Integrative Analysis of Methylation and Expression Data

Juan Carlos Company

Master en Bioestadística y Bioinformática

Área de Estadística y Bioinformática

Alexandre Sánchez Pla

26 diciembre 2016

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

lectual.

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Integrative analysis of methylation and Expression data*
Nombre del autor:	*Juan Carlos Company*
Nombre del consultor/a:	*Alex Sánchez Pla*
Nombre del PRA:	*Alex Sánchez Pla*
Fecha de entrega (mm/aaaa):	12/2016
Titulación:	*Master en Bioestadística y Bioinformática*
Área del Trabajo Final:	*Estadística y Bioinformática*
Idioma del trabajo:	*Castellano*
Palabras clave	*Metilación, Expresión, L-pattern*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

La metilación del DNA es un mecanismo epigenético directamente relacionado con la regulación génica, asociado al silenciamiento génico. Esté tipo de regulación es especialmente importante en el desarrollo o en determinadas enfermedades como cáncer. En concreto se ha observado que la ganancia de metilación está relacionada con la regulación en Cáncer de colón (CRC). Así pues, el aumento de la metilación y la disminución de la expresión forman un patrón en forma de L al representar los valores en un *scatterplot*. En este proyecto buscamos determinar la detección de genes potenciales en CRC en base al análisis de estos gráficos y a la clasificación en base a este patrón. Para ello, utilizaremos un concepto nuevo pero poco explotado *scagnostics[4]* y lo compararemos con estudios previos similares mediante la generación de nuevos set de datos de Gene Expression Omnibus (GEO). Como resultado de este trabajo hemos obtenido la implementación de un nuevo método en la detección de patrones en L, mediante diferentes funciones y algoritmos de clasificación basados en SVM que será incorporado a una aplicación de análisis basada en la plataforma Shinny.

Abstract (in English, 250 words or less):

DNA methnylation is an epigenetic mecanism related with gene regulation. This type of genomic regulation is specially important in development or in some diseases such as Cancer. In particular, it has been observed an increase in the methylation levels related with the etiology of Colorectal Cancer(CRC) . The rise in the methylation levels followed by a decrease of the expression creates a L-pattern in the scatterplot when those values are represented per gene. Our study is focus on the analysis of this pattern by the scagnostics[4]. We compared our findings with previous studies by using same datasets as well as increase the number of datasets related with CRC seaching at the GEO database.

As a result of this work, we have implemented a new application to the detecction of biomarkes in CRC based on SVM and the scagnostic concept which methods were added to a shinny based app. Additional to this project, we have introduced new functionalities that will help the user to have more freedom in analyses their data.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

La metilación del DNA constituye un mecanismo epigénico importante en la regulación génica[1]. Se trata básicamente de la adición de un grupo metil sobre la citosinas en el genoma. Generalmente, estas bases metiladas están asociadas a zonas enriquecidas en es un contexto CpG llamadas islas de CpG. La ganancia se considera hipermetilación y una perdida Hypometilación. Ambos procesos son independientes (diferentes maquinarias enzimáticas) y ocurren en diferentes situaciones en el genoma. De esta manera, cuando la metilación del DNA ocurre en las regiones promotoras se produce una irrupción de la transcripción y como consecuencia un silenciamiento del gen. Este mecanismo tiene conocida importancia en el desarrollo o en la diferenciación celular, pero además se ha visto que directamente relacionado en enfermedades humanas como el Cáncer[1].

Por lo tanto, una ganancia de metilación se puede traducir en una pérdida de expresión (no siempre tiene que ser así) y cuando se interpreta este patrón en u grafico en conjunto de la metilación, forma un patrón en forma de L característico[2]. La detección de estos patrones no es sencilla debido a que la metilación, aun siendo un proceso estable, ocurre con diferentes intensidad a lo largo del genoma y de la isla CpG (shelve, shore, island). Este sesgo puede volverse importante cuando se utilizan técnicas de High-Throughput dan valores por diferentes posiciones del gen. Además este tipo de análisis es crucial cuando el análisis implica la búsqueda de genes potencialmente regulados en enfermedades como el Cáncer Colon rectal (CRC).

Nuestro estudio se basa en la detección de patrones en forma de L para genes hipermetilados que potencialmente regulen los procesos en CRC. Para el estudio del patrón utilizaremos el concepto de *scagnostics [4]* y lo compararemos con otros estudios previos[3]. Este es un concepto en el cual se miden diferentes parámetros de un *scatterplot* y permiten determinar la forma y comportamiento a través de valores numéricos. Es un concepto antiguo propuesto por los hermanos Tukey en los años 80 [4], pero que ha sido poco explotado en la bibliografía y que es particularmente interesante en el caso que nos ocupa. Este hace diferentes mediciones: Outlying, Skewed, Convex, Skinny, Stringy, clumpy, sparse, striated, monotonic (mantendremos los términos en inglés para mantener la coherencia para la lectura del código). De los cuales, los más importantes para la forma son Convex, Skinny, Stringy [5]

En nuestro trabajo primero generaremos un set de análisis de datos de Cáncer (en concreto CRC) a partir de diferentes sets de datos públicos de GEO. Este conjunto no tiene por que excesivamente coherente para su estudio , ej. mismo tipo celular , ya que nos servirá principalmente para poder implementar el método. Otros datos más curados, serán utilizados al final para la detección de genes con cáncer de colon en un set de datos comparativo de expresión y metilación relacionados por el valor del gen. Para poder integrar todos estos datos utilizaremos el paquete de *scagnostics* [4], además de otros previamente utilizados para elaborar un análisis exhaustivo de los resultados mediante la utilización de SVM y un set de genes con importancia en cáncer obtenido mediante técnicas de *data-mining* [6].

