A COMPARATIVE STUDY OF MOTIF DETECTION TOOLS

Alin Voina, Petre G. Pop, Ligia Chiorean

Technical University of Cluj-Napoca, Comm.Dept., G. Baritiu str., 26-28, Cluj-Napoca, 400027, Romania,
alin.voina@com.utcluj.ro; petre.pop@com.utcluj.ro ; chiorean.ligia@com.utcluj.ro ;

ABSTRACT

A major challenge in biology is to understand the mechanism that regulate gene expression. An important step in this challenge is to find regulatory elements like binding sites in DNA for transcription factors. These binding sites are formed from short DNA segments that are called motifs. In the last decades, the advances in genome sequence have allowed the development of new computational methods for motif finding. In this study we analyzed some of the most popular computational methods and we present a performance analysis of the existing approaches.

Keywords: *motifs finding algorithms, gene expression, regulatory elements, binding site, computational methods*

1. INTRODUCTION

The fundamental unit of inherited information in deoxyribonucleic acid (DNA) is called *gene* and is defined as a section of base sequences that is used as a template for the copying process called transcription. Gene expression is referring to the fact that every gene contains the information to produce a protein. Gene expression begins with transcription factors, to enhancer and promoter sequences. The process of regulating gene expression is done with the help of transcription factors by activating or inhibiting the process of transcription. A major challenge in biology is to completely understand the mechanism that regulate gene expression. The identification of the regulatory elements especially the binding sites in DNA for transcription factors it's a major task in this challenge [1]. The prediction of such regulatory elements is an issue where computational methods are expected to offer a great solution and as a consequence, biologists have invested considerable effort into solving this problem. In its simplest form, the problem of detection of regulatory elements can be formulated as follows: give a set of N sequences find a pattern M of length l that occurs frequently. If the pattern M of length l appears in every sequence from the set of N sequences, then a simple enumeration of the l letters of the pattern M gives the regulatory element. The main issue when we are dealing with DNA sequences it is that the patterns may have mutations, insertions or deletions of nucleotides.

In the analysis of gene regulation, sequence motifs are becoming increasingly important. A DNA motif is defined as a short, recurring pattern in DNA that is presumed to have some biological significance such as being DNA binding sites for a regulatory protein, i.e., transcription factor (TF). Some of the motifs are involved in important processes at the RNA level, including ribosome binding, mRNA processing and transcription termination [1]. Normally, the pattern is relatively short (5 to 20 base-pairs (bp) long) and is known to recur in different genes or several times within a gene.

DNA motifs are often associated with structural motifs found in proteins and can occur on both strands of DNA, while transcription factors bind directly on the double-stranded DNA. A DNA sequence may have zero, one or multiple copies of a single motif. Besides the common forms of DNA motifs there are two special types of DNA motifs that are recognized: palindromic motifs and space dyad (gapped) motifs. A palindromic motif is a subsequence that is exactly the same as its own reverse complement, e.g., TCTCGCGAGA. A space dyad motif is formed of two smaller conserved sites separated by a spacer (gap). Because the transcription factor binds as a dimer, the spacer is localized in the middle of the motif. As a consequence, the transcription factor is formed from two subunits that have to separate contact points with the DNA sequence. The positions where the transcription factor binds to the DNA are conserved and are typically of short length (3-5 bp).

In the past, binding sites were typically determined through DNase footprinting and gel-shift or reporter construct assays [1]. Nowadays computational methods are used to generate the regulatory elements, by searching for overrepresented DNA patterns upstream of functionally related genes. Therefore, a sensible approach to detect regulatory elements is to search for overrepresented motifs. A statistically overrepresented motif is a motif that occurs more often than one would expect by chance. Most of the motif finding algorithms have great results for yeast and other lower organisms, but their performance is poor in higher organisms. To bypass this difficulty recent motif finding algorithms are taking advantage of phylogenetic footprinting. The main idea in phylogenetic footprinting is that selective pressure causes that functional elements evolve at a slower rate than nonfunctional sequences. This means that well conserved sites among a set of promoter regions are very good candidates for functional regulatory elements or motifs. Most recently, algorithms that integrate phylogenetic footprinting have significantly improved motif detection from genomic sequences [2]. Also, significant efforts have focused on developing algorithms that incorporate parameters that are used for motif detection in higher organisms [3].

In the last years, a large number of algorithms for finding DNA motifs have been developed. In this study we focus on those computational tools designed for the discovery of novel regulatory elements where nothing is known a priori about transcription factors. With a considerable number of tools for detection of DNA motifs, biologists have been offered little guidance in the choice among these tools. The main purpose of this assessment it's to provide some guidance regarding the accuracy of available motif detection tools for various set of sequences.

