samples [30]. In addition, with the exploding amounts of data submitted to the biological databases, there is an increasing possibility of finding erroneous data. In such conditions, it is important to make sure that the motif extraction methods, which are the start point of any mining process, are robust enough to detect even slight variations in input data like does any good sensor when describing its context environment.

The goal of this study is to propose new metrics to measure the robustness of motif extraction methods in terms of their ability to reveal any change occurring in the input data and also its ability to target the interesting motifs. We experiment these measures on protein data.

This paper is organized as follows. In the next section, we present an overview on some works that dealt with the concept of stability and present the motivation behind our work. In section 3 we define the various terms used in this paper. Section 4 describes our experimental protocol then we discuss the obtained results in section 4. A conclusion and some prospects are presented in section 5.

2 Background.

The topic of stability with respect to motif extraction methods has not been studied in the literature. However, this aspect was slightly studied in a very close field to motif extraction which is the feature selection [20] [35] [10] [13] [28] [34].

In [20], authors propose a measure which assesses the stability of feature selection algorithms with respect to random perturbation in data. In this work, the stability of feature selection algorithms can be assessed through the properties of the generated probability distributions of the selected feature subsets. The interest is, of course, in feature selection algorithms that produce probability distributions far from the uniform and close to the peak one. Given a set of features, all possible feature combinations of size k are considered achieving n feature subsets. The frequencies F of selected feature subsets are recorded during data perturbation in a histogram. For a size k, the stability S_k is measured based on the Shannon entropy:

$$S_k = -\sum_{i=1}^{n} F_i \log F_i \ . \tag{2.1}$$

In [35], authors perform an instance sub-sampling to simulate data perturbation. Feature selection is performed on each of the n sub-samples, and a measure of stability is calculated. The output f of the feature selection applied on each sub-sample is compared to the outputs of the other sub-samples using Pearson correlation coefficient, the Spearman rank correlation coefficient and the Jaccard index as similarity measure noted S. The more similar all outputs are, the higher the stability measure will be. The overall stability can then be defined as the average over all pairwise similarity comparisons:

$$S_{total} = \frac{2 \times \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S(f_i, f_j)}{n(n-1)} \ . \tag{2.2}$$

In [10], different sub-samples (or training sets) are created using the same generating distribution. Stability quantifies how different training sets affect the feature selection output. Authors take into account three types of representations for feature subsets. In the first type a weight or score is assigned to each feature indicating its importance. The second type of representation, ranks are assigned to features. The third type consists only of sets of features in which no weighting or ranking is considered. Measuring stability requires a similarity measure for feature representations. This obviously depends on the representation used by a

given feature selection algorithm to describe its feature subset. The authors used three similarity measures: the Pearson's correlation coefficient, the Spearman rank correlation coefficient and the Tanimoto distance.

3 Robustness of motif extraction methods.

3.1 Motivations. The application of the above-presented measures of stability is not convenient in our case (motif extraction). This comes from the nature of input data used by feature selection methods [18] [21] [16] [25].

In fact, these methods use an original set of features (motifs) as input and try to merely select a subset of relevant features. The perturbation of data is applied to the original set of features. In the case of motif extraction, the input data are still in rough state and the data perturbation step is applied directly to row data (before motif extraction step)(figure 1).

There are different kinds of data perturbation (introducing noise in data) [27] [26], especially in the context of biological data where it is possible to consider also deletion, mutation and insertion of nucleotides or amino acids into the sequences. In what follows, we will consider only one kind of noise consisting of removing a subset of sequences from the initial set. This case is the one that often happens in the context of distributed environment as described in section 1.

In our work, the motivation behind exploring the motif extraction method stability is to provide evidence that even slight changes in the data must also be followed by changes in the output results (extracted motifs). These changes must concern the motifs that are no longer significant for the perturbed input data; which means that the motifs that have been conserved must prove to be interesting i.e. help with better classification. Since the set of features are not known a priori we can not apply the measures quoted in the feature selection related works. For our purposes let be the two following assumptions:

Assumption 1. We suppose that a motif extraction method allows a reliable description of input data if any variation within these data affects the set of the generated motifs. That is to say that it reveals any change occurring in the input data.

Assumption 2. After changes in the set of generated motifs, the motifs that are conserved, are interesting i.e., help to better classify unknown sequences.

In the next subsection we define and describe the terms we use to formally express our assumptions and to evaluate the robustness of motif extraction methods.

3.2 Terminology. Based on assumption 1, we introduce the concept of sensibility . This concept reflects the ability to produce a different set of motifs i.e., a different data description, whenever we make a variation within the input data. The sensitivity criterion can be studied by means of the conserved motifs called stable motifs. It is also interesting to test the assumption 2, that is to say the quality of the stable motifs, by assessing their benefits in an artificial learning task.

Below we formally define the terms used in this paper. Consider the following:

- A dataset D, divided into n subsets D_1, D_2,.., D_n.
- A motif extraction method M applied to D on one side and to D_1, D_2,.., D_n on the other side and respectively generating the sets of motifs SM from D and SM_1, SM_2, .., SM_n from D_1, D_2,.., D_n.
- An artificial learning task T and a quality metric Mtr of T. Let $Mtr^T(SM)$ denote the value of the metric obtained if T is performed using the set SM as a feature space.

