33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

MEET: Motif Elements Estimation Toolkit

Erola Pairó*, Joan Maynou *, Montserrat Vallverdú, Pere Caminal,
Santiago Marco and Alexandre Perera

Abstract— **MEET is an R package that integrates a set of algorithms for the detection of transcription factor binding sites (TFBS). The MEET R package includes five motif searching algorithms: MEME/MAST(Multiple Expectation-Maximization for Motif Elicitation), Q-residuals, MDscan (Motif Discovery scan), ITEME (Information Theory Elements for Motif Estimation) and MATCH. In addition MEET allows the user to work with different alignment algorithms: MUSCLE (Multiple Sequence Comparison by Log-Expectation), ClustalW and MEME. The package can work in two modes, training and detection. The training mode allows the user to choose the best parameters of a detector. Once the parameters are chosen, the detection mode allows to analyze a genome looking for binding sites. Both modes can combine the different alignment and detection methods, offering multiple possibilities. Combining the alignments and the detection algorithms makes possible the comparison between detection models at the same level, without having to care about the differences produced during the alignment process. The MEET R package can be downloaded from `http://sisbio.recerca.upc.edu/R/MEET_1.0.tar.gz`**

I. INTRODUCTION

Gene expression is a highly regulated process. It initiates with the information transfer from DNA to mRNA by means of transcription. This process is modulated by the binding of some specific proteins called transcription factors (TF), to the correspondent binding sites (BS) [1]. The transcription factor binding sites (TFBS), also known as cys-regulatory elements, constitute the regulatory regions of a gene [2]. TFBS are short sequences that show a high variability because the same TF has the capacity to bind to different positions and sequences along the genome. This intrinsic variability makes impossible to establish a consensus sequence for binding site detection. Because of this, many computational methods have been developed to detect DNA sequence motifs [3].

Motif discovery algorithms can be classified according to the used model [4]. MEME/MAST (Multiple Expectation-Maximization for Motif Elicitation) [5], ITEME [6] and MATCH [7] are algorithms based on probabilistic models. Given a set of unaligned sequences, MEME uses the maximum probability to determine the number of parameters of the algorithm using the expectation-maximization (EM) model. MAST [8] (Motif Alignment and Search Tool) is based on Q-FAST algorithm and calculates the statistical significance of a sequence to belong to a certain motif. ITEME algorithm is based on information theory. The detection is performed doing an analysis of the variation of information contained in the group of sequences when a new sequence of study is added into this group. This algorithm allows the detection of TFBS considering independence or dependence within the binding sites positions, depending on the use of Rényi entropy [9] or divergence [10]. MDscan is an algorithm based on a deterministic model, enumeration of combined words, and a probabilistic model of Bayesian networks. Within the probabilistic models, MATCH is a tool based on the construction of a position weighting matrix (PWM) to look for TFBS in DNA sequences [7]. The detection is done by means of two measures: the similarity score and the core similarity score, which only takes into account the five more conserved positions of the PWM. The core similarity score allows a preselection of the candidate sequences and then the similarity score provides the quality of the selected sequences. Finally, the last method of detection in MEET R package is Q-residuals [11], based on a numerical model. Specifically, the detector converts each DNA sequence into a numeric sequence using a three dimensional representation where each nucleotide is placed at the vertex of a regular tetrahedron. A principal components analysis (PCA) is then applied to the numerical sequences. The hypothesis used in detection is that Q-residuals of the binding site sequences would be smaller than Q-residuals of genomic sequences.

The MEET R package, includes not only a wide range of computational methods for detection but also different multiple alignment algorithms. The user can use the alignment tools MUSCLE (Multiple Sequence Comparison by Log-Expectation) [12], ClustalW [13] and MEME [5] to obtain the different nucleotides in each position. MUSCLE is based on an iterative algorithm of sequences. The iterative multiple alignment algorithm divides the alignment in two phases. In the first one MUSCLE carries out a pair alignment of the sequences. In the second process, a multiple alignment is performed adding the sequences progressively and realigning the pair of sequences established at the beginning. On the other hand, ClustalW is based on a progressive model. Progressive models work in a similar way than iterative

*These two authors have to be considered first author of the paper

E. Pairó and S. Marco are with Institut of BioEngineering of Catalonia, IBEC balidiri REixach 4-6, 08028 Barcelona,Spain and the Electronics Department in the University of Barcelona (UB) epairo@ibecbarcelona.eu, smarco@el.ub.es

J. Maynou, M. Vallverdú, P. Caminal y A. Perera are at the Dep. ESAII, Centre Recerca en Enginyeria Biomèdica (CREB), Universitat Politècnica de Catalunya (UPC), Barcelona, Gargallo, 5, 08028 Barcelona, Spain(Catalonia). http://www.creb.upc.es, http://www.upc.edu. joan.maynou, montserrat.vallverdu, pere.caminal, alexandre.perera@upc.edu

This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2010-20886-C02-02 and the CIBER-BBN

E.P wants to thank IBEC for supporting her PhD financially.

models but when sequences are added sequentially, the first pair of sequences is not aligned again. Finally, MEET R package allows us to work with the MEME alignment and with sequences previously aligned, provided by the user himself.

