

Identificación de regiones diferencialmente metiladas utilizando un modelo oculto de Markov (HMM-DM)

Jenny Johana Gallo Franco

Pontificia Universidad Javeriana Cali

Email: jennygallof@javerianacali.com

Seminario permanente FIC, 2019

Contenido

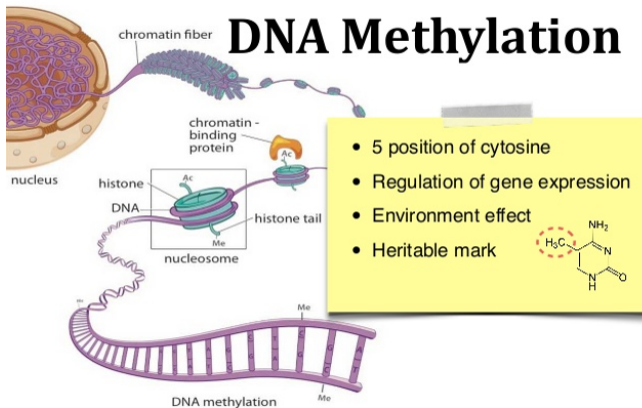
- 1 La epigenética en la respuesta a estrés
- 2 Modelo oculto de Markov
- 3 Modelo oculto de Markov en epigenética
- 4 Uso del programa HMM-DM en la estimación de DMRs
- 5 Resultados-Ejercicio práctico

Epigenética

La **epigenética** es el estudio de cambios hereditarios estables en la expresión génica sin alteraciones en la secuencia de ADN. (Conrad Waddington).

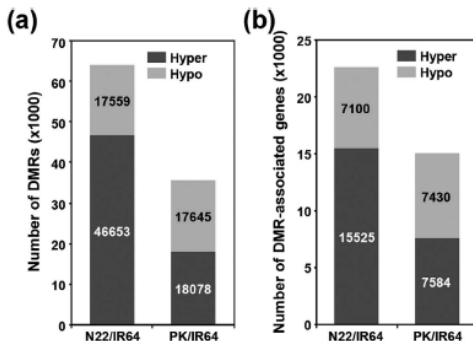
Los mecanismos subyacentes de la regulación epigenética incluyen la **metilación del ADN**, las **modificaciones y variantes de histonas**, **remodelación de la cromatina** y la regulación de la expresión génica mediada por **ARN no codificante**.

Epigenética



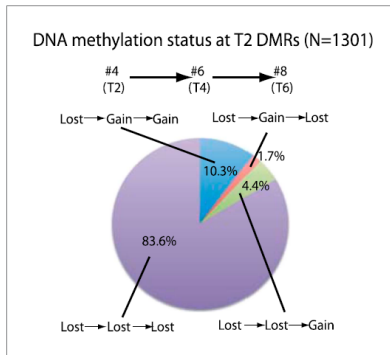
Respuesta a condiciones de estrés

Varios estudios han demostrado el papel crucial de la metilación del ADN en diversos procesos de desarrollo de las plantas y mas recientemente en respuesta a condiciones de estrés.



Respuesta a condiciones de estrés

Varios estudios han demostrado el papel crucial de la metilación del ADN en diversos procesos de desarrollo de las plantas y mas recientemente en respuesta a condiciones de estrés.

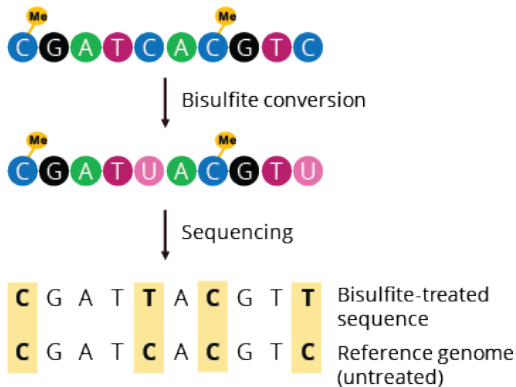


Que estamos haciendo en ÓMICAS

Caracterizar los patrones de metilación de variedades de arroz con respuesta contrastante a estrés por aluminio, a través de secuenciación de genoma completo por bisulfito.