2. MOTIF DETECTION ALGORITHMS

Motif detection algorithms can be classified into three major classes (depending on the type of the DNA sequence information involved in detection algorithm):

(1) algorithms that are using promoter sequences from coregulated genes of a single genome;
(2) algorithms that are using phylogenetic footprinting;
(3) a combination of (1) and (2).

In earlier literature, motif detection algorithms were classified in two major groups:

- word- based (string-based) methods based on exhaustive enumeration, i.e., counting and comparing oligonucleotide frequencies;
- probabilistic models based on maximum likelihood principle, Bayesian inference.

The first group (word-based methods) are more suitable for finding short motifs such as those encountered in eukaryotic genomes. String-based enumerative methods can be very fast when are implemented with optimized data structures like building a suffix tree of the sequences and are very useful when searching for identical instances of a motif. For typical transcription factor motifs that usually have several constrained position, word-based methods are not so efficient and in most of the cases additional post-processing needs to be done. One of the first algorithms based on the word-based approach, was Oligo-Analysis, developed by van Helden *et al.* [4]. Oligo-Analysis has proven efficient especially in the case of the yeast (*Saccharomyces cerevisiae*). The results obtained with the above algorithm had been previously found by biologists through laboratory experimental analysis. The initial algorithm proposed by van Helden *et al.* was limited to short motifs characterized by relatively simple patterns. One of the main drawbacks of the algorithm was that within an oligonucleotide, no variations were allowed. This problem was addressed later by Tompa [5], when he proposed an exact word-based method for detecting short motifs. The novelty of the algorithm consists in that it creates a table that for each *k*-mer (sequence of length k) *s*, records the number of sequences N_s –containing an occurrence of s, where an occurrence allows for *c* substitution residues in *s*. Sinha and Tompa [6] have used a similar approach in the development of Yeast Motif Finder (YMF) algorithm. YMF uses *a consensus model* of motifs . A motif for this algorithm is a string over the alphabet {A, C, G, T, R, Y, S, W, N} with consecutive N's only at the center, and a limited number of R (A or G), Y (C or T), S (C or G), and W (A or T) characters. YMF searches the entire space of motifs and reports the motifs sorted by their z-scores. In the comparison of Sinha et Tompa, with other two algorithms MEME [7] and AlignACE [8], YMF proved to be more accurate on the yeast data sets and concluded also that each of the algorithms has an exclusivity in the accuracy depending on the datasets.

In the case of probabilistic approaches, motif model is usually represented with the help of position weight matrix (PWM). The elements of positional weight matrix (PWM) correspond to scores which are reflecting the likelihood of a particular nucleotide at a particular position. The main advantage of probabilistic approaches is that they require a few search parameters and they rely on probabilistic models which can be very sensitive even at small changes of the input data. Among the first algorithm implementations that are based on probabilistic approaches was a greedy probabilistic sequence model developed by Hertz *et al* [9]. The algorithm was used to identify a common motif one in each analyzed sequence and has been later improved by providing a method to estimate the statistical significance of given information content score based on large deviations statistics.

Another motif finding algorithm –NestedMICA- developed by Down and Hubbard [10], uses a sequence model based on an independent component analysis framework to learn models for multiple motifs simultaneously. Authors reported that NestedMICA was more sensitive than MEME [7] and in one case it successfully extracted a target motif from background sequence four times longer than could be handled by MEME.

Bailey and Elkan [7] developed MEME algorithm by extending the expectation maximization (EM) approach. Three novel ideas were incorporated in MEME:

- subsequences that actually occur in the biopolymer sequences are used as starting points for EM algorithm;
- the assumption that each sequence contains only one instance on the shared motif, is removed;
- is incorporated a method for probabilistically erasing shared motif after they are found.

Most of the algorithms based on probabilistic approaches are more appropriate for finding motifs in prokaryotes, where the motifs are generally longer than eukaryotes. Also, most of these algorithms make use some form of local search, like expectation maximization (EM), Gibbs sampling, greedy algorithms which are leading to a locally optimal solution and not a global one.

In our study we have compared five of the most used tools: AlignACE, MEME, Improbizer [11], Weeder [12] and YMF.

3. EXPERIMENTS AND RESULTS

A major challenge in this assessment was to find good data sets so as not to disadvantage any of the analyzed tools. There were several possibilities for choosing datasets:

- we could use real sequences containing real annotated TFB (Transcription Factor Binding Sites) but there could be unannotated binding sites and the applications that predict these would be unfairly penalized;
- we could use synthetic DNA sequences and to artificially implant at random positions, instances of a

known binding site; the drawback of this approach is that in this way we can favor some tools over others due to stochastic process that nature uses.