II. MEET PACKAGE

The Motif Elements Estimation Toolkit (MEET) R package, integrates different algorithms for motif detection (MEME/MAST , Q-residuals, ITEME, MATCH and MDscan) and different algorithms for the alignment of nucleotide sequences (MUSCLE, ClustalW and MEME). The alignment and detection algorithms can be chosen independently, giving the user a wide range of possibilities as it can be seen in figure 1. The MEET package has two different working modes with different input parameters that are summarized in table I, the training mode and the detection mode.

The training mode has as an input, the alignment (alignment algorithm) and detection (detection method) algorithms, the specific parameters for each algorithm, the sequences used to construct the model (TF in fasta format), the sequence to analyze with a BS in a known position (DNA sequence in fasta format) and this position (position), the organism background and also a vector with the parameters that the user wants to estimate. This training mode allows to compare alignments, detectors and specially to choose the best parameter for a given detector. The comparison can be done directly from the output of the MEET program, the ROC curves and the Area under ROC curve (AUC). Because of the fact that a leave-one-out cross validation (l.o.o) is performed in the training mode, both ROC and AUC for each one of the parameters are a vector corresponding of the results in each step of the l.o.o. The mean and the error for AUC and ROC can be easily extracted from this vector and the user can choose the criteria to optimize the parameter. The output also includes the consensus sequence for the studied motif and a summary of the input parameters. This output can be seen in table II.

The detection mode allows the user to detect TFBS within a large DNA sequence. In order to work in this mode, the user must provide as an input the parameter of the detector, the p-value used as a threshold in detection and all the specific parameters for each detector or alignment algorithm. In this way, the user can work in the training mode and, once the detector and the parameter are chosen, use them to detect binding sites. The output, that is also shown in table II, consists of the detected sequences, the p-value corresponding to each sequence and the position where the sequence is found, as well as the consensus and the input summary.

The main advantage of MEET R package is that this package allows the direct comparison between detectors and between the parameters of a detector, allowing the user the option of optimizing the BS detection in each case. In addition, it integrates different alignment algorithms and allows combinations between the detectors and the alignment methods. In this way, the differences in detection due to the different alignment of the TFBS sequences are avoided and all methods can be compared in the same conditions.

TABLE I
INPUT PARAMETERS OF MEET PACKAGE

Parameters	Training	Detection
TF(.fasta)	X	X
DNA sequence (. fasta)	X	X
Alignment algorithm	X	X
Alignment parameters	X	X
Detection Method	X	X
Organism Background	X	X
P-value threshold		X
Detector parameters		X
Leave-one-out cross validation sequence	X	
TFBS positions	X	
Number of motifs (MEME and MDscan)	X	X
Direction	X	X
Missing values percentage	X	X
Vector of parameters to study	X	X
Call external programs	X	X

TABLE II
OUTPUT PARAMETERS OF MEET PACKAGE

Training	Detection
ROC	Detected sequences
Area under ROC	Initial position
X	P-value
X	Direction
Consensus	Consensus
Summary	Summary

III. IMPLEMENTATION

The MEET R package is an open access tool that can be executed using the open source statistical software R [14] and is available from `http://sisbio.recerca.upc.edu/R/MEET_1.0.tar.gz`. The package integrates detection algorithms (MEME/MAST 4.4.0, Q-residuals, ITEME, MATCH 1.0 Public and MDscan) and alignment algorithms (MUSCLE Version 3.8, ClustalW and MEME 4.4.0). Furthermore, it includes a wide documentation and some examples.

IV. EXAMPLE

To execute MEET R package the user has to download it and install into the computer, together with the needed programs. The working modes are defined using the system parameter. In the next example, the validation of the Q-residuals detector, based on a PCA analysis is performed for a range of principal components. The parameter to optimize is the number of components $npcs = 1 - 10$. The sequences used to construct the model belong to the *ABF1* binding sites from *Saccharomyces cerevisiae* organism, and have been aligned using ClustalW.

```
>Output<-MEET(TF="SqDNA.fa", seqin="DNA4.afa" ,alg=
    "ClustalW" ,method="PCA" ,system="validation",
    org="Saccharomyces cerevesiae", vector=c(1:10),
     sentit="f", position=c(501), mv=50, gapopen
    =-500, maxiters=16, call.clustalw="clustalw")
```

