# A new feature selection strategy for K-mers sequence representation

**Conference Paper** · June 2014

CITATIONS
0

READS
618

**2 authors:**

Giosuè Lo Bosco
Università degli Studi di Palermo
**142** PUBLICATIONS **1,750** CITATIONS
SEE PROFILE

Luca Pinello
Massachusetts General Hospital / Harvard Medical School
**236** PUBLICATIONS **10,802** CITATIONS
SEE PROFILE

# A NEW FEATURE SELECTION STRATEGY FOR K-MERS SEQUENCE REPRESENTATION

Lo Bosco, Giosué[(1),(2)], Pinello, Luca [(3),(4)]

(1) Universitá di Palermo
Dipartimento di Matematica e Informatica, Palermo, ITALY, giosue.lobosco@unipa.it

(2) I.E.ME.S.T.
Istituto Euro Mediterraneo di Scienza e Tecnologia, Palermo, ITALY

(3) Harvard School of Public Health
Department of Biostatistics, Boston, MA, USA, lpinello@jimmy.harvard.edu

(4) Dana-Farber Cancer Institute
Department of Biostatistics and Computational Biology, Boston, MA, USA

*Keywords*: k-mers, DNA sequence similarity, feature selection, DNA sequence classification.

**Abstract.** DNA sequence decomposition into k-mers (substrings of length k) and their frequency counting, defines a mapping of a sequence into a numerical space by a numerical feature vector of fixed length. This simple process allows to compute sequence comparison in an alignment free way, using common similarities and distance functions on the numerical codomain of the mapping. The most common used decomposition uses all the substrings of length $k$ making the codomain of exponential dimension. This obviously can affect the time complexity of the similarity computation, and in general of the machine learning algorithm used for the purpose of sequence classification. Moreover, the presence of possible noisy features can also affect seriously the classification accuracy. In this paper we propose a feature selection method able to select the most informative k-mers associated to a set of DNA sequences. Such selection is based on the *Motif Independent Measure (MIM)*, an unbiased quantitative measure for DNA sequence specificity that we have recently introduced in the literature. Results computed on three public datasets using the Support vector machine classifier, show the effectiveness of the proposed feature selection method.

## 1 Scientific Background

A fundamental biological question is to understand the function of the genome, and nowadays, despite the development of computational models for functional annotation, it still remains a daunting task. In particular, initial biological hypotheses about sequence similarity, has been traditionally generated by using sequence alignment methods, such as BLAST [1] and FASTA [2]. Although the power of alignment-based methods has been well-demonstrated in a broad range of applications, and despite the recent efforts in improving their computational efficiency, their applications are not unlimited. The main assumption of alignment based methods is that functional elements share common sequence features and the relative order of these elements is conserved between different sequences. Unfortunately there are cases showing that this assumption is violated e.g. the cis-regulatory elements sequences, where there is little evidence suggesting the order between different elements would have any significant effect in regulating gene expression. As such, the recently developed alignment-free methods [3] have emerged as a promising approach to investigate the regulatory genome. One of

the methods belonging to this latter class is based on substring counting of a sequence, and is generally named as *k-mers* or *L-tuple* representation. Informally, k-mers representation associates a sequence with a feature vector of fixed length, whose components count the frequency of each substrings belonging to a finite set of words. The main advantage is that the sequence is represented into a numerical space where a particular distance function between the vectors can be adopted to reflect the observed similarities between sequences. K-mers have shown their effectiveness in several in-silico analysis applied to different genomics and epigenomics studies. In particular they have been used in several computational methods to characterize nucleosome positioning [4], functional regions such as enhancers [5], epigenetic variability [6], in sequence alignment and transcriptome assembly [7] and in gene prediction [8]. The interested reader can find the basic ideas of k-mer based methods to different biological problems in the following review [12]. Anyway, their use involve practical computational issues: they uses all the substrings of length $k$ making the numerical biological data of exponential dimension [13]. Note that this represents a key problem, bioinformaticians frequently face the challenge of reducing the number of features of high dimensional biological data for improving the models involved in sequence analysis. To this purpose, *feature selection algorithms* can be succesfully applied. Their main goals are (1) speeding up the response of the model used for the analysis and (2) eliminate the presence of possible noisy features that could affect seriously the accuracy of the model. For sure, every feature selection method is based on the generation of a proper subset of features, which has been shown to be computationally intractable [14] and for this reason, they belong to the class of heuristics. The main classsification of feature selection methods is into wrapper approaches which uses a predictive model to evaluate the feature subset, and filter approaches which score the subset just by looking at the intrinsic properties of data [15]. In this paper we propose a filter feature selection method able to select the most informative k-mers associated to a set of DNA sequences. Such selection is based on the weight given to each feature by a measure called *Motif Independent Measure (MIM)*. The effectiveness of the method has been tested on three datasets with the purpose of sequence nucleosome classification, using the Support vector machine as classification paradigm.

