CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS
DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD ZACATENCO
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA
SECCIÓN DE BIOLECTRÓNICA

Análisis de la viabilidad del empleo de mediciones de alteración en vías metabólicas como alternativa al uso de la expresión genética para clasificar muestras de cáncer de mama

Tesis que presenta

Jonathan Alejandro Delijorge Ramírez

para obtener el grado de

Maestro en Ciencias

en la Especialidad de

Ingeniería Eléctrica

Director de tesis: Dr. David Elías Viñas

México, D.F. febrero, 2016

RESUMEN

Hoy en día no hay rama de la ciencia que no está involucrada en la lucha contra el cáncer. Una de las formas más mortíferas de esta enfermedad es el cáncer de mama. Se ha logrado detectar que el cáncer de mama puede dividirse en subtipos, y que cada subtipo responde mejor a ciertas terapias sin embargo, el diagnóstico de un paciente dentro de uno de los subtipos no es fácil, por lo que el uso de técnicas computacionales ahora es común para este propósito. La mayoría de las técnicas computacionales empleadas en la actualidad han basado su funcionamiento en mediciones de la expresión genética de las células afectadas pero, esto representa algunos problemas como la dimensionalidad de las bases de datos, además de que los resultados reportados aún no logran la exactitud en la clasificación deseada. Es por esos problemas que en el presente trabajo se busca probar si el uso de mediciones de alteración en los pathways metabólicos de los enfermos de cáncer de mama es una alternativa viable al uso de las mediciones de expresión.

ABSTRACT

Today, there is not branch of science unrelated with the battle against the cancer. Breast cancer is one of the deadliest forms of this disease. It has been demonstrated that breast cancer can be divided in subtypes, and each one responds better to certain therapies nevertheless, the diagnosis of a patient within one subtype is not easy, so the use of computational techniques is now common for this purpose. Most of computational techniques base their performance on measurements of gene expression of affected cells but, this represents some disadvantages such as the base data dimensions. Furthermore, the reported results have not achieved the desired exactitud of classification. For this reasons, the present investigation wants to probe if the use of measurements of deregulation in metabolic pathways in patients with breast cancer is a valid option to the use of expression measurements.

A toda mi familia, en especial a mis padres, a mis hermanos y a Grecia, por su confianza, cariño y apoyo.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología por confiarme el apoyo de la beca CONACYT para realizar mis estudios de maestría.

Al Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV), en especial al doctor David Elías Viñas y al laboratorio 12 de la sección de Bioelecrónica; y de igual forma, al Instituto Nacional de Medicina Genómica (INMEGEN), especialmente al doctor Enrique Hernández Lemus y al consorcio de Genómica Computacional, por su brindarme su apoyo para la realización de la presente investigación.

Contenido General

Lista de figuras

Lista de tablas

Nomenclatura

$\mathbf{A}$	Matriz.
$\mathbf{A}^{-1}$	Inversa de $\mathbf{A}$.
$\mathbf{A}^{H}$	Hermitiano de $\mathbf{A}$.
$\mathbf{A}^{T}$	Transpuesta de $\mathbf{A}$.
AUC	*Area under the curve.*
CDF	*Cumulative distribution function.*
$det(\mathbf{A})$	Determinante de $\mathbf{A}$.
$diag(\mathbf{A})$	Diagonal de $\mathbf{A}$.
$eig(\mathbf{A})$	Eigenvalores de $\mathbf{A}$.
k	Número de *clusters*.
$Tr(\mathbf{A})$	Traza de $\mathbf{A}$.

Capítulo 1

Introducción

1.1 Planteamiento del problema

La presente investigación surge de la siguiente cuestión: teniendo una base de datos que contiene observaciones de casos de cáncer de mama y sus respectivas mediciones de expresiones genéticas, ¿la clasificación en subgrupos de estos casos basándose en las mediciones de los genes individuales es coherente con una clasificación basada en mediciones hechas sobre conjuntos de esos genes?

1.2 Importancia de la investigación

El cáncer de mama es el tipo de cáncer más común en las mujeres del mundo. Se estima que en 2011, 508 mil mujeres murieron en el planeta a causa de éste [1]. Este cáncer es el más mortífero para las mujeres mexicanas, Cada dos horas una mujer muere en este país debido a dicha afección [2].

El cáncer de mama, como todos los tipos de cáncer, es un afección sumamente compleja que altera de forma distinta a cada persona que lo padece. El hecho de que cada caso de cáncer sea único explica la complejidad de diseñar un tratamiento general para combatirlo.