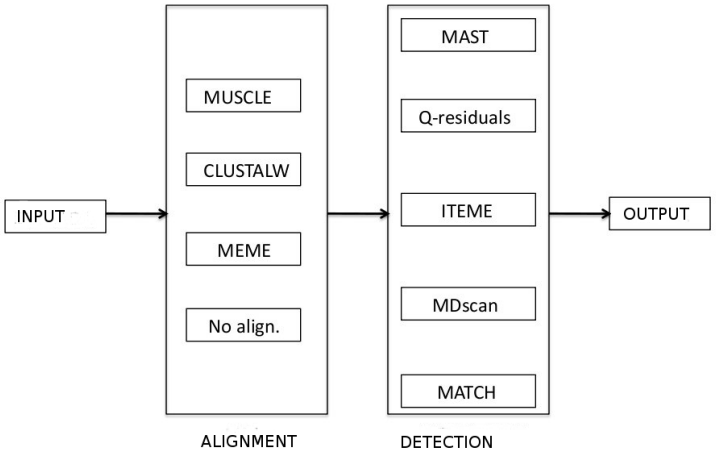

Fig. 1. Architecture of MEET R package. MUSCLE, ClustalW and MEME are the alignment programs , while MAST, Q-residuals, ITEME, MDscan and MATCH correspond to the binding sites detection programs. Each detection algorithm can use a matrix of binding sites, aligned with any of the alignment algorithms.

It is important to note that the input parameters TF (binding sites) and seqin (DNA sequence) have to be in *.fasta* format. The other parameters are used to set the studied organism, the alignment method, the position of the binding sites and the sequence to validate.

The output, in validation mode, contains the consensus sequence, the input parameters (Summary) and the results of validation. As it is said above, both ROC and AUC consists on a vector for each parameter. In figure 2 the AUC and its variance are plotted for the ABF1 example and the number of principal components going from 1 to 10. The mean an the error for a given number of components can be easily computed, for example for 3 components:

```
>names(Output)
```

[1] "Cosensus" "Summary" "Results"

```
>mean(Output$Results$Area[[3]])
```

[1] 0.999

```
>sd(Output$Results$Area[[3]])
```

[1] 0.002

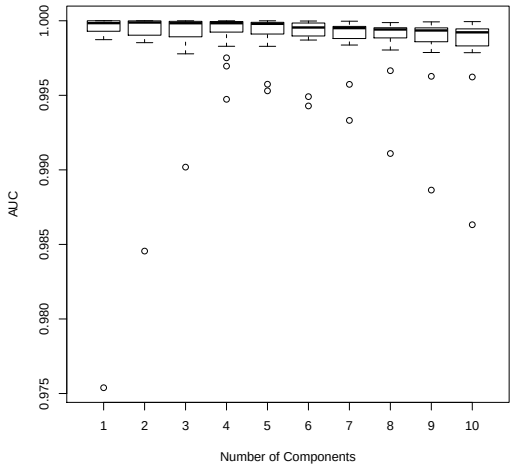

Fig. 2. Boxplot of the AUC for a range of components from 1 to 10 using the Q-residuals detector. The studied TFBS is ABF1 from *Saccharomyces cerevisiae*, and the background sequence is a 1000 nucleotide sequence extracted randomly from *Saccharomyces cerevisiae* chromosome 1.

V. RESULTS

The training mode of MEET R package allows the comparison between binding site detection performed using the different algorithms integrated in the package. As the AUC calculated is returned as a vector, it is easy to calculate the mean and the standard deviation. With this calculation it is easy to observe the differences in detection using the available algorithms, and also to see if they are significant. For each detection method MEET R package allows to optimize a characteristic parameter, in ITEME this parameter is the $q - Rnyi$, in Q-residuals it is the number of principal components, in MDScan and MEME the parameters are

TABLE III

RESULTS FOR ROX1 DETECTION IN *Saccharomyces cerevisiae* PROMOTER SEQUENCE, USING ALL THE AVAILABLE ALGORITHMS IN MEET R PACKAGE

Algorithm	AUC	Error	Parameter
MATCH	0.9997	0.0006	CoreSimilarity=0.85
ITEME (entropy)	0.9992	0.0009	RényiOrder=1.3
Q-residuals	0.9999	0.0001	nPCs=8
MEME	0.9937	0.0018	length=12, motif=1
MDscan	0.9675	0.005	length=12