We define the following:

Definition 3.1. (Motif Stability) *A motif x is said to be stable if and only if its occurrence rate in all SM_i, $i = 1..n$, exceeds a threshold . The occurrence rate is simply the ratio of the number of SM_i, $i = 1..n$, where x appears to n. Formally:*

$$\frac{Number\ of\ SM_i / x \in SM_i}{n} \geq \tau, with\ i = 1..n\ . \tag{3.3}$$

Definition 3.2. (Rate of stable motifs) *The rate of stable motifs (RSM) of a method M is the ratio of the number of stable motifs to the number of distinct motifs of all SM_i, $i = 1..n$. Formally:*

$$RSM = \frac{Number\ of\ stable\ motifs}{|\bigcup_{i=1}^{n} SM_i|}\ . \tag{3.4}$$

Definition 3.3. (Method sensibility) *A method M_1 is more sensible than another method M_2 if and only if for the same changes within the same dataset, the rate of stable motifs of M_1 is lower than that of M_2. Thus, the sensibility S of a method is complementary to its rate of stable motifs. It may be noted:*

$$S = 1 - RSM\ . \tag{3.5}$$

Definition 3.4. (Conservation) *A motif extraction method M conserves the quality metric value of a datamining task T if the use of the set of stable motifs SSM in T preserves the quality metric values for this task as when we use the set of motifs SM generated from*

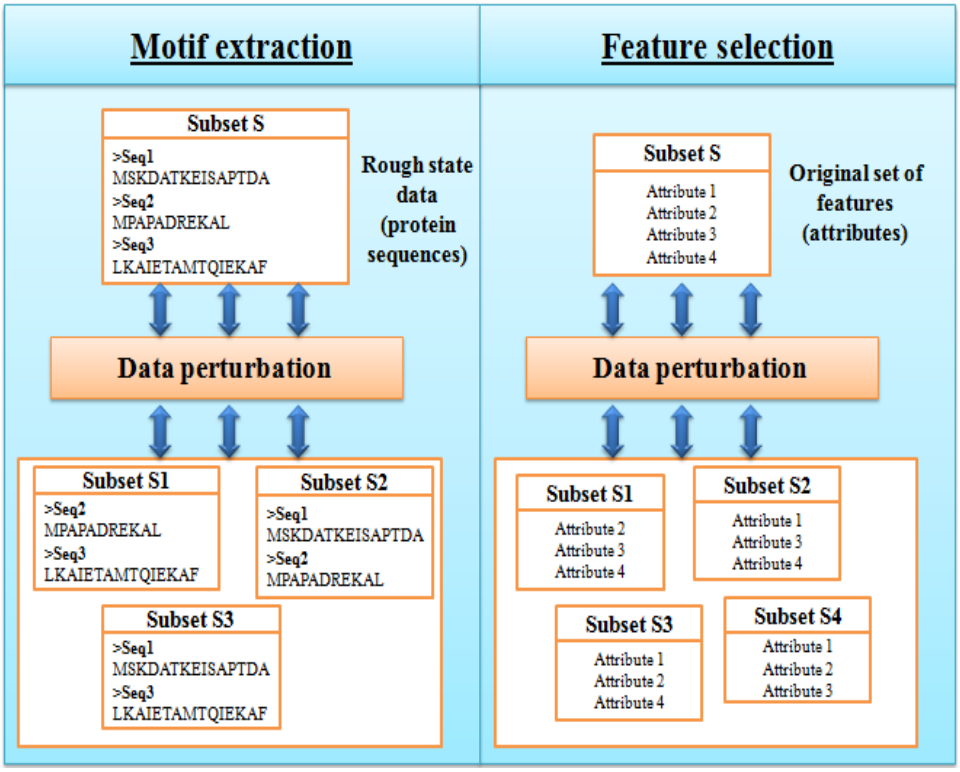

Figure 1: Data perturbation in feature selection and motif extraction

the original dataset D. However, it is noteworthy that we can not judge that conservation unless the method is already sensible. Indeed, an insensible method tends to generate the same motifs even after perturbations in the input data indicating that its extraction approach is rigid and does not adopt a concept of "choice". This conservation C can be measured by:

$$C = 1 - |Mtr^T(SM) - Mtr^T(SSM)| \,. \tag{3.6}$$

Definition 3.5. (Interest) *A set of stable motifs SSM is considered to be interesting if it allows interesting values of conservation and sensibility. Formally, we can measure this interest I by:*

$$I = 2 \times \frac{S \times C}{S + C} \,. \tag{3.7}$$

This measure is inspired from the F1-Score which is a statistical measure of a test's accuracy that combines Precision and Recall. The F1-score can be interpreted as a weighted average of the precision and recall, where an F1-score reaches its best value at 1 and worst score at 0. In our case, we combine conservation and sensibility to quantify the interest of stable motifs.