## 2 Materials and Methods

A generic DNA sequence $s$ can be represented as a string of symbols taken from a finite alphabet. We can think to a particular mapping function that project $s$ into a vector $x_s$ (the feature vector), allowing to represent $s$ into a multi-dimensional space (the feature space) where a particular distance function between the vectors can be adopted to reflect the observed similarities between sequences. One of the most common ways of defining such mapping, is to consider a feature vector $x_s$ that enumerates the frequency of occurrence of a finite set of pre-selected words $W = \{w_i, .., w_m\}$ in the string $s$. The simplest and most common definition of $W$ is by using *k-mers*, i.e. a set containing any string of length $k$ whose symbols are taken in the nucleotide alphabet $\Sigma = \{A, T, C, G\}$. In this case, each sequence $s$ is mapped to a vector $x_s \in \mathbb{Z}^m$ with $m = 4^k$.

The idea behind the proposed feature selection method is to assign a weight to each k-mer, and use this weights for their selection.

### 2.1 k-mers weigthting

In general, each numerical component $x_s^j$ of the feature vector is set to the value $f_s^j$ that represents the frequency of the $j-th$ k-mer $w_k^j$ in $s$ counted by a sliding window of length $k$ that is run through the sequence $s$, from position 1 to $L-k+1$. Another possible choice is to set $x_s^j$ to the empirical probabilities $p_s^j = f_s^j/(L-k+1)$ .

Specifically, let $\mathbf{P}_S = (p_{s_i}^j)$ be the k-mer probability distributions corresponding to a

set of $n$ target sequences $S = \{s_i\}$ for a fixed length $k$, where $i = 1, .., n$, $j = 1, .., m$. Let us assume to have a process to generate a set of $n$ *background* sequences $B = \{b_i\}$ (analogously $b_i$ represents a sequence in the set $B$). In the most simple case, the background set of sequences corresponds to random sequences that can be generated for example by randomly shuffling each one of the sequences belonging to the target set $S$.

Let $\mathbf{Q}_B = (q^j_{b_i})$ be the k-mer probability distributions corresponding to $B$ for a fixed length $k$. For each $j$, we can calculate the symmetrical Kullback-Leibler divergence between the empirical probabilities $P_j$ and $Q_j$:

$$d_{kl}(P_j, Q_j) = \frac{\sum_i p^j_{s_i} log_2 \frac{p^j_{s_i}}{q^j_{b_i}} + \sum_i q^j_{b_i} log_2 \frac{q^j_{b_i}}{p^j_{s_i}}}{2} \quad (1)$$

We recall that this divergence is able to measure the difference between two probability distributions. The *Motif Independent Measure* (MIM) value corresponding to a k-mer $w_j$ is defined as the expected value $d_{kl}(P_j, Q_j)$, which is estimated by averaging over a finite set $N > n$ of background sequences, and is indicated as $MIM(w_j)$.

### 2.2 k-mers selection

We can compute the $MIM$ values for each k-mer $w_j$, obtaining a list $L$ of $m = 4^k$ numerical values. Here we use their ranking as a guide to identify the most informative k-mers, in particular we sort $L$ in ascending order, resulting in a ordered list of k-mers $w_{j_1}$,..,$w_{j_m}$. The criteria used to select the most informative k-mers among the possible $4^k$ is based on Z-scores of the computed MIM values. Finally, let $A_\alpha = \{w_{j_i} | abs(Z(MIM(w_{j_i}))) > \alpha\}$ where $Z$ indicate Z-score, the adopted selection criteria consists in selecting a number of k-mers equal to

$$r = max(|A_\alpha|, \beta * m) \quad (2)$$

with $\alpha, \beta < 1$.