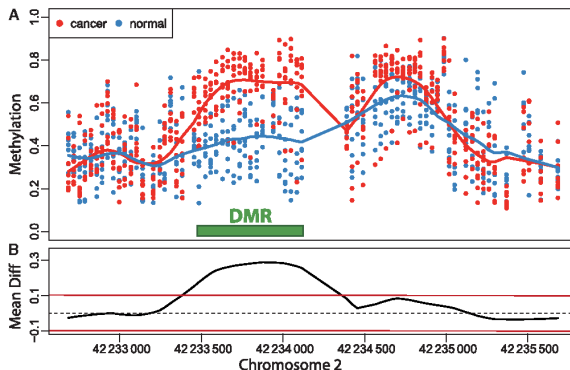


Secuenciación bisulfito



Regiones diferencialmente metiladas (DMRs)

Cómo metrica para determinar las diferencias en la metilación de plantas con diferentes capacidades de respuesta a estrés se utilizan las regiones diferencialmente metiladas (DMRs)



Estimación de Regiones diferencialmente metiladas

La identificación de las DMRs ha sido abordada utilizando dos pasos generales:

- Se detectan las citosinas diferencialmente metiladas (DMCs) mediante la comparación de los resultados de alineación entre muestras
- Los DMCs en las posiciones vecinas se agrupan como DMRs contiguos por un determinado criterio de distancia.

Ventajas de los métodos HMM para el cálculo de DMRs, específicamente el modelo HMM-DM (Yu & Su 2016):

- Puede identificar de forma robusta las DMRs con varias longitudes y singletons
- la variación de la metilación a través de muestras en el mismo grupo está bien explicada
- Es adecuado tanto para el genoma completo como para los datos de secuenciación de metilación de bisulfito dirigido.

Conceptos generales de las cadenas de Markov

Definición

Una cadena de Markov es un proceso estocástico a tiempo discreto $\{X_n : n = 0, 1, \dots\}$ y que satisface la propiedad de Markov, esto es, para cualquier entero $n \geq 0$, y para cualesquiera estados $x_0, \dots, x_{n+1}$, se cumple:

$$P(X_{n+1} | X_0, \dots, X_n) = p(X_{n+1} | X_n)$$

La distribución de probabilidad del estado del proceso al tiempo futuro $n + 1$ depende únicamente del estado del proceso al tiempo n , y no depende de los estados en los tiempos pasados $0, 1, \dots, n - 1$.

Conceptos generales de las cadenas de Markov

Definición

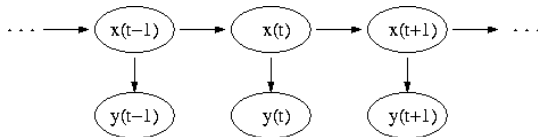
Una cadena de Markov es un proceso estocástico a tiempo discreto $\{X_n : n = 0, 1, \dots\}$ y que satisface la propiedad de Markov, esto es, para cualquier entero $n \geq 0$, y para cualesquiera estados $x_0, \dots, x_{n+1}$, se cumple:

$$P(X_{n+1} | X_0, \dots, X_n) = p(X_{n+1} | X_n)$$

La distribución de probabilidad del estado del proceso al tiempo futuro $n + 1$ depende únicamente del estado del proceso al tiempo n , y no depende de los estados en los tiempos pasados $0, 1, \dots, n - 1$.

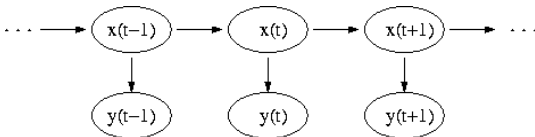
Modelo oculto de Markov

- El modelo oculto de Márkov, es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos.
- El objetivo es determinar los parámetros desconocidos (u ocultos, de ahí el nombre) de dicha cadena a partir de los parámetros observables.



Modelo oculto de Markov

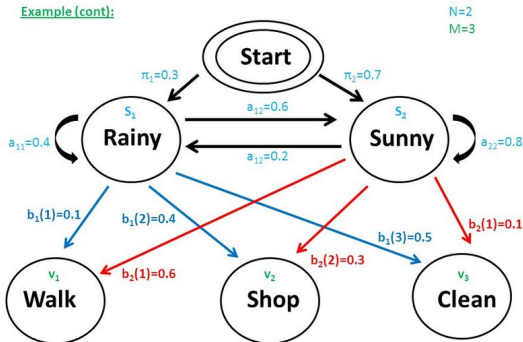
- El modelo oculto de Márkov, es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos.
- El objetivo es determinar los parámetros desconocidos (u ocultos, de ahí el nombre) de dicha cadena a partir de los parámetros observables.



