

# Tarea: Peak-calling choices

Amaranta Manrique de Lara y Ramirez, Claudia Saraí Reyes Ávila, Valeria Erendira Mateo Estrada

24 de febrero de 2016

## Introducción

El método de **ChIP-seq** es utilizado para localizar sitios de unión para factores de transcripción o marcas en la cromatina a lo largo de todo el genoma<sup>[1]</sup>. Éste se caracteriza por combinar dos poderosas herramientas: la inmunoprecipitación de la cromatina y las nuevas tecnologías de secuenciación masiva<sup>[2]</sup>. Posterior a la secuenciación se mapean las lecturas contra un genoma de referencia. Los picos de ChIP-Seq son las regiones ricas en lecturas y pueden obtenerse mediante el uso de algoritmos peak-calling, como por ejemplo **MACS** y **SWEMBL**. Los picos de ChIP-seq proporcionan información precisa sobre la localización de los sitios de unión de promotores y reflejan la diversidad de los sitios.

La amplitud y número de los picos dependen del peak-caller utilizado. La evaluación de los picos encontrados requiere de la selección de varios criterios. Entre los criterios se encuentran: significancia, enriquecimiento del motivo de referencia, concentración del motivo de referencia y de los motivos encontrados en el centro de los picos y relevancia biológica. RSAT es un sitio web que provee herramientas diseñadas específicamente para la detección de elementos regulatorios en secuencias no codificantes<sup>[3]</sup>.

A pesar de ser un método ampliamente utilizado en la investigación de organismos modelo, regulación génica, desarrollo, enfermedades y vías biológicas; puede llegar a tener grandes variaciones en la interpretación de resultados de acuerdo a la metodología que se haya empleado, los criterios utilizados y sus valores asignados. No obstante, también hay una gran diversidad de formas en la que se puede evaluar la calidad del experimento<sup>[4]</sup>.

En este trabajo se abordarán criterios de control de calidad para los resultados de estos experimentos realizados para el factor **FNR** en *Escherichia coli*: significancia, consistencia, enriquecimiento y comparación con un golden standard.

La **significancia** consiste en comparar los motivos encontrados en picos experimentales de ChIP-Seq contra controles negativos que son motivos encontrados en secuencias aleatorias. **Consistencia** se refiere a qué tan conservados están los resultados de diferentes análisis y con diferentes parámetros. En el **enriquecimiento** se realiza una comparación contra el comportamiento de motivos de referencia. Finalmente, la comparación con un **golden standard** es el uso de un set de datos disponibles y evaluar la sensibilidad del experimento, es decir, los verdaderos positivos o cuántos motivos inferidos realmente son motivos.

## Métodos

El conjunto de datos fue obtenido de los de *Escherichia coli* disponibles en el servidor *tepeu*. Ya que se pretendía comparar los métodos, se eligió al menos un archivo FASTA de cada método: **MACS** y **SWEMBL**. Se tomaron muestras representativas de cada método según los parámetros en los que divergían: los valores *R* en el caso de SWEMBL, y los valores *q* para MACS.

Además de los datos de *E. coli* descargados, se corrieron dos controles: uno negativo para MACS, uno negativo para SWEMBL.

A continuación se presentan los pasos para el análisis de criterios de medida:

1. Significancia:

- Uso de la herramienta *peak-motifs*. +Visualización de los resultados en la sección **Discovered motifs**. +Descarga de matrices en formato **TRANSFAC**.
2. Consistencia:
    - Uso de la herramienta *matrix-clustering* con las matrices obtenidas en el paso anterior.
    - Visualización de los clústeres de motivos.
  3. Enriquecimiento: +Uso de la herramienta *matrix-quality* con las matrices obtenidas en el paso anterior.
  4. Golden standard:
    - Descarga de una Position-Weight Matrix de FNR en **RegulonDB**.
    - Generación de un consenso degenerado para el motivo utilizando **convert-matrix**

Para los controles negativos, se hizo un paso previo con la herramienta *random genome fragments*.

## Resultados

A continuación se presentan los resultados de cada uno de los criterios.

### Significancia

A partir de correr *peak-motifs* obtuvimos los resultados de significancia los cuales se muestran en el material suplementario (Figuras 1.##.##). Ahora para comparar entre los métodos MACS y SWEMBL observamos sus valores de significancia; entre más grande sea el valor menos falsos positivos se tendrán en los motivos descubiertos.

Los mayores valores de significancia se encuentran en MACS  $q0.0001$  y en SWEMBL  $R0.05$ . Esto nos dice que la significancia es indistinta entre los métodos: En el caso de MACS el valor de  $q$  estima la tasa de falsos positivos descubiertos, por lo que a valores más pequeños se tendrá una mayor rigurosidad y menos falsos positivos; esto concuerda con lo observado en el experimento, ya que los mejores valores de  $q0.0001$  son los más pequeños estudiados. En el caso de SWEMBL no conocemos la representación más adecuada, pero con nuestros datos y las observaciones podemos inferir que funciona de manera similar que el valor  $q$ ; es decir, entre más pequeños más riguroso y menos falso positivos encontraremos.

### Consistencia

De los resultados de *matrix-clustering*, decidimos eliminar los clústeres en los que se anidaban motivos encontrados en los controles negativos. Asumimos que si un motivo predicho es similar a un control negativo, entonces es un falso positivo. De esta manera reducimos el número de clústeres a únicamente 12 que se presentan en el material suplementario (Figuras 2.1-12).

Para comparar entre MACS y SWEMBL un primer criterio es ver en cuántos de los clústeres significativos se encuentran representados los motivos de las diferentes condiciones. Pudimos observar que tanto MACS  $q0.0001$  como SWEMBL  $R0.1$  están representados en 8 de los 13 clústeres, mientras que MACS  $q0.01$  y SWEMBL  $R0.5$  se encuentran en 7. Finalmente, tanto MACS  $q0.1$  y SWEMBL  $R0.05$  aparecen únicamente en 6.

Otro punto de comparación es si hay clústeres donde haya sólo resultados de MACS o sólo resultados de SWEMBL. Por ejemplo, el clúster 3 (Figura 2.1) tiene únicamente motivos encontrados con MACS. Por otro lado, los clústeres 17 (Figura 2.8), 21 (Figura 2.10), 22 (Figura 2.11) y 23 (Figura 2.12) tienen sólo motivos de SWEMBL. Algo notable es cuando un método con un parámetro está representado más de una vez en un cluster. Por ejemplo, SWEMBL tanto con  $R0.1$  como con  $R0.5$  se agrupa dos veces cada uno en el

cluster 8 (Figura 2.5), al igual que MACS  $q0.0001$ . De igual manera, SWEMBL R0.5 también se encuentra representado dos veces en el cluster 5 (Figura 2.2) y en el 16 (Figura 2.7).

Creemos que estos resultados indican que SWEMBL es más sensible, porque detecta más motivos que MACS. Aunque R0.5 está representado en 7 clústeres mientras que R0.1 se encuentra en 7, consideramos que los resultados de R0.5 son más consistentes porque se agrupan entre ellos más veces.

## Enriquecimiento

Para analizar el enriquecimiento se compara el comportamiento de cada cluster contra el comportamiento de un estándar de la base de datos que se seleccionó al correr matrix-clustering.

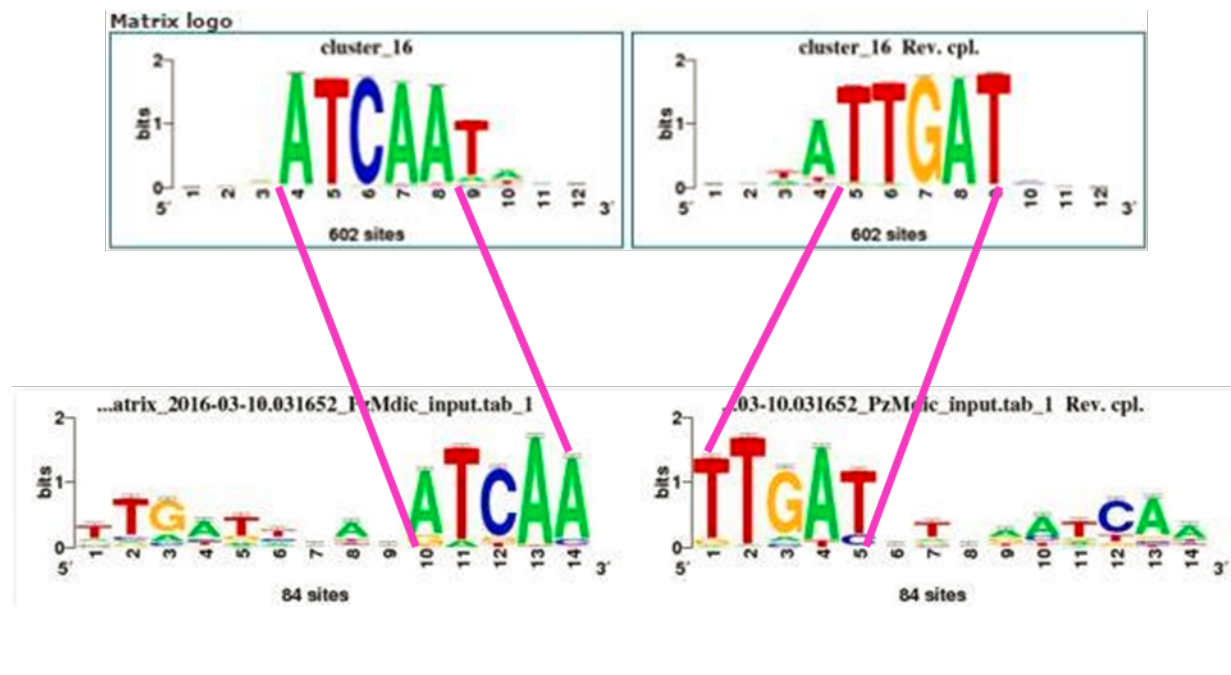
En este caso, se puede observar que en general las curvas son similares y no están tan dispersas. Por ejemplo, los clústeres del 3 al 8 (Figuras 3.1-6) tienen un comportamiento similar al de la base de datos RegulonDB. Sin embargo, también se considera un poco de ruido que se puede venir cargando desde los experimentos mismos o hasta del análisis bioinformático previo.

Por otro lado, los clústeres 9, 16, 17, 19, 21 y 23 no siguen la misma tendencia. En las gráficas logarítmicas de cada uno de estos clústeres individuales (Figuras 3.7-12.2) se puede observar que aumentan los números de lecturas; esto está relacionado con una disminución de especificidad.

## Golden standard

Como comparación con un golden standard, se indagó el motivo consenso (Figura 4.2) de unión a DNA de FNR según su matriz de posiciones disponible en RegulonDB.

Después comparamos con los consensos de los clústeres obtenidos en el criterio de consistencia para buscar un solapamiento y llegamos a la conclusión de que el cluster 16 tiene el consenso más similar al generado por cinco bases. Como figura significativa mostramos ambos logos juntos:



## Conclusiones y perspectivas

Del análisis de los cuatro criterios de calidad escogidos, podemos concluir que de los dos peak-callers utilizados en este trabajo, SWEMBL demostró ser el más apto para nuestro set de datos. Esto se debe a que fue más sensible, detectando motivos que no se detectaban a partir de los resultados de MACS. También podemos observar que según los valores que se le otorguen a R, la longitud de los motivos recuperados puede modificarse. Consideramos que el valor óptimo de R para nuestro set de datos es 0.5, ya que fue el más consistente en los clústeres encontrados.

Se podría discutir que R0.05 es un buen valor paramétrico para SWEMBL en este set de datos, ya que también observamos consistencia y se obtuvieron los mejores valores de significancia. Sin embargo, la comparación con el golden standard mostró que el motivo consenso del cluster 16 es el más cercano al obtenido de la matriz de posiciones anotada en la base de datos RegulonDB correspondiente al factor de transcripción. A pesar de que en el análisis de enriquecimiento el cluster 16 parecía no ser tan específico, sí se mostró consistente. Incluso en este cluster se ven representados dos motivos obtenidos de SWEMBL R0.5, lo cual valida nuestra hipótesis de que este método con este valor paramétrico son la mejor opción para el set de datos estudiado.

Finalmente, pensamos que para elegir el mejor método no se puede tomar un único criterio en cuenta, y sin embargo tampoco se pueden tomar todos porque empiezan a contradecirse. Lo ideal sería hacer repeticiones para ir calibrando los parámetros y equilibrar su exactitud.

---

## Material suplementario

Aquí se presenta una lista de fuentes de datos y comandos con afán de facilitar la reproducibilidad del estudio.

Además, se presentan imágenes de los resultados discutidos en el trabajo.

### Fuentes de datos

Acceso directo a las coordenadas de los picos del caso de estudio (GEO Accession):

[GSM1010219][<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1010219>]

[GSM1010224][<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1010224>]

### Recursos bioinformáticos utilizados

La siguiente tabla indica las herramientas bioinformáticas utilizadas en el análisis. .

Acronym	Description	URL
RSAT	Regulatory Sequence Analysis Tools	<a href="http://rsat.eu/">http://rsat.eu/</a>
RegulonDB	Base de datos de regulación transcripcional de <i>E. coli</i>	<a href="http://regulondb.ccg.unam.mx/">http://regulondb.ccg.unam.mx/</a>

## Lista de comandos y parámetros

A continuación se presentan los comandos empleados para producir los resultados.

### Significancia

Comandos utilizados con la herramienta peak-motifs.