Por último, como resultado de este estudio, definiremos unos valores de forma y una serie de funciones que iremos añadiendo gradualmente como un nuevo modulo en una aplicación shinny , mejorando en el proceso algunas funcionalidades del mismo, hasta el tiempo que el trabajo final de master permita.

1.2 Objetivos del Trabajo

El objetivo final del este trabajo es la implementación de métodos para la detección de Biomarcadores en CRG (o en otro tipo de cáncer que pueda ser relacionado con la metilación) en base al estudio de los diferentes parámetros que ofrecen una serie de estadísticos de correlación. Este trabajo, es un trabajo heredado, pero único en la implementación del concepto de *scagnostics[4]* [4] mediante el análisis de sets de datos artificiales, la clasificación de los métodos por machine learning y la comparación con estudios previos de data-sets públicos. Para ello el trabajo se divide en los siguientes objetivos:

- Minería de datos
    - Obtener una lista de genes con relevancia en CRC mediante data-mining
    - Obtener un set de datos de cáncer de colon para la metilación y expresión.

- Implementación técnica
    - Implementación de funciones que permitan descargar y anotar datos directamente desde GEO.
    - Implementar funciones introducir estos sets de datos en el análisis de *scagnostics[4]* y poder ser evaluados.
    - Implementar funciones de reconocimiento de patrones para su clasificación mediante SVM
    - Implementación funciones de análisis comparación sets de datos y enriquecimiento de términos GO.
    - Implementación de los métodos en una aplicación Shinny con una interfaz *user-friendly*

- Análisis de datos
    - Búsqueda de parámetros para detectar L-shape mediante el desarrollo de una función
    - Análisis de los datos en bases a métodos previos: mutual information, splines, correlation, dist correlation.
    - Clasificación de los datos en base al L-shape y su relación con el set de relevancia en CRC.

1.3 Enfoque y método seguido

El enfoque de este trabajo fue la división del tiempo en base de los diferentes apartados descritos en el apartado anterior y en base a recomendaciones del coordinador del proyecto con reuniones eventuales sobre la direccionalidad del proceso. Debido a este tipo de enfoque de toma de decisión, sobre todo en las primeras partes de generación de los sets, como se indicó en el resultado sobre el objetivo final puede variar y requerir de más tiempo de análisis. Sin embargo, debido a que este enfoque es modular y centrado en la biología y la implementación de métodos, permite aun en el supuesto de necesidad de más tiempo por errores o dificultades en la implementación la obtención como resultado de una aplicación reutilizables en otros estudios.

Nuestro primer punto de trabajo, fue la obtención de los sets de datos de manera manual mediante la búsqueda de datos de expresión y metilación en las bases de datos GEO. Al mismo tiempo que la obtención de diferentes genes relacionados en cáncer mediante minería de datos [6]

El siguiente paso, sería la implementación de métodos de análisis de previos mediante la instalación e implementación de los paquetes para Mutual Information, Dist Cor, Splines y Scagnostics[4]. Además, con la filosofía de mantener el sistema lo más automatizado posible, la creación de funciones de análisis y descarga como por ejemplo la desarrollada para la obtención de los datos anotados a partir de un GEO ID. Además de diversas herramientas auxiliares, como la combinación de las diferentes posiciones (sondas en un array) bajo el mismo gen o identificador, la representación de los valores de expresión/metilación en un scatterplot o funciones de clasificación o filtrado

Para comprobar el efecto de los parámetros estudiados mediante scagnostics[4] y comprobar así si presenta una forma de *L-shape* se plantean dos estrategias:

1) Utilización de una lista fiable de genes en forma de L. Necesariamente validados experimentalmente.

2) Implementación de "datos artificiales " que permitan determinar los valores de los parámetros de scagnostic para sets correlacionados.

Dado que no fue posible encontrar una lista lo suficientemente extensa y fiable que cumpla las características de la primera opción, se utilizará la segunda opción. Para ello generaremos una función que permite en base a un número dado (4..10,..100) la generación aleatoria de datos para su interpretación mediante scagnostics[4]. Estos valores nos servirán más tarde como valores default en la función de evaluación de la forma de la correlación expresión/metilación , así como en la aplicación shinny.

El último paso será el análisis de los genes del set de datos públicos. Primero, obtenemos los valores de shape de todos los genes presenten en los sets mediante los diferentes métodos. Utilizando los valores default de este patrón y la combinación de los mismos entre diferentes métodos para determinar aquellos que presenten L-shape, implementaremos una clasificación mediante un algoritmo de SVM (paquete caret)

Como resultado final se implementará la función de detección de *scagnostics[4],* con sus valores default , como una nueva función independiente (new tab) de un paquete previo con otros elementos para el análisis de *L-shape*.

1.4 Planificación del Trabajo

El trabajo se divide en dos puntos básicos como se describe en la figura debajo y se ha ido explicando durante la memoria. En el primero la evaluación de estos métodos servirá como objetivo para la detección de biomarcadores en CRC y el segundo la implementación de las funciones que sean posibles (seguro la detección por scagnostic) en una aplicación shinny.