Ejemplo de un modelo oculto de Markov

Hidden Markov Model

Example (cont):



HMM-DM: Modelo oculto de Markov para identificación de DMRs

Para la identificación de regiones diferencialmente metiladas (DMRs) se han diseñado modelos de Markov ocultos asumiendo que los estados de metilación diferencial ocultos en cada sitio CG dependen del estado de metilación diferencial del sitio CG inmediatamente anterior.

Definición de los parámetros del modelo

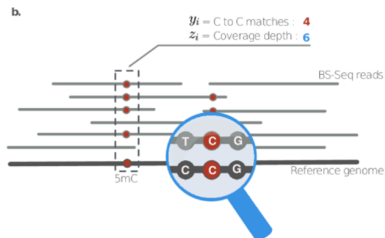
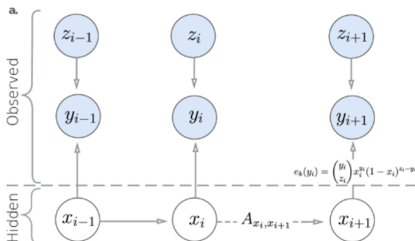
Observaciones

Matriz $P \times L$ de los valores de metilación para las muestras P en los sitios L de CG, $O = \{o_1, o_2, \dots, o_p, \dots, o_p\}$ con
 $o_p = \{o_{p,1}, o_{p,2}, \dots, o_{p,l}\}$

Cada observación varia entre 0 y 1, donde la señal de metilación se calcula como el ratio entre el número de lecturas con una C metilada y el número de lecturas que cubren este sitio CG

Definición de los parámetros del modelo

Comparación de los niveles de metilación.



Definición de los parámetros del modelo

Estados ocultos

$h_i \rightarrow$ Estado oculto de metilación diferencial

$$P(h_1 = \text{Hiper}) = P(h_1 = \text{IM}) = P(h_1 = \text{Hipo}) = 1/3$$

$$P(o_{1,1}, \dots, o_{L,p}) = P(o_{1,1}, \dots, o_{L,p} | h_1, \dots, h_L) \cdot P(h_1, \dots, h_L)$$

$$= \left[\prod_{p=1}^p \prod_{i=1}^L P(o_{pi} | h_i) \right] \cdot \left[P(h_1) \prod_{i=2}^L P(h_i | h_{i-1}) \right]$$

Definición de los parámetros del modelo

Probabilidades de transición

Describen la probabilidad de transferencia entre un estado de metilación diferencial a otro entre dos sitios CG consecutivos.

$$\text{Donde, } P(h_i|h_{i-1}) = t_{h_{i-1},h_i}, \quad i = 1, \dots, L$$

El vector θ denota las probabilidades de transición entre dos estados

$$\theta^t = \{t_{j,k}\} = \{t_{1,1}, t_{1,2}, t_{1,3}, t_{2,1}, t_{2,2}, t_{2,3}, t_{3,1}, t_{3,2}, t_{3,3}\}$$

Definición de los parámetros del modelo

Probabilidades de Emisión

Modelan la probabilidad de observar un nivel de metilación en un sitio CG dado un estado de metilación diferencial.

La distribución de $O_{p,i}$ condicional en h_i es:

$$O_{p,i}|h_i, \theta_i^e \sim \begin{cases} \beta(a_{i1}, 1) \text{ for } h_i = \text{Hiper}, & p = 1, 2, \dots, M \\ \beta(1, a_{i2}) \text{ for } h_i = \text{Hiper}, & p = M + 1, M + 2, \dots, P \\ \beta(a_{i3}, a_{i4}) \text{ for } h_i = \text{EM}, & p = 1, 2, \dots, M, M + 1, \dots, P \\ \beta(1, a_{i5}) \text{ for } h_i = \text{Hypo}, & p = 1, 2, \dots, M \\ \beta(a_{i6}, 1) \text{ for } h_i = \text{Hypo}, & p = M + 1, M + 2, \dots, P \end{cases}$$

Donde, $\theta_i^e = (a_{i1}, a_{i2}, a_{i3}, a_{i4}, a_{i5}, a_{i6})$

Estimación de los parámetros del modelo

Para estimar los parámetros de HMM-DM, se utilizan métodos bayesianos dándole a prioris a los parámetros desconocidos y utilizando las probabilidades posteriores. Todos los parámetros son muestreados de sus probabilidades condicionales.