SWEMBL R0.5

```
nice -n 19 $RSAT/perl-scripts/peak-motifs -v 1 -title 'coli_R0.5' -i $RSAT/public_html/tmp/apache/2016/03/09/
```

SWEMBL R0.1

```
$RSAT/perl-scripts/peak-motifs -v 1 -r_plot -title 'R.1' -i $RSAT/public_html/tmp/www-data/2016/03/10/
```

SWEMBL R0.05

```
nice -n 19 $RSAT/perl-scripts/peak-motifs -v 1 -title 'coli_R0.05' -i $RSAT/public_html/tmp/apache/2016/03/09/
```

MACS q0.0001

```
nice -n 19 $RSAT/perl-scripts/peak-motifs -v 1 -title 'coli_q0.0001' -i $RSAT/public_html/tmp/apache/2016/03/09/
```

MACS q0.01

```
$RSAT/perl-scripts/peak-motifs -v 1 -r_plot -title 'q0.01' -i $RSAT/public_html/tmp/www-data/2016/03/10/
```

MACS q0.1

```
nice -n 19 $RSAT/perl-scripts/peak-motifs -v 1 -title 'coli_q0.1' -i $RSAT/public_html/tmp/apache/2016/03/09/
```

### Consistencia

Comandos utilizados con la herramienta matrix-clustering

```
matrix-clustering -v 1 -max_matrices 300 -matrix_format transfac -i $RSAT/public_html/tmp/www-data/2016/03/09/
```

### Enriquecimiento

Comandos utilizados con la herramienta matrix-quality

```
nice -n 19 $RSAT/perl-scripts/matrix-quality -v 0 -ms $RSAT/public_html/tmp/apache/2016/03/09/matrix-qu
```

### Golden standard

Comandos utilizados con la herramienta conver-matrix a partir de una Position-Weight Matrix de FNR en RegulonDB.

```
$RSAT/perl-scripts/convert-matrix -from tab -to tab -i $RSAT/public_html/tmp/apache/2016/03/10/convert
```

Figuras

Significancia

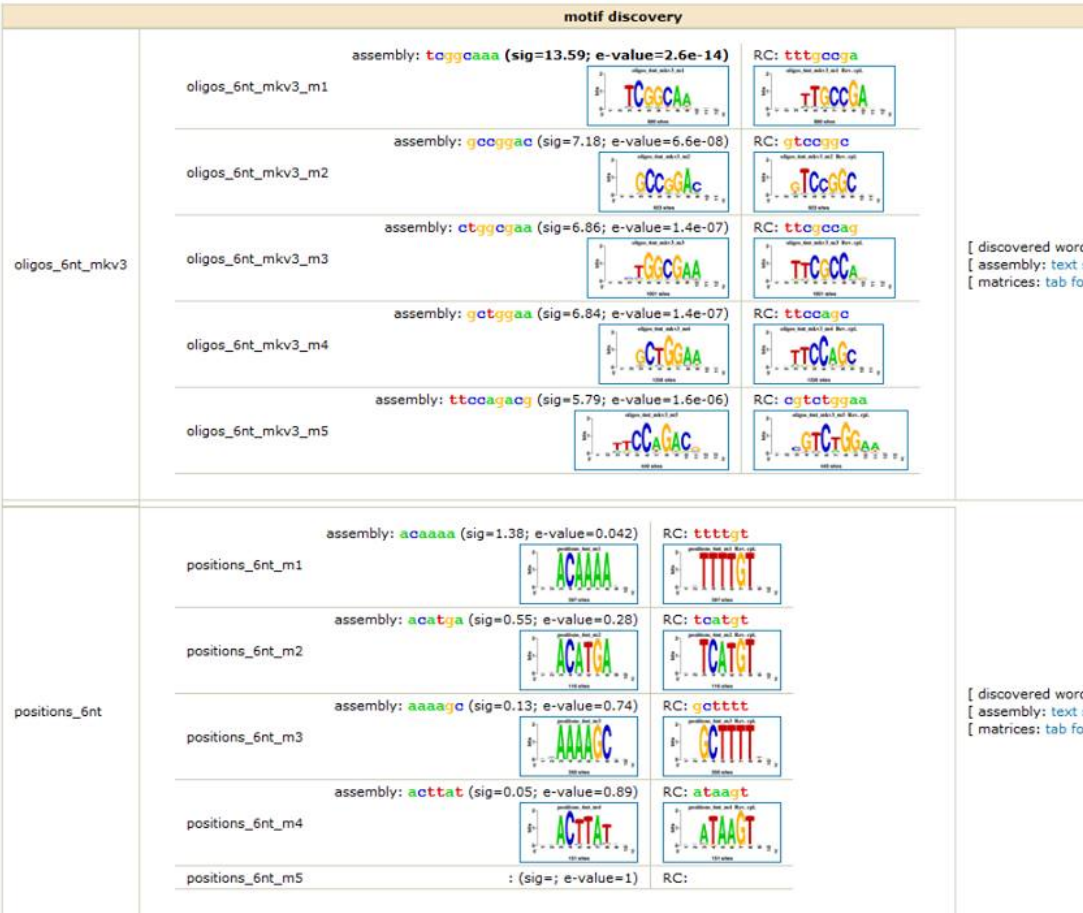


Figura 1.1.1 (SWEMBL R0.05)

oligos_7nt_mkv4	oligos_7nt_mkv4_m1	assembly: <b>caagcctt</b> (sig=5.79; e-value=1.6e-06)	RC: <b>aggcgctg</b>	[ discovered words: [ assembly: text sig [ matrices: tab fo
	oligos_7nt_mkv4_m2	assembly: <b>ctccggcac</b> (sig=4.05; e-value=8.9e-05)	RC: <b>gtccggag</b>	
	oligos_7nt_mkv4_m3	assembly: <b>tgcgcga</b> (sig=3.56; e-value=0.00028)	RC: <b>tcggcca</b>	
	oligos_7nt_mkv4_m4	assembly: <b>ccagctc</b> (sig=2.58; e-value=0.0026)	RC: <b>gagctgg</b>	
	oligos_7nt_mkv4_m5	assembly: <b>ggttcga</b> (sig=1.53; e-value=0.03)	RC: <b>tcgaacc</b>	
positions_7nt	positions_7nt_m1	assembly: <b>aacatga</b> (sig=1.84; e-value=0.014)	RC: <b>tcattgtt</b>	[ discovered words: [ assembly: text sig [ matrices: tab fo
	positions_7nt_m2	assembly: <b>agtaaat</b> (sig=1.63; e-value=0.023)	RC: <b>atttaact</b>	
	positions_7nt_m3	assembly: <b>acttttg</b> (sig=1.39; e-value=0.041)	RC: <b>caaaagt</b>	
	positions_7nt_m4	assembly: <b>acgtaat</b> (sig=0.58; e-value=0.26)	RC: <b>attaagt</b>	
	positions_7nt_m5	assembly: <b>aagttaa</b> (sig=0.33; e-value=0.47)	RC: <b>ataactt</b>	

Figura 1.1.2 (SWEMBL R0.05)







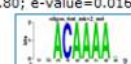













motif discovery				
oligos_6nt_mkv3	oligos_6nt_mkv3_m1	assembly: <b>acaaaa</b> (sig=2.37; e-value=0.0043)	RC: <b>ttttgt</b>	[ discovered words: [ assembly: text sig [ matrices: tab fo
	oligos_6nt_mkv3_m2	assembly: <b>ggtaga</b> (sig=1.27; e-value=0.054)	RC: <b>tcacc</b>	
	oligos_6nt_mkv3_m3	assembly: <b>gatcaa</b> (sig=0.86; e-value=0.14)	RC: <b>ttgatc</b>	
	oligos_6nt_mkv3_m4	assembly: <b>gtctga</b> (sig=0.24; e-value=0.58)	RC: <b>tcagac</b>	
	oligos_6nt_mkv3_m5	: (sig=; e-value=1)	RC:	
positions_6nt	positions_6nt_m1	assembly: <b>ccgcac</b> (sig=0.54; e-value=0.29)	RC: <b>gtccgg</b>	[ discovered words: [ assembly: text sig [ matrices: tab fo
	positions_6nt_m2	assembly: <b>atcaaa</b> (sig=0.44; e-value=0.36)	RC: <b>tttgat</b>	
	positions_6nt_m3	: (sig=; e-value=1)	RC:	
	positions_6nt_m4	: (sig=; e-value=1)	RC:	
	positions_6nt_m5	: (sig=; e-value=1)	RC:	

Figura 1.2.1 (SWEMBL R0.1)

oligos_7nt_mkv4	assembly: <b>ggttoga</b> (sig=4.51; e-value=3.1e-05)	RC: <b>togaacc</b>	[ discovered words: t [ assembly: text sig [ matrices: tab forma
	oligos_7nt_mkv4_m1		
	assembly: <b>tagctca</b> (sig=1.91; e-value=0.012)	RC: <b>tgagcta</b>	
	oligos_7nt_mkv4_m2		
	assembly: <b>cgaatcc</b> (sig=1.25; e-value=0.056)	RC: <b>ggattcg</b>	
positions_7nt	oligos_7nt_mkv4_m3		[ discovered words: t [ assembly: text sig [ matrices: tab forma
	assembly: <b>aggcctt</b> (sig=0.72; e-value=0.19)	RC: <b>aggcctt</b>	
	oligos_7nt_mkv4_m4		
	oligos_7nt_mkv4_m5	: (sig=; e-value=1)	
		RC:	
positions_7nt	assembly: <b>aggggaa</b> (sig=3.02; e-value=0.00095)	RC: <b>ttccctt</b>	[ discovered words: t [ assembly: text sig [ matrices: tab forma
	positions_7nt_m1		
	assembly: <b>acgccac</b> (sig=2.61; e-value=0.0025)	RC: <b>gtggcgt</b>	
	positions_7nt_m2		
	assembly: <b>ccaccga</b> (sig=2.48; e-value=0.0033)	RC: <b>tgggtgg</b>	
positions_7nt	positions_7nt_m3		[ discovered words: t [ assembly: text sig [ matrices: tab forma
	assembly: <b>tacctoa</b> (sig=2.35; e-value=0.0045)	RC: <b>tgaggta</b>	
	positions_7nt_m4		
	assembly: <b>cagtatg</b> (sig=2.11; e-value=0.0078)	RC: <b>catactg</b>	
	positions_7nt_m5		

Figura 1.2.2 (SWEMBL R0.1)



motif discovery			
oligos_6nt_mkv2	oligos_6nt_mkv2_m1	assembly: <b>tggtaga</b> (sig=3.57; e-value=0.0027) 	RC: <b>tetacca</b> 
	oligos_6nt_mkv2_m2	assembly: <b>gctggcaa</b> (sig=2.06; e-value=0.0087) 	RC: <b>tgcacagc</b> 
	oligos_6nt_mkv2_m3	assembly: <b>cgtac</b> (sig=1.85; e-value=0.014) 	RC: <b>gtacg</b> 
	oligos_6nt_mkv2_m4	assembly: <b>acaaaa</b> (sig=1.80; e-value=0.016) 	RC: <b>ttttgt</b> 
	oligos_6nt_mkv2_m5	assembly: <b>tctottca</b> (sig=1.43; e-value=0.037) 	RC: <b>tgaagaga</b> 
positions_6nt	positions_6nt_m1	assembly: <b>gatcaata</b> (sig=1.65; e-value=0.022) 	RC: <b>tattgato</b> 
	positions_6nt_m2	assembly: <b>atcaca</b> (sig=1.00; e-value=0.1) 	RC: <b>tgatgat</b> 
	positions_6nt_m3	assembly: <b>ctaaga</b> (sig=0.45; e-value=0.35) 	RC: <b>tcttag</b> 
	positions_6nt_m4	assembly: <b>agtata</b> (sig=0.12; e-value=0.76) 	RC: <b>tatact</b> 
	positions_6nt_m5	assembly: <b>acaaga</b> (sig=0.09; e-value=0.81) 	RC: <b>tcttgt</b> 

[ discovered words:  
[ assembly: text sig  
[ matrices: tab form

[ discovered words:  
[ assembly: text sig  
[ matrices: tab form

Figura 1.3.1 (SWEMBL R0.5)

oligos_7nt_mkv2	assembly: <b>cgaatcc</b> (sig=1.73; e-value=0.019)	RC: <b>ggattcg</b>	[ discovered words: [ assembly: text sig [ matrices: tab form
	oligos_7nt_mkv2_m1		
	assembly: <b>tagctca</b> (sig=1.60; e-value=0.025)	RC: <b>tcgagcta</b>	
	oligos_7nt_mkv2_m2		
	assembly: <b>tctaccaa</b> (sig=0.73; e-value=0.19)	RC: <b>ttggtaga</b>	
	oligos_7nt_mkv2_m3		
positions_7nt	assembly: <b>ctgaatc</b> (sig=0.65; e-value=0.22)	RC: <b>gattcag</b>	[ discovered words: [ assembly: text sig [ matrices: tab form
	oligos_7nt_mkv2_m4		
	assembly: <b>gaagaga</b> (sig=0.28; e-value=0.52)	RC: <b>tctcttc</b>	
	oligos_7nt_mkv2_m5		
	assembly: <b>atcaata</b> (sig=3.51; e-value=0.00031)	RC: <b>tattgat</b>	
	positions_7nt_m1		
positions_7nt	assembly: <b>gttataac</b> (sig=2.79; e-value=0.0016)	RC: <b>gttataac</b>	[ discovered words: [ assembly: text sig [ matrices: tab form
	positions_7nt_m2		
	assembly: <b>ctatatag</b> (sig=2.44; e-value=0.0036)	RC: <b>ctatatag</b>	
	positions_7nt_m3		
	assembly: <b>taaatga</b> (sig=1.27; e-value=0.054)	RC: <b>tcattta</b>	
	positions_7nt_m4		
positions_7nt	assembly: <b>tagatca</b> (sig=0.91; e-value=0.12)	RC: <b>tcagtcta</b>	[ discovered words: [ assembly: text sig [ matrices: tab form
	positions_7nt_m5		

Figura 1.3.2 (SWEMBL R0.5)