3.3 Illustrative example. Considering a dataset D in a supervised classification task T. The data perturbation of D generates three subsets D_1, D_2 and D_3. The application of a motif extraction method M to D on one side and to D_1, D_2 and D_3 on another side generates the sets of motifs SM from D and SM_1, SM_2 and SM_3 from D_1, D_2 and D_3 respectively :
SM ={$m1$, $m2$, $m3$, $m4$, $m5$, $m6$, $m7$, $m8$, $m9$, $m10$}
SM_1={$m1$, $m4$, $m5$, $m6$}
SM_2={$m1$, $m2$, $m3$}
SM_3={$m1$, $m6$, $m7$, $m8$}

Using τ such that $\tau = 65\%$, the motifs $m1$ and $m6$ are considered stable since they appear in more than 65% of the motifs subsets.
We can easily calculate the rate of stable motifs RSM = 0.25, which is two over the set of eight motifs. We consider that $m9$ and $m10$ are noise, and thus are not relevant for the classification task.

The sensibility S of M is equal to 0.75. In this case, the set of stable motifs SSM_1 is equal to {$m1$, $m6$}.

Using $\tau > 66\%$, only the motif $m1$ are considered stable since it appears in more than τ% of the motifs subsets. We can easily calculate the rate of stable motifs RSM as equal to 0.125. The sensibility S of M is equal to 0.875, and SSM_2 = {$m1$}.

Suppose we use sets of motifs SM and SSM_1 as variables space to measure the accuracy rate (Mtr^T)

of the supervised classification task T. Let consider the following obtained values with Mtr^T : $Mtr(SM) = 0.85$ and $Mtr(SSM) = 0.80$

The set of stable motifs SSM_1 enables a conservation of the quality metric value $C = 0.95$ Finally, we can measure the interest of stable motifs by $I = 0.83$.

4 Experiments.

In this section, we describe an experimental study conducted on four motif selection methods quoted in [22]. Calculations were run on a duo CPU 1.46GHz PC with 2GB memory, operating on Linux. The following is a presentation of the input datasets and the used tools.

4.1 Experimental data. We used four datasets containing 1327 protein sequences extracted from Swiss-Prot [3] and described in table 1. These datasets differ from one another in terms of size, number of class, class distribution, complexity and sequence identity percentage (IdP). The change in the nature of the datasets allows us to avoid specific outcomes to data and to have better interpretations.

Table 1: Experimental data

Dataset	IdP	Family / class	#sequences
DS1	48%	High-potential Iron-Sulfur Protein	19
		Hydrogenase Nickel Incorporation Protein HypA	20
		Glycine Dehydrogenase	21
DS2	28%	Human TLR	14
		Non-human TLR	26
DS3	48%	Chemokine	255
		Melanocortin	255
DS4	25%	Monomer	208
		Homodimer	335
		Homotrimer	40
		Homotetramer	95
		Homopentamer	11
		Homohexamer	23
		Homooctamer	5

4.2 Motif extraction methods. We compare the motif extraction methods quoted in [22], i.e., n-grams NG [15], active motifs AM [31], discriminative descriptors DD [19] and discriminative descriptors with substitution matrix DDSM [22]. In our experiments we use the same default settings as in [22].

NG is a simple method that generate the distinct words of fixed length by sliding a window of N characters on the whole sequences i.e., at every character i the word $[i, i + N]$ is extracted.

The AM method allows generating the commonly occurring motifs whose lengths are longer than a specified value in a set of biological sequences. The activity of a motif is the number of matching sequences given an allowed number of mutations [31]. The motif generation is based on the construction of a generalized suffix tree (GST) which is an extension of the suffix tree [2] and is dedicated to represent a set of n sequences indexed each one by $i = 1..n$.

The DD method allows building sub-strings that can discriminate a family of proteins from other ones [19]. This method is based on an adaptation of the Karp, Miller and Rosenberg (KMR) algorithm [14]. This algorithm identifies the repeats in character strings. The extracted repeats are then filtered in order to keep only the discriminative and minimal ones. A substring is considered to be discriminative between the family F and the other families if it appears in F significantly more than in the other families.

The DDSM method is an extension of the DD method that takes into account the phenomenon of mutation within proteins. It uses the substitution matrices [8] [12] to achieve a sort of clustering of motifs and keep, for each cluster, the most mutable motif. It is therefore a more concise representation than DD since it reduces significantly the number of generated motifs [22].

Comparisons made in [22] between these methods revealed that DDSM performs the best to help in problems of protein sequences classification even in difficult cases where other methods fail to produce reliable descriptors for an accurate classification. In this work, we try to find a relationship between this performance and the concepts introduced in section 3.

4.3 Experimental process. In our experiments, we perturb each input dataset in a systematic way and we observe the impact of this perturbation on the set of generated motifs. To do this, we use the 10-cross-validation (10CV) and leave-one-out techniques (LOO) [11]. Therefore, the variation of a dataset containing n sequences consists in removing a partition (one tenth with 10-cross-validation and a single sequence with leave-one-out) from the dataset and the rest is used to generate a set of motifs. This is done several times (ten times with 10-cross-validation and n times with leave-one-out). At each iteration, the number of occurrences of generated motifs is updated. Otherwise, a sequences-

