Análisis integrador de datos de metilación y expresión génica para la detección de genes regulados por metilación

Mercedes Monte Serrano
Máster Universitario Bioinformática y Bioestadística

Alexandre Sánchez Pla

30 de Junio, 2016

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis integrador de datos de metilación y expresión génica para la detección de genes regulados por metilación
Nombre del autor:	Mercedes Monte Serrano
Nombre del consultor:	Alexandre Sánchez Pla
Fecha de entrega:	05/2016
Área del Trabajo Final:	Estadística Bioinformática
Titulación:	*Máster Universitario en Bioinformática y Bioestadística UOC-UB*

Resumen del Trabajo (máximo 250 palabras):

La metilación del ADN constituye uno de los mecanismos epigenéticos más importantes en la regulación de la expresión génica, asociado al silenciamiento de la expresión génica. Aberraciones epigenéticas como la hipermetilación están relacionadas con distintos tipos de cáncer, incluyendo en cáncer colorrectal (CRC). Como consecuencia de esta dualidad alta metilación/baja expresión y baja metilación/alta-baja expresión, los gráficos de los pares metilación-expresión de genes regulados por metilación presentan unos patrones en forma de L. La detección de esos patrones es la estrategia a seguir para seleccionar genes regulados por metilación mediante el método basado en información mutua condicional (CMI) y el método de regresión basada en Splines. El objetivo de este trabajo es analizar los puntos críticos de ambos métodos y estudiar el efecto de modificar sus parámetros y condiciones de selección con el fin de optimizarlos, utilizando datos de metilación y expresión génica disponibles en el repositorio de dominio público Gene Expression Omnibus. Además, se crea el paquete R *lpattern*, que incluye las funciones requeridas para ejecutar ambos métodos, y la aplicación interactiva basada en *Shiny* que permite visualizar el funcionamiento del paquete así como el efecto de variar los diferentes criterios sobre el listado de genes seleccionados. La identificación de estos genes regulados por metilación e implicados en el avance del CRC sería útil en la determinación de nuevas dianas para el desarrollo de fármacos.

Abstract (in English, 250 words or less):

DNA methylation is one of the most important epigenetic mechanisms for regulation of gene expression associated with the silencing of gene expression. Epigenetic aberrations as hypermethylation are related to different types of cancer, including colorectal cancer (CRC). As a result of this duality high methylation/low expression and low methylation/high-low expression, graphics from methylation-expression pairs of genes regulated by methylation show L-sapped patterns. The detection of these patterns is the strategy to select genes regulated by methylation carried out by the Selection method based on conditional mutual information (CMI) and selection method based on Spline regression. The aim of this work is to analyze the critical points of both methods and study the effect of changing the parameters and selection conditions in order to optimize them. For this purpose we use methylation and gene expression data from the public domain repository Gene Expression Omnibus. In addition, the R *lpattern* package which includes functions required to run both methods and the interactive application based on *Shiny* that displays the package as well as the effect of modifying different criteria on the list of selected genes. The identification of these genes regulated by methylation and involved in CRC would be useful in identifying new targets for drug development.

Palabras clave (entre 4 y 8):

Metilación, expresión, regulación, cáncer colorrectal, biomarcadores

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

La metilación del ADN constituye uno de los mecanismos epigenéticos más importantes en la regulación de la expresión génica. Por lo general, la metilación se da en las islas CpG, unas zonas ricas en citosina y guanina localizadas en la zona promotora de los genes. Está ampliamente aceptado que la metilación del ADN está asociada al silenciamiento de la expresión génica, de modo que si la isla CpG está metilada la expresión del gen es baja o nula, y si la isla CpG no está metilada, la expresión del gen no está reprimida y por tanto puede estar expresándose o no[1]. Este mecanismo se observa con frecuencia en la regulación de la expresión de genes en tejidos normales, pero aberraciones epigenéticas como la hipermetilación están relacionadas con distintos tipos de cáncer, incluyendo en cáncer colorrectal (CRC)[2]. La hipermetilación de promotores de genes supresores de tumores proporciona a la célula una ventaja proliferativa, siendo de vital importancia en el desarrollo del cáncer.
Existen diferentes aproximaciones diseñadas para detectar genes regulados por metilación[3,4]. En concreto, el Departamento de Estadística y la Unidad de Estadística y Bioinformática de la Universidad de Barcelona junto con el Departamento de Oncología Molecular-CIBBIM del Instituto de Investigación de la Vall d'Hebrón (VHIR) se han basado en la dualidad alta metilación/baja expresión y baja metilación/alta-baja expresión, que da lugar a gráficos en forma de L de los pares metilación-expresión de genes regulados por metilación[5], para desarrollar métodos de selección de este tipo de genes[6]. En el presente trabajo se analiza la metodología desarrollada por el grupo de investigación para la detección de los patrones en L que permiten seleccionar genes regulados por metilación y se proponen mejoras para la optimización de estos métodos. Además, se presenta el paquete R[7] *lpattern* que contiene las funciones necesarias para la ejecución de los métodos originales y sus mejoras, así como las instrucciones para la visualización del funcionamiento del paquete mediante una aplicación interactiva basada en *Shiny*[8] con el fin de mostrar el efecto que tiene la modificación de los distintos parámetros o criterios de selección sobre el listado final de genes seleccionados. Los datos de metilación y expresión génica que se utilizan en este trabajo se han obtenido de la base de datos de dominio público *Gene Expression Omnibus*.
El interés en conocer qué genes están regulados por metilación en tejidos aislados de pacientes con CRC radica en la necesidad de encontrar nuevas dianas para fármacos antitumorales, así como localizar nuevos biomarcadores para una detección temprana de la enfermedad.

1.2 Objetivos del Trabajo

- Conocer los métodos desarrollados por el grupo de investigación, analizarlos y mejorarlos para crear un paquete en R que permita ejecutar estos métodos de selección de patrones en L con suficiente fiabilidad. Para cumplir con este objetivo se deberá:

- Ejecutar en R el método de selección basado en la Información Mutua Condicional y el método de selección basado en regresión basada en Splines, así como aportar mejoras con el fin de optimizar la calidad los resultados obtenidos.
- Implementar en un paquete R los métodos con las mejoras destacadas para la selección de genes que siguen un patrón en L.
- Comparar los resultados obtenidos en ambos métodos así como con un listado de genes reconocidos como regulados por metilación en CRC.

- Crear una aplicación interactiva basada en *Shiny* que permita al usuario comprender los métodos y manipular los diferentes puntos críticos, de forma que pueda observar el efecto de estas modificaciones sobre el listado de genes seleccionados. Para ello se necesitará:
    - Ejecutar la aplicación sobre un set de datos de metilación y expresión de dominio público.
    - Comparar los resultados obtenidos al analizar los mismos datos con los distintos métodos.