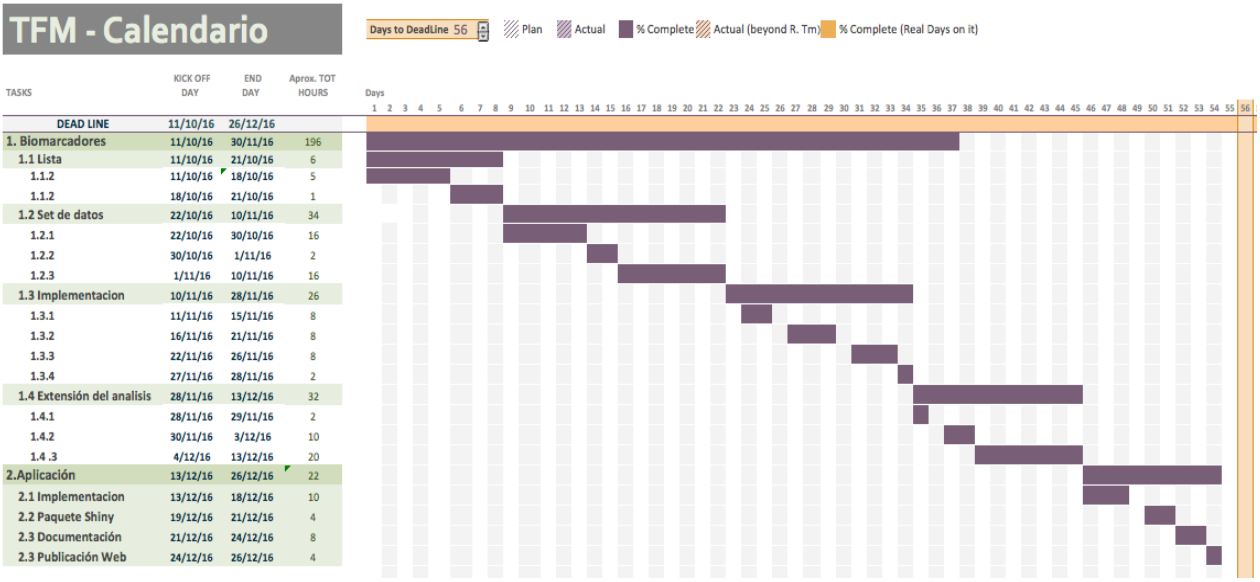


Fig 1: Plan de trabajo desarrollado al principio del proyecto. Empieza el día 11.10 y acaba el 26.12

1.5 Breve sumario de productos obtenidos

- Set de datos de genes con importancia biológica en CRC obtenidos por data-mining
- Set de datos públicos de CRC de la página de GEO
- Valores default para los métodos de *scagnostics[4]* en la detección de un patrón en forma de L
- Función Shinny con el módulo de *scagnostics[4]* añadido.

1.6 Breve descripción de los otros capítulos de la memoria

En el punto 2 de esta memoria se explicará los diferente apartados, procedimientos y funciones utilizadas durante el desarrollo de este trabajo. Como resultado final se

implementará la función de detección de *scagnostics[4]* como una nueva función a un paquete previo con otros elementos de análisis de *L-shape* además de realizar un estudio para la detección de biomarcadores en CRC.

Este punto se divide en estos dos capítulos básicos: Análisis y el desarrollo de la Aplicación. En el primero consta de un primer paso introducción y de recopilación de la información pública disponible para la detección de biomarcadores mediante data-mining. A continuación le seguirá un paso de análisis de los datos, el cual se describirá el desarrollo de diversas funciones para realizar el análisis y la implantación de las mismas en el set de datos previamente descargado. Por último en este punto, se explicará los procedimientos utilizados para el análisis e interpretación de resultados.

El último punto de la memoria está dedica a la explicación de la utilización de la aplicación shinny.

2. Resto de capítulos

1. Identificación de Biomarcadores en Cáncer Colon rectal.

Introducción

El objetivo principal de este trabajo , es la continuación de estudios previos en la detección de patrones identificables y correlacionadles para la detección de genes con importancia en cáncer colon rectal. Aun que hay muchas capas de regulación génica y epigénetica en el genoma que pueden definir el desarrollo de una enfermedad, este estudio se ha focalizado en aquellos genes regulados por metilación y en concreto en el patrón en forma de L característico que presentan en cuanto están regulados por hipermetilación (ganancia de metilación). Otras aproximaciones previas [1] a esta han sido llevado a cabo en la búsqueda e interpretación de la correlación entre la expresión y metilación, pero siempre se han basado en un enfoque de la cuantificación de la correlación como valor para la detección del patrón entre los valores. En nuestro caso este enfoque varía y nos centrarnos en la relación que tienen todos los puntos entre si de un conjunto sobre un plano.

Una ventaja de este método de estudio al respecto de los anteriores, es que la relación entre patrones entre una marca epigenética y la expressión como consecuencia de una regulación es extrapolable a otros capas o eventos en el genoma. De esta manera, este estudio podría aportar un nuevo punto de vista en tanto en cuanto incremente las opciones detección adicionales al enfoque tradicional.

Para cumplir este objetivo nos basaremos en estudios previos y en la comparación de diversos métodos en concreto Splines, DistCor y Mutual information del concepto de scagnostics[4]. Este se compone de 10 valores medidos a partir de la representación de los puntos en un plano y permite la detección de diferentes características, como la forma, la dispersión el sesgo al respecto del en el eje entre otros. Además, complementaremos estos métodos con la implementación de funciones que calculen métodos previos descritos utilizaremos los paquetes de R splines[7], energy[8], entropy [9], infotheo [10]. Todos estas funciones requieren de dos vectores que se corresponderían con los valores de expresión y metilación.