### 2.3 Support Vector Machine

Support Vector Machine (SVM) is a powerful classification method that has been widely used in the realm of bioinformatics. Differently from the other classification methods, there is a strong theory behind it that motivate its broad applicability. Since in the case of linear separability it offers the best generalization, the basic idea behind SVM is to transform the data into a high dimensional feature space by a nonlinear mapping, and there determines the optimal separating hyperplane by solving a quadratic optimization problem. The data transformation is done by the so called *kernel function* whose choice represents one of the parameter of the method. In this work we have adopted a SVM which uses a quadratic kernel. It allows to project the data by a nonlinear function, and differently from other kernels, does not require to estimate any additional parameter.

### 2.4 Dataset description

In this study we have considered three datasets of DNA sequences underlying nucleosomes from the following three species: (i) *Homo sapiens (HM)*; (ii) *Caenorhabditis elegans (CE)* and (iii) *Drosophila melanogaster (DM)*. The nucleosome is the primary repeating unit of chromatin, which consists of $147$ bp of DNA wrapped $1.67$ times around an octamer of core histone proteins [9]. Several studies have shown that nucleosome positioning plays an important role in gene regulation and that distinct DNA sequence features have been identified to be associated with nucleosome positioning [10]. Details about all the step of data extraction and filtering of the three datasets can

be found in the work by Guo et al [11] and in the references therein. Each of the three datasets is composed by two classes of samples: the nucleosome-forming sequence samples (positive data) and the linkers or nucleosome-inhibiting sequence samples (negative data). The $HM$ dataset contains $2,273$ positives and $2,300$ negatives, the $CE$ $2,567$ positives and $2,608$ negatives and the $DM$ $2,900$ positives and $2,850$ negatives. The length of a generic sequence is $147$ bp.

## 3 Results

Giving a dataset of $n$ sequences $S$, the SVM classifier uses $R \subset S$ as training set, while the remaining as test set using a $10$ fold cross validation schema. We have used as numerical dataset $D_S$ for the SVM classifier a matrix of size $n \times 4^k$ that contains the empirical probabilities $p_s^j$ of a sequence $s$ belonging to a selected species for a fixed length $k$. We have computed the experiments for different $k$ ranging from $5$ to $7$. Such range has been chosen due to the used classifier, since it has been noted that the SVM with the quadratic kernel does not lead the optimization to converge for $k < 5$. It is important to point out that the literature motivate the use of a k-mer length equal to $6$ as a good choice to capture dependencies between adjacent nucleotides. We have computed a total of $3$ metrics to measure the performance of the classifier: *Sensitivity (Se), Specificity (Sp) and Accuracy (A)*. In the following, we recall their definitions:

$$Se = \frac{TP}{TP + FN},\ Sp = \frac{TN}{FP + TN},\ A = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