Aún con lo heterogéneo que parece ser el universo de casos de cáncer de mama, se han logrado observar grupos de casos que son similares en sus perfiles genéticos, y lo más importante es que también son similares en sus respuestas a los tratamientos. Lo anterior explica el esfuerzo de la comunidad científica en años recientes para clasificar al cáncer de mama en subtipos más homogéneos, para posteriormente poder diseñar tratamientos enfocados a cada subtipo.

A partir de las investigaciones enfocadas a la clasificación del cáncer de mama, se ha logrado detectar un número limitado de genes, cuya actividad o inactividad generalmente se asocia de forma directa al desarrollo de cáncer de mama.

1.3 Objetivo general

El objetivo de este trabajo es probar si, empleando inteligencia artificial, la clasificación de casos de cáncer de mama tomando como factor de discriminación las características observadas sobre conjuntos de genes relacionados entre sí, tiene coherencia con las clasificaciones basadas en mediciones individuales sobre los genes por separado.

1.4 Estructura general

Se comienza con la presentación de los temas principales del marco teórico que fundamenta a los métodos empleados. Se enlistan las investigaciones existentes referentes al tema de esta tesis y se hace un análisis de las mismas.

A continuación, se explican los fundamentos teóricos y técnicos, así como la secuencia de los métodos y técnicas empleadas.

Posteriormente se describen los parámetros finales elegidos para cada algoritmo empleado.

Después se muestran, a forma de comparación, los resultados derivados tanto para el empleo de mediciones de expresión como base, como para el uso de mediciones de alteraciones en *pathways*.

Finalmente, se muestra la interpretación y discusión de los resultados así como la conclusión de la investigación.

Capítulo 2

Antecedentes

2.1 Clasificadores computacionales existentes

Los clasificadores de cáncer de mama más comunes actualmente son aquellos en donde cada muestra (caso de tumor cancerígeno) se agrupa con otras muestras con las que tiene similitud en sus mediciones de expresión genética, separándose a su vez de las muestras con las que no tiene similitud. La expresión genética es el proceso por el cual la información codificada en los genes es convertida en estructuras funcionales en la célula, en su mayoría proteínas. Algunos ejemplos recientes de métodos que clasifican a los casos de cáncer de mama basándose en la medición de la expresión de sus genes son presentados de [3] a [10].

Basándose en la expresión genética, la forma más aceptada de dividir al cáncer de mama en subgrupos es la llamada clasificación molecular, la cual distingue cuatro subtipos de cáncer: luminal A, luminal B, HER2+ y basal [11]. La mayoría de los clasificadores adopta esta división, entonces cada observación de cancer que sea ingresado al clasificador es encasillado al final dentro de uno de los subtipos mencionados.

2.2 Comparación de técnicas no supervisadas y supervisadas

El empleo de técnicas computacionales para separar un conjunto de observaciones en grupos o clases tiene dos enfoques: la clasificación[1] no supervisada y la supervisada [12].

2.2.1 Técnicas no supervisadas

Cuando se tiene un conjunto de observaciones y mediciones realizadas sobre algunos atributos de éstas, las técnicas no supervisadas se emplean para la partición del conjunto en subconjuntos. Las observaciones que son colocadas en un mismo subconjunto (o *cluster*) tienen características (generalmente estadísticas) similares. Las técnicas de este tipo obedecen dos principios fundamentales:

- no se indica al programa el número de clases en las que debe dividir a las observaciones.
- no hay una etapa de entrenamiento previa a la clasificación en donde el programa aprenda las características de las observaciones pertenecientes a cada clase.

2.2.2 Técnicas supervisadas

En este enfoque, antes de la etapa de clasificación, el programa ya cuenta con información acerca de las clases y sus características. Esta información es provista por el usuario, quien interviene directamente en este tipo de técnicas. En estos clasificadores, al contrario que en los no supervisados:

- el usuario conoce e indica al programa el número de clases en las que debe dividir a las observaciones.

[1]En inglés, suele hacerse una distinción, en donde la palabra clasificación (*classification*) es usada exclusivamente para las técnicas supervisadas, mientras que para las no supervisadas, se usa la palabra *clustering*.

- previo a la etapa de clasificación, hay una etapa de entrenamiento, en donde el usuario introduce observaciones e indica al programa a qué clase pertenecen. De esta manera el programa aprende las características generales de las observaciones propias de cada clase.