Data Mining

Sets de datos con importancia en cáncer

La implementación de los métodos de detección formará una segunda parte del proceso. En la primera necesitaremos obtener una lista de datos para poder realizar el estudio que nos ocupa y una lista de genes que nos permita evaluar su importancia biológica. Para ello centraremos la búsqueda de estos datos en tres elementos principales: lista de genes con importancia biológica, lista de data sets de expresión y/o metilación de CRC y datos de un estudio previo de expresión-metilación [1]

Obtenemos como resultado de la búsqueda de archivos GEO 25 sets de datos, de los cuales 12 corresponden con metilación y 13 corresponden con expresión. De manera manual añadiremos los identificadores de la base de datos de GEO [11] una vez completemos la búsqueda con CRC/Colon rectal/Metilación en el browser. Aun que hemos seguido este enfoque , al igual que la descarga y anotación las búsquedas en GEO también pueden implementadas. Esto datos en formato .csv (Nombre del fichero) se añadirá en un formato fácilmente agregable al entorno de programación R [12] que más se adecua al tipo de trabajo que vamos a desarrollar en el trabajo fin de master. Como se ha dicho anteriormente, la importancia de estos sets es el generar un set de datos lo suficientemente extenso para poder implementar las diferentes funciones de clasificación, no tanto reducir el sesgo biológico.

A continuación la integración de estos sets de datos en un entorno de R de manera manual al igual que su búsqueda puede ser larga y puede retrasarnos el los puntos de entrega. Por esta razón y ha sugerencia del coordinador, en este punto se implementa una función para la lectura-descarga-anotación de estos sets de datos **`getGEO.anno`**. Para ello, utilizaremos el paquete geoQuery [13] que permite la descarga de aquellos datos disponibles en GEO y transfórmalos en una matriz de expresión con los datos ya normalizados. Dado, que los sets en esta base de datos son diversos y sus plataformas diversas también, la implementación de un método de adicional de normalización en cada caso es complicado y puede llegar a utilizar el corto tiempo del proyecto. Por este motivo se decide utilizar directamente las tablas de normalización. Otro elemento de dificultad al respecto de automatización de la obtención de los datos públicos es la anotación de estos sets de manera automática o manual. Para ello, se utiliza solamente aquellos sets de datos que contiene una plataforma con anotación presente en bioconductor y que nos permitirá no

tener que generar una base de datos por nosotros mismos (otra vez por restricción en los tiempos de entrega). Esto reduce la lista de nuestro set de 25 a 8 data sets: 4 sets de expresión y 4 de metilación relacionados con cáncer, más los dos estudios adicionales relacionados con CRC. Aunque resulta un descenso significativo de los sets, el número elevado sigue cumpliendo nuestro objetivo y nos permitirá añadir una nueva funcionalidad a la aplicación en el futuro.

Type	Description	GEOid
EXP	ERbeta modulation of NFkB pathway in colon cancer cell lines SW480 and HT29	GSE65979
EXP	Expression data at each site in colon cancer	GSE65480
EXP	LATS2 regulated genes in human colon cancer cell line	GSE51715
EXP	Gene expression profiling of human colon cancer cell lines stimulated with dsDNA90	GSE75205
METH	Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers	GSE57342
METH	Colon cancer profiling	GSE2138
METH	DNA methylation differences between multiple sclerosis and controls in frontal lobe white matter	GSE40360
METH	5-hydroxymethylcytosine marks promoters in colon that resist hypermethylation in cancer [Methylation450 Array]	GSE63421

Tabla 1: Tabla con los sets de datos una vez filtrados por anotación.

Búsqueda de listas de genes

El siguiente punto es la creación de una lista de genes con relevancia en cáncer. La motivación es tener un set evaluar y si no fuera posible un set de datos que permita determinar la importancia biológica de nuestro análisis. Para este punto existen dos estrategias, la primera es la búsqueda de aquellos genes en la bibliografía de manera manual y la segunda es la utilización de herramientas de data mining [6] Como se ha dicho anteriormente y con el objetivo de mantener este estudio lo más automatizado posible , utilizaremos varias herramientas de data-mining que generará una lista de 67 genes con importancia biológica en este contexto (nombre del fichero.). Por ejemplo entre ellos se encontraría TET2 y TP53 que participan en los procesos de metilación [3] y cáncer [14].

Análisis de datos

Función de predicción de L-shape

Una vez completada la parte de data-mining lista de sets de datos, implementamos una serie de funciones que se basan en la interpretación de los parámetros de resultados. Dado que nos focalizamos en los parámetros de *scagnostics[4]* para definir aquellos con L-pattern y no L-shape el algoritmo de clasificación, únicamente valoraremos los resultados de este paquete mientras que los otros diferentes métodos pero no serán analizados en profundidad. Para el primer paso será la implementación de una función, *shape.param()* , que dado una serie de valores por gen (en realidad comparará dos sets diferentes) mediante diversas técnicas las devolverá como un data frame con las diferentes medidas de correlación asociadas. Está función será la que se implemente en la aplicación shiny más tarde para la interpretación de resultados. Aun que no para la aplicación, si no que para el análisis, esta función tendrá asociadas diversas funciones auxiliares que permitirán la integración de los valores de los diferentes sets de datos por gen como la función *merge.gene.expMeth()* la cual realizará la función recién explicada. Utilizaremos estas funciones auxiliares para integrar, evaluar y filtrar nuestros data.sets (10 data.sets) dando como resultado una tabla con todos los valores por gen de los diferentes métodos. Habiendo integrado en ellos todos los sets de datos, como resultado obtenemos una tabla con 22.000 genes de humano de la versión de assembly hg19 con los diferentes métodos de correlación.