1.3 Enfoque y método seguido

Para cumplir con los objetivos propuestos en el apartado anterior, el tutor del proyecto facilita el código en R que contiene las funciones necesarias para la ejecución de ambos métodos de selección, además de un set de datos para trabajar. Se trata de adaptar las funciones existentes con el fin de optimizar la selección de genes regulados por metilación.
El ajuste de los parámetros que se incluyen en los algoritmos se puede llevar a cabo mediante, al menos, dos formas diferentes:

a) Validación cruzada utilizando el paquete R *caret*, de gran utilidad en minería de datos ya que permite probar todas las combinaciones de valores de parámetros en lo que se llama una búsqueda en rejilla.
b) Asignación de valores de forma manual y comprobación de los resultados mediante análisis visual de los gráficos de los genes seleccionados.

Aunque mucho más atractiva y completa la opción a), la ausencia de conocimiento del paquete *caret* y el poco tiempo disponible para poder comprender su funcionamiento fueron los motivos por los cuales se escogió la opción b).

La creación del paquete R debía incluir las funciones básicas así como las mejoras propuestas, además del código necesario para poder ser llamado a través de la aplicación *Shiny* y así poder mostrar el funcionamiento de los métodos desarrollados.

1.4 Planificación del Trabajo

A continuación, la figura 1 y la tabla 1 incluyen los hitos y las tareas entregables a lo largo de este TFM.

Figura 1. Escala de tiempo.

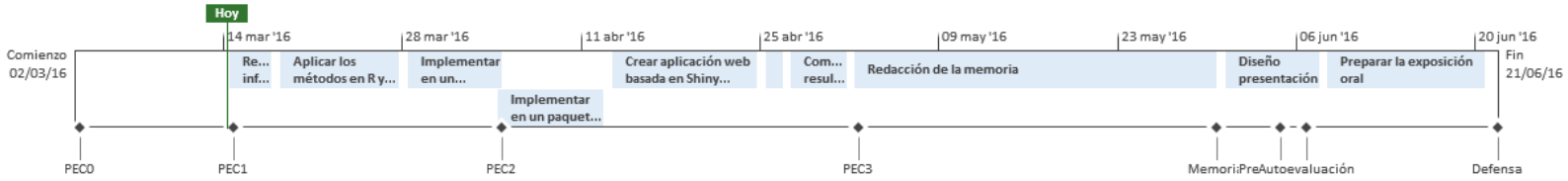


Tabla 1. Fechas de entrega y dedicación de hitos y tareas.

	Fechas		Dedicación	
Hitos y Tareas	**Inicio**	**Fin**	**horas**	**%**
PEC0 - Selección área TFM	24/02/2016	02/03/2016	3	1
PEC1 - Plan de trabajo	07/03/2016	14/03/2016	37,5	10
PEC2 - Desarrollo del trabajo Fase I	14/03/2016	04/04/2016	70,5	18
Recopilar información sobre los métodos	14/03/2016	17/03/2016	8	2
Aplicar los métodos en R y analizarlos	18/03/2016	27/03/2016	25	6
Implementar en un paquete R los métodos destacados (I)	28/03/2016	04/04/2016	37,5	10
PEC3 - Desarrollo del trabajo Fase II	04/04/2016	02/05/2016	128	34
Implementar en un paquete R los métodos destacados (II)	04/04/2016	12/04/2016	37,5	10
Crear una aplicación interactiva basada en *Shiny* mediante *RStudio*	13/04/2016	24/04/2016	50	13
Seleccionar la base de datos que contenga los datos de interés	25/04/2016	26/04/2016	3	1
Identificar los genes regulados por metilación para cada método y comparar	27/04/2016	01/05/2016	37,5	10
Memoria del trabajo final	02/05/2016	30/05/2016	75	20
Redactar la memoria	02/05/2016	30/05/2016	75	20
Presentación del trabajo	30/05/2016	06/06/2016	30	8
Diseñar de la presentación	30/05/2016	06/06/2016	30	8
Autoevaluación del trabajo	30/05/2016	06/06/2016	3	3
Completar cuestionario de autoevaluación	30/05/2016	06/06/2016	3	1
Defensa pública - Tribunal TFM	14/06/2016	21/06/2016	30	8
Preparar la exposición oral	08/06/2016	20/06/2016	30	8
			377	**100**

1.5 Breve sumario de productos obtenidos

- Tablas con los valores de sensibilidad de los métodos al utilizar diferentes combinaciones de valores asignados a los parámetros. Aunque este producto no estaba incluido en los objetivos del trabajo, la sensibilidad del método según las modificaciones propuestas ha sido de elevada importancia a la hora de escoger entre las diferentes opciones de mejora de los métodos.
- Dos paquetes en R:
    - lpattern.01: contiene las funciones proporcionadas por el profesor además de dos sets de datos de dominio público sobre los que poder aplicar las funciones.
    - lpattern.02: además de lo mencionado en el paquete lpattern.01, se incluye una nueva función (`InitialSelection2`) para llevar a cabo la selección inicial de genes que aumente la calidad de los genes seleccionados en el método de selección basado en Splines. También se incluyen las instrucciones que son llamadas a través de la aplicación interactiva.
- Aplicación interactiva basada en *Shiny* que permite evaluar el efecto de variar los diferentes parámetros sobre el listado de genes seleccionados por cada método.

1.6 Breve descripción de los otros capítulos de la memoria

- Selección y procesamiento de los datos a analizar: Se incluyen las dos bases de datos propuestas y el motivo de elección de una de ellas para la obtención de los datos a analizar, así como el procesamiento de los mismos.
- Optimización del método de selección basado en información mutua condicional (CMI): Incluye una revisión del método en su estado original, las modificaciones propuestas y un resumen de los resultados obtenidos con cada combinación ensayada.
- Optimización del método de selección basado Splines: Incluye una revisión del método en su estado original, las modificaciones propuestas y un resumen de los resultados obtenidos con cada combinación ensayada.
- Comparación de resultados obtenidos en ambos métodos: Incluye un gráfico que compara el total de genes coincidentes seleccionados por ambos métodos. También se comparan estos resultados con un listado de genes conocidos regulados por metilación obtenidos a partir de la bibliografía.
- Creación del paquete R *lpattern*: Se presenta el contenido incluido en las dos versiones creadas del paquete R *lpattern* y el motivo por el cual se decide crear la segunda versión.
- Creación de la aplicación interactiva basada en *Shiny*: con la intención de mostrar el funcionamiento del paquete *lpattern* se crea una aplicación interactiva que permita seleccionar genes regulados por metilación con cada uno de los métodos, modificando los parámetros y observando el efecto que tienen estas modificaciones sobre el listado final de genes seleccionado.

2. Resto de capítulos

2.1 Selección y procesamiento de los datos a analizar

En un principio se trabajó con los datos de metilación y expresión facilitados por el tutor para comprobar que al ejecutar el código se obtenían los mismos resultados. Tras entregar las diferentes actividades, fueron sugeridas modificaciones y se advirtió que los datos utilizados hasta el momento eran confidenciales, lo que implicó replantear la decisión escogida. Tras una intensa búsqueda de datos públicos analizados mediante métodos diferentes a los propuestos en este trabajo, no se encontró la información deseada y se decidió trabajar con datos públicos no analizados con anterioridad para este mismo propósito.

