Computer Science Journal of Moldova, vol.22, no.2(65), 2014

A New Algorithm for Localized Motif Detection in Long DNA Sequences

Invited Article

Alin G. Voina, Petre G. Pop, Mircea F. Vaida

Abstract

The evolution in genome sequencing has known a spectacular growth during the last decade. One of the main challenges for the researchers is to understand the evolution of the genome and in particular to identify the DNA segments that have a biological significance. In this study we present a new algorithm – ADMSL – optimized for finding motifs in long DNA sequences and we emphasize some experiments done in order to evaluate the performance of the proposed algorithm in comparison with other motifs finding algorithms.

Index Terms: motifs search algorithms, motifs identification, transcriptions factor binding site, biological data analysis.

1 Introduction

The identification of novel cis-regulatory motifs in DNA sequences experienced a spectacular development in the recent years. As a consequence, an important number of algorithms have been developed with the scope to detect transcriptional regulatory elements from genes that belong to a specific genome [1].

The main scope of these algorithms is to identify the transcriptional regions and to find the motifs which are repeating most because those are good candidates for functional elements in genome. Phylogenetic footprinting is a particular method that is used to identify transcription factor binding sites in a set of orthologous noncoding DNA sequences.

The algorithms elaborated so far are capable of analyzing multiple DNA sequences and some of them can perform also over an entire gene. The process of regulating gene expression is an important challenge in molecular biology. The main task in this challenge is to identify the DNA binding sites for transcription factors. Computational methods have a special place in researcher's studies as are expected to offer the most promising results.

The problem of motifs detection can be formulated as: having a group of S sequences, search for a pattern M of length l which is spread more often. If the pattern M of length l is present in each sequence from the group of S sequences, then by enumerating the l letters of the pattern we obtain the regulatory element. The mutations of the nucleotides can affect the identification of transcription factor binding sites from a set of DNA sequences.

The identification of sequence motifs is an important step for understanding the process behind gene expression. A DNA motif is a short, well conserved pattern that usually has a biological significance [2]. Some of the motifs are included in complex RNA processes like transcription termination, mRNA processing, ribosome binding [3]. The length of the motif can vary from five base pairs (bp) to twenty (bp) and can be identified within the same gene or in different genes. Motifs can be classified based on their length but can be split also in palindromic motifs and gapped (space dyad) motifs [4]. We classify a motif as palindromic if its complementary read backwards is identical with the motif itself (e.g. '*AGAGCGCTCT*' is a palindromic motif). Space dyed (gapped) motifs are usually formed from two sites of relatively short length, well conserved and usually separated by a spacer. The gap is usually located in the middle of the motif due to the fact that transcription factor (TF) usually binds as a dimer. The length of the sites where TF binds to the DNA varies from three to five bp which are usually well conserved.

In the past, binding sites determination was performed with gel-shift and footprinting methods or reported construct assays [3].

In the recent years, for determining motifs in a sequence or a set of sequences, computational methods are used increasingly more.

The development of DNA motifs search algorithms was materialized into more than seventy elaborated methods for motifs identification. A good part of these methods are based on phylogenetic footprinting and/or probabilistic models.

The algorithms dealing with motifs identification can be organized into three main groups:

- algorithms that use promoter sequences from co-regulated genes of a single genome;
- algorithms that use phylogenetic footprinting;
- a combination of the above algorithms.

In this study we present a new algorithm (ADMSL) for motifs identification in long DNA sequences and we compare the results with the ones obtained with six popular tools: MEME, Weeder, AlignACE, YMF, Scope and Improbizer which are presented in the table below (Table 1).

2 Motifs localization in long sequences

The detection of motifs in case of long-range regulatory sequences became a requirement in ChIP experiments [5] – especially when searching for vertebrate promoters. If we refer to long DNA sequences, some recent studies [6] [7], reported that stochastic patterns may behave as real motifs. This can lead to false positive motifs which can eclipse the motifs identified as real. The length of the analyzed DNA sequence has a large influence over memory and time requirements for algorithms that search for motifs.

The binding sites are specifically bound by one or more DNA-binding proteins and are usually localized in specific positions [5]. Most of the Transcription Factor Binding Sites-TFBS are positioned relative to TSS to allow the transcription factors to anchor at specific positions with respect to each other and the TSS [8]. For this particular situation, the detection of the motif can be performed by searching into

Table 1. Analyzed Tools-Operation Principles

Analyzed tool	Principle of functionality	Observations
AlignAce	It uses an iterative masking procedure together with Gibbs sampling.	The detection of motifs is accomplished using an iterative masking procedure [6].
MEME	Uses statistical modeling techniques.	Motif detection consists in performing expectation maximization from starting points derived from each subsequence occurring in the input dataset [15].
Improbizer	Uses Expectaction Maximization.	In particular, Improbizer is using a variation of the expectation maximization (EM) algorithm [16].
Weeder	Consensus-based method.	It has options for "post-processing" i.e. analysis of location and significance of the motifs [17].
YMF	Finds motifs with the greatest z-score.	Identifies candidates for binding sites by searching for statistically over-represented motifs.
SCOPE	Uses three programs behind the scenes to identify different kind of motifs.	Utilizes three algorithms to identify sequence motifs: BEAM-finds non degenerate motifs, PRISM-finds degenerate motifs and SPACER – finds bipartite motifs [18].

an appropriate interval after the sequence is aligned relative to an anchor point. In this way, the regions that are not containing any motif are removed and the probability of reporting false positive motifs is decreased.