Sin embargo, para determinar cual es mejor valor de L-pattern implementamos una función basada en la creación de un set de set de datos en forma de L para diferentes valores, *lshape.diagnostics()*. Está utilizará como input un numero determinado de puntos en de correlación y utilizando la misma función de detección para los sets de datos (*shape.param()*) nos dará una tabla con los valores. Por lo tanto, implementamos esta función para diversos tamaño de muestra (tabla 2) . Observamos que conforme añadimos número de posiciones a comparar, los parámetros de Convexity, Stringy y Skinny que corresponden a la interpretación de la forma por parte del paquete de scagnostics[4] varían sustancialmente. Observamos que con un 100 de datos obtenemos que skinny se encuentra en un rango de 0.28 hasta 0.73, convex en un valor cercano a 0 y stringy en un valor que va desde 0.63, pero que aumenta aproximándose a 1.

	Outlying	Skewed	Clumpy	Sparse	Striated	Convex	Skinny	Stringy	Monotonic
N=10	0	1	0.01	0.26	0.44	0	1	0.63	0.16
N=20	0	0.71	0.01	0.11	0.74	0.02	0.28	0.83	0.26
N=50	0	0.66	0.01	0.04	0.84	0.00	0.32	0.94	0.39
N=100	0.17	0.76	0.09	0.03	0.92	0.00	0.23	0.97	0.48
N=500	0	0.50	0.02	0.02	0.95	0.00	0.28	0.97	0.47
N=1000	0	0.55	0.03	0.02	0.95	0.00	0.73	0.97	0.46

Tabla 2: Tabla con los valores de scagnostics[4] para diferentes tamaños del set artificial.

Cuando comparamos con trabajos previos obtenemos que los valores representados se aproximan, incluso utilizando sets de datos diferentes, así utilizando funciones previas se obtiene que por ejemplo el gen GREM1 presenta un patrón en forma de L . Tras comprobar que si tiene un patrón similar en nuestros, vemos que los valores de *Convex* es de 0.018, valor de *Skinny* de 0.61 y el valor de *Stringy* de 0.72. Observamos que estos valores están en el rango que se ha mostrado en la tabla anterior.

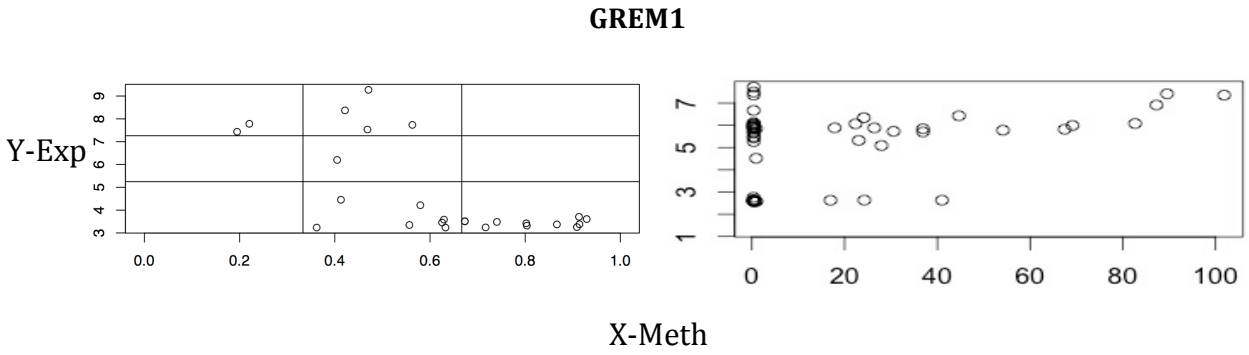


Fig2: Scatterplot del gen GREM1. A la izquierda los valores de para un estudio previo, a la derecha la representación utilizando nuestros datos.

Implementación de la detección en el set de datos

Para la detección de los diferentes valores, utilizamos por cada gen todas las sondas disponibles por set de datos y los agruparemos en un vector que contendrá los valores de expresión y metilación, por separado agrupado de cada sonda por gen de la misma manera

que hemos explicado en el punto anterior. Este un punto crítico, ya que las sondas/las posiciones pueden presentar patrones distintos. Por esta razón el set de datos grande se considera el más ruidoso (este efecto se puede observar en la imagen anterior) , pero nos permitirá gracias a los otros dos sets (menos sesgo) una mejor implementación de, algoritmo.

Una vez agrupados en un vector los valores de expresión y metilación (por separado) se agruparan en un matriz que contenga cada valor por gen y por set de datos. Aun que más computacionalmente intensivo, utilizaremos todos los genes de Humano (*org.Hs.eg.db*) para su detección (25.000 posiciones) . De esta manera, evitamos seleccionar genes y ver como es el input general eliminando este error. Más tarde, podremos seleccionar en base a diferentes parámetros el set y clasificar los genes. Adicionalmente, está funcionalidad estará más reducida para que se pueda utilizar correctamente a nivel de la aplicación web. La realización de esta tarea implica técnicas de paralización en R utilizando el paquete Snow y parallel que nos permitirá disminuir el tiempo de computación.