To avoid the drawbacks described above, we used TRANSFAC database to choose real transcription factors (TF). Each transcription factor gives rise to a set of data sequences for which we know the binding sites, their position and orientation. From TRANSFAC database we selected only the TF for which the database is listing also a consensus sequence. Also, for each TF, we removed duplicated binding sites instances, those that were not having position information and also we removed binding sites that contradicts each other. We've obtained datasets which contains transcription factor binding sites from human, mouse, yeast and fly genes.

In the figures below we represented the position and the length of the motifs reported by each of the tools (AlignACE, MEME, Improbizer, Weeder and YMF) in case of a human, mouse, yeast, and fly dataset.

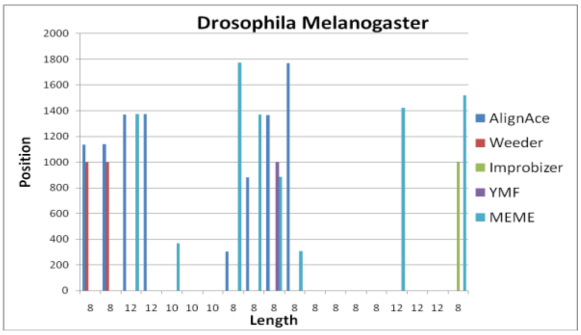

Figure 1: *Position and length of the detected motifs for Drosophila Melanogaster dataset*

Detected Motif	Length	Interval
TAAGACGC	8	[1135…1141]
GTTGGCAA	8	[306..310], [1770…1773]
CTGGTTTTCCC	10	[365…367]
CGGAGACCCGAG	12	[1371…1373]
AAATGTTTATTT	12	[1418…1421]

Table 1: *Detected motifs for Drosophila Melanogaster dataset*

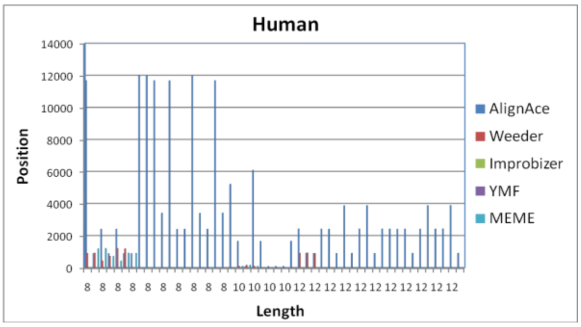

Figure 2: *Position and length of the detected motifs for Human body dataset*

Detected Motif	Length	Interval
TGGACCCA	8	[449…740], [910…941], [2430…3438], [11705…12028]
TGTGAGTCTC	10	[103…123], [1672…1683]
TAGTGCCTGACA	12	[913…952], [2436…2450]

Table 2: *Detected motifs for Human body dataset*

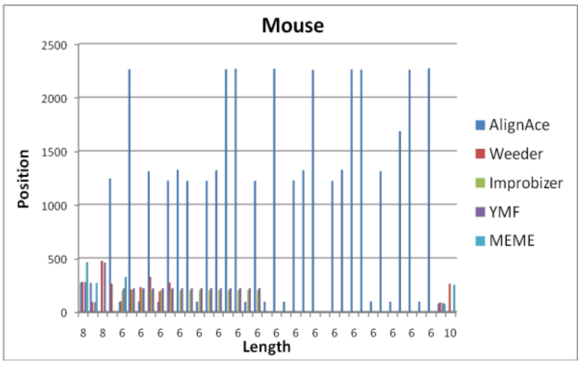

Figure 3: *Position and length of the detected motifs for Mouse dataset*

Detected Motif	Length	Interval
GGACCA	6	[93…273], [1222…1325], [2260…2272]
CACCATCC	8	[260…280]
GCGCCTAA	8	[75…87]
GGCTAGCGAT	10	[261…1245]
GCTAGCGATG	10	[251..262]