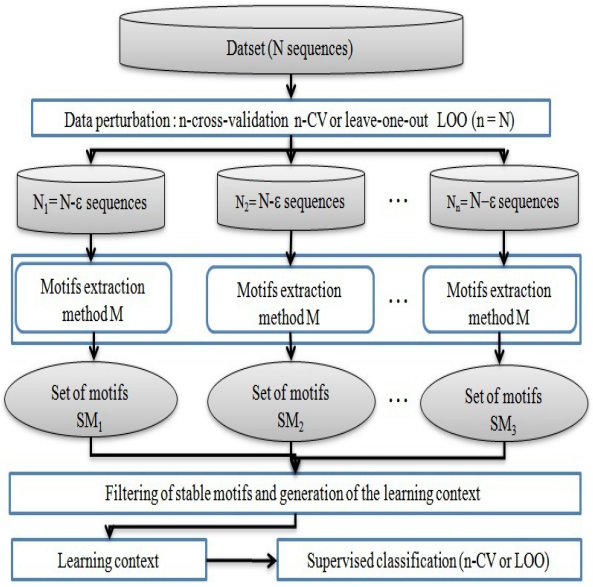

Figure 2: Experimental process

based perturbation can be applied here,

As already defined in Section 2, the technique we adopt to measure the sensibility of motif extraction methods from protein sequences is based on the rate of stable motifs. Whereas the sensibility is related to the amount of stable motifs, the interest of stable motifs is related to their quality. In other words, if these motifs are generated by the extraction method to appear often enough then they should be "interesting". We measure the interest in our experiments by their usefulness in a supervised classification task. Once the stable motifs are generated, they are used to convert protein sequences into binary vectors where the value '1' denotes the presence of motif in the sequence and '0' its absence, all these binary vectors compose what is called a learning context.

Thus the classification of proteins in this new format is now possible with different classifiers. To do this, we use the support vector machine classifier SVM of WEKA workbench [32]. The effect of this choice should be insignificant in the process of measuring the robustness of motif extraction methods. However we put our focus on a well-known and efficient classifier for our experiments. In addition SVM classifier has shown its efficiency in a previous study [22].

The classification is performed based on 10-cross-validation (10CV) and leave-one-out (LOO) techniques Hence, our experiments are conducted using the following four combinations for data variation and classification: (LOO; LOO), (LOO; 10CV), (10CV; LOO) and (10CV; 10CV).

5 Results and discussion.

We show in table 2 the classification results of our datasets using the motif extraction methods quoted in section 4.2.

The classification is performed without making any perturbation on our datasets using the SVM classifier of WEKA [32] based on 10-cross-validation (10CV) and leave-one-out (LOO). Comparing these results with those obtained using the stable motifs allow us to better evaluate the studied methods and test assumptions 1 and 2.

The experimental results are presented in tables 3 and 4. Table 3 contains the results obtained with a LOO based variation and table 4 with 10CV based variation. For each dataset and for each value of τ, we note the rate of stable motifs and their corresponding accuracy if we use these motifs to classify protein sequences of that dataset based on 10-cross-validation and leave-one-out tests. As shown in figure 3 and 4, we can notice that the classification test technique i.e., 10CV or LOO does not affect the obtained results (the classification accuracy rates are almost the same). Using results from tables 2, 3 and 4, we draw the interest of stable motifs histogram corresponding to dataset (figure 5 and 6).

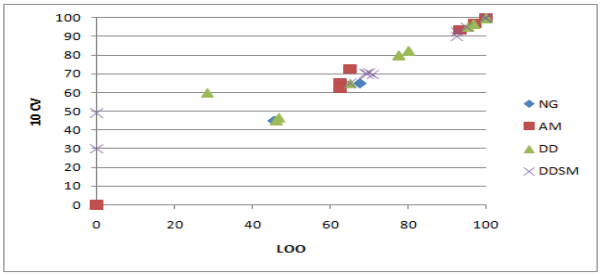

Figure 3: Effect of classification test technique (LOO and 10CV) to the classification accuracy rates using LOO-based variation

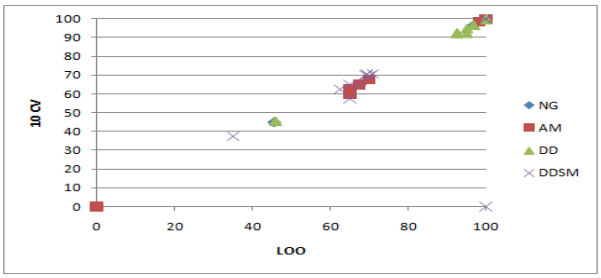

Figure 4: Effect of classification test technique (LOO and 10CV) to the classification accuracy rates using 10CV-based variation

We notice that NG is virtually insensible to varia-

Table 2: Accuracy rate of the studied methods using datasets without modification

Method	DS1		DS2		DS3		DS4	
	10CV	LOO	10CV	LOO	10CV	LOO	10CV	LOO
NG	96.7	96.7	67.5	67.5	100	100	44.9	45.5
AM	100	100	72.5	65	100	100	-	-
DD	96.7	96.7	82.5	80	100	100	43.5	43.5
DDSM	96.7	96.7	95	95	100	100	82.5	87.5

tions in data. Indeed its rate of stable motifs is often equal or very close to 100%. Therefore, the variation of input data has no bearing on the generated motifs. In other words, we often obtain the same motifs even in the presence of variations in input data. In this case, we can not evoke the interest of stable motifs (see figures 5 and 6).