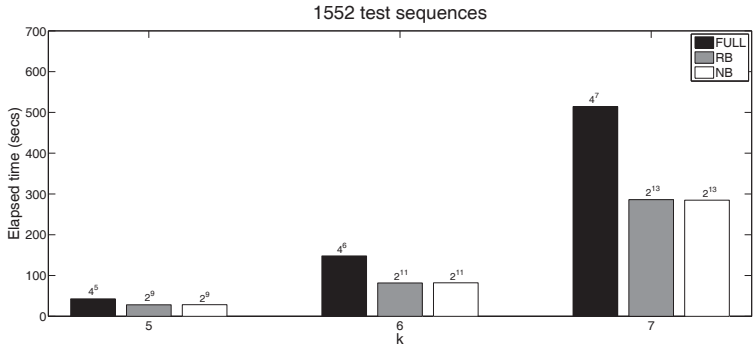
where the prefix $T$ (true) indicates the number of correctly classified sequences, $F$ (false) the uncorrect ones, $P$ the positives class and $N$ the negatives class. The three metrics have been computed by using the classifier on the full datasets with $4^k$ features (indicated by *FULL*), and on two different schema of feature selections. The first one, indicated by *random background selection (RB)* generates the k-mers background distribution by estimating first the probability of each single nucleotide in $R$, and then calculating the probability of the j-th k-mer $w_j$ as the product of the probability of the single nucleotides in $w_j$. The latter, indicated by *negative background selection (NB)* uses as background the set of negative sequences $B = \{s \in R \mid s \text{ is neagtive}\}$. For the selection criterion, we have decided to set $\alpha = 0.7$ and $\beta = 0.5$. This assures that at least one half of the best features are selected. In order to have a fair and complete comparison, we have also considered a *random feature selection (WR)* that consider the first $r$ elements of a random permutation of the features where $r$ indicate the number of features established by $RB$ and $NB$. In Table 1 we report the results of mean ($\mu$) and standard deviation ($\sigma$) of the three metrics for all the three datasets, computed on $10$ folds.

Table 1: In column, for each k in the range 5,..,7 the mean and standard deviation values of Specificity (Sp), Sensitivity (Se) and Accuracy (A) values computed on 10 folds in the cases of the Caenorhabditis elegans (CE), Drosophila melanogaster (DM) and Homo sapiens (HM) full (Full) and reduced (WR,RB,NB) datasets. In bold, the values with the best values for each dataset.

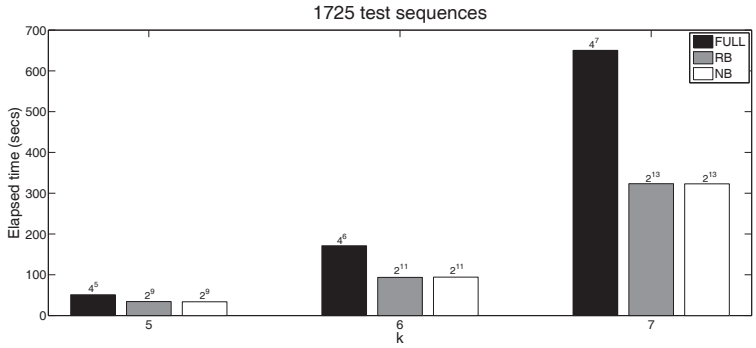
| | K=5 | | | | | | K=6 | | | | | | K=7 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | Se | | Sp | | A | | Se | | Sp | | A | | Se | | Sp | |
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| **CE-FULL** | 79,36 | 0,02 | 75,89 | 0,03 | **82,78** | 0,02 | 62,94 | 0,02 | 31,01 | 0,03 | **94,36** | 0,02 | 50,51 | 0 | 0,31 | 0 | **99,92** | 0 |
| **CE-WR** | 76,37 | 0,02 | 73,86 | 0,04 | 78,84 | 0,02 | 69,97 | 0,02 | 49,04 | 0,03 | 90,57 | 0,01 | 50,51 | 0 | 1,36 | 0,01 | 98,89 | 0,01 |
| **CE-RB** | 78,17 | 0,01 | 79,90 | 0,02 | 76,46 | 0,02 | 74,82 | 0,02 | 63,14 | 0,04 | 86,31 | 0,02 | 51,98 | 0,01 | **5,10** | 0,01 | 98,12 | 0,01 |
| **CE-NB** | **80,08** | 0,02 | **80,76** | 0,02 | 79,41 | 0,03 | **80,49** | 0,02 | **73,94** | 0,02 | 86,93 | 0,02 | **50,67** | 0,01 | 2,07 | 0,01 | 98,50 | 0,01 |
| **DM-FULL** | **76,73** | 0,01 | 70,59 | 0,03 | **82,98** | 0,02 | 74,52 | 0,02 | 62,21 | 0,03 | **87,05** | 0,02 | 52,30 | 0,01 | 14,31 | 0,02 | 90,95 | 0,01 |
| **DM-WR** | 73,72 | 0,01 | 69,24 | 0,02 | 78,28 | 0,02 | 73,08 | 0,02 | 66,28 | 0,02 | 80,00 | 0,02 | 53,53 | 0,01 | 14,41 | 0,02 | 93,33 | 0,03 |
| **DM-RB** | 75,51 | 0,02 | **73,03** | 0,03 | 78,04 | 0,03 | 75,93 | 0,02 | 66,17 | 0,04 | 85,86 | 0,02 | 52,96 | 0,01 | 9,41 | 0,02 | **97,26** | 0,01 |
| **DM-NB** | 76,09 | 0,01 | 72,55 | 0,03 | 79,68 | 0,02 | **78,16** | 0,02 | **72,86** | 0,03 | 83,54 | 0,03 | **54,26** | 0,02 | **18,31** | 0,03 | 90,84 | 0,02 |
| **HM-FULL** | **84,38** | 0,02 | **91,72** | 0,02 | 77,13 | 0,04 | 84,81 | 0,02 | 91,63 | 0,02 | **78,09** | 0,03 | 48,6 | 0 | 0,09 | 0 | **96,48** | 0,01 |
| **HM-WR** | 82,36 | 0,02 | 87,67 | 0,02 | 77,13 | 0,03 | 83,61 | 0,01 | 88,55 | 0,01 | 78,74 | 0,03 | 63,28 | 0,03 | 32,42 | 0,07 | 93,74 | 0,02 |
| **HM-RB** | 83,00 | 0,02 | 89,25 | 0,04 | 76,83 | 0,02 | 84,11 | 0,01 | 92,60 | 0,01 | 75,74 | 0,03 | 73,68 | 0,03 | 73,79 | 0,10 | 73,57 | 0,12 |
| **HM-NB** | **84,38** | 0,02 | 90,70 | 0,02 | **78,13** | 0,03 | **85,03** | 0,01 | **93,17** | 0,01 | 77,00 | 0,03 | **73,98** | 0,02 | **93,96** | 0,02 | 54,26 | 0,03 |