Se llevó a cabo una búsqueda en la base de datos de dominio público *The Cancer Genome Atlas* (TCGA)[9] de datos de metilación y expresión que estuviesen relacionados entre sí. Al no haber trabajado nunca con este repositorio, la obtención de datos llevó más tiempo del esperado y finalmente se optó por buscar directamente en la bibliografía una publicación que hiciese referencia a datos útiles para realizar este trabajo. Como alternativa, se decidió obtener los datos del repositorio público de datos genómicos *Gene Expression Omnibus* (GEO)[10]. Finalmente, se obtuvieron datos de metilación y expresión génica de tejidos de CRC[11]. Los datos de metilación se habían obtenido mediante hibridación con la plataforma *Illumina Infinium 27k Human Methylation Beadchip v1.2*, y estaban almacenados con el código GSE25062. Los datos de expresión se obtuvieron mediante tecnología *Illumina Ref-8 whole-genome expression BeadChip*, almacenados con el código GSE25070. Se seleccionaron los datos pertenecientes a los pacientes para los cuales existían tanto valores de metilación como de expresión. Para cada tipo de dato se creó un documento en el cual se iban incluyendo los datos de los pacientes por columnas. Una vez los dos archivos estuvieron completos, se eliminaron los genes que contenían observaciones con valores nulos, ya que impedían la correcta ejecución de las funciones. Tras el procesado se obtuvieron dos *data frames* con 25 observaciones para 11191 genes.

2.2 Optimización del método de selección basado en información mutua condicional (CMI)

2.2.1 Estado original del algoritmo

El algoritmo incluye los siguientes pasos:

(1) Asignación de los distintos valores posibles de t (*threshold*). t es el punto de corte a partir del cual se binariza la variable continua metilación de forma que, los valores de metilación $<t$ se consideran no metilados y los valores de metilación $>t$ se consideran metilados. El valor de t va de 0 a 1 porque es en este rango en el que se mueven los valores de metilación.

(2) Cálculo de del valor de CMI(t) para cada gen.
(3) Identificación del valor de *t* que origina el mínimo CMI para cada gen.
(4) Cálculo del ratio (r = mincMI(t)/cMI(0)) para cada gen.
(5) Cálculo de la expresión media en las observaciones que no están metiladas y expresión media en las observaciones que están metiladas para cada gen.
(6) Selección de los genes que cumplen las siguientes condiciones:
    a. ratio suficientemente bajo
    b. cMI(0) suficientemente alto
    c. expresión media a la izquierda de *t*> expresión media a la izquierda de *t*

En base a los pasos que componen el algoritmo, en este método se distinguen 4 parámetros:

- t: en el método original se incluyen 101 valores posibles de 0 a 1 con un paso de 0.01.
- h: este parámetro se encuentra incluido en la función que calcula los distintos valores de CMI en función de *t*, ya que el método kernel se utiliza para estimar la función de densidad en torno a los valores muestrales y el valor *h* representa el área de influencia que se le asigna a cada valor muestral. *h* se conoce también como parámetro de suavizado o ancho de ventana y está establecido por defecto en 0.3.
- r: en la selección original se seleccionan genes con un $r<0.25$.
- cMI(0): en la selección original se seleccionan genes con un cMI(0)>0.1.

Los valores asignados de *r* y cMI(0) se determinaron mediante permutación aleatoria[12].

2.2.2 Propuestas de mejora del algoritmo

La optimización de los métodos podía hacerse mediante validación cruzada utilizando el paquete R *caret* o modificando manualmente cada uno de los parámetros, ejecutando el método cada vez y comprobando los resultados visualmente a partir de los gráficos de los genes seleccionados. Como ya se mencionó en el apartado 1.3, la utilización del paquete *caret* se descartó por desconocimiento y falta de tiempo para poder poner a punto el método.

De este modo, además de los valores asignados en el método original, se propusieron las diferentes modificaciones para cada uno de los parámetros (tabla 2). Esta elección se hizo escogiendo valores cercanos a los ya establecidos por defecto.

Tabla 2. Modificaciones propuestas en los parámetros del método basado en CMI.

Parámetro	Valores
t	201 valores posibles, de 0 a 1 con un paso de 0.005
	101 valores posibles, de 0 a 1 con un paso de 0.01
	21 valores posibles, de 0 a 1 con un paso de 0.05
	11 valores posibles, de 0 a 1 con un paso de 0.1
h	0.2
	0.3
	0.4
r	<0.20
	<0.25
cMI(0)	>0.10
	>0.15
	>0.20

Combinando las diferentes posibilidades, el método se ejecutó 33 veces sobre los mismos sets de datos. Se obtuvieron 33 listas de genes junto con los gráficos de los pares metilación-expresión de los genes seleccionados.

2.2.3 Resultados

Tras ejecutar el método modificando los diferentes parámetros, se obtuvieron los gráficos de los pares metilación-expresión para cada gen seleccionado y, tras valoración visual, se contabilizó el total de genes seleccionados con verdadera forma de L. Este valor se utilizó para determinar la sensibilidad del método a la hora de detectar genes regulados por metilación (tabla 3).

Tabla 3. Sensibilidad del método en función de los valores establecidos para cada parámetro.

Combinación	h	r	cMI(0)	Nº genes seleccionados	Nº genes con patrón en L	Sensibilidad
1	0.3	<0.25	>0.10	75	15	20.00%
2	0.2	<0.25	>0.10	263	23	8.75%
3	0.4	<0.25	>0.10	29	5	17.24%
4	0.2	<0.20	>0.10	149	22	14.77%
5	0.3	<0.20	>0.10	43	10	23.26%
6	0.4	<0.20	>0.10	21	3	14.29%
7	0.2	<0.25	>0.15	155	21	13.55%
8	0.3	<0.25	>0.15	23	7	30.43%
9	0.4	<0.25	>0.15	3	0	0.00%
10	0.2	<0.20	>0.15	80	8	10.00%
11	0.3	<0.20	>0.15	11	1	9.09%
12	0.4	<0.20	>0.15	3	0	0.00%

En la tabla 3, la combinación 1 contiene los parámetros con los valores asignados por defecto. Con esta combinación se obtuvo un número bajo de genes con una sensibilidad del método también baja. Sin embargo, ninguna de las combinaciones evaluadas ofreció posibilidad de mejora del método. No se muestran los resultados obtenidos al modificar el valor de *t* con el fin de simplificar los resultados, ya que tampoco ofrecieron mejoras destacables en la sensibilidad del método.
Es posible que la estrategia de modificación manual de los parámetros no fuese la más apropiada para mejorar el método de selección basado en CMI o bien, que los valores probados no fuesen la mejor elección. Esta aproximación empírica incluía este riesgo.
A pesar de no conseguir el resultado esperado, fue la combinación 2 la escogida en los pasos posteriores. El motivo por el cual se seleccionaron estos valores fue porque aumentaba el total de genes con verdadero patrón en L, sacrificando el número de falsos positivos, con el fin de confirmar la selección mediante el segundo método de selección que se presenta a continuación.