Evaluación de los resultados

Una vez realizada la parte más complicada y tediosa que es la generación de las tabla de datos y valores generada para cada set de datos obtenido. Utilizamos los valores del paquete scagnostics[4] para evaluar el número de genes con un patrón L-shape y clasificarlos al respecto. Dado, que la combinación de los valores de scagnostics[4] puede ser muy complicada (10^{10} en un rango de 0-1) , necesitamos un método que nos permita clasificar los datos teniendo en cuenta no solamente skiny, convex y stringy, si no todos los demás valores Para ello decidimos utilizar SVM mediante el paquete caret. La justificación de este paquete, es que es una técnica potente que nos permite la clasificación de los diferentes valores. Una limitación es el uso de una buen set de datos, y es aquí donde es importante la creación previa del set de datos grande. De está manera introducimos las funciones de la aplicación dentro de una función que además analice los sets de datos en base al cambio de los parámetros en todo el conjunto de genes.

Sin embargo, para ello primeramente necesitamos definir aquellos que son L-shape. Dado que anteriormente hemos definido estos valores “default” para este método para la detección mediante el análisis de estos parámetros utilizando un set artificial con diferentes valores en el scatter plot. Testaremos primero estos valores y lo haremos con

diferentes valores de cross-validation para observar si el precisión, la sensibilidad, sensibilidad y probando diferentes combinaciones para los 3 elementos que hemos definido como claves en la detección de L-shape: Skinny, Convex, Stringy.

- **Valores default con mas y menos cross validation** : Probamos valores para SVM con cross-validation 2,3,4,10 con los valores default . Observamos un precisión de ~95% y la detección de 199 valores detectados y predichos con L-pattern. Sin embargo, observamos que el aumento o no de la cv no cambia el resultado.

- **Valores variados de skinny (CV=3) :** Probamos diferentes combinaciones de Skinny. Dado que el valor de Skinny fluctúa desde 0.2 hasta 0.7 en el default, calcularemos la capacidad de SVM de clasificar los datos como L en base a menor 0-0.3, alto 0.6-1.0 y el global (sin restricción) 0-1. Observamos que la precisión es similar en todos los rangos. Sin embargo la sensibilidad, la predicción el valor de positivos predichos y negativos predichos es mayor en el rango de 0.6-1-0. Curiosamente, si tenemos en cuenta más factores como correlación, el mutual information, dist cor o splines, observamos que la sensibilidad y sensibilidad aumentan también. Ocurre también que cuando disminuimos por debajo del valor de 30, la mayoría de muestras desaparecen.

  Por lo que concluimos que la relajación de los parámetros en skinny hace difícil distinguir entre buenos o no buenos resultado. Pero un rango pequeño no lo mejora. Nos mantendríamos en el default.

- **Valores variados convex (CV=3):** Aplicamos el mismo control a los valores de convex. En este caso, el default esta cercano al 0 por lo que aumentaremos gradualmente los valore para observar si este aumenta o no (0.1, 0.5,0.7). Vemos que si aumentamos mucho el valor de convex todos los elementos desaparecen. Por lo tanto nuestro valor de datos bajo y cercano a 0.1 es el correcto

- **Valores variados stringy (CV=3):** Por último, probamos la combinación de valores para Stringy. En este caso valores altos indicaban la forma en el patrón de L. Por lo que mediremos sus valores una vez disminuyamos. Generaremos diferentes resultados

Conclusiones que la variación de los parámetros de shape en scagnostics[4] : Stringy, convex y skinny disminuyen la detección de genes con este tipo de patrón. Una manera de comprobar este efecto es añadir información biológica a esta comparación. Para ello, utilizaremos el set de genes con importancia biológica generados por tools data-mining para ver la correlación con los valores.

Una vez cruzados los datos de las diferentes observaciones anteriores, comprobamos que aun que con importancia biológica en cáncer, este tipo de genes para no estar muy correlacionado en forma de L. Este tipo de resultado puede ser esperado por dos razones: La primera es el sesgo inherente a la integración de diferentes sets de datos para su comparativa sobretodo en los datos de expresión (correlación mínima de -0.5) y la detección y la segunda la propia biología de la metilación. Como se ha dicho en la introducción, la hipermetilación se caracteriza por un silenciamiento génico, esto se traduce en mas metilación y menos expresión. Sin embargo, los niveles de metilación varia en las diferentes puntos de la isla de CpG, por lo que esta variabilidad también podría explicar que no se detectaran estos genes.

Otra manera de detectar la importancia biológica de estos genes es la utilización de categorías funcionales y su enriquecimiento en los diferentes sets. Por lo que implementamos una función (*go.enrichment()*) que permita evaluar estas categorías. Sin embargo, para set de genes tan pequeño como el obtenido por los datos default no presenta enriquecimiento suficiente para que supere el threshold (0.05)

2. Implementar la función en una aplicación web.

Diseño

Utilizamos una aplicación previa como estructura base para la introducción de datos de expresión y metilación directamente desde un fichero.