The AM method follows almost the same fluctuating behavior for all datasets (except for DS4 where we could not conduct our experiments due to lack of memory). In fact, below $\tau = 0.7$, AM is insensible (RSM is equal or close to 100%). Beyond this value, AM becomes sensible. This sensibility varies depending on dataset and the variation technique (10CV or LOO). It is very significant for DS1, average for DS2 and slight for DS3. For example in table 3, for $\tau = 0.7$, the rate of stable motifs are 32.5, 83.6 and 98.3% respectively for DS1, DS2 and DS3. Similarly, the interest of stable AM motifs is very fluctuating and varies as well depending on the dataset (see figures 5 and 6). This method is sometimes completive to DD and DDSM. But, we note that it is greedy in memory and can not handle large datasets as it is the case with the dataset DS4 (see tables 3 and 4).

The approach adopted by the DD method offers it a sensible nature. In fact, according to this method, each motif must satisfy the conditions of discrimination and minimality (see section 4.2). Therefore, it is likely that a disruption of input yields not meeting these conditions and thus the elimination of some existing motifs and/or addition of new ones. At the same time, this method generates sets of interesting stable motifs with all the data samples and different values of τ . Indeed, it generally allows better interest rates than NG and AM (see figures 5 and 6).

The DDSM method is an extension of DD, which adopts a competitive approach among the motifs to generate. Indeed, to be chosen, a motif must be the most mutable among other ones of equal size. This constraint remarkably increases the sensibility of the method vis-a-vis the changes in the input data. This can be noticed by the decreasing rates of stable motifs compared to the DD method. In addition, this high sensibility is always accompanied by a set of very interesting stable motifs manifested by generally allowing the highest interest rates. However we note that for $\tau = 1$, DDSM does not often have the best rates of interest especially with 10CV based variation (we recall that this value of τ means that the stable motifs are those that appear in all variations of data). This is because that the substitution, which is a fundamental criterion in the process of DDSM, is not taken into account in the construction of the set of stable motifs. Hence, similar forms of a given motif may be ignored. But by relaxing the condition of $\tau = 1$ and moving to smaller values of τ we see that the interest rates get improved considerably. This method reveals both the property of sensibility and interest of its stable motifs (see figures 5 and 6), which allows it to redescribe well the input data, which is in accordance with results quoted in [22] showing the efficiency of this method for feature extraction in protein sequences.

6 Conclusion.

In this paper, we introduced the notions of stability and sensibility as new criteria to compare motif extraction methods from biological sequences. The sensibility of a method is its ability to produce a different set of motifs, so a different description, whenever a perturbation is made in the dataset. This criterion must be accompanied by a set of interesting stable motifs. This concept of interest arises when a method eliminates certain motif and conserves others following a change in the input data and that the conserved motifs are useful if used in a data mining task.

The experimental study shows that the DDSM method is more sensible compared to the other methods. This sensibility is usually accompanied by sets of stable interesting motifs. This confirms the results found by [22] that show the contribution of the DDSM method in supervised classification tasks.

As future works, the proposed feature extraction approach is announced to be generic since it can be coupled with any data mining task T. Hence, we plan to apply it to other types of tasks such as clustering [17] [37]. We will also explore its extension to text mining supervised classification tasks [24].

References

[1] A. Al-Ani, *A dependency-based search strategy for feature selection*, Expert Systems with Applications, 36 (2009), pp. 12392–12398.

[2] A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, eds., *Combinatorial Pattern Matching, Third Annual Symposium, CPM 92, Tucson, Arizona, USA, April 29 - May 1, 1992, Proceedings*, vol. 644 of Lecture Notes in Computer Science, Springer, 1992.

[3] A. Bairoch and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000*, Nucleic Acids Research, 28 (2000), pp. 45–48.

[4] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh, *The universal protein resource (uniprot)*, Nucleic acids research, 33 (2005), pp. 154–159.

[5] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, *Genbank*, Nucleic acids research, 37 (2009), pp. D26–31.

[6] J. Bernardes, J. Fernandez, and A. Vasconcelos, *Structural descriptor database: a new tool for sequence based functional site prediction*, BMC Bioinformatics, 9 (2008), p. 492.

[7] H. Bhaskar, D. C. Hoyle, and S. Singh, *Machine learning in bioinformatics: A brief survey and recommendations for practitioners*, Computers in Biology and Medicine, 36 (2005), pp. 1104–1125.

[8] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *A model of evolutionary change in proteins*, Atlas of protein sequence and structure, 5 (1978), pp. 345–351.

[9] D. Debroas, J. F. Humbert, F. Enault, G. Bronner, M. Faubladier, and E. Cornillot, *Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (lac du bourget - france)*, Environ. Microbiol, 11 (2009), pp. 2412–2424.

[10] K. Dunne, P. Cunningham, and F. Azuaje, *Solutions to Instability Problems with Sequential Wrapper-Based Approaches To Feature Selection*, Tech. Report TCD-CD-2002-28, Department of Computer Science, Trinity College, Dublin, Ireland, 2002.

[11] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems)*, Morgan Kaufmann, 2 ed., Jan. 2006.

[12] S. Henikoff and J. G. Henikoff, *Amino acid substitution matrices from protein blocks*, Proceedings of the National Academy of Sciences, 89 (1992), pp. 10915–10919.

[13] A. Kalousis, J. Prados, and M. Hilario, *Stability of feature selection algorithms: a study on high-dimensional spaces*, Knowledge and Information Systems, 12 (2007), pp. 95–116. 10.1007/s10115-006-0040-8.