One solution would be to divide the long sequences into short overlapping sequences of the same length and to analyze each subsequence with a motif finding algorithm. But this approach can lead us into several problems:

- in most of the situations we have no prior information regarding the regulatory region where motifs may be localized;
- it is a big challenge to localize the motifs which are most significant for the whole DNA sequence when a considerable number of motifs were reported over a range of intervals;
- the length of the subsequences has a big influence over the motif identification process – in case of a short length the motif may not be visible and in case of a long length, the motif may be eclipsed;
- the analyzed sequence must be divided automatically; otherwise it will take considerable time and also may be predisposed to errors.

In the proposed algorithm of this research, we've taken the decision to not use subsequences of the original DNA sequence and to make the analysis over the entire sequence as we get it from genome repositories.

The problem of motif detection is well defined in the literature. One of the most common definitions is the one described in [9]. So, the main task is to determine all the instances of the pattern M of length l with d substitutions that occur into the set of analyzed sequences. The pattern M is known as a *motif* and each instance of the motif M represents a *binding site.*

Positional weight matrix (PWM) is another representation that can be used for motif detection, especially for the motifs that have particular instances localized over DNA sequences. For initial motif detection, the consensus representation (l, d) proved to be more efficient, in particular for the motifs which are not having a consistent instance across the sequences [9].

The definition of motif referenced above is taking into consideration the fact that instances of a motif can be distributed over the entire sequence which is true, in particular for short sequences. For long sequences it is considered that most of the motif instances are found into a specific interval, relative to an anchor point (Figure 1).

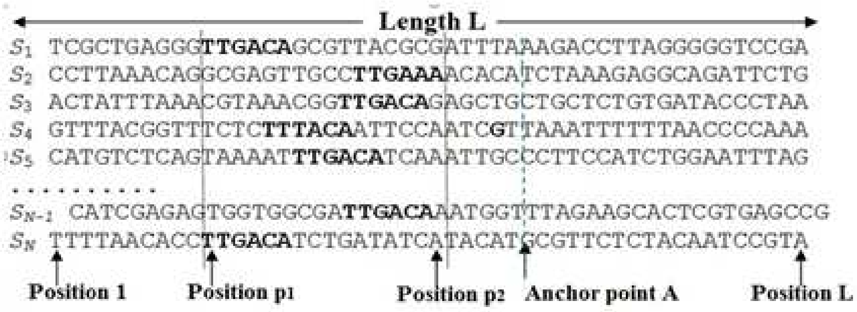
Length L
S1 TCGCTGAGGGTTGACAGCGTTACGCGATTTAAAGACCTTAGGGGGTCCGA
S2 CCTTAAACAGGCGAGTTGCCTTGAAAACACATCTAAAGAGGCAGATTCTG
S3 ACTATTTAAACGTAAACGGTTGACAGAGCTGCTGCTCTGTGATACCCTAA
S4 GTTTACGGTTTCTCTTTACAATTCCAATCGTTAAATTTTTTAACCCCAAA
S5 CATGTCTCAGTAAAATTTGACATCAAATTGCCCTTCCATCTGGAATTTAG
....
SN-1 CATCGAGAGTGGTGGCGATTGACAAATGGTTTAGAAGCACTCGTGAGCCG
SN TTTTAACACCTTGACATCTGATATCATACATGCGTTCTCTACAATCCGTA
Position 1
Position p1
Position p2
Anchor point A
Position L

Figure 1. Motifs detection of pattern (6,1) into a set of N sequences, each one of length L. The random pattern TTTAAA hides the real motif TTGACA

The problem of motif detection, in fact, is a variation of the above definition:

- for a set of S sequences $S = S_1, S_2, \ldots, S_N$, each of length L, we have to find all instances of pattern M of length l across the interval (p_1, p_2) of the sequences S;
- known values: S – the set of sequences, l – length of the pattern, d – maximum number of substitutions.

3 ADMSL – Algorithm description

The scope of the ADMSL algorithm is to identify candidate motifs from different intervals of the analyzed sequence and to report the ones which have the best score. An exhaustive enumeration strategy will require the computation of the score functions for 4^l patterns from all possible intervals of the sequences which gets to an increased complexity. One of the goals of the ADMSL algorithm is to process faster long data sequences.