Teorema de Bayes en términos de funciones de probabilidad

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

Estimación de los estados de metilación diferencial

Los tres estados de metilación diferencial ocultos son estimados por el **método de Markov Chain Monte Carlo (MCMC)**, específicamente utilizando el **muestreador de Gibbs** de su distribución de probabilidad condicional.

Para cada sitio de GC, se proporciona una probabilidad posterior para cada uno de los tres estados posibles, lo que muestra la **posibilidad de que el sitio de CG sea uno de los tres estados**. El estado con mayor probabilidad posterior se asigna como estado del sitio CG i .

Identificación de las regiones diferencialmente metiladas

Los sitios adyacentes de CG-DM se agrupan en **regiones teniendo en cuenta su estado de metilación diferencial**, la distancia y la cobertura de secuenciación. Solo los sitios de CG con los mismos estados se pueden incluir en la misma región. Por lo tanto, este método genera dos tipos de regiones DM, regiones hipermetiladas y regiones hipometiladas.

Generalidades de HMM-DM

La identificación de regiones CG diferencialmente metiladas utilizando el paquete `HMM.DM()` se realiza en 4 pasos:

- Control de calidad basado en cobertura
- Identificación de sitios CG metilados diferencialmente (MD) usando el método HMM-DM
 - Estimar los estados de metilación diferencial (Hipermetilado (Hiper) igualmente metilados (IM) e hipometilados (Hipo) para todos los sitios CG con HMM
 - Filtrar los sitios MD (Hiper e Hipo)

Generalidades de HMM-DM

La identificación de regiones CG diferencialmente metiladas utilizando el paquete `HMM.DM()` se realiza en 4 pasos:

- Control de calidad basado en cobertura
- Identificación de sitios CG metilados diferencialmente (MD) usando el método HMM-DM
 - Estimar los estados de metilación diferencial (Hipermetilado (Hiper) igualmente metilados (IM) e hipometilados (Hipo) para todos los sitios CG con HMM
 - Filtrar los sitios MD (Hiper e Hipo)

Generalidades de HMM-DM

La identificación de regiones CG diferencialmente metiladas utilizando el paquete `HMM.DM()` se realiza en 4 pasos:

- Control de calidad basado en cobertura
- Identificación de sitios CG metilados diferencialmente (MD) usando el método HMM-DM
 - Estimar los estados de metilación diferencial (Hipermetilado (Hiper) igualmente metilados (IM) e hipometilados (Hipo) para todos los sitios CG con HMM
 - Filtrar los sitios MD (Hiper e Hipo)

Generalidades de HMM-DM

La identificación de regiones CG diferencialmente metiladas utilizando el paquete `HMM.DM()` se realiza en 4 pasos:

- Organizar los DM CG filtrados en regiones DM, teniendo en cuenta sus estados DM, Distancia entre CGs y probabilidades posteriores.
- Definir las regiones diferencialmente metiladas

Generalidades de HMM-DM

La identificación de regiones CG diferencialmente metiladas utilizando el paquete `HMM.DM()` se realiza en 4 pasos:

- Organizar los DM CG filtrados en regiones DM, teniendo en cuenta sus estados DM, Distancia entre CGs y probabilidades posteriores.
- Definir las regiones diferencialmente metiladas

Generalidades de HMM-DM

Como caso de estudio se evaluaron ocho muestras provenientes de dos tratamientos contrastantes (4 y 4) con un total de 5811 sitios CG, correspondientes a un fragmento del genoma completo.