[14] R. M. Karp, R. E. Miller, and A. L. Rosenberg, *Rapid Identification of Repeated Patterns in Strings, Trees and Arrays*, in Proc. of the 4th ACM Symposium on Theory of Computing, Denver, Colorado, May 1972, pp. 125–136.

[15] C. Leslie, E. Eskin, and W. S. S. Noble, *The spectrum kernel: a string kernel for SVM protein classification.*, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, (2002), pp. 564–575.

[16] H. Liu and H. Motoda, *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*, Chapman & Hall/CRC, 2007.

[17] H. Liu and L. Yu, *Toward integrating feature selection algorithms for classification and clustering*, Knowledge and Data Engineering, IEEE Transactions on, 17 (2005), pp. 491–502.

[18] S. Ma and J. Huang, *Penalized feature selection and classification in bioinformatics.*, Briefings in bioinformatics, 9 (2008), pp. 392–403.

[19] M. Maddouri and M. Elloumi, *Encoding of primary structures of biological macromolecules within a data mining perspective*, Journal of Computer Science and Technology (JCST), 19 (2004), pp. 78–88.

[20] K. Pavel, K. Josef, and H. Václav, *Improving stability of feature selection methods*, in Proceedings of the 12th international conference on Computer analysis of images and patterns, CAIP'07, Berlin, Heidelberg, 2007, Springer-Verlag, pp. 929–936.

[21] Y. Saeys, I. n. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics*, Bioinformatics, 23 (2007), pp. 2507–2517.

[22] R. Saidi, M. Maddouri, and E. Mephu Nguifo, *Protein sequences classification by means of feature extraction with substitution matrices.*, BMC bioinformatics, 11 (2010), pp. 175+.

[23] R. Saidi, M. Maddouri, and E. M. Nguifo, *Comparing graph-based representations of protein for mining purposes*, in Proceedings of the Kdd-09 Workshop on Statistical and Relational Learning in Bioinformatics, 2009, pp. 35–38.

[24] F. Sebastiani, *Machine learning in automated text categorization*, ACM Comput. Surv., 34 (2002), pp. 1–47.

[25] M. Sebban and R. Nock, *A hybrid filter/wrapper approach of feature selection using information theory*, Pattern Recognition, 35 (2002), pp. 835 – 846.

[26] R. Sloan, *Types of Noise in Data for Concept Learning (Extended Abstract)*, in Proceedings of the 1988 Workshop on Computational Learning Theory, MIT, Aug. 1988, ACM Press, pp. 91–96.

[27] R. Sloan, *Four types of noise in data for PAC learning*, Information Processing Letters, 54 (1995), pp. 157–162.

[28] P. Somol and J. Novovicov, *Evaluating the stability of feature selectors that optimize feature subset cardinality*, in Structural, Syntactic, and Statistical Pattern

Recognition, N. da Vitoria Lobo, T. Kasparis, F. Roli, J. Kwok, M. Georgiopoulos, G. Anagnostopoulos, and M. Loog, eds., vol. 5342 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, pp. 956–966.

[29] G. Stoesser, W. Baker, A. van den Broek, E. Camon, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. L. Q. Lin, V. Lombard, R. Lopez, N. Redaschi, P. Stoehr, M. A. Tuli, K. Tzouvara, and R. Vaughan, *The embl nucleotide sequence database*, Nucleic acids research, 30 (2002), pp. 21–26.

[30] T. Topaloglou, *Biological data management: research, practice and opportunities*, in Proceeding of the 30th VLDB conference, Toronto, 2004, pp. 1233–1236.

[31] J. T. L. Wang, T. G. Marr, D. Shasha, B. A. Shapiro, and G. W. Chirn, *Discovering Active Motifs in Sets of Related Protein Sequences and Using Them for Classification*, Nucleic Acids Research, 22 (1994), pp. 2769–2775.

[32] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, second ed., June 2005.

[33] Z. Xu, R. Jin, J. Ye, M. R. Lyu, and I. King, *Non-monotonic feature selection*, in Proceedings of the 26th Annual international Conference on Machine Learning (ICML'09), vol. 382, New York, 2009, ACM, pp. 1145–1152.

[34] L. Yu, C. Ding, and S. Loscalzo, *Stable feature selection via dense feature groups*, in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08, New York, NY, USA, 2008, ACM, pp. 803–811.

[35] S. Yvan, A. Thomas, and P. Yves, *Robust feature selection using ensemble feature selection techniques*, in Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II, ECML PKDD '08, Berlin, Heidelberg, 2008, Springer-Verlag, pp. 313–325.

[36] Y. Zhang and M. Zaki, *EXMOTIF: efficient structured motif extraction.*, Algorithms for molecular biology : AMB, 1 (2006), p. 21.

[37] Z. Zhao and H. Liu, *Spectral feature selection for supervised and unsupervised learning*, in ICML '07: Proceedings of the 24th international conference on Machine learning, New York, NY, USA, 2007, ACM, pp. 1151–1157.

[38] Z. Zhao, J. Wang, H. Liu, J. Ye, and Y. Chang, *Identifying biologically relevant genes via multiple heterogeneous data sources*, in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, USA, 2008, pp. 839–847.