In order to make judgments if a candidate pattern is a motif or not we've used several score functions. The motifs are expected to be distinct from the general nucleotide composition of the regulatory sequences – known as *background* – since the transcription factors can distinguish them from other neighborhood nucleotide patterns. One score function that we've used in order to measure the difference between the motif M and the background model B is the relative entropy score [10] [11] measured as Kullback-Leibler divergence:

$$D\_norm(M\|B) = \frac{1}{l\ln 4}\sum_{i=1}^{L}\sum_{b} f_{b,i}\ln(f_{b,i}) - \frac{1}{\ln 4}\sum_{b}\overline{f_b}\ln p_b \ , \qquad (1)$$

where

$$\overline{f_b} = \frac{1}{l}\sum_{i=1}^{L} f_{b,i}. \qquad (2)$$

$f_{b,i}$ – represents the average frequency of occurrence on each nucleotide $b \in \{A, C, G, T\}$ at each position $i = 1, 2, ..., l$. To measure the statistical deviation between the observed and expected occurrences of a motif we've used the Z-score function [12]:

$$Z - score = \frac{(\frac{n}{NL}) - e}{\delta} \ , \qquad (3)$$

where n is the number of observed instances, N – is the total number of input sequences, L – is the average length of input sequence, e – is the probability to generate a motif instance according to the background model and δ – represents the standard deviation.

In order to make decisions regarding the distribution and localization of motifs into a certain interval (p_1, p_2) we've used the following score function defined by the mathematical definition from relation (4):

$$D(\hat{p}\|p_0) = \hat{p}\ln(\frac{\hat{p}}{p_0}) + (1-\hat{p}\ln(\frac{1-\hat{p}}{1-p_0}), \qquad (4)$$

where $\hat{p}$ – represents the observed proportion of the motifs that is found in (p_1, p_2) interval, $(1-\hat{p})$ – the observed proportion of the motifs that lies outside of (p_1, p_2); p_0 and $1-p_0$ are the proportions that correspond to uniform distribution.

The combined score function may be calculated as a sum of the above scoring functions (Hamming measure) or as an Euclidean measure – root mean square of the above score functions.

The algorithm contains several optimizations which are presented in the next paragraphs.

One of the optimizations that we've done is to create a position dictionary. The main role of the dictionary is to optimize the computation of the number of candidate pattern instances from a specific position interval of the sequence. The dictionary is formed from all unique character arrays, with the length l, identified in input sequences. One of the particularities of the dictionary is the fact that the patterns which overlap are excluded: e.g. if the array 'ACACACAC' is found in each input sequence and we are interested to find just the patterns of 4 nucleotides length, then into the dictionary we'll have just two instances for the pattern 'ACAC' instead of three. Another particularity of the proposed dictionary is the fact that the patterns which are having a Hamming distance d or lower than d are interconnected. This interconnection allows a fast enumeration of all instances for each pattern of a specific length.

Another optimization that we've used in the algorithm is to accelerate the calculation of the score functions that we've used. Score functions calculations for each candidate pattern in all positions intervals $(p_1, p_{2,})$, where $0 \leq p_1 \leq p_2 \leq L$, will be ideal. In the current algorithm implementation we've taken into consideration the intervals $(p_1, p_{2,})$; $p_1 < p_2$, $p_1, p_2 \in \{0, i, 2i, 3i, \ldots, L\}$, where i represents the step size of the search. The score functions are being determined individually for each position interval. The score for a long interval can be directly determined from the scores of the shorter intervals from which it is formed. The necessary computations are made in two steps: the score functions for all intervals of size i are being computed in the first step, then in the second step, the scores for longer intervals are computed from the scores of the constituent intervals obtained in first step. The most time consuming is the first step; in the second step the time and complexity are significantly reduced due to the fact that it is just a direct computation from the results obtained previously. This is why the proposed computational method is efficient also in case of long sequences.

The filtering of the similar patterns is another optimization that accompanies the proposed algorithm. As the scores of the candidate patterns are being determined for different intervals, the algorithm is maintaining a list with the scores in descending order. The similar patterns which are having a relative low score, and the ones which have position intervals which overlap, are removed from the list of possible motifs. In this way we maintain only the n motifs where n is user defined and represents a percentage from the total number of candidate motifs. These filters are leading to an important reduce of memory requirements for ADMSL algorithm. The similarity between two patterns of length l is evaluated by using the Needleman-Wunsch algorithm for global alignment. The similarity score is evaluated based on length l.

At each run, the ADMSL algorithm finds motifs for specific values of l and d. To combine the results at each run of the algorithm, for different (l, d) values, a post processing algorithm is needed. Since the score functions used in ADMSL algorithm don't depend on l or d, the motifs with different values for length l and substitutions number d, can be compared directly based on their scores. The motifs which are having a similar pattern can be determined using Needleman-Wunsch alignment algorithm. In this way, if we build motif groups with a similarity greater than 65% (relatively measured for the shortest motif), the motif with the lowest score is being removed. If two motifs have a high similarity (greater than 90%) and localization intervals are overlapping, these are combined into a single motif which has as localization interval the union of the two intervals.