Generalidades de HMM-DM

Archivos de entrada para el programa HMM-DM

total.reads.txt

Posición	Muestra1	Muestra2	Muestra3	Muestra4	Muestra5	Muestra6	Muestra7	Muestra8
516596	1	0	0	0	0	0	0	0
516626	1	0	0	0	0	0	0	0
528534	0	0	1	0	1	0	0	0
528569	0	0	1	0	2	0	0	0
533879	0	0	0	0	0	0	1	0
534581	0	0	0	0	0	0	9	0
534602	0	0	0	0	0	0	9	0
534606	0	0	0	0	0	0	9	0
534613	0	0	0	0	0	0	9	0
534615	0	0	0	0	0	0	9	0
535890	50	30	67	291	27	65	149	101
535893	50	30	68	291	27	65	150	101
535897	50	30	68	291	27	65	150	101
535899	50	30	68	291	27	65	150	101
535921	50	30	68	289	26	65	149	101
535924	50	28	68	290	27	64	149	101
535958	69	45	31	105	17	61	175	60
535961	69	45	31	105	17	61	175	60
535965	69	45	31	105	17	61	175	60
535989	69	45	31	104	17	61	172	60

Generalidades de HMM-DM

Archivos de entrada para el programa EHHM-DM

meth.reads.txt

Posición	Muestra1	Muestra2	Muestra3	Muestra4	Muestra5	Muestra6	Muestra7	Muestra8
516596	1	0	0	0	0	0	0	0
516626	0	0	0	0	0	0	0	0
528534	0	0	1	0	1	0	0	0
528569	0	0	1	0	2	0	0	0
533879	0	0	0	0	0	0	0	0
534581	0	0	0	0	0	0	1	0
534602	0	0	0	0	0	0	0	0
534606	0	0	0	0	0	0	0	0
534613	0	0	0	0	0	0	0	0
534615	0	0	0	0	0	0	0	0
535890	50	30	66	291	27	65	149	100
535893	49	26	65	284	25	65	142	98
535897	46	30	59	279	27	63	116	95
535899	50	30	68	291	27	65	150	101
535921	48	28	64	277	23	65	143	95
535924	50	22	67	282	26	61	137	100
535958	69	45	31	105	17	61	175	60
535961	68	44	30	101	17	58	174	59
535965	68	43	31	105	16	60	169	57
535989	68	41	31	100	17	56	132	56

Resultados

mC.matrix.txt Niveles de metilación para un total de 5811 sitios CG utilizados en este trabajo.

all.CG.txt Estados de metilación para los 5811 sitios CG evaluados.

DM.CG.txt Estados de metilación diferencial para los 201 sitios identificados como CG diferencialmente metilados.

DMRs.txt Detalles de los 71 DMRs identificados.

De los 201 sitios CG identificados como diferencialmente metilados 62 fueron hipermetilados y solamente 9 fueron hipometilados.

Resultados

- La longitud promedio de estas DMRs fue de 23 pb, con un valor mínimo de 1pb y un valor máximo de 145 pb.
- El promedio de las probabilidades posteriores de las CGs dentro de las regiones.

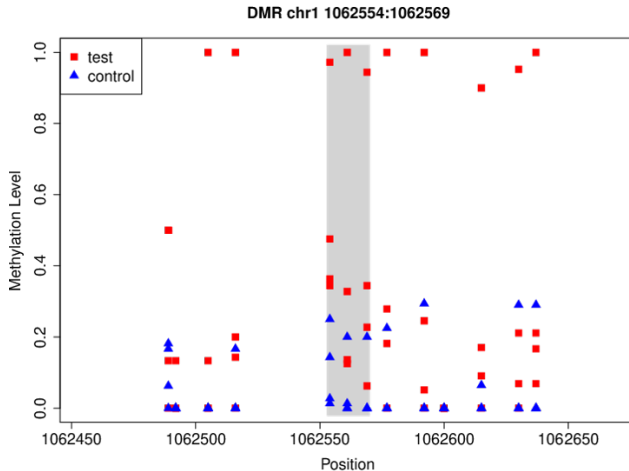
Una de las ventajas mas importantes de los métodos HMM es que se pueden obtener DMRs de longitud variable, sin pre-establecer tamaños poco objetivos.