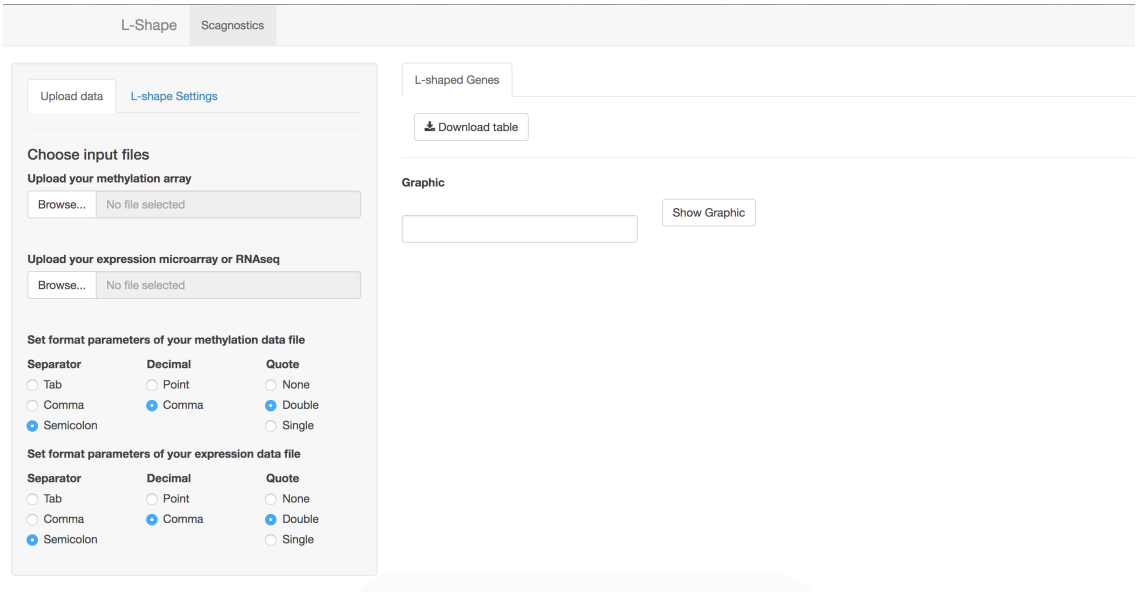
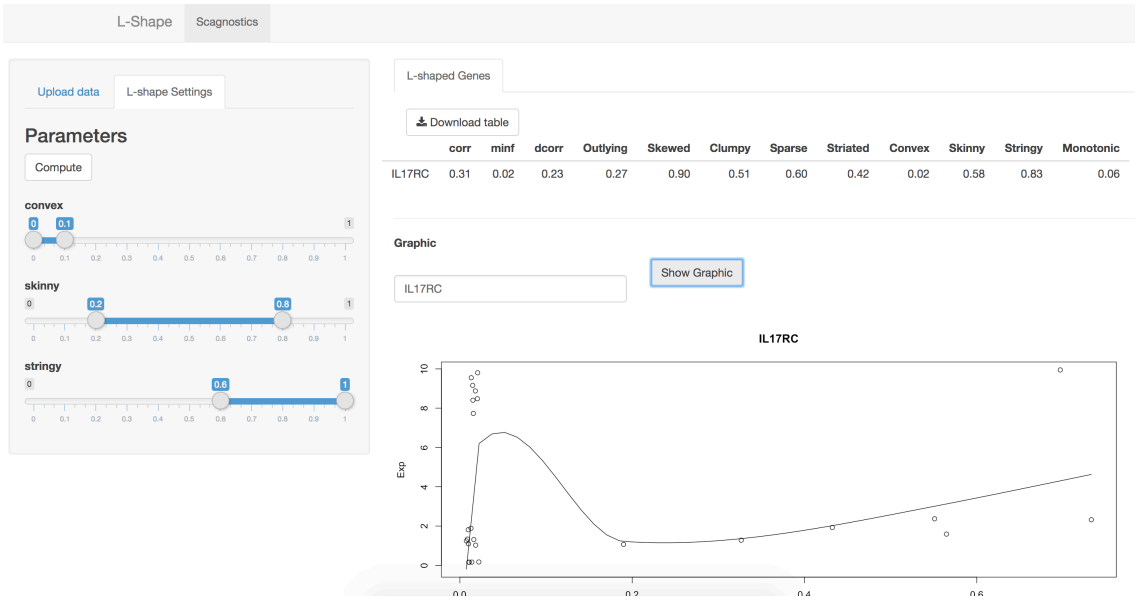


Fig 3: Captura de pantalla de la introducción de sets de datos a la aplicación.

Estos datos son leídos y analizados mediante la función de shape.param(). Además se le han añadido diversas funcionalidad que permitirán al usuario modificar las características de scagnostics[4] que definen la forma. Al modificar estas características obtendremos una lista de genes diferentes que podremos descargar, así como observar el patrón de un gen que queramos en la lista mediante la introducción de su id en el cuadro



Implementaciones futuras

Como se ha ido describiendo a lo largo de la memoria, diversas funciones que permiten evaluar la forma de los *scatterplot* al mismo tiempo que permite determinar la correlación entre dos diferentes sets de datos han sido introducidas en la aplicación *shiny*. Sin embargo, otras por problemas con el tiempo no han podido ser correctamente implementadas para el final del proyecto. A pesar de que se han cumplido los objetivos propuestos en el plan de trabajo, un futuro de implementación de estas funciones sería la descarga directa de los datos de GEO, el análisis mediante *limma* de las matrices normalizadas, la implementación de la predicción de elementos *L-shape* en base al uso de SVM podrían ser introducidas.

3. Conclusiones

Aun que es cierto que este método por si solo no podría arrojar un resultado seguro sobre la importante de la regulación sobre la regulación de su expresión, Mediante este trabajo hemos comprobado que el método del estudio de las correlaciones entre expresión y metilación implementando *scagnostics[4]* genera unos buenos resultados. Esto ha quedado retratado como hemos podido valor en la comparación con otros métodos y la observación de elementos individuales.

La ventaja de este método sobre otros, es que no solamente permite determinar la interacción en forma de patrón L al respecto de dos valores. Si no que permite extrapolar este tipo de estudios a otras relaciones y formas. En concreto, el hecho de implementar una función de creación de elementos individuales y la combinación con otros métodos para su clasificación utilizando SVM permite obtener buenos resultados con una precisión muy alta.

