MEET: Motif Elements Estimation Toolkit

Erola Pairo, Joan Maynou *, Montserrat Vallverdú, Joan-Josep Gallardo-Chacón,
Pere Caminal, Santiago Marco y Alexandre Perera e-mail: epeiroibec.pcb.ub.es, smarco@el.ub.es

***Abstract*— MEET (Motif Elements Estimation Toolkit) es un paquete en R que integra un conjunto de algoritmos para la detección computacional de los puntos de unión de los factores de transcripción (TFBS). El paquete en R MEET incluye cinco programas de búsqueda de motivos: MEME/MAST (Multiple Expectation-Maximization for Motif Elicitation), Q-residuals, MDscan (Motif Discovery scan), ITEME (Information Theory Elements for Motif Estimation) y Match. Además, permite al usuario trabajar con diferentes algoritmos de alineamiento múltiple: MUSCLE (Multiple Sequence Comparison by Log-Expectation), ClustalW y MEME. El paquete puede trabajar en dos modos diferentes, entrenamiento y detección. El modo entrenamiento permite escoger los parámetros óptimos del detector escogido. Y el modo detección permite, una vez escogidos los parámetros, analizar un genoma en busca de puntos de unión. Además, ambos modos pueden combinar los diferentes métodos de alineamiento y de detección, permitiendo al usuario un amplio abanico de posibildades. Esta característica permite comparar los diferentes métodos computacionales al mismo nivel, sin realizar ningún agravio comparativo debido al alineamiento.**

I. MOTIVACIÓN

La regulación de la expresión génica es un proceso altamente regulado. Se inicia con la transferencia de información del ácido desoxirribonucleico, DNA, a ácido ribonucleico mensajero, mRNA, mediante la transcripción. Dicho proceso se modula mediante la asociación de determinadas proteínas, factores de transcripción (TF), con la correspondiente secuencia de unión (BS) [1]. Los puntos de unión de los factores de transcripción (TFBS), también conocidos por elementos cis-regulatory, constituyen las regiones regulatorias de un gen [2]. TFBS son secuencias cortas que presentan una alta variabilidad debido a que un mismo TF tiene la capacidad de unirse a diferentes posiciones y secuencias a lo largo del genoma. Esta variabilidad intrínseca que presentan los TFBS hace imposible establecer una secuencia consensus para su detección, por ese motivo se ha originado un gran conjunto de métodos de detección de patrones en secuencias de DNA[3].

Los algoritmos de descubrimiento de motivos pueden clasificarse según el modelo utilizado [4]. MEME/MAST (Multiple Expectation-Maximization for Motif Elicitation) [5], ITEME [6] y Match [7] son algoritmos basados en modelos probabilísticos. Dado un conjunto de sequencias no alineadas, MEME utiliza la máxima verosimilitud para determinar el máximo número de parámetros libres del modelo, utilizando el algoritmo de EM, Expectation-Maximization. MAST [8], Motif Alignment and Search Tool, es un algoritmo de búsqueda de secuencias homologas basado en el algoritmo Q-FAST para calcular la significancia estadística de las secuencias estimadas de un grupo de motivos característico. El algoritmo ITEME es un algoritmo basado en la teoría de la información. La detección se realiza mediante el análisi de la variación de la información contenida en el conjunto de secuencias de entrenamiento cuando se añade una secuencia de estudio. Dicho algoritmo permite realizar la detección de los TFBS considerando dependencia o independencia entre posiciones, según si se trabaja con un modelo basado en la entropía de Rényi [9] o en diveregencias [10]. MDscan es un algoritmo basado en un modelo determinístico, enumeración de palabras combinadas, y un modelo probabilístico, redes Bayesianas. Dentro de los modelos probabilísticos, Match es una herramienta basada en la matriz de pesos para buscar posibles TFBS en secuencias de DNA [7]. La detección se realiza mediante la consideración de dos puntajes: la matriz de similitud y el núcleo de similitud. El núcleo de similitud permite la preselección de los posibles TFBS, y la matriz de similitud es un puntaje de la calidad de la secuencia. Finalmente, el último método de detección existente en el paquete MEET es Q-residuals, el cual está basado en un modelo numérico. Concretamente, dicho detector transfoma cada base nitrogenada a una representación tridimensional donde cada nucleótido es colocado en el vértice de un tetraedro regular. A partir de las secuencias númericas se realiza un análisi de componentes principales (PCA). La hipótesis utilizada en Q-residuals es que los residuos de las secuencias del punto de unión modelado serán pequeños, mientras que los residuos de las secuencias del genoma serán mayores [11].