2.3 Optimización del método de selección basado en Splines

2.3.1 Estado original del algoritmo

El algoritmo incluye los siguientes pasos:

(1) Filtrado inicial para seleccionar genes con una correlación Spearman negativa significativa.
(2) Ajuste de cada par expresión-metilación a una curva de regresión B-Spline cúbica.
(3) Clasificación de las curvas obtenidas en base a los coeficientes de los B-Splines resultantes.
(4) Selección de los *clusters* que responden a patrones en L.

En base a los pasos que componen el algoritmo, en este método se distinguen 2 partes: en la primera parte se realiza un filtrado inicial de forma que se escogen genes con una correlación negativa entre las variables continuas metilación y expresión con un p-valor < 0.05 y que comprenden rangos de metilación suficientemente amplios, mientras que en la segunda parte se calcula la curva de regresión basada en B-Splines de los genes seleccionados y se clasifican las curvas según los coeficientes de las mismas, con el fin de agrupar genes con patrones similares y seleccionar aquellos que sigan un patrón en forma de L.

La función `InitialSelection,` diseñada para realizar la selección inicial, contiene los siguientes parámetros: QInf=25, metInf, QSup=75, metSup, Adjust=FALSE, pAdj=0.05. De este modo, si QInf = 25, metInf = 0.33, QSup = 75, metSup = 0.66 y pAdj = 0.05, significa que se seleccionan los genes que tienen valores de metilación por debajo de 0.33 en el percentil 25, valores de metilación por encima de 0.66 en el percentil 75 y correlación negativa de Spearman con un p-valor inferior a 0.05.

2.3.2 Propuestas de mejora del algoritmo

Se proponen mejoras en las dos partes del algoritmo: filtrado inicial y clasificación de patrones según los coeficientes de los B-Splines. Al igual que en el método anterior, se valoró entre la posibilidad de utilizar el paquete R *caret* para optimizar los parámetros mediante validación cruzada o la manipulación manual de los parámetros. Además, se planteó la posibilidad de modificar completamente la función del filtrado inicial con el fin de aumentar la calidad de la selección de los genes. Para optimizar la clasificación de los patrones según los coeficientes de los B-Splines, se optó por probar modificaciones sobre las funciones ya existentes.

Finalmente, con respecto al filtrado inicial, se decidió llevar a cabo una asignación manual de los valores de los parámetros en las funciones del método original.
Al aplicar las condiciones del apartado anterior sobre los sets de datos, únicamente 4 genes fueron seleccionados para el cálculo de Splines. Por este motivo se decidió ser más permisivo aumentando el nivel de significación (tabla 4).

Tabla 4. Total de genes clasificados según el nivel de significación de la correlación negativa de Spearman.

pAdj	nº genes seleccionados
0.05	4
0.1	8
0.5	36

Tras analizar los gráficos resultantes, solo un 2,8% de los genes seleccionados presentaba un patrón en L. La evaluación del resultado obtenido, que no ofrecía ninguna mejora, junto con las diferentes correcciones sugeridas, exigieron un cambió de estrategia. Se optó por añadir una nueva función que sustituyese a la función `InitialSelection` y mejorase la calidad de la selección inicial de genes.

Se diseñó una nueva función para binarizar los datos de metilación y expresión de cada gen, de forma que se descartaran genes que tuviesen valores de alta metilación que correspondiesen con una alta expresión. Esta binarización se planteó de tres modos diferentes según la forma de calcular los puntos de corte:

- Estableciendo el mismo punto de corte para todos los genes.
- Calculando la media de los valores de la variable para cada gen.
- Calculando la mitad del valor máximo de la variable para cada gen.

A partir de estos modos de cálculo, se probaron las siguientes combinaciones para establecer los puntos de corte en las dos variables y dividir cada gráfico de pares metilación-expresión en 4 cuadrantes:

a) Media de la expresión y punto de corte de metilación prefijado.
b) Mitad del valor máximo de expresión y punto de corte de metilación prefijado.
c) Mitad del valor máximo de expresión y mitad del valor máximo de metilación.

Con cada una de estas combinaciones, se calculó el número de observaciones en tres de los cuatro cuadrantes:

- nSI: número de observaciones en el cuadrante superior izquierdo (baja metilación-alta expresión).
- nID: número de observaciones en el cuadrante inferior derecho (alta metilación-baja expresión).
- nSD: número de observaciones en el cuadrante superior derecho (alta metilación-alta expresión).

Además, para asegurar que el rango de valores de metilación del gen era suficientemente amplio como para poder observar el comportamiento típico de un gen regulado por metilación, la nueva función incluía el cálculo de:

- met.max: valor máximo de los datos de metilación para cada gen.
- dif.met: diferencia entre el valor máximo y el mínimo de los datos de metilación para cada gen.

La selección de genes se ejecutó combinando los diferentes valores de los parámetros calculados a partir de las distintas opciones (tabla 5), llevándose a cabo 66 ejecuciones del método, 33 de ellas sobre el bruto de genes, independientemente de la correlación entre las dos variables, el 33 sobre el listado de genes resultante al aplicar la función `InitialSelection` con los los siguientes parámetros: QInf=0, QSup=100, metInf=0.25, metSup=0.75, pAdj=0.5.

Tabla 5. Valores asignados a los parámetros de la nueva función de filtrado inicial.

Método calculo cuadrantes	Punto corte metilación	nSI	nID	nSD	met.max	dif.met
a	0.5 0.6 0.66	>3	>3 -	- <1 <2	>0.25	>0.40
b	0.5 0.6 0.66	>3	>3 -	- <1 <2	>0.25	>0.40
c	met.max/2	>3	>3 >2	- <1	>0.25 >0.05	>0.35 >0.40

Con respecto a la optimización de la clasificación de patrones según los coeficientes de los Splines, el objetivo era hacer una clasificación más eficiente para seleccionar mejor los componentes de los *clusters* con forma de L, por lo que se propuso utilizar P-Splines en lugar de B-Splines. Al utilizar esta nueva estrategia aumentaría el número de rectas calculadas para ajustarse a los datos de cada gen y por tanto el número de coeficientes para llevar a cabo el agrupamiento.

2.3.3 Resultados

Los resultados obtenidos con cada método y con los diferentes valores de los parámetros se evaluaron mediante la visualización de los gráficos de los genes seleccionados, calculando el porcentaje de verdaderos positivos para cada combinación. Los niveles más altos de sensibilidad se obtuvieron con el cálculo de los puntos de corte calculados a partir de la mitad del valor máximo de expresión y la mitad del valor máximo de metilación utilizando los 11191 genes, llegando a obtenerse hasta un nivel de sensibilidad del 92.77% sobre un 83 de genes seleccionados. Las condiciones de selección fueron:

- nSI>3
- nID>2
- nSD<1
- met.max>0.05
- dif.met>0.35

Con el fin de simplificar los resultados obtenidos, no se han incluido en este trabajo los resultados de sensibilidad de las diferentes combinaciones evaluadas, con resultados de sensibilidad inferiores al 58% en todos los casos.