Results shows that for $k = 5$, in the case of $CE$ and $DM$ datasets the accuracy obtained by the $RB$ and $NB$ feature selection methods is comparable or superior to

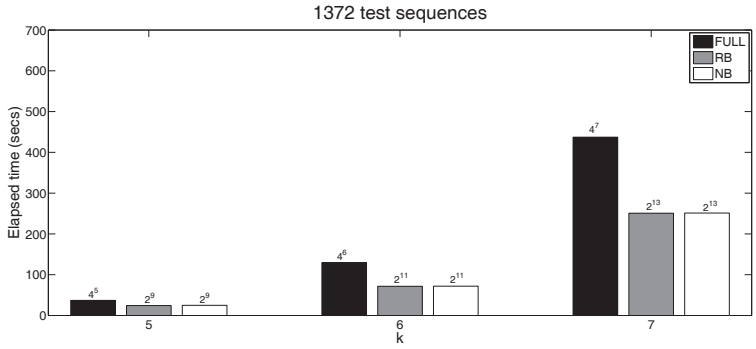
the $FULL$ and $WR$ cases. Moreover, their sensitivity with respect to the $FULL$ case is improved (at least $4\%$ for $CE$ and at least $2\%$ for $DM$). In the case of $k = 6$ it is observable a significant increase in sensitivity (at most $40\%$) and accuracy (at most $18\%$). Such improvements are not observable for the $HM$ dataset, but both sensitivity and accuracy are comparable to the $FULL$ case. Finally, the use of $k = 7$ seems not the right choice for every considered dataets, but this is more visible for $CE$ and $DM$. Finally, note that the proposed feature selection method can decrease slightly the specificity. In order to show the computational complexity advantage of the feature selection, we have also computed the empirical computation time of the classifier for the test sequences, in the case of $FULL$, $RB$, and $NB$. This is shown on Figure 1a,b,c. In the same figure, the number of features $r$ used by the classifier for the three cases is also shown on top of each bar. Note that the used values of $\alpha = 0.7$ and $\beta = 0.5$ always reduce the number of features of $RB$ and $NB$ by a factor of $0.5$, resulting in a significant reduction of computation time.