El paquete MEET, además de incluir un espectro amplio de métodos computacionales para la detección de motivos, incluye un conjunto de programas de alineamiento múltiple de secuencias. Concretamente, el usuario puede utilizar las herramientas MUSCLE (MUltiple Sequence Comparison by Log-Expectation) [12], ClustalW [13] y MEME [5] para obtener los diferentes nucleótidos involucrados en cada posición. MUSCLE está basado en un alineamiento iterativo de las secuencias. El método de alineamiento múltiplo iterativo se caracteriza por realizar el alineamiento en dos estadios. En

*CoAutor

Erola Pairo y Santiago Marco pertencen al Institut de Bioenginyeria de Catalunya, IBEC,y al Departament d'Electrònica, Universitat de Barcelona, Avinguda Diagonal, 647 08028 Barcelona, España.

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

J.J. Gallardo pertenece al CIBER de Bioingeniería, Biomateriales y Nanomedicina. http://www.isciii.es/htdocs/redes/ciber.jsp e-mail:joan.josep.gallardo@upc.edu

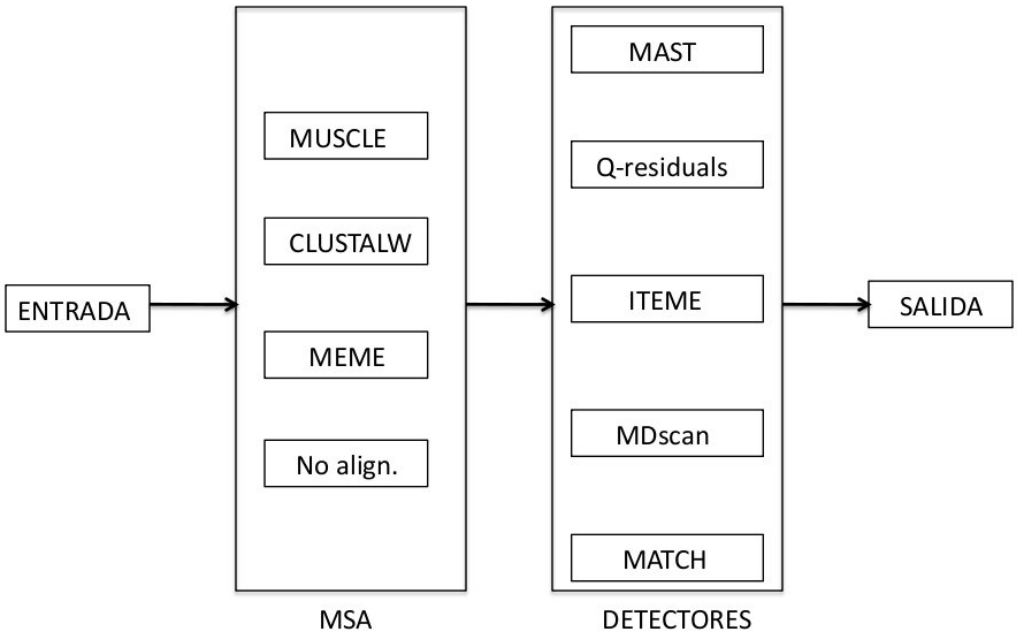

Fig. 1. Arquitectura del paquete en R MEET. MUSCLE, ClustalW y MEME son los programas de alineamiento utilizados, mientres que MAST, Q-residuals, ITEME, MDscan y MATCH son los correspondientes programas de detección de puntos de unión. Cada programa puede utilizar una matriz de secuencias de puntos de unión alineada con cualquier de los métodos enunciados.

el primero se realiza un alineamiento de pares. En el segundo estadio, se realiza el alineamiento multiple añadiendo progresivamente las secuencias, realineando el par de secuencias establecidas inicialmente. En cambio, ClustalW es un programa de alineamiento multiple de secuencias basado en un modelo progresivo. Dichos programas trabajan de forma similar que los programas basados en modelos iterativos, pero cuando se añade secuencialmente las secuencias no realinean el par de secuencias establecidas incialmente. Finalmente, el paquete MEET nos permite también trabajar con el alineamiento establecido por MEME.

II. PAQUETE MEET

El paquete MEET, Motif Elements Estimation Toolkit, integra diferentes algoritmos de detección de motivos (MEME/MAST , Q-residuals, ITEME, MATCH) y de alineamiento de secuencias de nucleótidos (MUSCLE, ClustalW y MEME). El algoritmo de alineamiento y el de detección se pueden escoger independientemente, permitiendo al usuario un amplio espectro de posibilidades, tal y como se puede ver en la figura I. El paquete tiene dos modos de trabajo distintos, con distintos parámetros de entrada, que se pueden ver resumidos en la tabla I, el modo entrenamiento y el modo detección.