Adicional a este estudio , se han implementado una aplicación shinny que recoge este tipo de análisis y que permitiría con el tiempo la adición de mayores funcionalidades , que por restricción de tiempo en el desarrollo del trabajo ha sido imposible ser planificadas.

Aun que el nivel técnico del trabajo podemos considerar que se ha implementado un nuevo método gracias a la combinación *artifical set – scagnostics[4]* – SVM., al nivel de detección de genes biomarcadores para CRC (un objetivo del proyecto) no hemos conseguido grandes resultados. Así pues hemos podido observar que aquellos genes seleccionados que pudieran representar un patrón en forma de L en el set general y así como la selección de genes obtenidos mediante minería de datos no aparecen relaciones entre si. Sin embargo, la implementación de este método sobre otros estudios mas depurados podría arrojar mejores resultados y sugiere que existe un sesgo en el set de datos utilizado. Aun que útil para la implementación del método como paso inicial, pero presenta demasiado ruido en la detección potencial de biomarcadores, asumiendo el mismo sesgo biológico al respecto de la metilación y la expresión.

En conclusión, este estudio ha cumplido con los objetivos propuestos a principio del plan de trabajo, así como al cumplimiento de los diferentes puntos de control planteados durante el mismo. De los dos grandes tareas, en la primera la detección de genes ha permitido la implementación de un método cambiando el enfoque directamente a la representación de la correlación , así como la generación de dos sets de datos que ha falta de depurarse más podrían utilizarse para futuras investigaciones. La segunda tarea ha permitido la generación de una aplicación *shiny* modular que permita añadir otros tipos de análisis a la misma para su correcto uso por investigadores que no tengan conocimiento de programación para la detección de diferentes tipos de patrones de correlación.

4. Glosario

- **SVM:** Suport Vector Machine, algoritmo de machine learning utilizado como modelo de caja negra que se basa en representación de un hiper plano de las características y su correcta separación

- **CRC:** Colonrectal Cancer, un tipo de cáncer agresivo de colon.

- **R:** Programming language R , lenguaje de programación estadístico. Tiene su origen el lenguaje de programación S.

- **Bioconductor**: Repositorio de paquetes R para el análisis biológico

- **Scagnostics[4]**: Algoritmo de definición de los parámetros para definir un *scatterplot*

- **Metilación**: Proceso de adición de un grupo metilo sobre una citosina en el genoma. Se trata de un proceso de regulación epigenético.

- **Expresión Génica:** Valor de expresión de un gen en el genoma. Esta fluctúa y cambia en función de la actividad de la célula.

- **Gene Ontology:** Ontología de términos biológicos. En este estudio se ha utilizado la de procesos biológicos(BP), pero otras podrían definirlo como la localización de los componentes o la función molecular

- **GEO**: Gene Expression Onmybunus: Repositorio mantenido por el NCBI en EEUU en el que se pueden encontrar diversos sets de publicaciones

- **Shinny**: Librería de R que permite el diseño y la implementación de aplicaciones web utilizando el lenguaje de programación R
- **Biomarcador**: Molécula/gen que sirve como indicador de una situación o condición como es el caso de una enfermedad.

5. Bibliografía

[1] Yamazaki J, Jelinek J, Lu Y, Cesaroni M, Madzo J, Neumann F, He R, Taby R, Vasanthakumar A, Macrae T, Ostler KR, Kantarjian HM, Liang S, Estecio MR, Godley LA, Issa JP. TET2 Mutations Affect Non-CpG Island DNA Methylation at Enhancers and Transcription Factor-Binding Sites in Chronic Myelomonocytic Leukemia. Cancer Res. 2015 Jul 15;75(14):2833-43.

[2] Wang, K.-S. Integrative Analysis of Genome-wide Expression and Methylation Data. J. Biom. Biostat. 4, (2013).

[3] Sarah Bazzocco, Ha d Alazzouzi, M. Carme Ruiz de Villa, Alex Sanchez-Pla, John M. Mariadason, Diego Arango (2013) Genome-Wide Analysis of DNA Methylation in Colorectal Cancer. Submitted.

[4] JW Tukey, Tukey PA: Computer graphics and exploratory data analysis: An introduction . In: National Computer Graphics Association (ed.): Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics85 . III. Fairfax, VA. 1985.

[5] Scagnostics[4] Wikipedia Page: https://de.wikipedia.org/wiki/Scagnostics[4]

[6] Data-Mining coremine aplication. http://www.coremine.com/

[7] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[8] Maria L. Rizzo and Gabor J. Szekely (2016). energy: E-Statistics: Multivariate Inference via the Energy of Data. R package version 1.7-0. https://CRAN.R-project.org/package=energy

[9] Jean Hausser and Korbinian Strimmer (2014). entropy: Estimation of Entropy, Mutual Information and Related Quantities. R package version 1.2.1. https://CRAN.R-project.org/package=entropy

[10] Patrick E. Meyer (2014). infotheo: Information-Theoretic Measures. R package version 1.2.0. https://CRAN.R-project.org/package=infotheo

[11] GEO: https://www.ncbi.nlm.nih.gov/geo/

[12] R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.14.0. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org

[13] Davis S and Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." Bioinformatics, 14, pp. 1846–1847.

[14] Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–767 (1990).

6. Anexos

- Tabla de genes con importancia en Cáncer (gene.target.csv)
- Tabla con los set de datos de Cáncer Colon rectal (geo.id.colon-cancer.txt)
- Código en .R con las funciones del análisis (TFM_analysis.R)
- Aplicación Shinny en un archivo Comprimido (ScagnosticsShinny.zip)