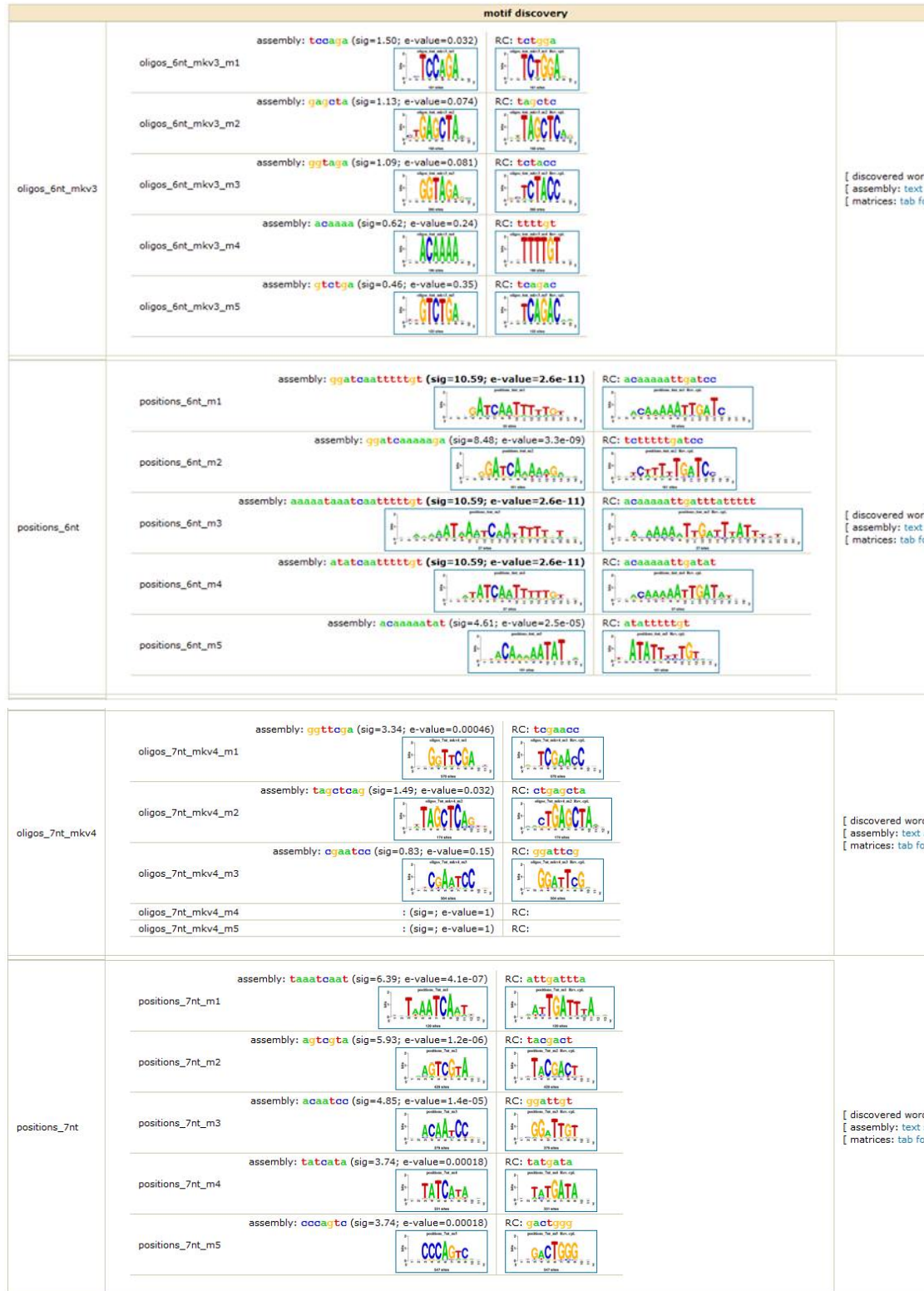


Figura 1.4.1 (MACS q0.0001)

Figura 1.4.2 (MACS q0.0001)



Figura 1.5.1 (MACS q0.01)

Figura 1.5.2 (MACS q0.01)



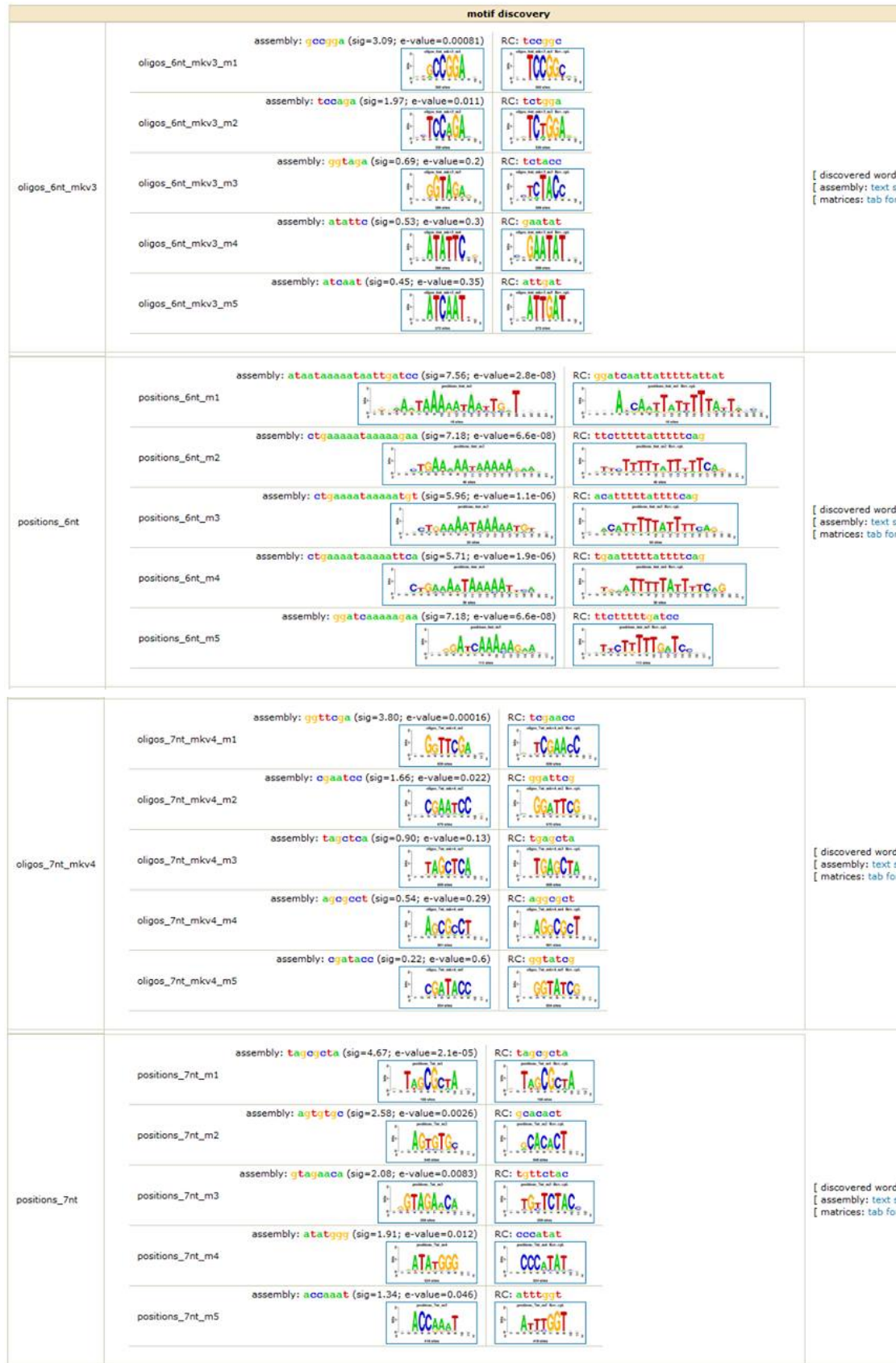


Figura 1.6.1 (MACS q0.0001)

Figura 1.6.2 (MACS q0.0001)