4 Experiments and Results

The first test that we've done with the scope to get an overview of the ADMSL algorithm was by generating with [13] a dataset which contains 50 DNA sequences, each of them of 3000 nucleotides length. Randomly, we've inserted the motif GCATG (5,1) in 75% of the sequences at different positions. The obtained sequences were analyzed using ADMSL configured to search for motifs of length $l = 5$ and a maximum of $d = 1$ substitutions. The motif instances have been determined by ADMSL algorithm as localized in [900, 1500] interval.

From the analysis of other researchers [7] [14], the motif (5,1) is a subtle motif and is almost impossible to detect through a sequence of 3000 nucleotides because there actually are like a few thousands possible random motifs. The first ten motifs detected by ADMSL (together with the afferent scores) are presented in Table 2.

Table 2. The first 10 motifs reported by ADMSL algorithm when running over a dataset of 50 DNA sequences of 3000 nucleotides each

Pattern	Interval	SER	SSR	SIS	Score
GCATG	[900, 950]	0.469	0.345	0.432	1.246
CGCGA	[400, 450]	0.471	0.325	0.423	1.219
GTCGA	[900, 950]	0.424	0.342	0.359	1.125
ATCGT	[1200, 1250]	0.425	0.297	0.398	1.12
CTTCG	[2100, 2150]	0.378	0.432	0.295	1.105
TACGC	[2850, 2900]	0.421	0.305	0.292	1.018
CCGAT	[2650, 2700]	0.397	0.297	0.291	0.985
TACCG	[1800, 1850]	0.345	0.348	0.287	0.98
CGTCG	[900, 950]	0.451	0.276	0.251	0.978
CGATC	[950, 1000]	0.411	0.324	0.237	0.972

The pattern (5,1) was correctly identified as the most prominent motif and the localization interval was detected with accuracy.

This first test was performed to get an overview of the ADMSL performance before getting to more representative tests.

In the next paragraphs we'll present the ADMSL performance in case of short sequences, long sequences and real sequences.

4.1 Short DNA sequences

The tests performed on short DNA sequences have the role to evaluate the detection accuracy of ADMSL algorithm and to emphasize the robustness of the algorithm. Each set of sequences was having N sequences of nucleotides, each of them with a length $L < 1000$, randomly generated using [13]. All of the sequences were artificially implanted with a motif M which has the characteristics $l = 6$ and $d = 1$ along of a randomly position interval (p_1, p_2). We have generated 10 datasets by varying the number of sequences N and the length of the sequence L. The parameters and their values are presented in Table 3.

Table 3. The value of the parameters used in performance analysis over short DNA sequences

Parameter	N	L	l	d
Value	10..50	200-1000	6	1

In Figure 2 it is presented the detection accuracy of the ADMSL algorithm in case of short DNA sequences (randomly generated) implanted with motif $M =' CGATGC'$.

The ADMSL algorithm was configured to report the first 50 possible motifs for each DNA sequence. From the reported motifs, we've chosen the motif most closely of the implanted motif M and we had retained – based on the score – the position occupied in the list of reported motifs.

The motifs reported in this case are presented in Table 4.

From Figure 2 we can observe that the detection accuracy is decreasing while the length of the sequence is increasing but the average detection accuracy value was around 83.6%. So, we can observe that the detection accuracy of the ADMSL algorithm is relatively high. This is because the ADMSL algorithm is not dependent upon the length of

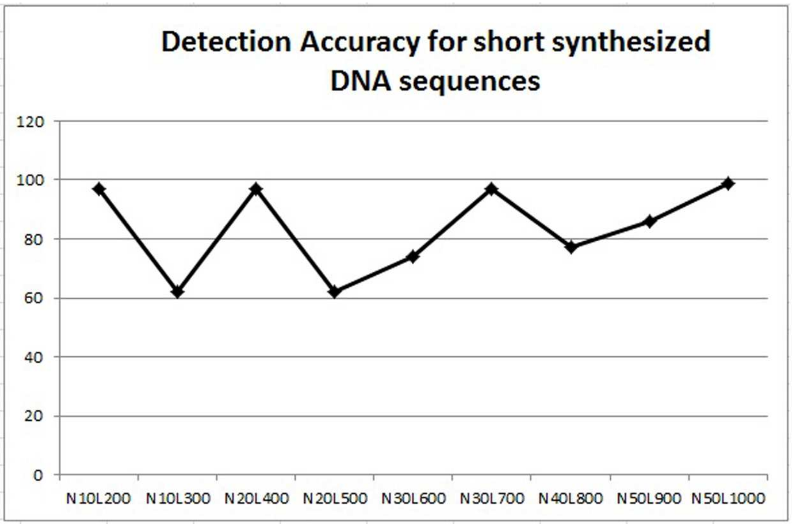

Figure 2. ADMSL detection accuracy in case of short DNA sequences randomly generated