En las figuras 2 y 3 se muestran los patrones al clasificar las curvas obtenidas para cada gen en función de las medias de los coeficientes o de las medianas, respectivamente. Se aprecia como todos los patrones responden a un comportamiento típico de genes regulados por metilación.

Figura 2. Patrones en L en base a las medias de los coeficientes de regresión basada en B-Splines.

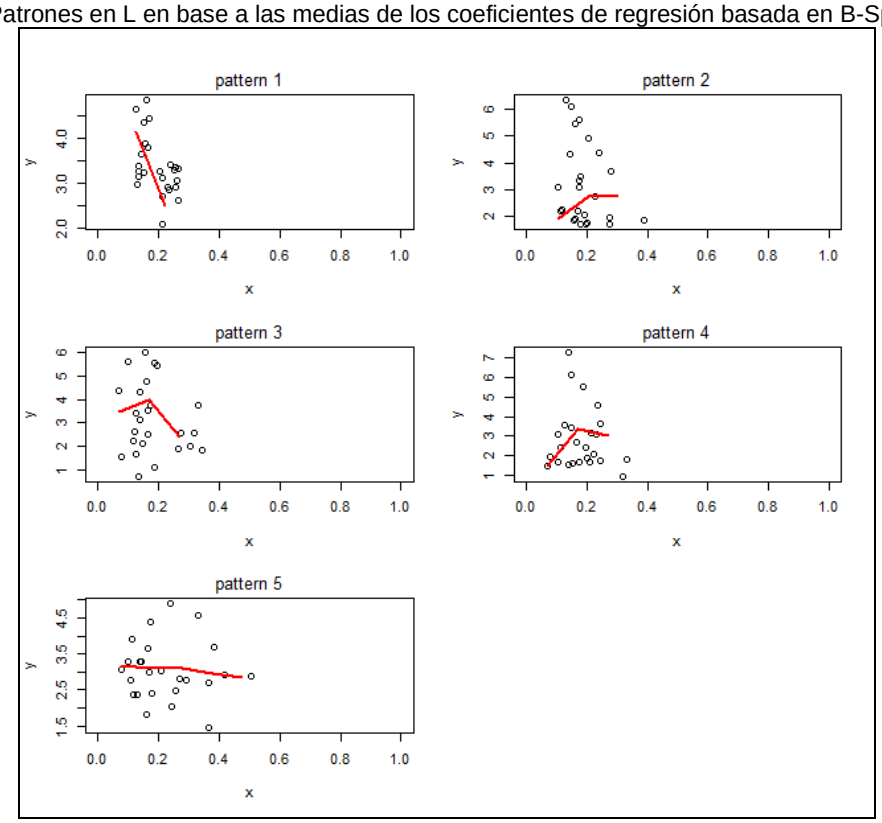
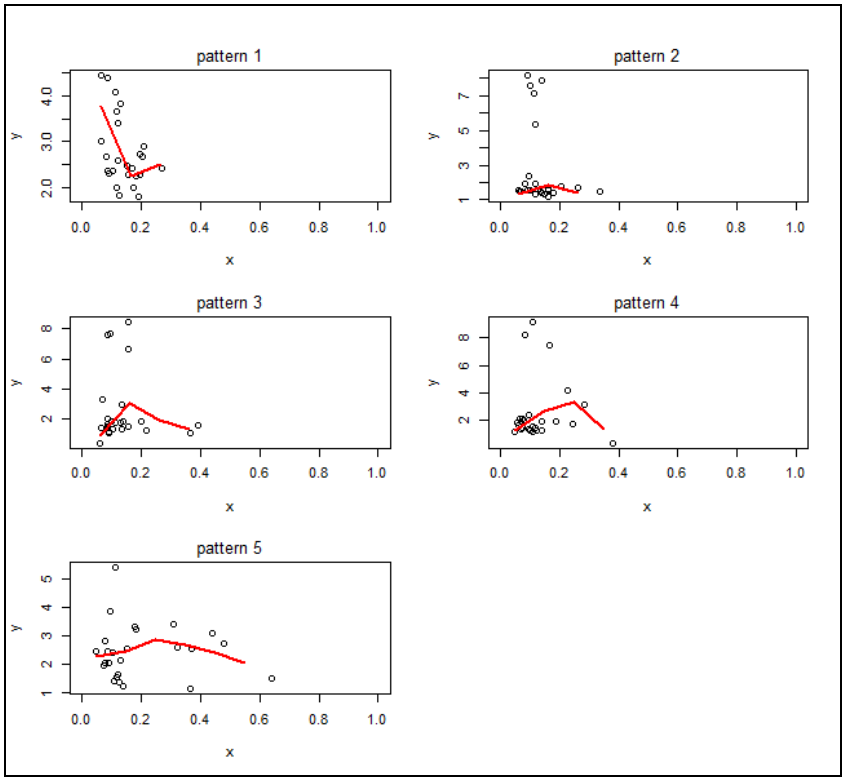


Figura 3. Patrones en L en base a las medianas de los coeficientes de regresión basada en B-Splines.



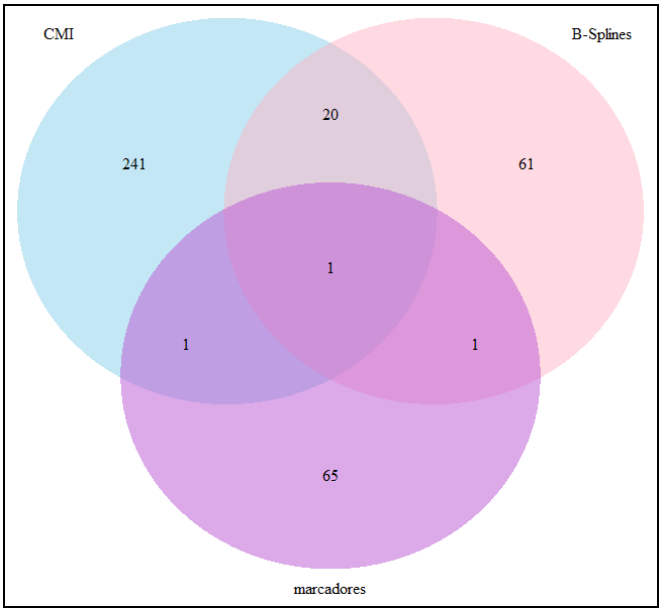
La clasificación de los patrones no consiguió mejorarse mediante la utilización de P-Splines ya que, para un elevado número de genes, el cálculo de los coeficientes dio lugar a valores nulos y fueron eliminados para poder ejecutar la clasificación. Solo un total de 15 genes de los 83 seleccionados pasaron a ser clasificados. Por este motivo, se descartó la utilización de P-Splines para optimizar la clasificación de los patrones en forma de L.