## Consistencia



Figura 2.1

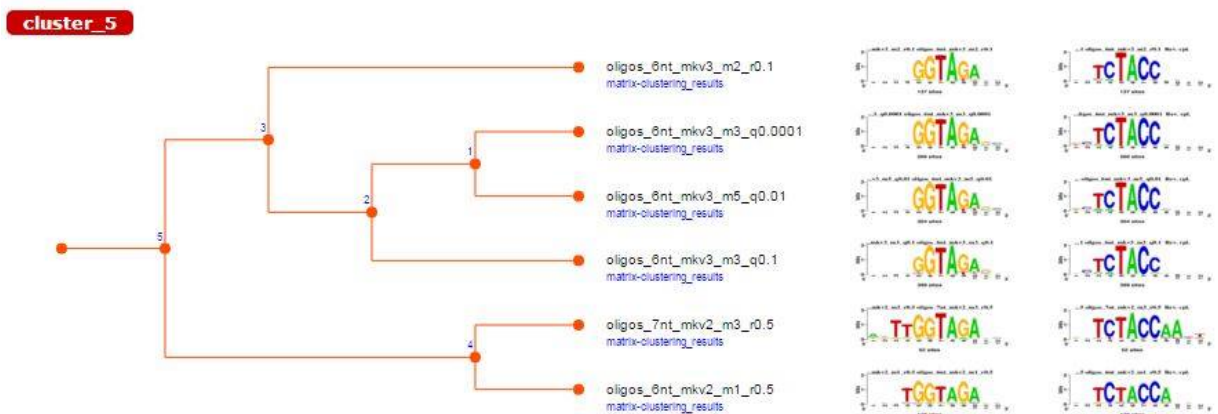


Figura 2.2

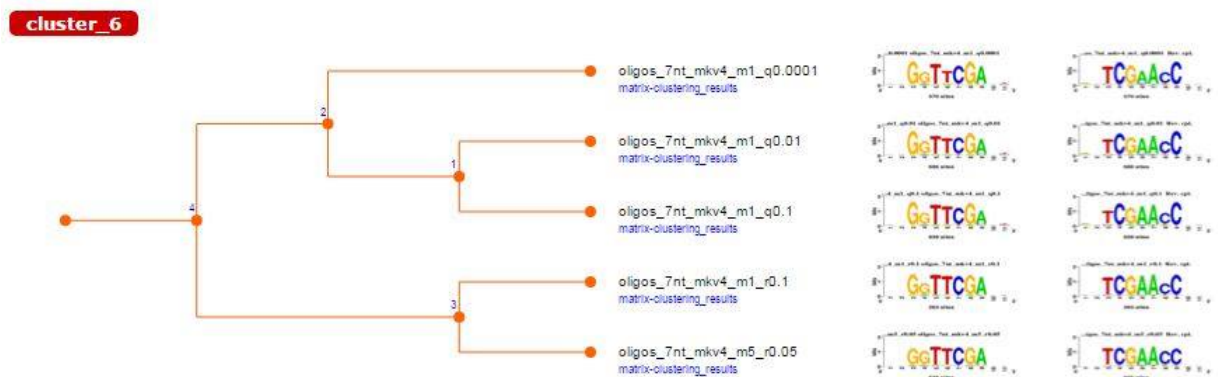


Figura 2.3

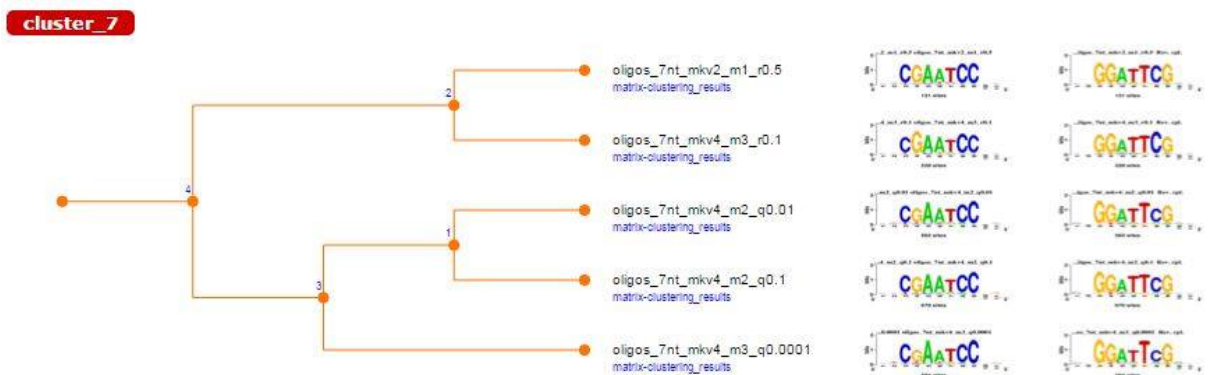


Figura 2.4

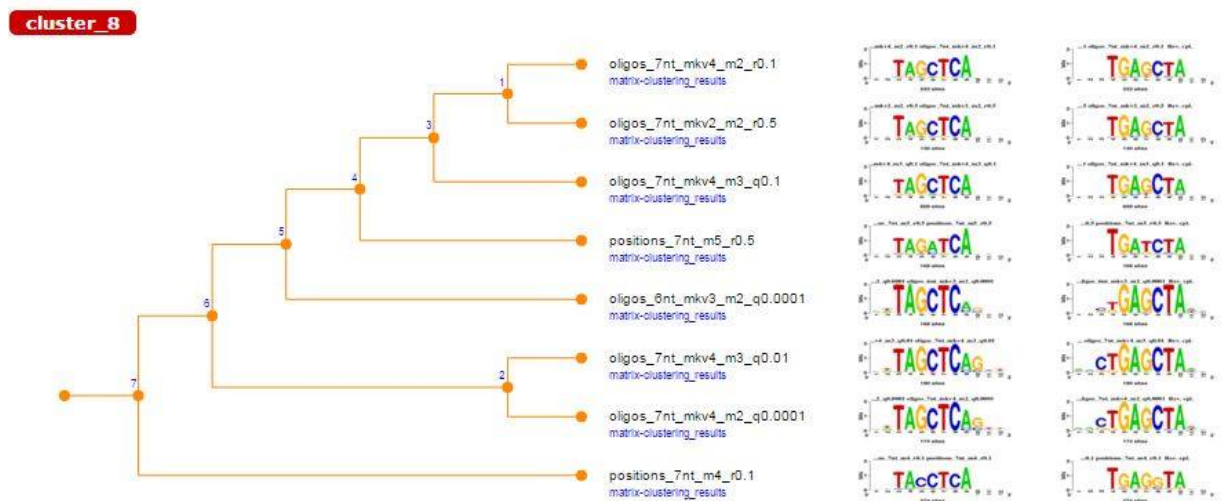


Figura 2.5



Figura 2.6

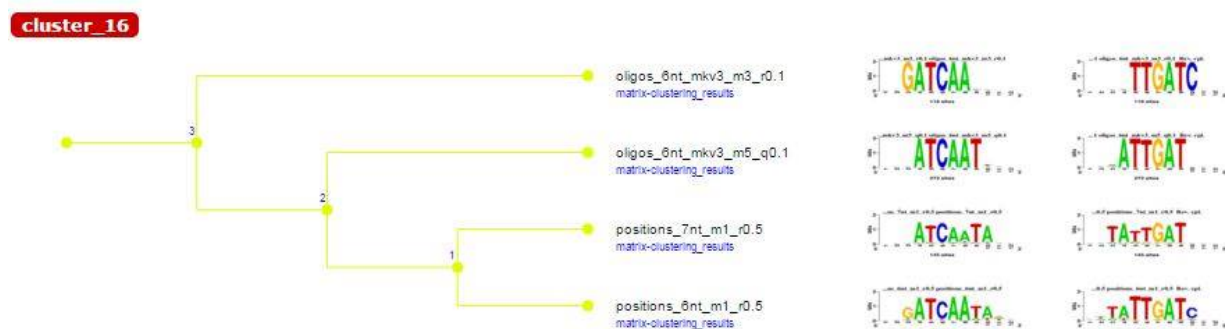


Figura 2.7



Figura 2.8



Figura 2.9



Figura 2.10

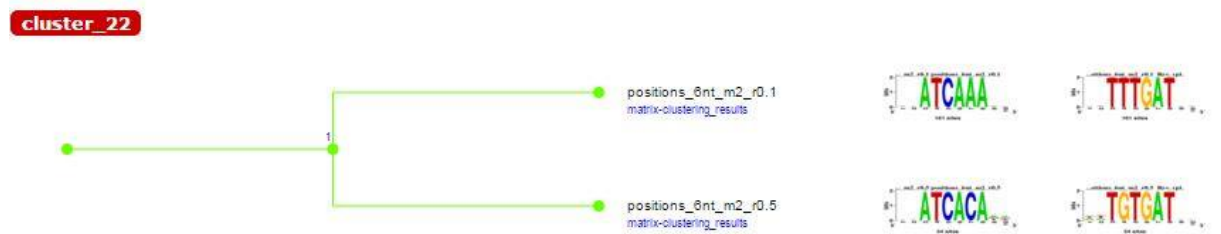


Figura 2.11



Figura 2.12