El modo entrenamiento tiene como entrada, aparte de los algoritmos de alineamiento y detección y sus parámetros específicos, las secuencias a modelizar, la secuencia a analizar con la posición de los TFBS conocida, el background del organismo a estudiar y también un vector con los parámetros que se quiere explorar. Este modo de trabajo permite comparar detectores, métodos de alineamiento múltiple y sobretodo escoger los parámetros óptimos de un detector. Para realizar estas comparaciones se puede utilizar directamente la salida del paquete en este modo, que consta de las curvas ROC, la AUC en función del parámetro estudiado, la AUC máxima y el parámetro óptimo junto con un resumen de las opciones de entrada del programa y la secuencia consensus. Esta salida se encuentra resumida en la tabla II.

El modo de detección permite detectar factores de transcripción dentro de una secuencia de ADN. El usuario, para trabajar en este modo, debe especificar el parámetro del detector que quiere utilizar y el umbral de p-valores a partir del cual las secuencias son detectadas como factores

TABLE I

PARÁMETROS DE ENTRADA DEL PAQUETE MEET

Parámetros de entrada	Entrenamiento	Detección
TF(.fasta)	X	X
Secuencia de DNA (. fasta)	X	X
Algoritmo de Alineamiento	X	X
Parámetros de Alineamiento	X	X
Método de detección	X	X
Background Organismo	X	X
Umbral del p-valor		X
Parámetro del detector		X
Secuencia Leave-out-cross validation	X	
Posición del TFBS	X	
Número de motivos (MEME y MDscan)	X	X
Sentido	X	X
Porcentaje de Missing values (PCA)	X	X
Vector(ITEME)	X	X
Parámetros representación gráfica	X	

TABLE II

PARÁMETROS DE SALIDA DEL PAQUETE MEET

Entrenamiento	Detección
ROC	Secuencia/s detectada/s
Área ROC	Posición inicial
Parámetro óptimo	Posición final
Gráfica AUC Vs Parámetro	p-valor secuencia
Consensus	Consensus

de transcripción. De este modo el usuario que lo desee puede optimizar el detector escogido antes de estudiar una secuencia con TFBS desconocidos, o simplemente puede trabajar en este modo sin tener el detector optimizado. La salida de este modo, que también se puede ver en la tabla II son las secuencias detectadas como BS junto con su posición y el p-valor correspondiente, además del resumen de las opciones de entrada y la secuencia consensus del TFBS estudiado igual que en el modo entrenamiento.

La prinicipal ventaja del paquete MEET es que permite la comparación directa entre detectores y entre los diferentes parámetros de un detector, dejando al usuario la opción de optimizar la detección de BS en cada caso. Además, al integrar diferentes algoritmos de alineamiento y permitir combinaciones entre todos los detectores y todos los algoritmos de alineamiento se evitan los agravios comparativos entre los métodos de detección establecidos, debido al diferente alineamiento de la secuencias de TFBS realizado en cada uno de los métodos computacionales de detección.

III. IMPLEMENTACIÓN

El paquete MEET, Motif Elements Estimation Toolkit, es una herramienta de libre acceso que se ejecuta sobre el software de computación estadístico R[1]. Integra diferentes algoritmos de detección de motivos (MEME/MAST Version 4.4.0, Q-residuals, ITEME, MATCH Version 1.0 Public y MDscan) y de alineamiento de secuencias de nucleótidos (MUSCLE Version 3.8, ClustalW y MEME Version 4.4.0). Además, incluye una amplia documentación y ejemplos de consulta. El paquete MEET está disponible en http://sisbio.recerca.upc.edu/R/MEET.1.0.tar.gz

[1]http://www.r-project.org/

IV. EJEMPLO

Para ejecutar el paquete MEET, en cualquiera de sus modos de trabajo, se requiere el siguiente conjunto de librerías de R.

```
>library(seqinr)
>library(bio3d)
>library(aaMI)
>library(seqLogo)
>library(fields)
```

Los modos de trabajo se definen a partir del parámetro *system*. En el siguiente ejemplo, se realiza la validación del método de detección ITEME, basado en la entropía de Rényi, para un cojunto de secuencias de unión del factor de transcripción *ABF1*, alineadas mediante *ClustalW*, para el organismo *Saccharomyces cerevisiae*.

```
>Output<-MEET(TF="SqDNA.fa",seqin="DNA4.afa",
+alg="ClustalW",method="Entropy",system="validation",
+org="Saccharomyces cerevesiae", vector=c(1.0,1.3),
+sentit="f",position=c(501), xlim=1,errorbarby=0.1,
+avg="threshold")
```

Es importante remarcar que los datos de entrada, *TF (secuencias de unión),* seqin (secuencia de DNA), tienen que estar en formato .fasta. Los demás parámetros caracterizan el organismo de estudio, el alineamiento, el método de detección y la posición de los puntos de unión dentro de la secuencia a validar.