2.4 Comparación de resultados obtenidos en ambos métodos

La comparación de los resultados obtenidos se podía hacer comparando entre los métodos descritos, así como comparando con resultados obtenidos en diferentes publicaciones. Los motivos expuestos en el apartado 2.1 limitaron la comparación entre los resultados obtenidos ejecutando los métodos presentados en este trabajo así como con un listado de genes conocidos regulados por metilación[2,13] que se llamarán genes marcadores.

Para el método de selección basado en CMI, aunque no se dio con una combinación de valores que supusiese una mejora, se escogió la combinación que seleccionó 263 genes, el mayor número de genes (263): h=0.2, r<0.25 y cMI(0)>0.1 (sensibilidad 8.75%). Mientras que para el método de selección basado en regresión basada en Splines, se utilizó el listado de 83 genes obtenido con los parámetros indicados en el apartado anterior (sensibilidad 92.77%). Al comparar las dos listas junto con el listado de 68 genes marcadores, se obtuvo el diagrama de Venn que se muestra en la figura 4.

Figura 4. Diagrama de Venn. Genes compartidos por los diferentes métodos y con la lista de marcadores.



Como el número de genes marcadores seleccionados no fue tan alto como se esperaba, se plantearon tres posibilidades:

a) Que los genes no estuviesen presentes en los datos analizados.
b) Que los genes no estuviesen regulados por metilación en los tipos de CRC analizados.
c) Que los genes estuviesen presentes y regulados por metilación pero no hubiesen sido detectados por ninguno de los dos métodos.

De los 68 genes incluidos en el listado de marcadores, un total de 50 estaban presentes en los sets de datos analizados. Además, tras comprobar los gráficos de los pares metilación-expresión de los 47 genes marcadores que no habían sido seleccionados por ninguno de los métodos, se vio como ninguno seguía un patrón en L ya que en todos los casos había observaciones con un elevado nivel de metilación asociado a un elevado nivel de expresión. Por ello, se asumió que los genes no se seleccionaron porque no estaban regulados por metilación en los tipos de CRC analizados y no por un defecto en los métodos.

2.5 Creación del paquete R *lpattern*

El completo desconocimiento de la creación de un paquete en R, como de las necesidades que podrían aparecer a lo largo del trabajo hicieron que en un principio se crease el paquete R *lpattern* v0.1, incluyendo las funciones originales proporcionadas por el tutor. Conforme se iban preparando los distintos entregables se hizo evidente la necesidad de realizar numerosas modificaciones. Aunque fuera de plazo, con el fin de poder continuar con los objetivos del trabajo se construyó una segunda versión del paquete *lpattern*. En este nuevo paquete se incluyeron tanto las funciones originales como la función propuesta para la mejora del método de selección basado en B-Splines

(`InitialSelection2`). A continuación se listan las funciones documentadas en el paquete:

- cMI: cálculo de la información mutua condicional entre las variables metilación y expresión, en función del punto de corte *t*.
- matAllCorrs: cálculo de la matriz de correlaciones entre las variables metilación y expresión.
- InitialSelection: selección de genes con correlación de Spearman negativa entre las variables metilación y expresión.
- InitialSelection2: selección de genes sin valores con alta expresión asociados a una alta metilación.
- CalculaSplines: ajuste de cada par metilación-expresión a una regresión pasada en B-Splines.
- plotWithSplines: gráfico de dispersion de cada par metilación-expresión y curva ajustada a partir de los B-Splines.

Además, se incluyeron los sets de datos de dominio público utilizados a lo largo del presente trabajo y que se utilizaron también para la creación de la viñeta:

- MethData: *data frame* con los datos de metilación
- ExpData: *data frame* con los datos de expresión
- CorrData: *data frame* con los datos de correlación. Producto obtenido al ejecutar la función `matAllCors`.

Por último, surgió la necesidad de incluir el código que sería llamado a través de la aplicación basada en *Shiny*, como se amplía en el siguiente apartado.

2.6 Creación de la aplicación interactiva basada en *Shiny*

En este punto se valoraron diferentes posibilidades: crear una única función para cada método de forma que con una única llamada se obtuviese el listado de genes seleccionados, ejecutando un método cada vez y ofreciendo la posibilidad de ir modificando los criterios de selección según el método elegido, o conseguir el resultado anteriormente expuesto pero trabajando sobre diferentes sets de datos, de forma que el usuario pudiese cargar y analizar sus propios datos. Por último, se pensó en incluir al listado de genes la opción de representar gráficamente los pares metilación-expresión de los genes seleccionados.

La falta de experiencia creando paquetes en R y aplicaciones interactivas con *Shiny*, hicieron complicado establecer un criterio para decidir qué opción era la más adecuada para alcanzar el objetivo de demostrar el funcionamiento del paquete. Como en el resto de apartados, se probaron todas las alternativas disponibles, y fue la falta de tiempo lo que marcó la opción a escoger.

Se añadió al paquete R *lpattern* el código necesario para ejecutar cada uno de los métodos llamando a una única función. Estas funciones no se documentaron en el paquete porque únicamente se utilizarían a través de *Shiny*.
El funcionamiento de la aplicación se planteó muy sencillo:

- Panel lateral: únicamente se incluyó, una opción para seleccionar entre los dos métodos que según la selección se activaría dos paneles condicionales. Si el método escogido era el método de selección basado en cMI se activaría un apartado que permitiese escoger el ancho de ventana para el cálculo de cMI y un segundo apartado para determinar las condiciones de selección de genes (ratio y cMI(0)). Sin embargo, si el método escogido era el método de selección basado en regresión basada en Splines, el panel condicional activado ofrecería la posibilidad de modificar los criterios de selección utilizados en la optimización del filtrado inicial del método (nSI, nID, nSD, met.max y dif.met).

- Panel principal: listado de genes seleccionados con un formato en tabla numerada.

Los datos utilizados para mostrar el funcionamiento de la aplicación fueron los mismos utilizados en este trabajo, así como los incluidos en el paquete R.
El resultado obtenido al ejecutar la aplicación con cada uno de los métodos se puede observar en las figuras 5 y 6.

Figura 5. Ejecución en *Shiny* del método basado en cMI.