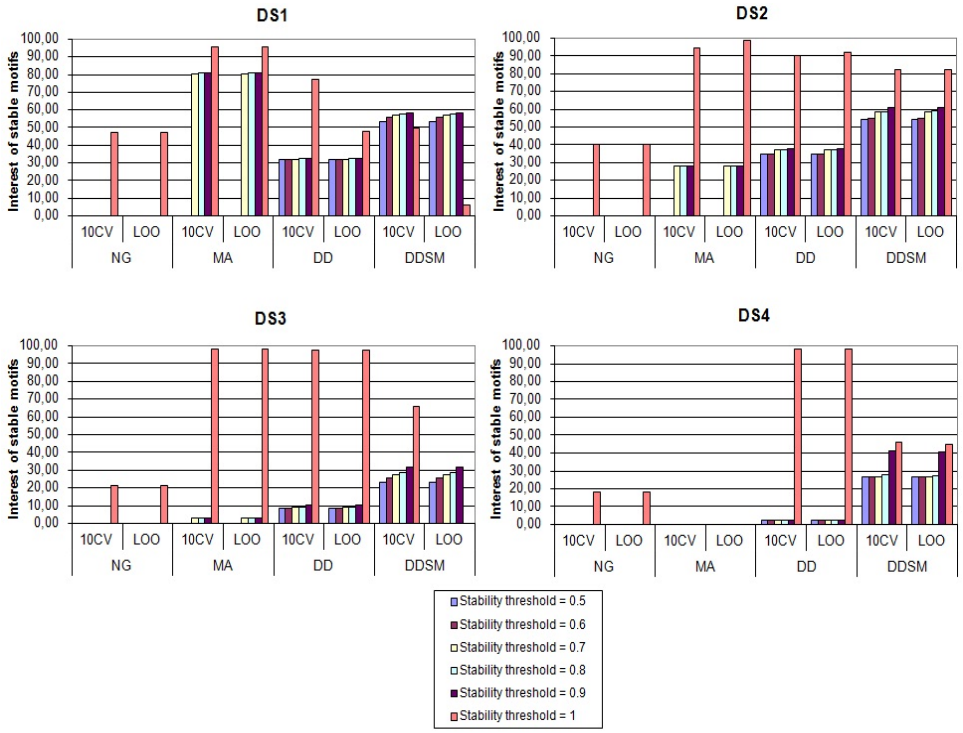

Figure 5: Interest of stable motifs using LOO-based variation

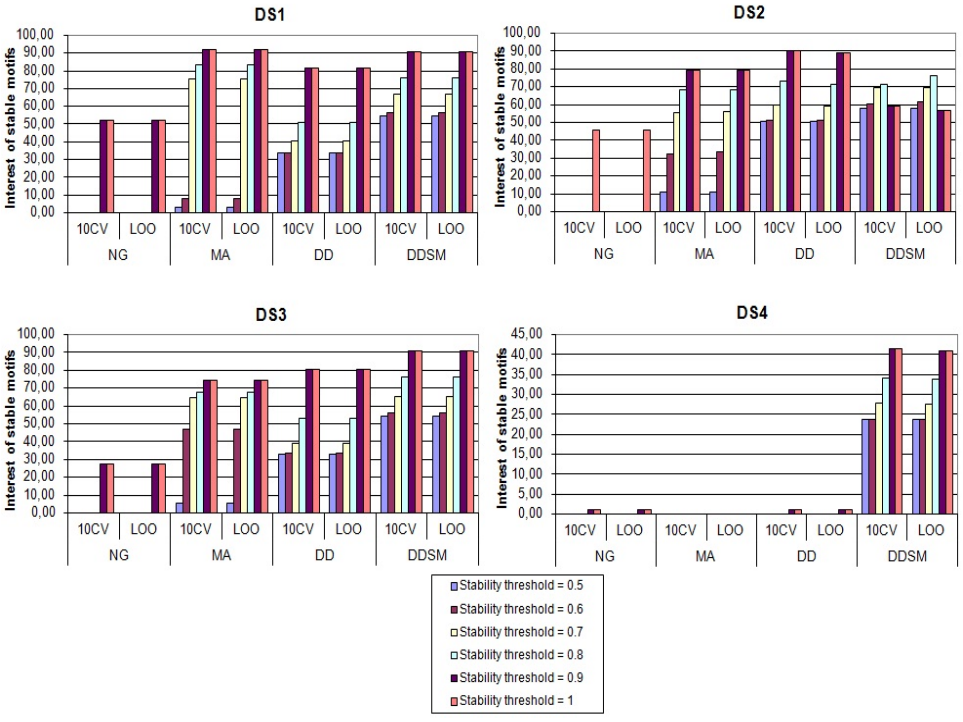

Figure 6: Interest of stable motifs using 10CV-based variation

Table 3: Rate of stable motifs and their classification accuracy using LOO-based variation