Table 4. Detected motifs by ADMSL algorithm in case of short DNA sequences. Detection accuracy for the considered dataset

DNA Sequence	Motif	Position (in the list of 50 motifs reported)	Detection accuracy
N10L200	GCATGC	3	97%
N10L300	GCATGC	39	62%
N20L400	TCATGC	4	97%
N20L500	ATGCTT	39	62%
N30L600	CATGCG	27	74%
N30L700	GCATGC	5	97%
N40L800	GTGCTA	24	77%
N50L800	CATGTA	16	85%
N50L900	CCATGC	15	86%
N50L1000	ATGCGT	2	99%

the sequence but rather of the motif localization interval. The localized search reduces the number of concurrent random patterns and increases the possibility of comparing motifs.

4.2 Long DNA sequences

The analysis of detection accuracy in case of ADMSL algorithm for long sequences was performed using data sequences as it follows:

- we've generated using [13] ten data sets of 30 random sequences by varying the length of the sequences from 1000 to 6000 of base pairs;

- in each data set we've randomly inserted, in the interval position [200-800], the motif *CATGCT.*

The ADMSL algorithm was executed directly on the sequences previously obtained, with a maximum length of the interval set to 500 nucleotides. We must specify that the fragmentation of the analyzed sequences was not needed (even if their length hit almost 180000 nucleotides), because the ADMSL algorithm automatically determines the localization interval of the motif.

In case of other motif detection algorithms (like MEME, Weeder) there is necessary a fragmentation of the long sequences and to maintain the accuracy, these fragments need to have an overlapping rate of about 50%.

Each run of the ADMSL algorithm was performed using the parameters specified in Table 5.

Table 5. The value of the parameters used in performance analysis over long DNA sequences

Parameter	N	L	l	d
Value	30	1000-6000	6	1

The detection accuracy of the randomly implanted motif, in case of the long sequences is presented in Figure 3. The detection accuracy was evaluated as the detection sensitivity based on the combined score function of the reported motif.

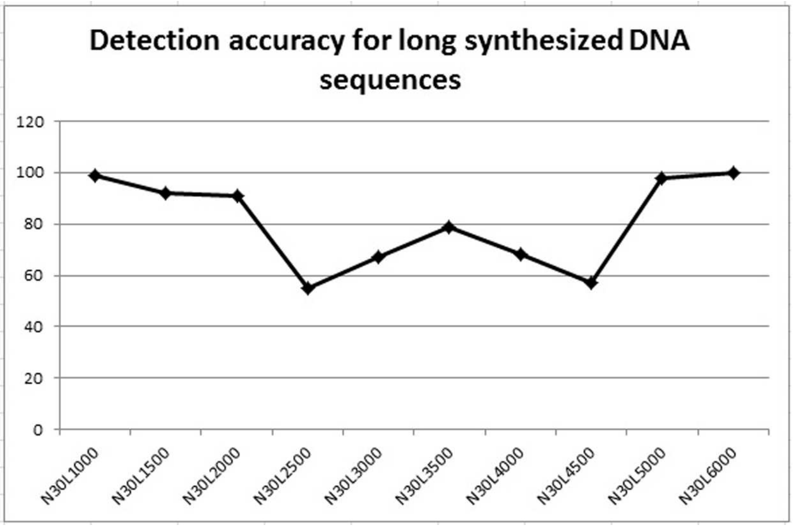

Figure 3. ADMSL detection accuracy in case of long DNA sequences randomly generated

As we can observe, the detection accuracy is maintained over 60% also in case of the long sequences. It is interesting to observe the fluctuation of the accuracy detection once the length of the analyzed sequences grows – we can notice that the accuracy value is increasing and decreasing randomly for the sequences that have a length between 200 and 4500 nucleotides. These fluctuations appeared due to the fact that we had randomly inserted the implanted motif and also because the analyzed sequences were randomly generated.

The motifs detected by ADMSL algorithm, as being the most closest to the implanted motif, are presented in Table 6.

In Table 6, we can observe that motifs similar to the implanted motif were detected and reported in localization intervals where the in-

Table 6. Detected motifs by ADMSL algorithm in case of short DNA sequences. Localization interval

DNA Sequence	Motif	Position (in the list of 50 motifs reported)	Localization interval
N30L1000	CATGCG	2	[300,550]
N30L1500	GCATGC	9	[250,300]
N30L2000	CATGCG	10	[250,600]
N30L2500	CATGCT	48	[450,500]
N30L3000	ATGCTC	34	[1050,1100]
N30L3500	ATGCTG	22	[2700,2750]
N30L4000	GCATGC	33	[1550,1600]
N30L4500	ACATGC	44	[300,350]
N30L5000	GCATGC	3	[400,450]
N30L6000	CATGCA	1	[250,350]

sertion of the random motif did not occurred – the motif was implanted only in the positions interval [200, 800]. Those reported motifs have been detected as valid motifs because they were present in the initial sequences, randomly generated.