## Enriquecimiento

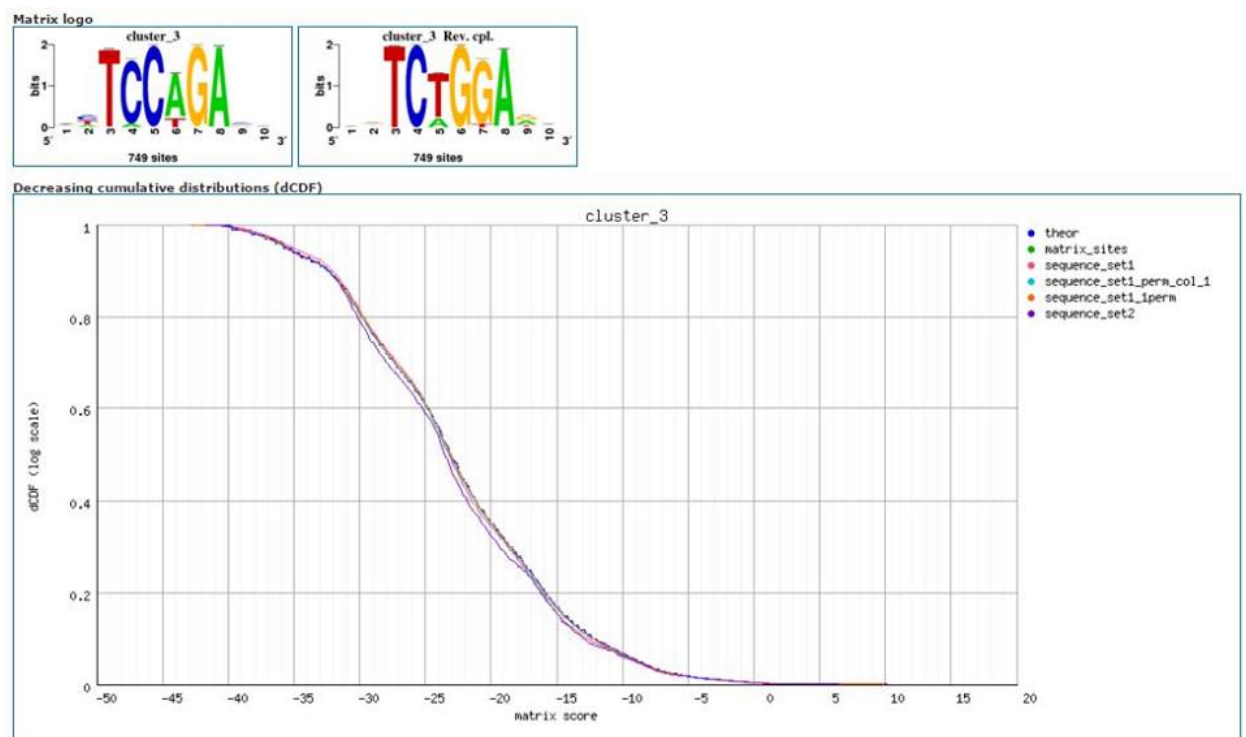


Figura 3.1.1



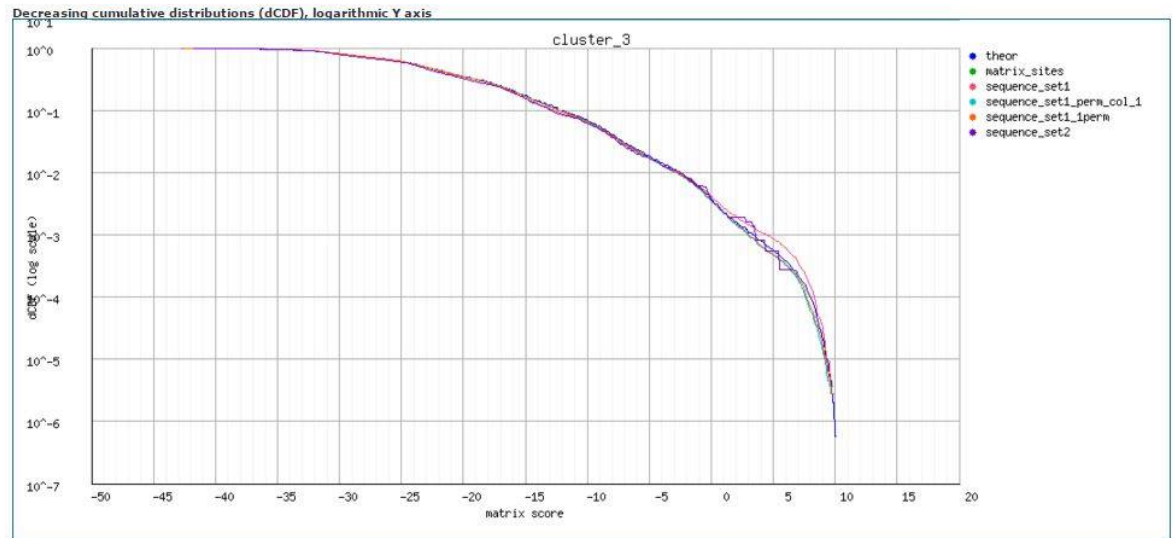


Figura 3.1.2

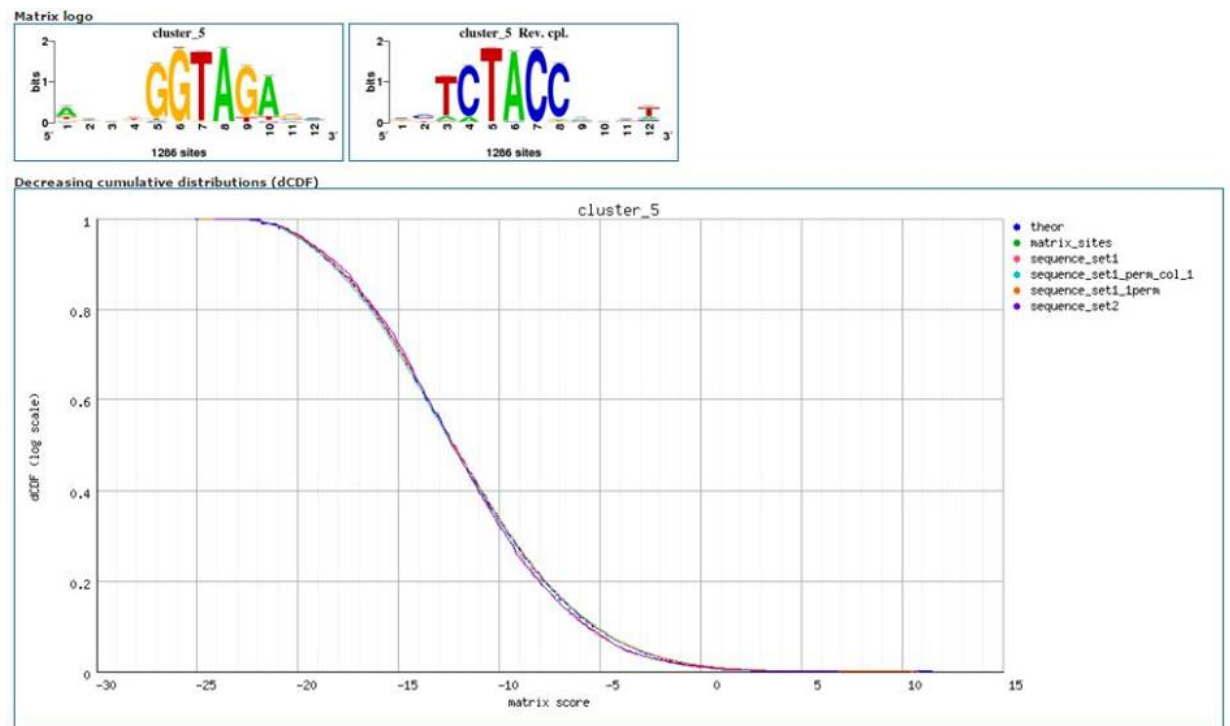


Figura 3.2.1

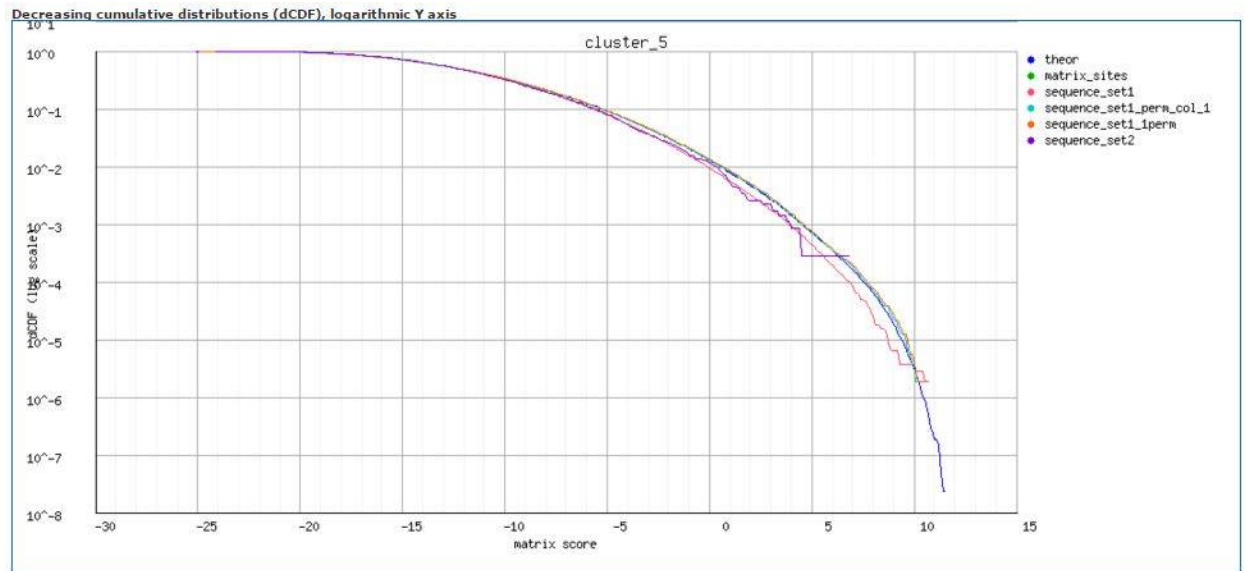


Figura 3.2.2

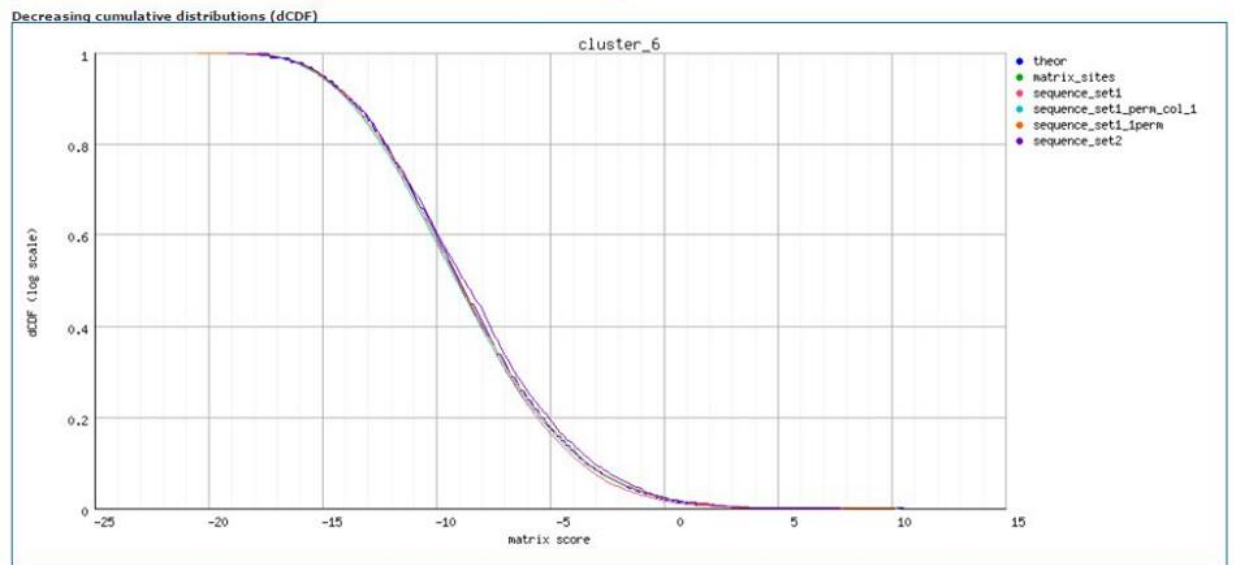
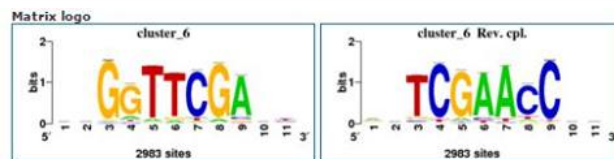


Figura 3.3.1

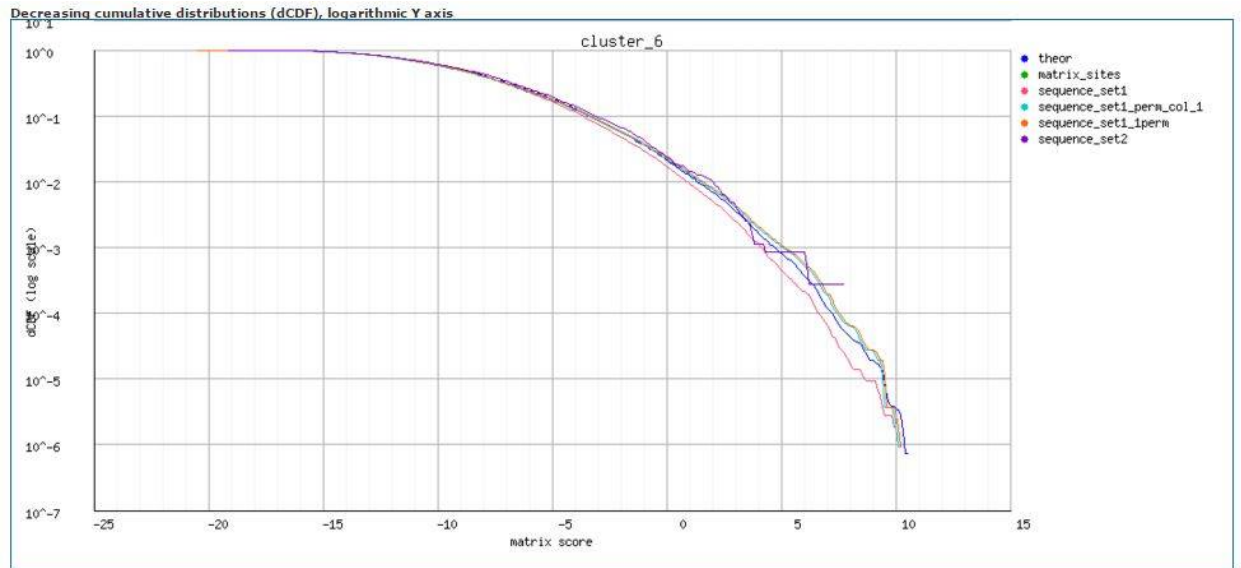


Figura 3.3.2

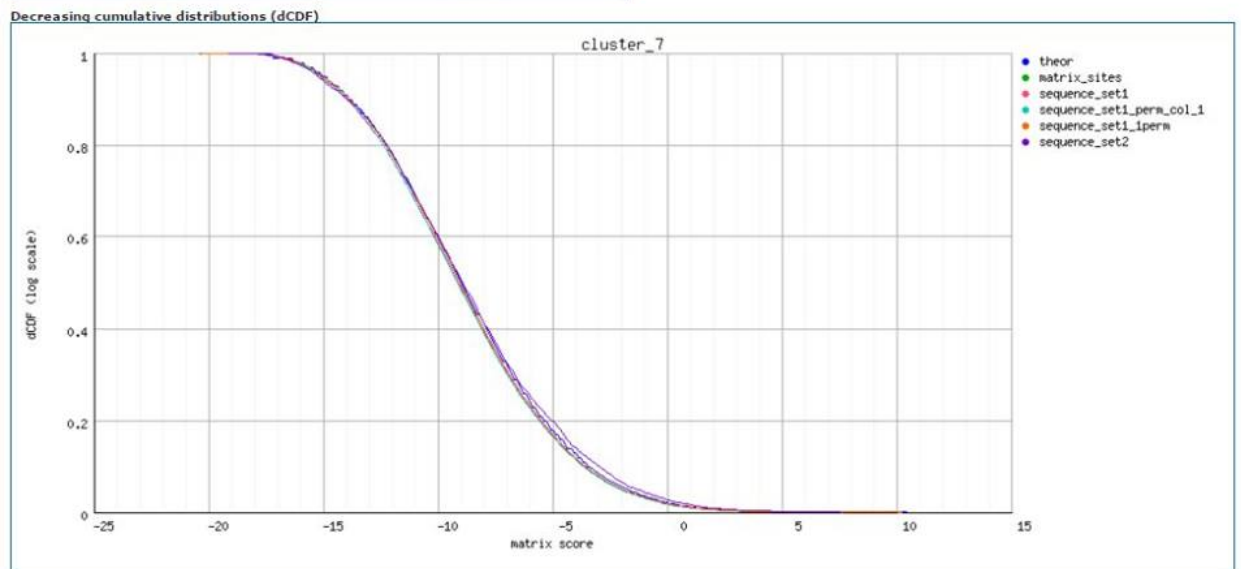
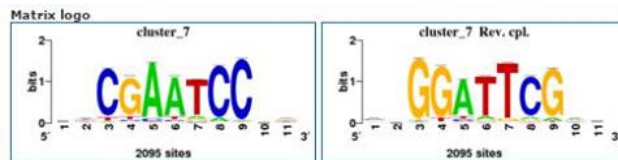