Table 3: *Detected motifs for Mouse dataset*

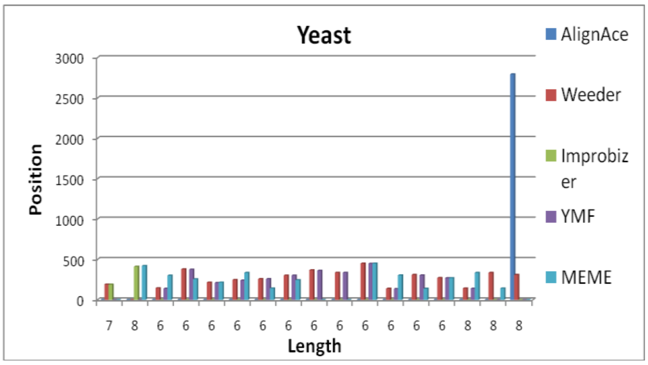

Figure 4: *Position and length of the detected motifs for Yeast dataset*

Detected Motif	Length	Interval
GAATAAT	7	[188…190]
TGACTC	6	[140…144], [336…379]
CTGCGC	6	[135…138], [304..309], [448...450]
ACTTTCTA	8	[410…420]
ATTGACTC	8	[140…334]

Table 4: *Detected motifs for Yeast dataset*

The tables above (Tables 1-4) contains the common motifs detected by the analyzed tools, accompanied by the length and position intervals for each reported motif.
As we can see from the figures above (Figure 1-4) each tested tool performs different on each data set. AlignAce and MEME were the only ones which reported motifs on all datasets. Also, we observed that MEME and AlignAce were very similar in terms of length and position of the reported motifs. Improbizer, Weeder and YMF perform better in the case of low organisms (especially in the case of yeast).

Also, from Tables 1-4 we can observe that the position intervals are tighter in case of short motifs length which leads us to the conclusion that the analyzed tools have better accuracy for short motifs.

4. CONCLUSIONS

Despite the considerable effort in computational biology and genome sequencing, prediction of regulatory elements remains a wonderful and complex challenge for biologists.

The first motif search algorithms were based on co-regulated genes and the search was focus on overrepresented motifs. Recent algorithms use also the overrepresentation of motifs and their conservation along DNA sequences. Because of the existence of a large number of tools/applications specialized in motif detection, to the user would be more useful to have some guidelines in choosing the tool/application that fits their needs. However, performance evaluation of algorithms is relatively difficult. This is primarily due to the fact that the underlying biology of regulatory mechanisms is very incompletely understood and we don't have an absolute standard against which to measure the correctness of tools. In addition, in evaluating the performance of the tools we should take into consideration that each predicted set of motif instances was done by human choices of parameters and this can also lead to a loss of accuracy.

Most of the algorithms have better results for low organisms (including yeast). Recent algorithms which integrate overrepresentation of motifs and their conservation among species proved to perform better in high organisms, including here also the human body. Also, in our assessment we observed that there are tools that perform better on some specific datasets and we can use multiple tools for the same dataset as complementary source of information. Therefore, is better to use in combination more tools rather than rely on a single one and to take into consideration the top few predicted motifs rather than the most significant one.

ACKNOWLEDGEMENTS

This paper was supported by the project "Doctoral studies in engineering sciences for developing the knowledge based society-SIDOC" contract no. POSDRU/88/ 1.5/S/60078, project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

REFERENCES

[1] Patrik D'haeseleer: "What are DNA motifs?", *Nature Biotechnology* 24, 423 - 425 (2006)

[2] Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M: "Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis". *Genome Re*s 2001, 11:1175-1186.

[3] Hon LS, Jain AN: "A deterministic motif finding algorithm with application to the human genome". *Bioinformatics* 2006, 22:1047-1054.

[4] van Helden J, Andre B, Collado-Vides J: "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies". *Journal of Molecular Biology* 1998, 281:827-842.

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

[6] Sinha S, Tompa M: "A statistical method for finding transcription factor binding site". *Proceedings of the Eighth International Conference on Intelligent Systems on Molecular Biology,* San Diego, CA 2000:344-354.

[7] Bailey TL, Elkan C: "Unsupervised learning of multiple motifs in biopolymers using expectation maximization". *Machine Learning* 1995, 21:51-80.

[8] Roth FP, Hughes JD, Estep PW, Church GM: "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation". *Nature Biotechnology* 1998, 16:939-945.

[9] Hertz GZ, Hartzell GW, Stormo GD: "Identification of consensus patterns in unaligned DNA sequences known to be functionally related". *Comput Appl Biosci* 1990, 6:81-92.

[10] Down TA, Hubbard TJ., "NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence", *Nucleic Acids Res*. 2005 Mar 10;33(5):1445-53.

[11] Ao, W., Gaudet, J., Kent, W.J., Muttumu, S. & Mango S.E, "Environmentally Induced foregut remodelling by PHA-4/FoxA and DAF-12/NHR", *Science* 305, 1743-1746 (2004)

[12] Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G., "Weeder Web: discovery of transcription factor binding sites by statistical overrepresentation. *Nucleic Acid Res.* 31, 3586-3588 (2003).