DS	τ	Rate of stable motifs				Classification accuracy rate							
		NG	AM	DD	DDSM	NG		AM		DD		DDSM	
						10CV	LOO	10CV	LOO	10CV	LOO	10CV	LOO
DS1	**0.5**	100	100	81	63	96.7	96.7	100	100	96.7	96.7	100	100
	0.6	100	100	81	60.5	96.7	96.7	100	100	96.7	96.7	100	100
	0.7	100	32.5	81	59.7	96.7	96.7	100	100	96.7	96.7	100	100
	0.8	100	31.6	80.8	59.1	96.7	96.7	100	100	96.7	96.7	100	100
	0.9	100	31.6	80.7	58.6	96.7	96.7	100	100	96.7	96.7	100	100
	1	69.3	1.8	0.01	0.02	96.7	96.7	93.3	93.3	60	28.3	30	0
DS2	**0.5**	100	100	79	62.8	67.5	67.5	72.5	65	82.5	80	95	95
	0.6	100	100	79	62.4	67.5	67.5	72.5	65	82.5	80	95	95
	0.7	100	83.6	76.9	58.4	67.5	67.5	65	62.5	80	77.5	92.5	92.5
	0.8	100	83.6	76.8	57.7	67.5	67.5	65	62.5	80	77.5	90	92.5
	0.9	100	83.6	76.7	55.5	67.5	67.5	65	62.5	80	77.5	92.5	92.5
	1	74.6	0.4	0.05	0. 01	65	67.5	62.5	62.5	65	65	65	65
DS3	**0.5**	100	100	95.6	86.7	100	100	100	100	100	100	99.8	99.8
	0.6	100	100	95.5	85.5	100	100	100	100	100	100	99.8	99.8
	0.7	100	98.3	95.2	83.9	100	100	100	100	100	100	99.8	99.8
	0.8	100	98.3	95.1	83.1	100	100	100	100	100	100	99.8	99.8
	0.9	100	98.3	94.7	81	100	100	100	100	100	100	99.8	99.8
	1	87.9	1.23	0.01	0.1	100	100	97	97.1	95.3	95.3	49	0
DS4	**0.5**	100	-	98.8	84.1	44.9	45.5	-	-	45	46	70.2	69.2
	0.6	100	-	98.8	84.1	44.9	45.5	-	-	45	46	70.2	69.2
	0.7	100	-	98.8	84.1	44.9	45.5	-	-	45	46	70.2	69.2
	0.8	100	-	98.8	83.6	44.9	45.5	-	-	45	46	69.9	69.5
	0.9	100	-	98.7	73.2	44.9	45.5	-	-	45	46	69.8	70.9
	1	90	-	0.01	69	44.9	45.5	-	-	46.7	46.7	70.6	70

Table 4: Rate of stable motifs and their classification accuracy using 10CV-based variation

DS	τ	Rate of stable motifs				Classification accuracy rate							
		NG	AM	DD	DDSM	NG		AM		DD		DDSM	
						10CV	LOO	10CV	LOO	10CV	LOO	10CV	LOO
DS1	**0.5**	100	98,4	79,9	62,1	96.7	96.7	98.3	98.3	96.7	96.7	100	100
	0.6	100	96	79.7	59.9	96.7	96.7	98.3	98.3	96.7	96.7	100	100
	0.7	100	38,5	74,5	49,2	96.7	96.7	98.3	98.3	96.7	96.7	100	100
	0.8	100	28.1	66.1	37.3	96.7	96.7	100	100	96.7	96.7	100	100
	0.9	64,6	14,8	30,8	14,5	96.7	96.7	100	100	96.7	96.7	100	100
	1	64.6	14.8	30.8	14.5	96.7	96.7	100	100	96.7	96.7	100	100
DS2	**0.5**	100	94,3	64,6	49,1	65	67.5	62.5	65	92.5	92.5	62.5	62.5
	0.6	100	80	64	46.6	65	67.5	60	65	92.5	92.5	65	67.5
	0.7	100	60,6	54,4	31,2	65	67.5	65	67.5	95	95	65	65
	0.8	100	46.5	38.2	16.7	65	67.5	67.5	70	92.5	95	57.5	65
	0.9	100	32,1	9,5	4,1	65	67.5	67.5	70	92.5	92.5	37.5	35
	1	70.1	32.1	9.5	4.1	65	67.5	67.5	70	92.5	92.5	37.5	35
DS3	**0.5**	100	97,1	80,3	62,8	100	100	100	100	100	100	99.8	99.8
	0.6	100	69.3	80	60.9	100	100	100	100	100	100	99.8	99.8
	0.7	100	52,4	75,8	51,8	100	100	100	100	100	100	99.8	99.8
	0.8	100	48.5	63.8	38	100	100	100	100	100	100	99.8	99.8
	0.9	84,3	40,5	33	17,1	100	100	100	100	100	100	99.8	99.8
	1	84.3	40.5	33	17.1	100	100	100	100	100	100	99.8	99.8
DS4	**0.5**	100	-	100	86.2	44.89	45.47	-	-	45.19	45.89	70.2	69.2
	0.6	100	-	100	86.2	44.9	45.5	-	-	45.2	45.9	70.2	69.2
	0.7	100	-	100	83.5	44.9	45.5	-	-	45.2	45.9	70.4	69.3
	0.8	100	-	100	78.8	44.9	45.5	-	-	45.2	45.9	69.5	70
	0.9	99,4	-	99,4	72.8	44.9	45.5	-	-	45.2	45.9	70.6	71
	1	99.4	-	99.4	72.8	44.9	45.5	-	-	45.2	45.9	70.6	71