Figura 3.4.1

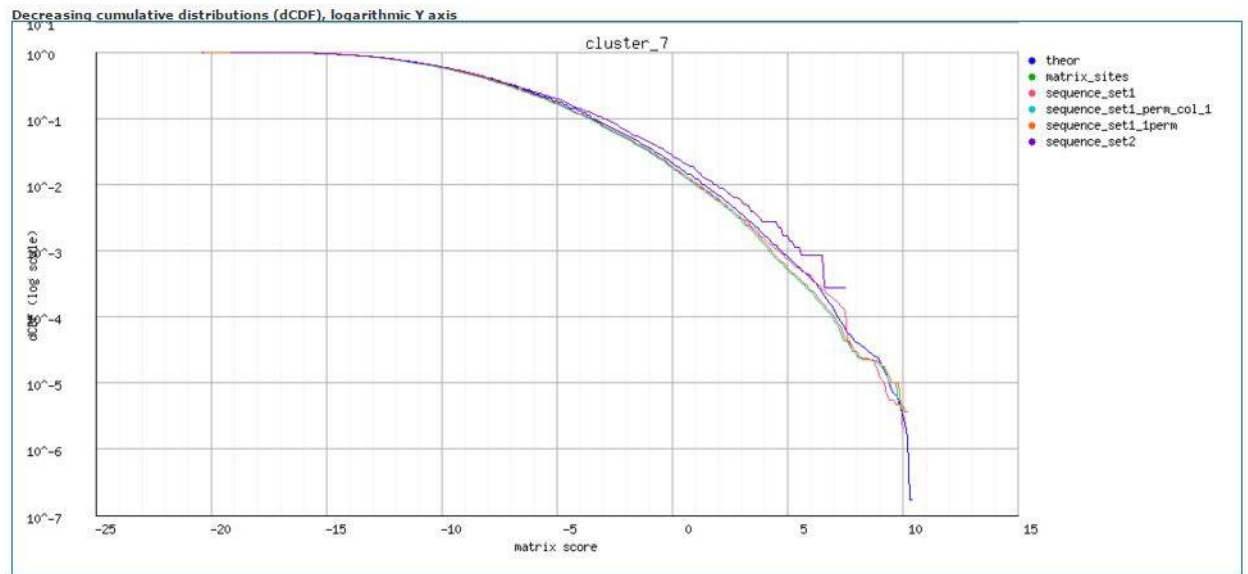


Figura 3.4.2

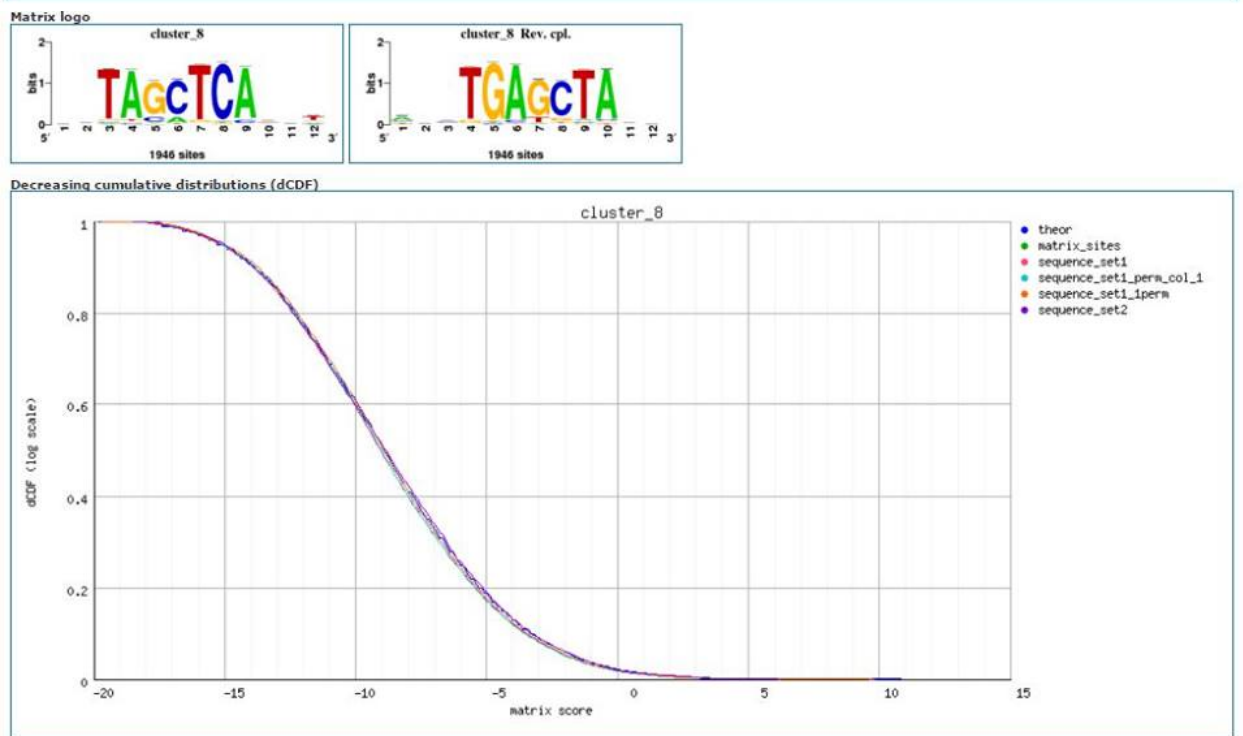


Figura 3.5.1

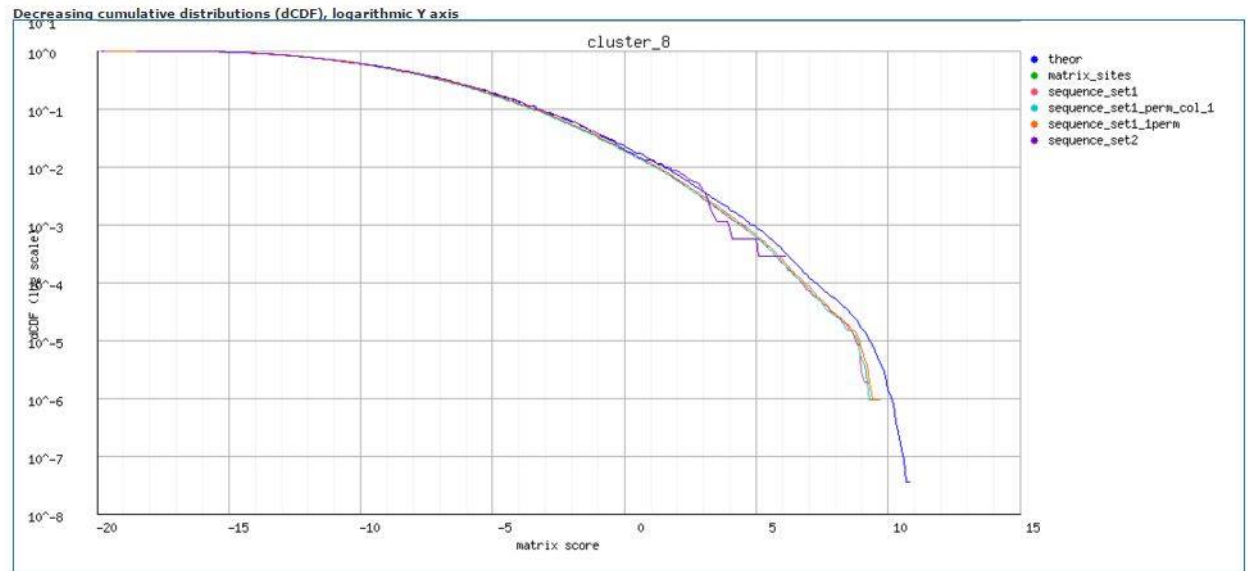


Figura 3.5.2

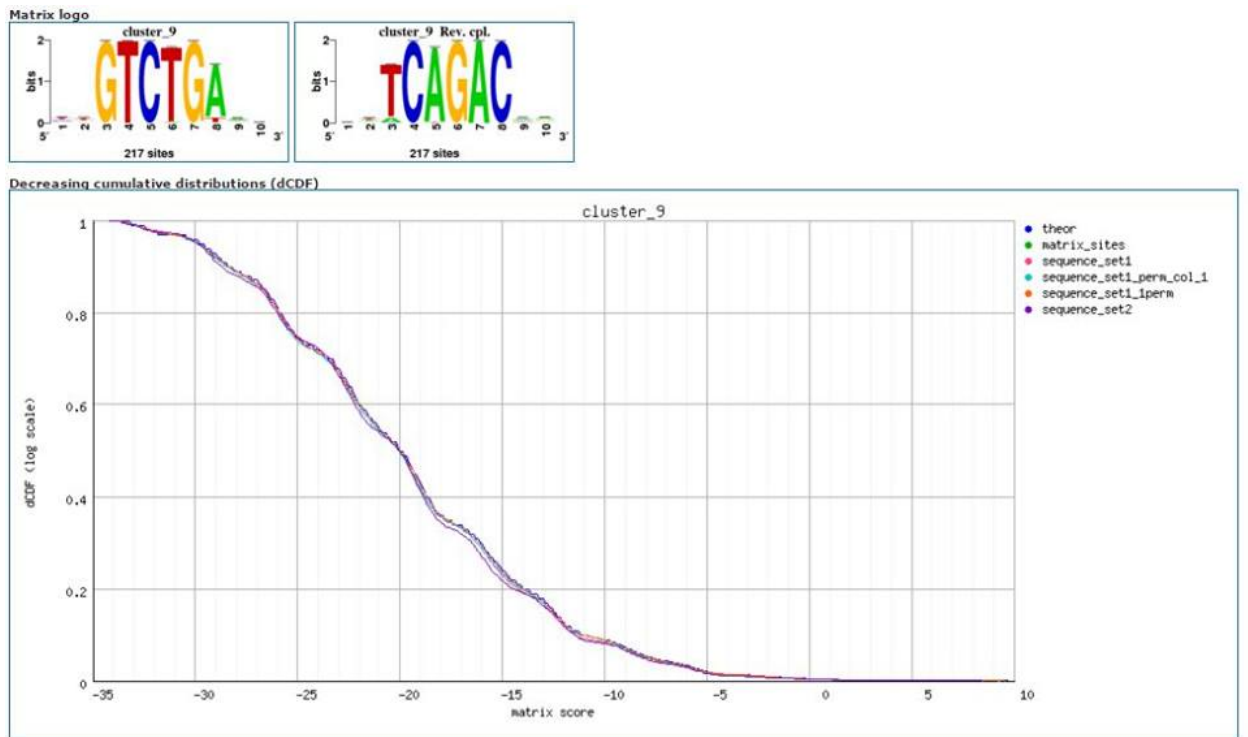


Figura 3.6.1

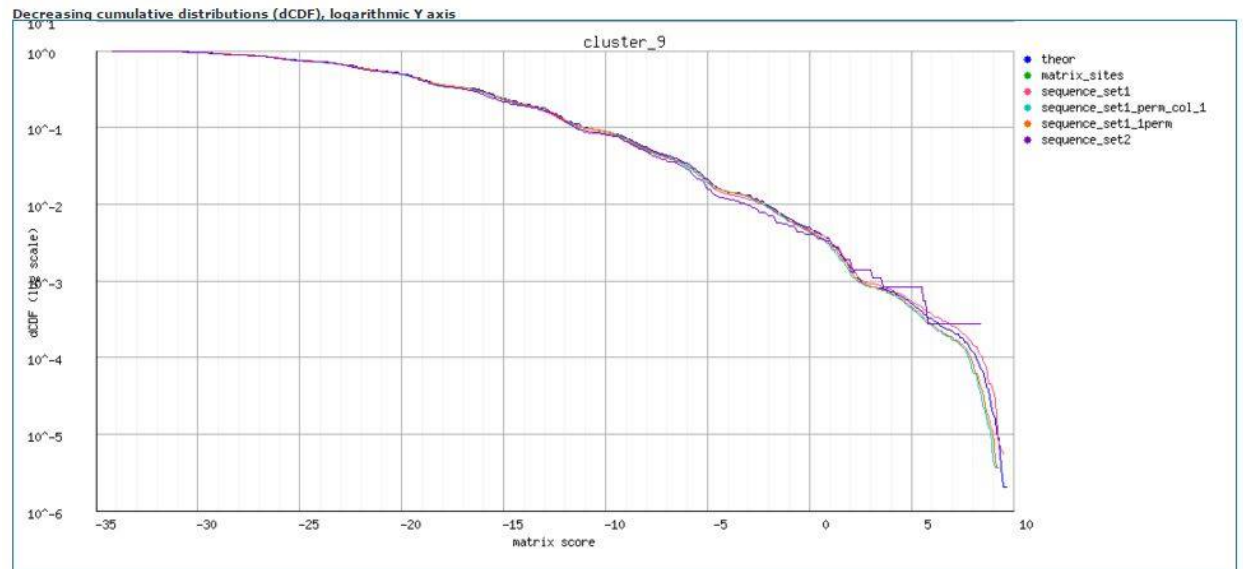


Figura 3.6.2

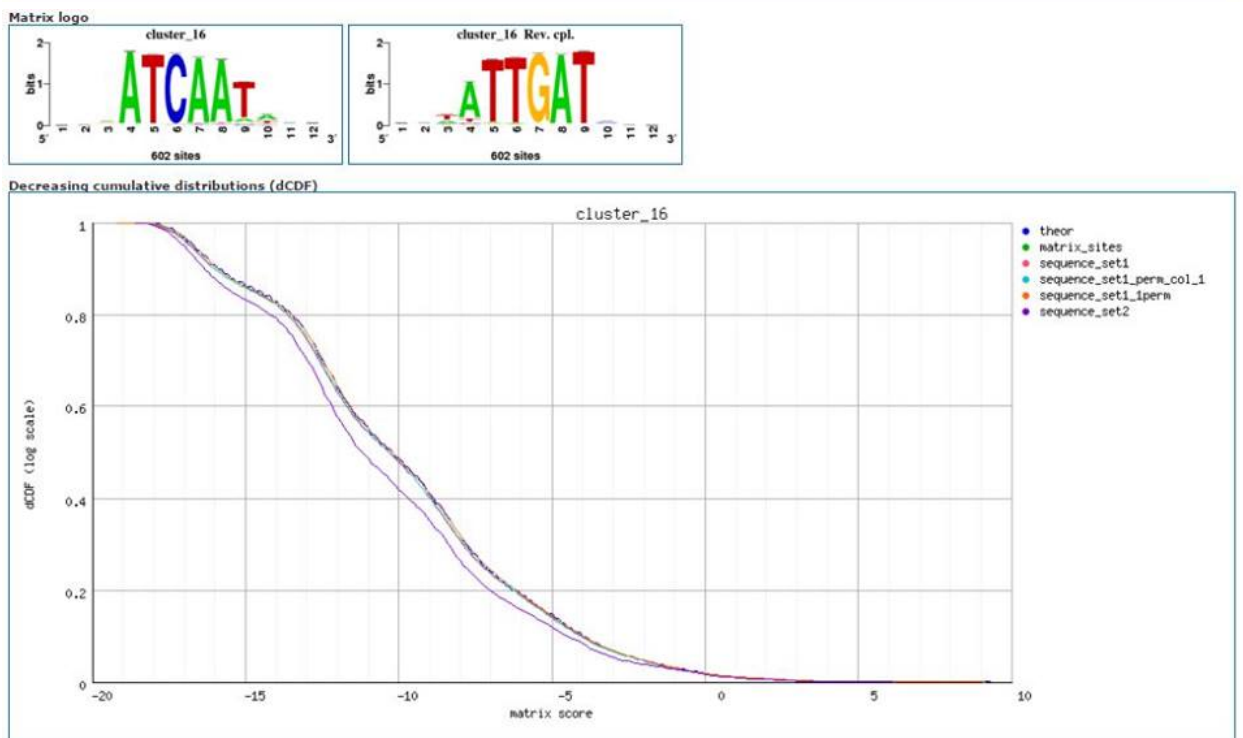


Figura 3.7.1

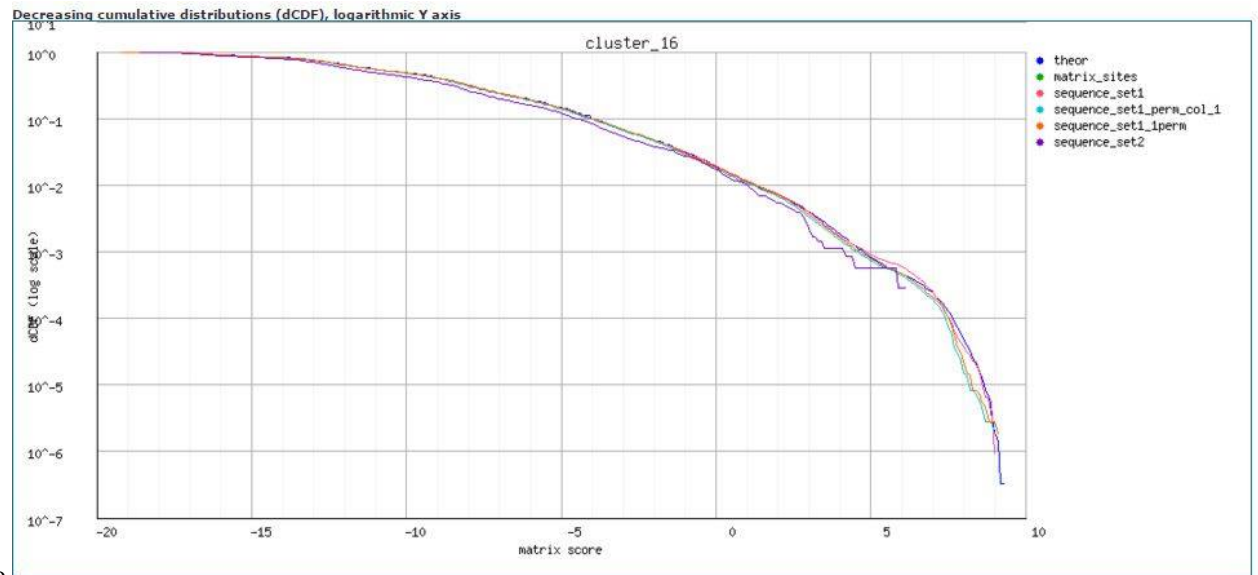


Figura 3.7.2

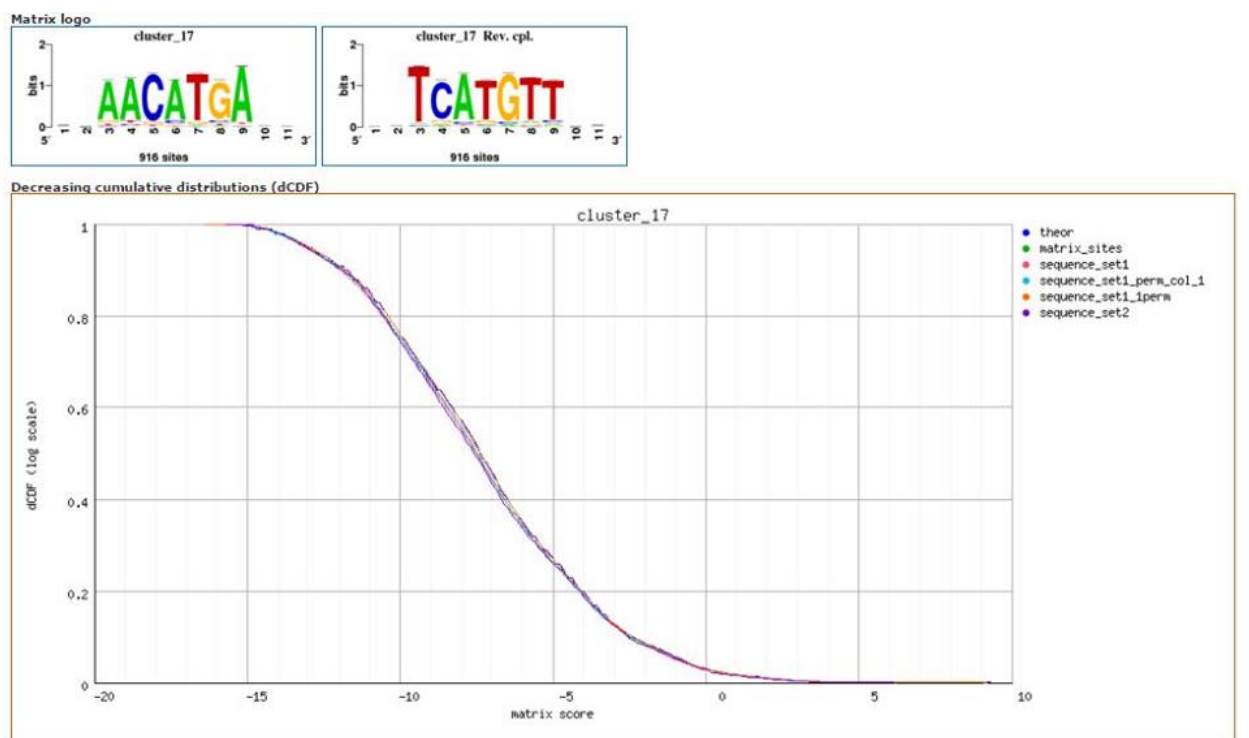


Figura 3.8.1

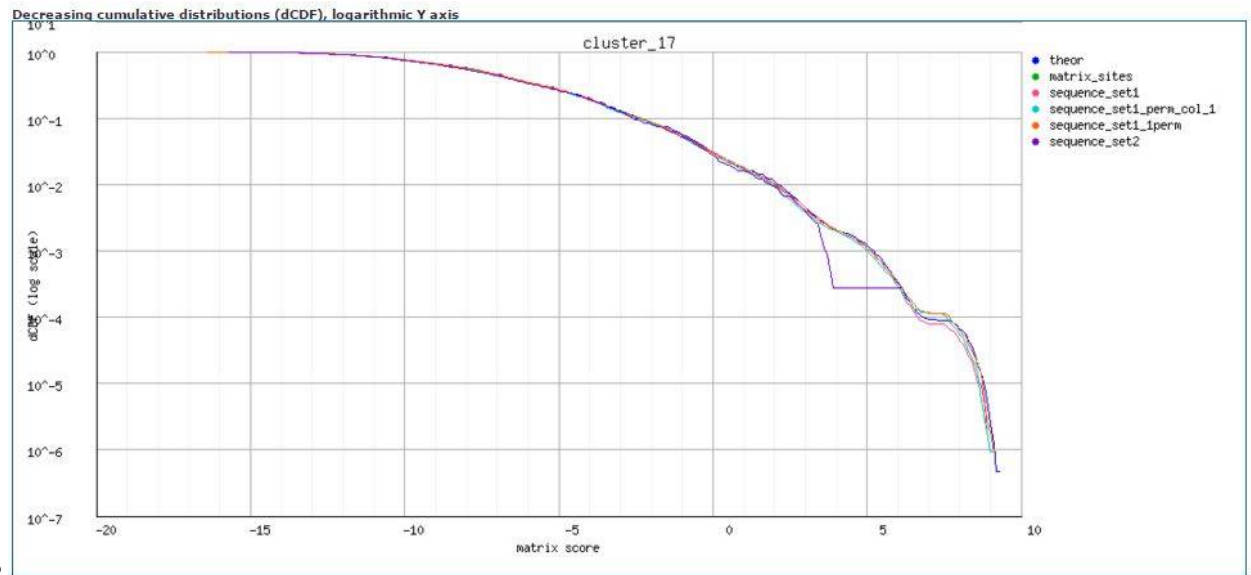


Figura 3.8.2

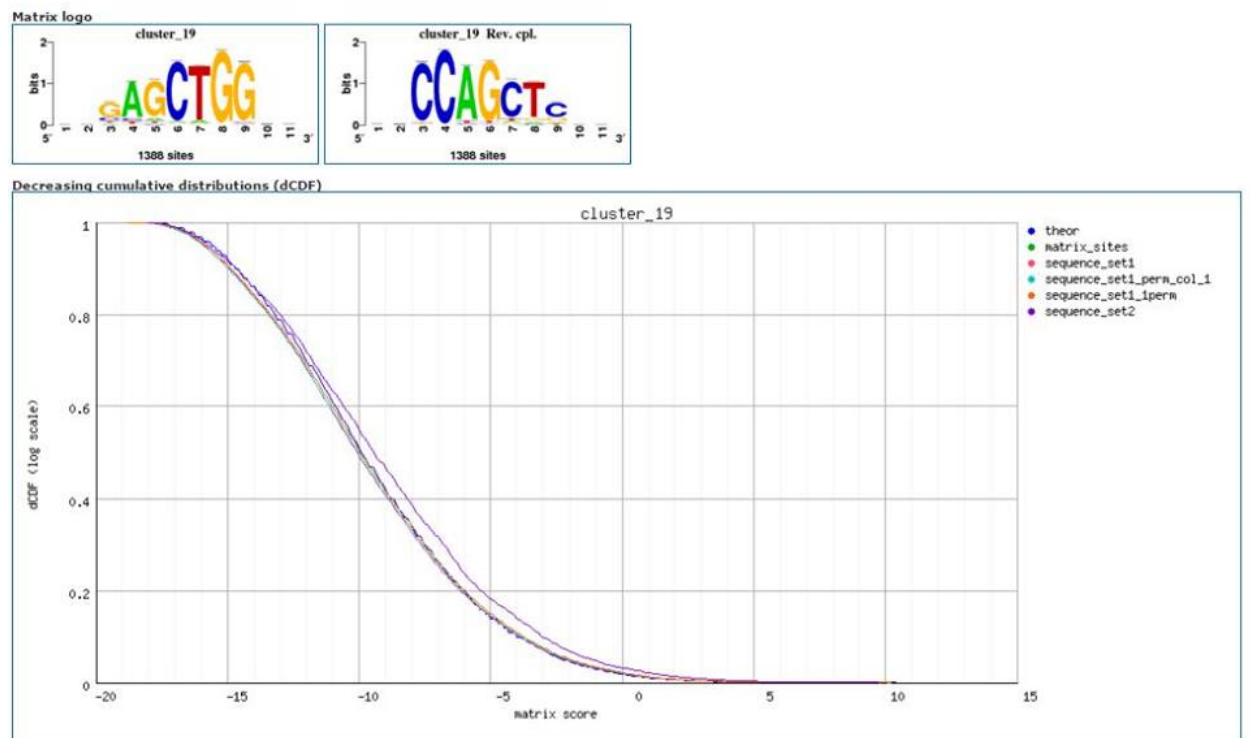


Figura 3.9.1



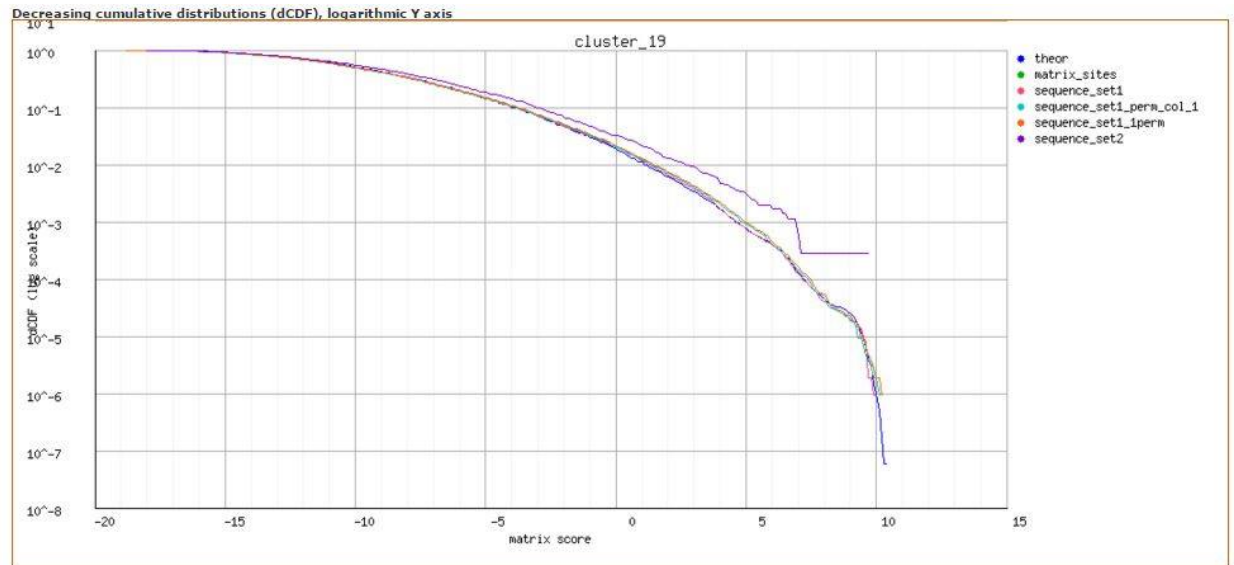


Figura 3.9.2

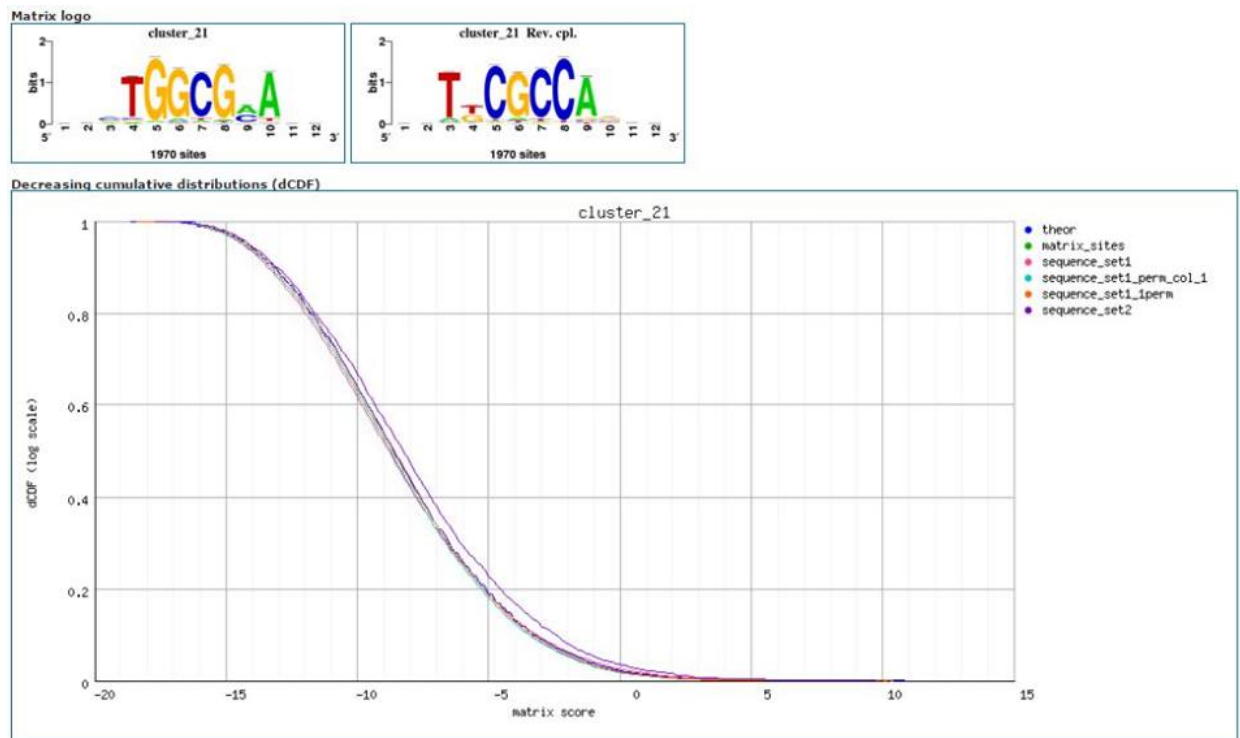


Figura 3.10.1

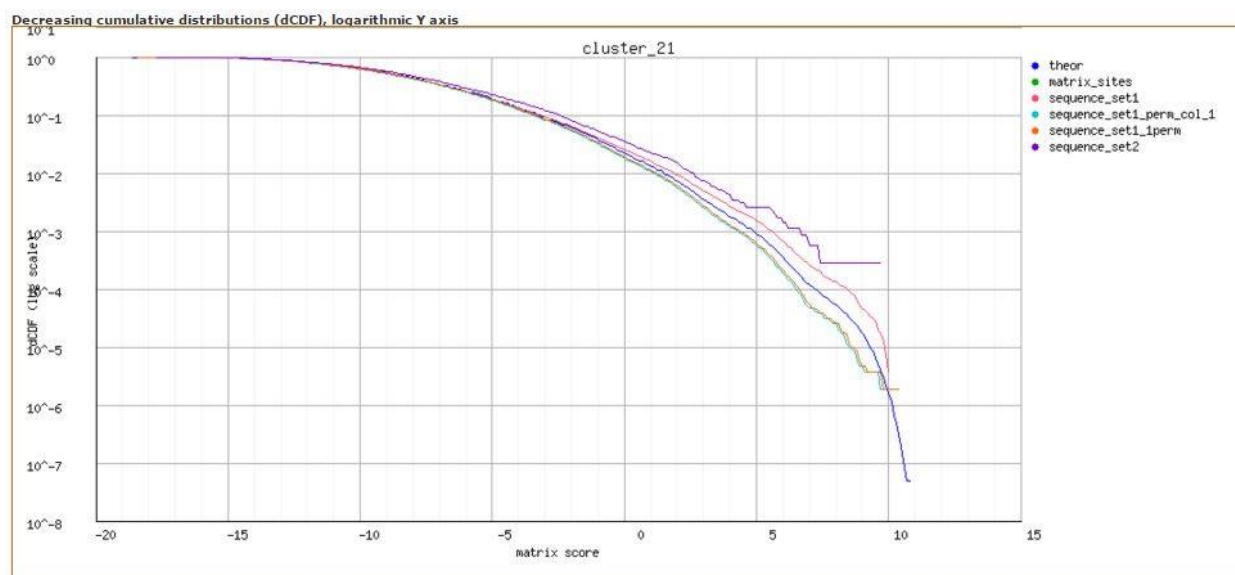


Figura 3.10.2

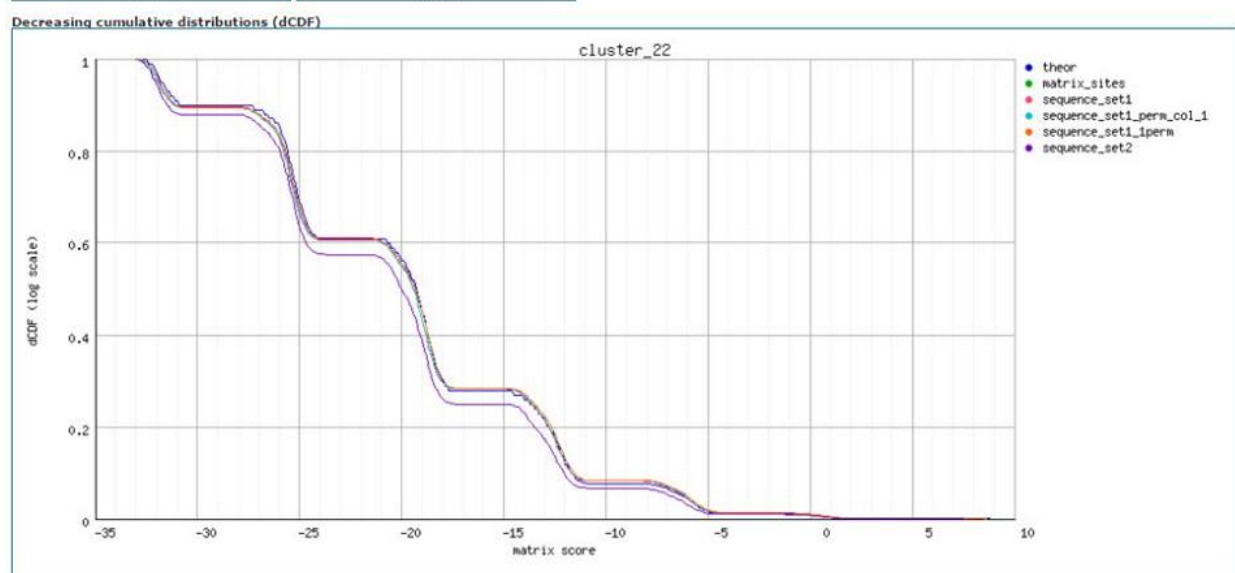
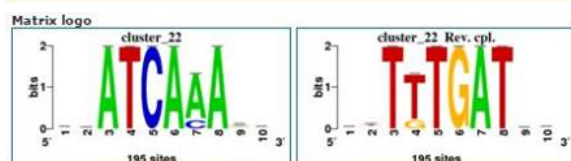


Figura 3.11.1

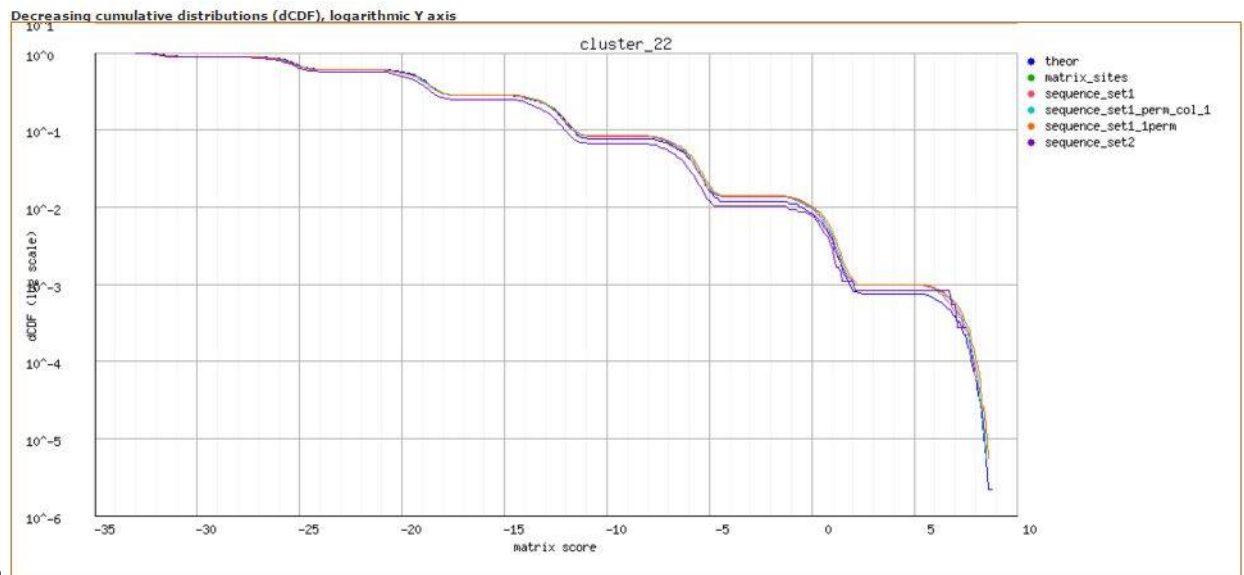


Figura 3.11.2

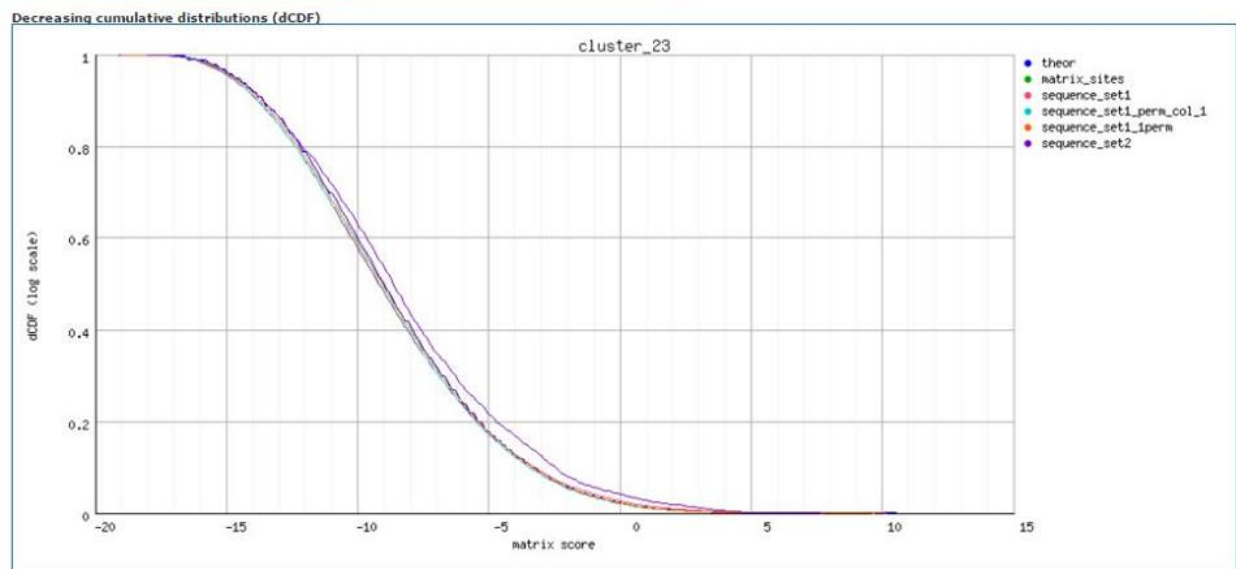
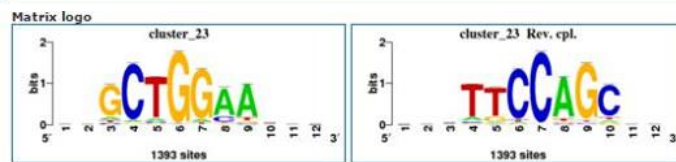


Figura 3.12.1

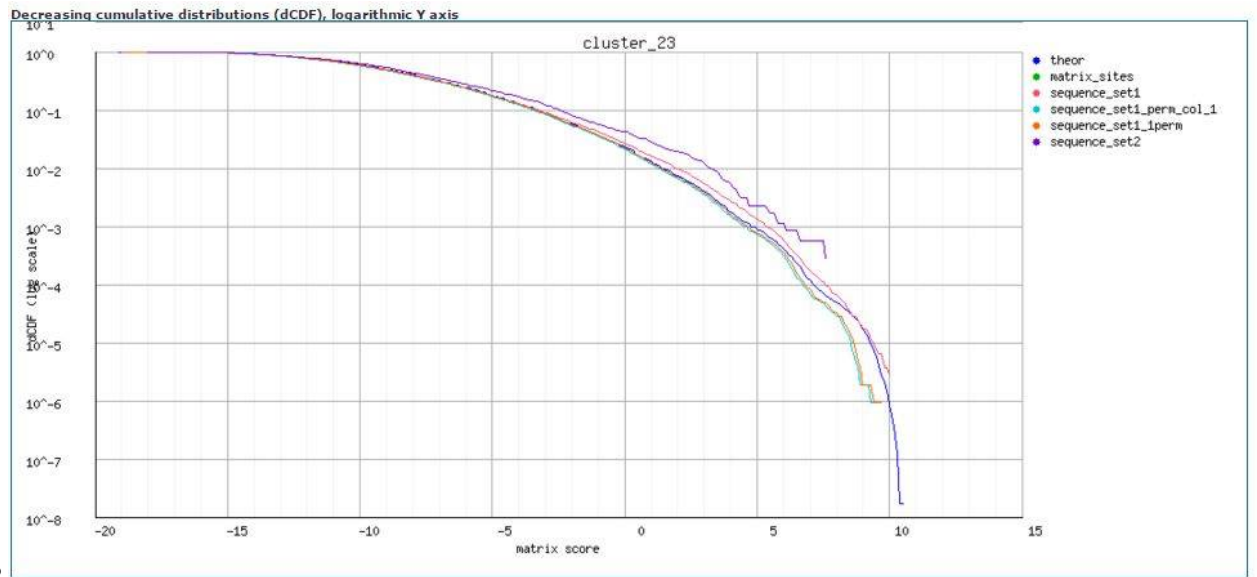


Figura 3.12.2

Golden standard

```

consensus      tTGatywayATCAA