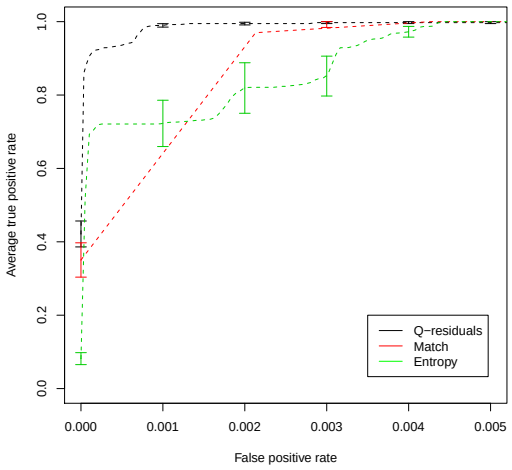

Fig. 3. ROC curve for the parameters that give the best AUC mean in ITEME, in green, MATCH, in red, and Q-residuals in black. The binding sites detected correspond to ROX1, within a *Saccharomyces cerevisiae* promoter sequence.

length and number of motifs and in MATCH it is the Core Similarity. The AUC vector for a range of parameters has been computed using all the available detection methods. The best one has been chosen as the parameter with a mean AUC closer to one. In table III it can be seen the detection of ROX1 transcription factor in a promoter sequence of *Saccharomyces cerevisiae* using ClustalW as the alignment algorithm. The Area shown corresponds to the best mean AUC obtained for each detection method, and the parameter is the detector parameter that returns this Area.

The results indicate in this case, and with the exception of MDscan and MEME, that the differences in AUC between methods are within the error of AUC. The ROC curve and its error for the best parameter of the detection methods ITEME (entropy), MATCH and Q-residuals can be visualized in figure 3. The other methods are not shown for visualization purposes, to give more visibility at the differences.

VI. CONCLUSIONS

MEET is an R package constituted of a group of computational methods for the detection of motifs and a set of multiple alignment algorithms. Specifically, MEET includes five detection programs: MEME/MAST, Q-residuals, MATCH, ITEME and MDscan and three multiple alignment algorithms: MUSCLE, ClustalW and MEME. The package allows each one of the detection methods to work independently with each one of the alignment algorithms. This allows the user a wide range of possibilities in detection. Given alignment, the detection can be performed with all the available algorithms and vice versa. Due to the great versatility of the package it is easy to compare directly all the detectors using the same alignment method, allowing a comparison at the same level of all algorithms. Moreover, when working in training mode, the package allows to select the optimal parameter for each one of the detectors.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of National Research Organization and reviewers' comments.

REFERENCES

[1] R. Mutihac, A. Cicuttin, and R. Mutihac, "Entropic approach to information coding in dna molecules," *Materials Science & Engineering C*, vol. 18, no. 1-2, pp. 51–60, 2001.

[2] T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. M. Hannett, C. T. Harbison, M. Thompson, I. Simon, J. Zeitlinger, E. Jennings, H. Murray, D. Gordon, B. Ren, J. Wyrick, J. Tagne, T. Volkert, E. Fraenkel, D. Gifford, and R. A. Young, "Transcriptional regulatory networks in saccharomyces cerevisiae," *Science*, vol. 298, pp. 799–804, 2002.

[3] W. Wei and X.-D. Yu, "Comparative analysis of regulatory motif discovery tools for transcription factor binding sites," *Geno. Prot. Bioinfo.*, vol. 5, 2007.

[4] M. K. Das and H.-K. Dai, "A survay of dna motif finding algorithms," *BMC Bioinformatics*, vol. 8(Suppl 7), p. S21, 2007.

[5] T. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, August 1994, pp. 28–36.

[6] J. Maynou, J.-J. Gallardo-Chacon, M. Vallverdu, P. Caminal, and A. Perera, "Computational detection of transcription factor binding sites through differential rényi entropy," *Information Theory, IEEE Transactions on*, vol. 56, no. 2, pp. 734 –741, feb. 2010.

[7] A. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. Kel-Margoulis, and E. Wingender, "MATCHTM: a tool for searching transcription factor binding sites in DNA sequences," *Nucl. Acids Res.*, vol. 31, no. 13, pp. 3576–3579, 2003.

[8] T. Bailey and G. Michael, "Combining evidence using p-values: application to sequence homology searches," *Bioinformatics*, vol. 14, pp. 48–54, 1998.

[9] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Simposium on Mathematics, Statistics and Probability*, 1961, pp. 547–561.

[10] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann.Math. Stat*, vol. 22, pp. 79–86, 1951.

[11] E. Pairo, S. Marco, and A. Perera, "A subspace method for the detection of transcripction factor binding sites." in *Proc. of the 1st International Conference IEEE International Conference on BioInformatics*, 20–23 Jan. 2009, pp. 1–5.

[12] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, 2004. [Online]. Available: http://dx.doi.org/10.1093/nar/gkh340

[13] J. Thompson, D. Higgins, and T. Gibson, "Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, pp. 4673–4680, 1994.

[14] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org