4.3 Real DNA sequences

Motifs detection in long regulatory sequences it is an actual requirement especially in ChIP experiments for determining the promoters for vertebrates [5]. Some recent studies [7] [14] are highlighting that random patterns from DNA sequences may become remarkable as if the real motifs. In this specific case, the algorithms used for motif detection are returning false positives hiding the real motif. For most of the algorithms, the necessary resources – memory requirements and execution time – are proportionally increasing with the size of the analyzed sequence.

In the literature, it is known the fact that the motif instances are found to be localized at specific positions, relatively to a reference position (anchor point) [5]. Most of the transcriptions factors are being localized relatively to a transcription start site to allow the transcription factors to be localized in specific positions. In these conditions, the motif detection can be done by searching into a specific interval after the alignment of the sequences relatively to the anchor point.

The localization of motifs has an important advantage by removing the regions which are not containing motifs and by decreasing the possibility of returning false positives.

One possibility is to divide the DNA sequences into short overlapping subsequences of the same size. Some problems may occur:

- in most of the cases we don't have prior information regarding the regions where the motifs are distributed;
- in case of a big number of reported motifs in a range of intervals it is really a challenge to identify and extract those motifs which have the greatest importance for the analyzed sequence;
- depending on the chosen length for the sequences the motifs might not be so obvious if the length is short and might be poorly demarcated if the length is too big;
- the division of the analyzed sequence in subsequences must be done automatically otherwise it will require time and it will be more susceptible to errors.

In the performance evaluation for real data sequences, to not disadvantage any of the algorithms, we've chosen to not split the sequence into subsequences. The analysis was performed on the entire sequence in order to make judgments regarding performance directly over long DNA sequences as we found them in genomic repositories.

A big challenge in this research was to choose the right datasets with the scope to not favor or disfavor any of the algorithms that we've used in comparison with ADMSL. Tompa [12] presents a few solutions for DNA datasets selection but each of them have several drawbacks. In order to pass these drawbacks we've used transcription factors reported as real in TRANSFAC repository. From the biological database previously mentioned, we had chosen only the transcription factors which were having also a consensus sequence defined.

We've executed tests on different sequences corresponding to the following species: *Saccaromyches Cerevisiae, Drosophila Melanogaster* and *Homo Sapiens.* All the algorithms used in this assessment (ADMSL, MEME, AlignAce, YMF, Improbizer, Weeder and SCOPE) have been configured to detect motifs that have a length in the range of six to ten nucleotides. In this performance evaluation we took into account the first ten motifs detected by each of the analyzed algorithm. In order to obtain an overview of each algorithm we've run the applications/algorithms over each dataset. Besides the proposed algorithm – ADMSL – all others have been used without modifying the source code of the applications and the evaluations were performed over their official web sites or by running the application locally.

In the next figures we present the detection accuracy of the considered algorithms.

If we consider the *Drosophila Melanogaster* dataset (Figure 4), a big majority of the reported motifs had a length between 3bp (Improbizer) and 10bp (MEME, YMF).

Most of the motifs reported had a corresponding real transcription factor in TRANSFAC database (the motifs reported by Improbizer were not found in TRANSFAC database – that's why the accuracy is set to 0). From the performance point of view we can confirm that ADMSL had reported the most motifs for which we had found a corresponding transcription factor in TRANSFAC database. Also, we've noticed that YMF and SCOPE had good performances.

For *Homo Sapiens* dataset (Figure 5) we've used sequences with more than 36000 nucleotides. We've observed that the algorithm had reported motifs that we had found as transcription factors in biological database (TRANSFAC). Over 90% of the motifs reported by ADMSL were identified as real transcription factors in TRANSFAC genome repository. Once the length of the analyzed sequence had increased, the number of false positives had also increased.