consensus.rc    TTGATRTWRATCAA

Matrix parameters
Columns        14
Rows           4
Alphabet        a|c|g|t
Prior           a:0.25|c:0.25|g:0.25|t:0.25
program         tab
matrix.nb       1
min.prior       0.25
alphabet.size   4
max.bits        2
total.information 7.41971
information.per.column 0.529979
max.possible.info.per.col 1.38629
consensus.strict tTGattaatATCAA
consensus.strict.rc TTGATATTAATCAA
consensus.IUPAC  tTGatywayATCAA
consensus.IUPAC.rc TTGATRTWRATCAA
consensus.regexp tTGat[ct][at]a[ct]ATCAA
consensus.regexp.rc TTGAT[AG]T[AT][AG]ATCAA
residues.content.crude.freq a:0.3793|c:0.1684|g:0.1327|t:0.3197
G+C.content.crude.freq 0.30102
residues.content.corrected.freq a:0.3777|c:0.1693|g:0.1340|t:0.3189
G+C.content.corrected.freq 0.303361
min(P(S|M)) 1.08205e-23
max(P(S|M)) 0.00186984
proba_range 0.00186984
Wmin -33.48
Wmax 13.13
Wrange 46.61

```

Figura 4.1

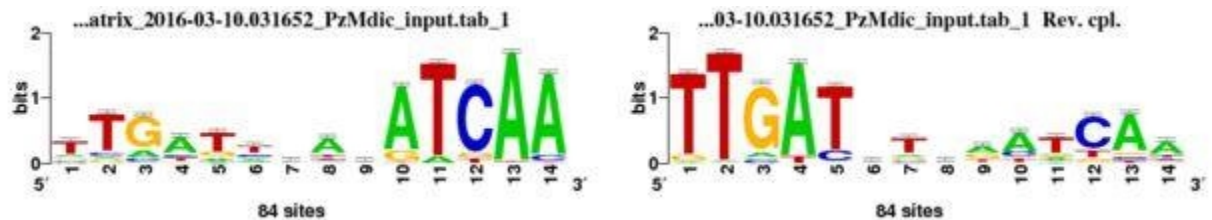


Figura 4.2

## Referencias

- [1] J Castro-Mondragon, C Rioualen, B Contreras-Moreira, J van Helden (2016) [RSAT::Plants: motif discovery in ChIP-seq peaks of plant genomes](#).