(a)




(b)




(c)

Figure 1: The empirical computation times in seconds for the C.Elegans (a), D. Melanogaster (b), H. Sapiens (c). On top of each bar the number of features used by the SVM classifier.

## 4 Conclusion

In this paper we have presented a feature selection method for DNA sequences based on an unbiased quantitative measure for DNA sequence specificity called Motif Independent Measure (MIM). It uses the k-mers counting representation, and selects the most informative k-mers associated to a target set of DNA sequences by computing the *Kullback-Leibler* divergence between the k-mers distributions computed in the target set and in an opportunely generated background set. Results carried out on three datasets of nucleosomes belonging to C. Elegans, D. Melanogaster and H. Sapiens species, using the Support vector machine classifier with quadratic kernel, have shown the advantage of the proposed feature selection method in terms of Sensitivity, Accuracy and computation time. In the future we plan to extend the experimental part on other dataset of sequences, also adopting other classification paradigms e.g. using other SVM kernels.

**Funding**

G. Lo Bosco was partially supported by Progetto di Ateneo dellUniversitá degli Studi di Palermo 2012-ATE-0298 *Metodi Formali e Algoritmici per la Bioinformatica su Scala Genomica*.

References

[1] S. Altschul, W. Gish, W. Miller et al. "Basic local alignment search tool". *JMol Biol*, vol.25, N.3, pp. 403410, 1990.

[2] D. Lipman, W. Pearson, "Rapid and sensitive protein similarity searches". *Science*, vol.227, N.4693, 1985.

[3] S. Vinga, J. Almeida, "Alignment-free sequence comparisona review". *Bioinformatics*, vol.19, N.4, pp.513523, 2003.

[4] G.-C. Yuan, J. S. Liu, "Genomic sequence is highly predictive of local nucleosome depletion". *PLoS Comput Biol*, vol.4, N.1, e13, 2008.

[5] D. Lee, R. Karchin, M. A. Beer, "Discriminative prediction of mammalian enhancers from DNA sequence". *Genome Research*, vol.21, N.12, pp.2167–2180, 2011.

[6] L. Pinello, J. Xu, S. H. Orkin, G.-C. Yuan, "Analysis of chromatin-state plasticity identifies cell-type specific regulators of H3K27me3 patterns". *Proceedings of the National Academy of Sciences*, vol.111, N.3, pp. 344–353, 2014.

[7] K. Paszkiewicz, D. J. Studholme, "De novo assembly of short sequence reads". *Briefings in bioinformatics*, vol 11, N.5, pp.457–472, 2010.

[8] Y. Liu, J. Guo, G.-Q. Hu, H. Zhu, "Gene prediction in metagenomic fragments based on the svm algorithm". *BMC Bioinformatics*, vol 14, S-5, S12, 2013.

[9] R.D. Kornberg and Y. Lorch, "Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome". *Cell*, vol.98, pp. 285294, 1999.

[10] K. Struhl and E. Segal, "Determinants of nucleosome positioning". *Nat StructMol Biol*, vol.20, N.3, pp. 267-273, 2013.

[11] S-H. Guo, E-Z. Deng, L-Q. Xu, H. Ding, H. Lin, W. Chen, K-C. Chou, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition", *Bioinformatics*, vol.30, n.11, pp.1522-1529, 2014.

[12] L. Pinello, G. Lo Bosco and G-C. Yuan, "Applications of alignment-free methods in epigenomics", *Briefings in Bioinformatics*, vol.15, N.3, pp.419-430, 2013.

[13] A. Apostolico, O. Denas, "Fast algorithms for computing sequence distances by exhaustive substring composition", *Algorithms for Molucular Biology*, vol.3, N.13, pp. 19, 2008.

[14] R. Kohavi, G.H. John, "Wrappers for feature subset selection", *Artificial Intelligence*, vol.97, N.1–2, pp.273324, 1997.

[15] Y. Saeys, I. Inza, P. Larrañaga, "A Review of Feature Selection Techniques in Bioinformatics", *Bioinformatics*, vol 23, N.19,pp. 2507–2517, 2007.