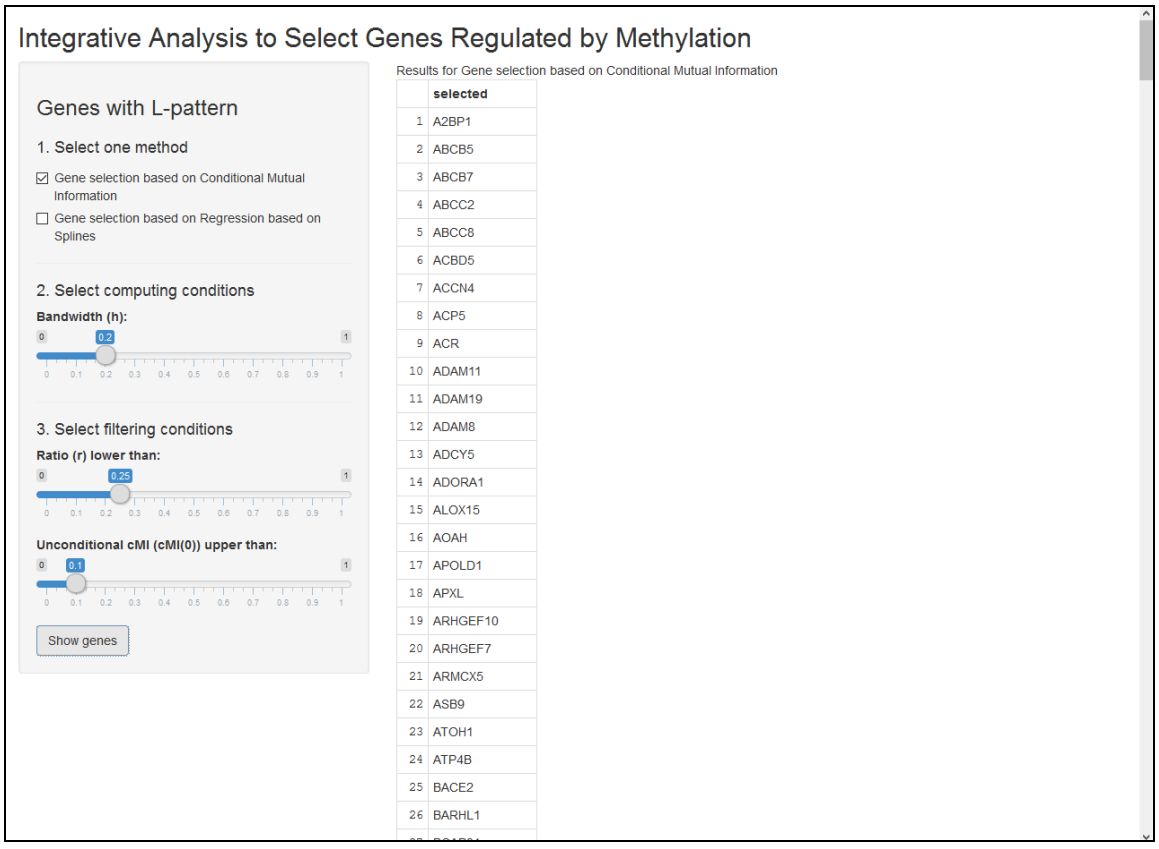
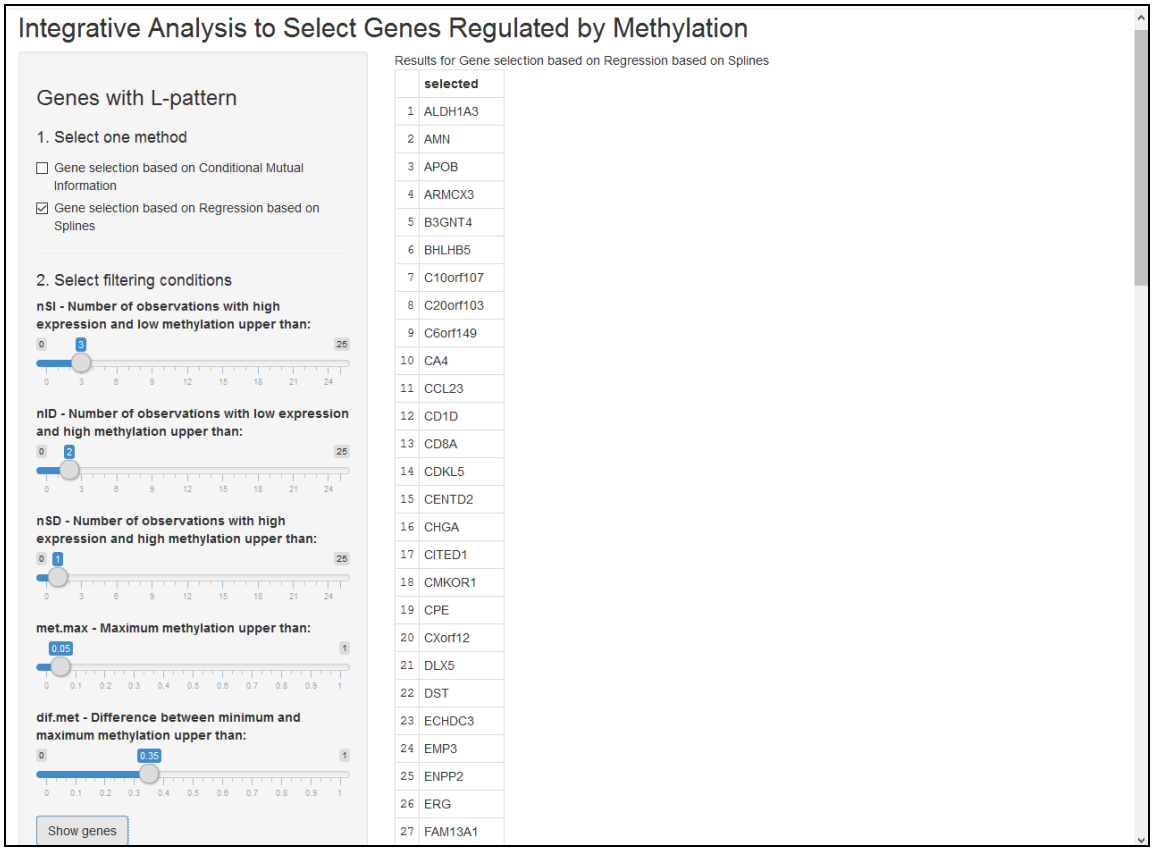


Figura 6. Ejecución en *Shiny* del método basado en regresión basada en B-Splines.



3. Conclusiones

En referencia a las competencias adquiridas en este trabajo, cabe destacar la mejora en la capacidad de aplicar los conocimientos y resolver problemas de programación en Bioinformática. El hecho de no conseguir la mejora de los métodos creó una situación de bloqueo, por lo que ha sido muy importante la dirección por parte del tutor a la hora de proporcionar pistas para continuar con el trabajo. Cabe citar también la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la Bioinformática y la Bioestadística. Además, el completo desconocimiento sobre la creación de un paquete en R así como de *Shiny* obligó a invertir gran parte del proyecto a gestionar información para la resolución de los problemas derivados.