Also in case of *Saccaromyches Cerevisiae* dataset (Figure 6) the

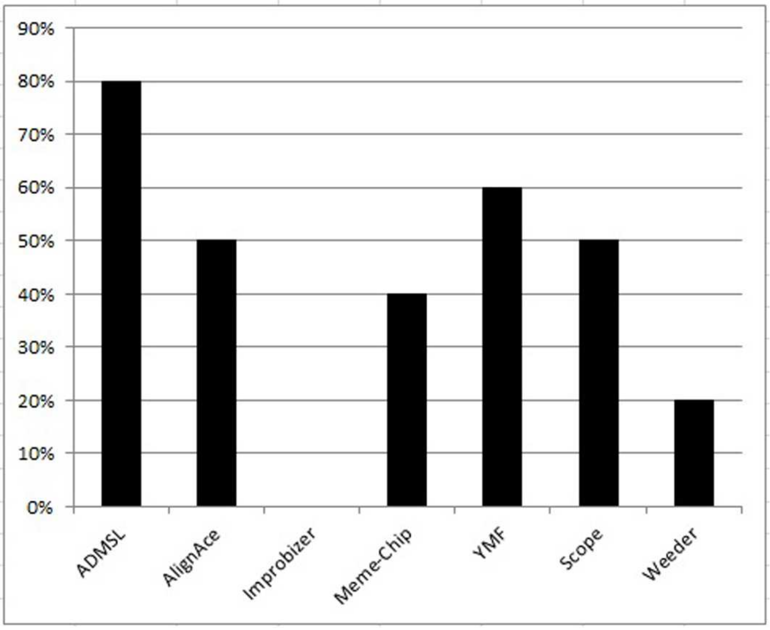

Figure 4. Detection accuracy for Drosophila Melanogaster dataset

ADMSL algorithm had proved to be more accurate than the other analyzed algorithms. MEME and YMF had accuracy close to ADMSL algorithm.

5 Conclusions

The main purpose of this research was to design and develop a new algorithm for detecting DNA motifs especially in long sequences where the performance of existing applications is relatively poor. The algorithm proposes an innovative way for detection and localization of DNA motifs by combining multiple score functions to evaluate the existence of a motif.

ADMSL had been optimized to fast process long DNA sequences. The results obtained on synthetic or real data confirmed us that

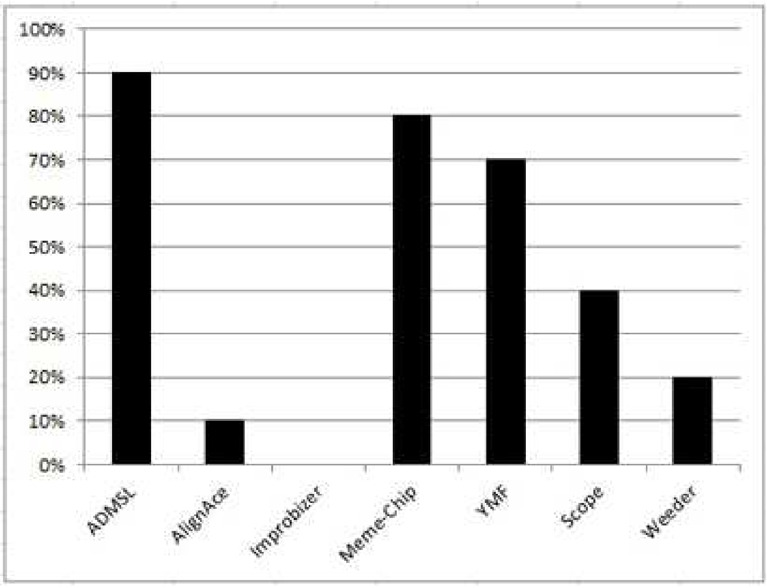

Figure 5. Detection accuracy in case of Homo Sapiens dataset

ADMSL has a definite advantage beside other algorithms due to the detection accuracy of the motifs in long DNA sequences.

In the recent years, considerable efforts were made in elaborating computational methods and more and more species have a complete DNA sequence. Nevertheless, the identification of the elements that are part of the cis-regulatory process continues to be an important challenge for scientists.

At the beginnings, the algorithms focused on motifs searching were combining the phylogenetic data with co-regulated genes in order to find regulatory motifs. In the present, most of the algorithms are oriented to computational methods and researchers are designing new approaches to better identify the motifs from the analyzed DNA sequences.

Due to the big number of algorithms and multitude of the methods designed for motif identification, for a user, it will be helpful a set of

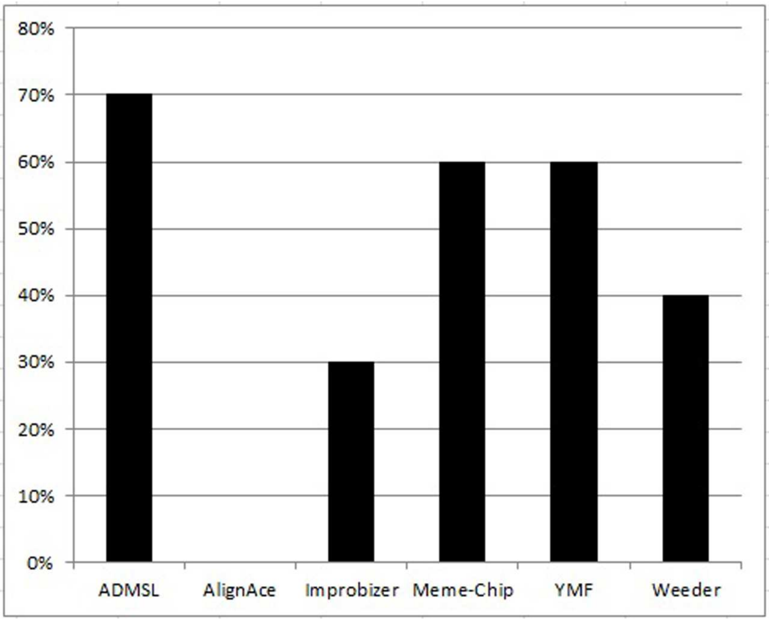

Figure 6. Detection accuracy in case of Saccaromyches Cerevisiae dataset

instructions for choosing the best algorithm/method before starting the analysis of the DNA sequence.