La variable de salida, en modo validación, contiene la sequencia consensus, los parámetros de entrada (*Summary*) y el resultado de validación. El cual, a su vez, contiene los parámetros óptimos del método considerado y la área bajo la curva ROC con su correspondiente error.

```
>names(Output)
[1] "Cosensus" "Summary" "Results"
>Output$Results
$Order 1.3 $Area 0.9992 $Areaerror 0.0009
```

V. RESULTADOS

El modo validación del paquete MEET permite comparar la detección de un factor de transcripción mediante los diferentes algoritmos utilizados. Al calcular no solamente el área, sino también el error de la misma, se puede ver si las diferencias entre los distintos métodos son significativas. En la tabla III se puede observar la detección del factor de transcripción ROX1 en una secuencia promotora del organismo *Saccharomyces Cerevesiae*. El algoritmo de alineación utilizado ha sido el CLUSTALW y el área corresponde a la mejor AUC obtenida para cada método. Además, se puede visualizar el parámetro que devuelve dicha AUC.

Los resultados indican que en este caso, y con excepción de MDScan y MEME, las diferencias en AUC entre los métodos de detección entran dentro del error de AUC, con lo cuál no son significativas. En la figura 2 se visualiza la curva ROC del mejor parámetro para los métodos ITEME (entropy), Match y Q-residuos con los correspondientes

TABLE III

RESULTADOS PARA LA DETECCIÓN DE ROX1

Algoritmo	Área	Error	Parámetro
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

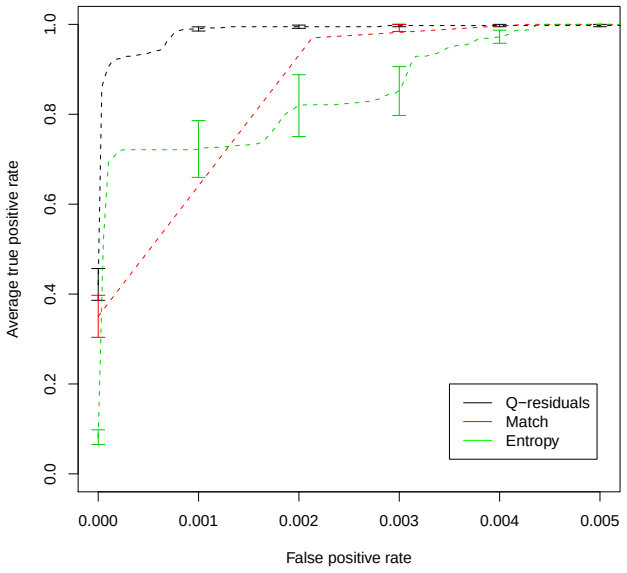

Fig. 2. Curva ROC de los mejores parámetros de los métodos ITEME Match y Q-residuals, con los errores correspondientes

errores. Los otros, al tener una AUC más baja quedaban en la parte inferior del gráfico, así que no los hemos representado para dar mayor visibilidad a las diferencias.

VI. CONCLUSIONES

A. Conclusiones

MEET es un paquete en R constituido por un conjunto de métodos computacionales para la detección de motivos y de algoritmos de alineamiento múltiple de secuencias. Concretamente, MEET incluye cinco programas de detección de motivos: MEME/MAST, Q-residuals, MATCH, ITEME y MDscan y tres algoritmos de alineamiento múltiple de secuencias : MUSCLE, ClustalW y MEME. El paquete permite la ejecución de cada uno de los métodes computacionales y de alineamiento de forma independiente y secuencial. Permitiendo un amplio abanico de posibilidades de detección. Para un mismo alineamiento, se puede realizar la detección con diferentes métodos computacionales, y a la inversa. Debido a esta gran versatilidad, es senzillo realizar la comparación directa entre los detectores para un mismo alineamiento. Permitiendo así una comparación al mismo nivel entre detectores. Además, el hecho de incorporar la posibilidad de trabajar en modo entrenamiento, MEET permite seleccionar el parámetro óptimo para cada uno de los detectores establecidos. Para futuras versiones, se implementará dicho paquete en C para disminuir el tiempo de cálculo de cada uno de los programas de detección.

VII. ACKNOWLEDGMENTS

Este trabajo ha sido parcialmente financiado por la CICYT TEC2007-63637/TCM del Ministerio de Ciencia y Tecnología, así como por el programa Ramón y Cajal del Ministerio de Educación y Ciencia. El CIBER de Bioingeniería, Biomateriales y Nanomedicina es una iniciativa del ISCIII.

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.