Se han alcanzado los objetivos planteados ya que se ha mejorado al menos uno de los métodos, se ha creado un paquete en R que contiene las funciones necesarias para la ejecución de los métodos y se ha diseñado una aplicación basada en *Shiny* para que el usuario pueda conocer el funcionamiento de la aplicación. En cambio, el modo en el que se han resuelto los diferentes apartados admite todavía numerosas mejoras y los productos obtenidos no tienen el nivel de calidad esperado en el planteamiento inicial del trabajo. A pesar de que este trabajo estaba perfectamente definido en sus objetivos, la falta de formación necesaria para desarrollar la mayoría de los apartados han dificultado el seguimiento del mismo y el resultado obtenido no puede considerarse completamente finalizado.

La planificación establecida se fue adaptando conforme a las necesidades, ya que todos los apartados han requerido más tiempo del que estaba previsto. Las correcciones de los distintos entregables han requerido revisión del trabajo y volver a ejecutar las funciones. Se han tenido en cuenta los comentarios y sugerencias a lo largo de todo el proyecto, añadiendo modificaciones siempre que ha sido necesario. Por este motivo, el tiempo invertido ha superado considerablemente al tiempo estimado de dedicación. Tanto el paquete en R como la aplicación se han ido modificando continuamente ya que, al estar todo conectado, la realización de un cambio en el método ha implicado ejecutar de nuevo el código, modificarlo, evaluar el paquete o crear la viñeta, aumentando de forma notable la inversión de tiempo. Es probable que un conocimiento previo en algunos de los aspectos tratados en este trabajo hubiese permitido demostrar esa implicación en el trabajo con una mejor calidad de los productos obtenidos.

A continuación se mencionan algunos puntos que no se han podido tratar a lo largo de este trabajo:

- Optimización de los parámetros de ambos métodos mediante validación cruzada utilizando el paquete R *caret* u otra aproximación más rigurosa.
- Utilizar un listado de genes de genes que no deberían estar presentes en los listados de genes seleccionados y que podrían ser útiles para evaluar la especificidad de los métodos.

- Optimizar la función `InitialSelection2` estudiando el efecto de dividir los gráficos en más de 4 cuadrantes y descartar aquellos que contengan observaciones en los cuadrantes más próximos a la parte derecha con el fin de aumentar la especificidad del método.
- Ejecución de la propuesta de filtrado inicial sugerida para el método de selección basado en B-Splines con el fin de comprobar si se obtienen los mismos resultados sobre otros sets de datos.
- Profundización en el cálculo de P-Splines para evitar la creación de valores nulos y comprobar si el cálculo de P-Splines mejora la clasificación de las curvas.
- Comparación de los resultados obtenidos con los métodos presentados en este trabajo con los de otros métodos disponibles en la literatura.
- Ampliación de la aplicación *Shiny* incluyendo un listado de genes compartidos al ejecutar ambos métodos, así como la posibilidad de visualizar los gráficos de los pares metilación-expresión de los genes seleccionados y de sus B-Splines ajustados.

4. Glosario

Epigenético: relativo a la epigenética, que es la ciencia que se refiere a los cambios heredables en el ADN e histonas que no implican alteraciones en la secuencia de nucleótidos y modifican la estructura y condensación de la cromatina, por lo que afectan la expresión génica y el fenotipo.

Hipermetilación: La metilación del ADN es un proceso epigenético que participa en la regulación de la expresión génica. Durante la replicación del ADN el carbono 5 de las citosinas de la cadena recién sintetizada puede ser metilado, de este modo se impide la unión de factores de transcripción y el ADN adopta una estructura propiciando la estructura "cerrada" de la cromatina que impide su lectura. La hipermetilación es un aumento de la metilación de citosina y adenosina residuos en el ADN.

Información mutua condicional (CMI): la información mutua entre dos variables aleatorias mide la dependencia entre ellas y es útil para detectar tanto relaciones lineales como no lineales. Si consideramos que estas variables son la metilación y la expresión génica, y llamamos t al punto de corte para la binarización de los datos de metilación, de modo que los valores de metilación $>t$ se consideren metilados y los valores de metilación $<t$ se consideren no metilados, CMI se puede considerar como la suma ponderada de MI en ambos lados de t.

Splines: funciones continuas formadas a partir de conectar segmentos lineales más cortos. Los puntos donde se conectan los segmentos se conocen como nudos. La diferencia entre B-Splines y P-Splines radica en que con estos últimos se tiene en cuenta una penalización adicional para aumentar la suavidad de la curva y evitar el sobreajuste.

Biomarcador: fracción de ADN que indica una característica diferencial entre dos individuos.

Validación cruzada: Determinación de la validez de un modelo basado en la eliminación, a partir de la muestra original, de una submuestra de datos con la que evaluar el modelo propuesto.

Sensibilidad: capacidad del método de seleccionar un que tenga un verdadero patrón en L. También se conoce como fracción de verdaderos positivos, ya que es el cociente entre verdaderos positivos y total de genes seleccionados.

Cluster: conjunto de genes agrupados en función de los coeficientes de sus Splines, y que por tanto muestran un comportamiento similar de sus pares metilación-expresión.

5. Bibliografía

1. Wang, K.-S. Integrative Analysis of Genome-wide Expression and Methylation Data. *J. Biom. Biostat.* **4,** (2013).
2. Mitchell, S. M. *et al.* A panel of genes methylated with high frequency in colorectal cancer. *BMC Cancer* **14,** 54 (2014).
3. VanderKraats, N. D., Hiken, J. F., Decker, K. F. & Edwards, J. R. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Res.* **41,** 6816–6827 (2013).
4. Jiao, Y., Widschwendter, M. & Teschendorff, A. E. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* **30,** 2360–2366 (2014).
5. Liu, Y., Ji, Y. & Qiu, P. Identification of thresholds for dichotomizing DNA methylation data. *EURASIP J Bioinforma. Syst. Biol.* **2013,** 8 (2013).
6. Alex Sánchez-Pla, M. Carme Ruíz de Villa, Francesc Carmona, Sarah Bazzoco & Diego Arango del Corro. Integrative Analysis to Select Genes Regulated by Methylation in a Cancer Colon Study. *BIOSTATNET Workshop on Biomedical (Big) Data. Research Perspectives CRM Barcelona, Trends in Mathematics* **7,** 1–4 (2015).
7. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2013).
8. RStudio, Inc. *Easy web applications in R*. (R Consortium, 2013).
9. The Cancer Genome Atlas - Data Portal. Available at: https://tcga-data.nci.nih.gov/tcga/. (Accessed: 28th May 2016)
10. Home - GEO - NCBI. Available at: http://www.ncbi.nlm.nih.gov/geo/. (Accessed: 28th May 2016)
11. Barat, A., Ruskin, H. J., Byrne, A. T. & Prehn, J. H. M. Integrating Colon Cancer Microarray Data: Associating Locus-Specific Methylation Groups to Gene Expression-Based Classifications. *Microarrays* **4,** 630–646 (2015).
12. Yihua Liu & Peng Qiu. Integrative analysis of methylation and gene expression data in TCGA. *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)* 2–4 (2012).
13. Naumov, V. A. *et al.* Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips. *Epigenetics* **8,** 921–934 (2013).