One of the drawbacks when providing instructions for deciding to a method or another is the number of settings and parameters which need to be chosen for each algorithm. The main advantage of ADMSL algorithm is the fact that the DNA sequences, even if they have a considerable length, don't need to be divided in order to obtain motif localization information. Another plus of ADMSL algorithm is the fact that needs just a few parameters (e.g. length of the search motif, allowed substitutions, size of interval search) which have also default values set in the application that runs ADMSL algorithm. In this way a user can obtain a first set of results with a minimum effort.

Performance evaluation of a motif search algorithm by comparing with other algorithms is especially problematic. This is because we don't have yet a complete understanding of the process that regulates gene activity and expression. Also, there is no standardized model against to evaluate the efficiency of an algorithm. In the tests done in this research we must consider the fact that in case of the other algorithms used in comparison with ADMSL, the parameters were set with values to reflect as much as accurate the configurations done for ADMSL algorithm. Because this was done through human interaction – it is susceptible to errors.

Most of the algorithms used in comparison to evaluate the performance of ADMSL algorithm, have good results in case of lower organisms, especially when they are set to report short motifs (of 6-8 nucleotides). The ADMSL algorithm, through the computed score functions, highlights the motifs conservation through different species. The performances of ADMSL algorithm have proved to be much better especially for long DNA sequences like the ones that we've analyzed from human genome.

References

[1] Martin Tompa Mathieu Blanchette. *Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting*, Genome Res, pp. 739–748, 2002.

[2] Patrik D'haeseleer. *What are DNA motifs?*, Nature Biotechnology, vol. 24, pp. 423–425, 2006.

[3] Hubbard TJ. Down TA, *NestedMICA: sensitive inference of overrepresented motifs in nucleic acid sequence*, Nucleic Acids Res., pp. 1445–53., 2005.

[4] Ho-Kwok Dai Modan K Das. *A survey of DNA motif finding algorithms*, BMC Bioinformatics, 2007.

[5] Ankush Mital, Wing-Kin Sung Vipin Narang. *Localized Motif discovery in gene regulatory sequences*, Bioinformatics, vol. 26, no. 9, 2010.

[6] F.P. Roth, J. D. Hughes, P. W. Estep, G.M. Church. *Finding DNA Regulatory Motifs within Unaligned Non-Coding Sequences Clustered by Whole-Genome mRNA Quantitation*, Nature Biotechnology, pp. 939–945, 1998.

[7] Martin Tompa Jeremy Buhler. *Finding Motifs using Random Projections*, Journal of Computational Biology, vol. 9, no. 2, pp. 225–242, 2002.

[8] S.T. Smale, J.T. Kadonaga. *The RNA polymerase II core promoter*, Annu Rev Biochem, vol. 72, pp. 449–479, 2003.

[9] P.A. Pevzner, M.Yu. Borodovsky, A.A. Mironov. *Linguistics of nucleotide sequences*, J Biomol Struct Dyn, vol. 6, pp. 1013–1026, 1989.

[10] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, Y. Moreau. *A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling*, Bioinformatics, pp. 1113–1122, 2001.

[11] Stormo G.D. *DNA binding sites: representation and discovery*, Bioinformatics, vol. 16, no. 23, 2000.

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

[13] http://www.bioinformatics.org/sms2/random_dna.html

[14] U. Keich, P.A. Pevzner. *Finding motifs in the twilight zone*, Bioinfomratics, vol. 18(10), no. 1382–1390.

[15] Elkan C Bailey TL. *Unsupervised learning of multiple motifs in biopolymers using expectation*, Machine Learning, pp. 51–80, 1995.

[16] W.Y. Ao, J. Gaudet, W.J. Kent, S. Muttumu, S.E. Mango. *Environmentally Induced foregut remodelling by PHA-4/FoxA and DAF-12/NHR*, Science 305, pp. 1743–1746, 2004.

[17] G. Pavessi, P. Mereghetti, G. Mauri, G. Pesole. *Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes*, Nucleic Acid Res. 32, pp. 199-203, 2003.

[18] Jonathan M. Carlson, Arijit Chakravarty, Charles E. DeZiel, Robert H. Gross. *Scope: a web server for practical de novo motif discovery*, Nucl. Acids Res, pp. 259–64, Jul 2007.

Alin G. Voina, Petre G. Pop, Mircea F. Vaida

Received May 19, 2014

Technical University of Cluj-Napoca,
Communications Department,
26-28 George Baritiu St., Room 364
400027 Cluj-Napoca, Romania,
Tel: +40-264-401226, Fax: +40-264-597083
E–mails: *alin.voina@com.utcluj.ro*
petre.pop@com.utcluj.ro
mircea.vaida@com.utcluj.ro