Resultados

DMRs.txt

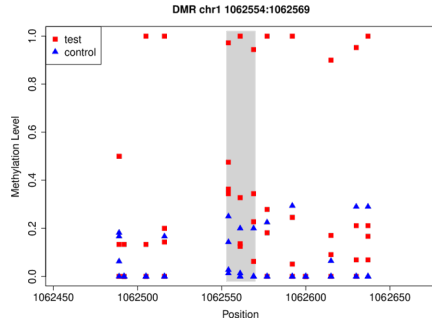
chr	start	end	len	DM	num.CG	total.CG	meanCov.test	meanCov.control	meanDiff.mC	meanPost
chr1	995053	995053	1	hyper	1	1	70.5	53.75	0.6325	0.9667
chr1	997296	997300	5	hyper	2	2	12.25	11.25	0.3883	0.7
chr1	997479	997495	17	hyper	5	5	8.05	23.8	0.465	0.7733
chr1	997716	997804	89	hyper	7	7	73.89	39.04	0.4941	0.9905
chr1	997954	998098	145	hyper	6	7	60.83	34.88	0.5283	0.9611
chr1	1048521	1048521	1	hyper	1	1	8.75	11.25	0.4884	0.7333
chr1	1050334	1050374	41	hyper	4	4	67.06	56.56	0.5509	0.9417
chr1	1058351	1058351	1	hyper	1	1	94.5	71.75	0.3862	0.6333
chr1	1061913	1062006	94	hyper	5	7	7.85	8.25	0.5109	0.88
chr1	1062332	1062420	89	hyper	10	10	40.05	52.48	0.456	0.93
chr1	1062554	1062569	16	hyper	3	3	43.25	42	0.3731	0.7667
chr1	1062709	1062709	1	hyper	1	1	34.5	31.5	0.3567	0.5667
chr1	1063404	1063404	1	hyper	1	1	62.75	38.75	0.6428	1

Resultados



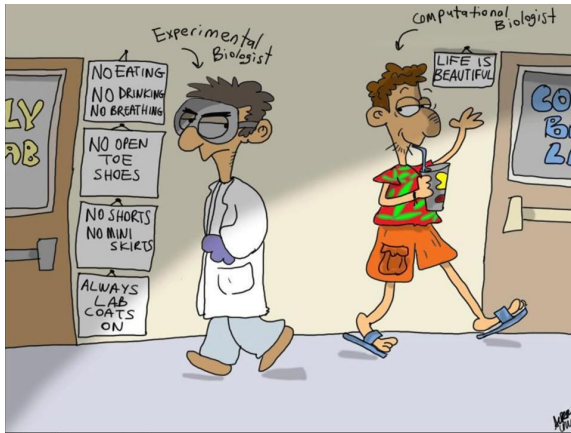
Resultados

chr	start	end	len	DM	num.CG	total.CG
chr1	995053	995053	1	hyper	1	1
chr1	997296	997300	5	hyper	2	2
chr1	997479	997495	17	hyper	5	5
chr1	997716	997804	89	hyper	7	7
chr1	997954	998098	145	hyper	6	7
chr1	1048521	1048521	1	hyper	1	1
chr1	1050334	1050374	41	hyper	4	4
chr1	1058351	1058351	1	hyper	1	1
chr1	1061913	1062006	94	hyper	5	7
chr1	1062332	1062420	89	hyper	10	10
chr1	1062554	1062569	16	hyper	3	3
chr1	1062709	1062709	1	hyper	1	1
chr1	1063404	1063404	1	hyper	1	1



Conclusiones

- Los cambios de los patrones de metilación diferencial en distancias cortas pueden ser capturados.
- HMM-DM permite que los usuarios se beneficien completamente de la resolución de medición de metilación de un solo CG proporcionada por las tecnologías de secuenciación BS de genoma completo.
- El HMM utilizado por HMM-DM puede manejar regiones genómicas con sitios CG densos o dispersos



GRACIAS...

Referencias

Rincón, L. 2011. Introducción a los procesos estocásticos. Departamento de Matemáticas Facultad de Ciencias UNAM. 1-29.

Shafi, A., Mitrea, C., Nguyen, T., & Draghici, S. (2017). A survey of the approaches for identifying differential methylation using bisulfite sequencing data. Briefings in bioinformatics, 19(5), 737-753.

Shokoohi, F., Stephens, D. A., Bourque, G., Pastinen, T., Greenwood, C. M., & Labbe, A. (2019). A hidden markov model for identifying differentially methylated sites in bisulfite sequencing data. Biometrics, 75(1), 210-221.

Yu, X., & Sun, S. (2016). HMM-DM: identifying differentially methylated regions using a hidden Markov model. Statistical applications in genetics and molecular biology, 15(1), 69-81.

Repositorio de Github <https://github.com/xy39/HMM-DM>