- [2] Illumina. [Precise analysis of DNA-protein binding sequences: Combining chromatin immunoprecipitation with NGS for genome-wide surveys of gene regulation](#).
- [3] [Regulatory Sequence Analysis Tools \(RSAT\)](#)

- A Medina-Rivera, M Defrance, O Sand, C Herrmann, J Castro-Mondragon, J Delerce, S Jaeger,

C Blanchet, P Vincens, C Caron, DM Staines, B Contreras-Moreira, M Artufel, L Charbonnier-Khamvongsa, C Hernandez, D Thieffry, M Thomas-Chollier, J van Helden J (2015) RSAT 2015: Regulatory Sequence Analysis Tools . Nucleic Acids Res.

- M Thomas-Chollier, M Defrance, A Medina-Rivera, O Sand, C Herrmann, D Thieffry, van Helden (2011) RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Res.
- M Thomas-Chollier, O Sand, JV Turatsinze, R Janky, M Defrance, E Vervisch, S Brohee, J van Helden (2008) RSAT: regulatory sequence analysis tools. Nucleic Acids Res.
- J van Helden (2003) Regulatory sequence analysis tools. Nucleic Acids Res.

[4] SG Landt, GK Marinov, A Kundaje, et al. (2012) [ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia](#). Genome Research 22(9): 1813-1831.

---