2.3 Teoría matricial

Tanto las técnicas supervisadas como las no supervisadas buscan clasificar un conjunto de objetos, cada uno con una un respectiva serie de mediciones hechas sobre atributos o características comunes a todos ellos. La forma estándar de organizar a los objetos con sus mediciones es una base de datos. La imagen 2.1 es una sección de una base de datos.

	row.names	A1CF	A2M	A4GALT	AAAS	AACS	AANAT	AARS	AASDHPPT
1	GSM107072.CEL	6.89271	11.42930	6.57599	8.53620	8.26943	6.07011	10.02890	8.719968
2	GSM107073.CEL	7.21704	10.39070	6.37221	9.04925	8.71750	6.19042	10.41290	9.195175
3	GSM107074.CEL	7.55959	10.61370	6.64089	8.73067	8.81086	6.44507	9.38137	8.735811
4	GSM107075.CEL	7.22727	11.72890	6.21051	8.55977	7.75690	6.12810	9.91653	8.616220
5	GSM107076.CEL	7.31756	11.80180	6.44041	8.84267	8.20941	6.16254	9.84520	8.597745

Figura 2.1 Extracto de una base de datos que contiene la medición de expresión genética de observaciones de cáncer de mama.

En la imagen 2.1 se aprecia que cada columna tiene un encabezado, indicando la información que ésta contiene. La primer columna (*row.names*) contiene a todas las observaciones (cinco), que en este caso son casos de cáncer de mama, identificados con nombre (con formato *GSM10707X.CEL*); para la segunda columna en adelante, el encabezado indica el nombre identificador de un gen (se observa que el primero de ellos es *A1CF*). Entonces, en el renglón correspondiente a cada caso, se encuentran las mediciones de sus atributos, que en este caso son mediciones de la expresión de sus genes.

Si ignoramos por un momento los encabezados de columna y los identificadores de los objetos (primera columna), lo que obtenemos es únicamente mediciones numéricas de los

atributos organizados en renglones y columnas. Es evidente que podemos analizar esta información como a una matriz de dimensión mxn, donde m es el número de observaciones (renglones) y n es el número de atributos o mediciones (columnas). Debido a lo anterior, no es sorpresa que los clasificadores, supervisados o no, basen su funcionamiento en operaciones matriciales.

Gracias a las técnicas computacionales, es posible aplicar las operaciones matriciales de forma instantánea y automática, por ello, muchas veces se deja de analizar la teoría de dichas operaciones y el significado de las mismas. Para facilitar la comprensión de las técnicas de clasificación, es conveniente recordar los conceptos básicos del álgebra matricial [13].

2.3.1 Propiedades básicas de las matrices

Las expresiones (2.1)-(2.7) definen las propiedades básicas de las matrices.

$$(\mathbf{ABC}...)^{-1} = ...\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} \tag{2.1}$$

$$(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T \tag{2.2}$$

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T \tag{2.3}$$

$$(\mathbf{ABC}...)^T = ...\mathbf{C}^T\mathbf{B}^T\mathbf{A}^T \tag{2.4}$$

$$(\mathbf{A}^H)^{-1} = (\mathbf{A}^{-1})^H \tag{2.5}$$

$$(\mathbf{A} + \mathbf{B})^H = \mathbf{A}^H + \mathbf{B}^H \tag{2.6}$$

$$(\mathbf{ABC}...)^H = ...\mathbf{C}^H\mathbf{B}^H\mathbf{A}^H \tag{2.7}$$

donde **A**,**B** y **C** son matrices.

2.3.2 Traza

La traza de una matiz está definida por la ecuación (2.14).

$$Tr(\mathbf{A}) = \sum_i \mathbf{A}_{ii} \tag{2.8}$$

2.3.3 Inversa

Siendo **A** una matriz cuadrada, su inversa $\mathbf{A}^{-1}$ está definida si se cumple la relación (2.9).

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = I \tag{2.9}$$

donde I es la matriz identidad de nxn.

Si $\mathbf{A}^{-1}$ no existe, el sistema en **A** es singular.

2.3.4 Transpuesta

Cuando convertimos los renglones de una matriz **A** en las columnas de otra, la matriz resultante es la transpuesta de la matriz original ($\mathbf{A}^T$). Esta operación se define por la ecuación (2.10).

$$(\mathbf{A})_{i,j} = \mathbf{A}_{j,i} \tag{2.10}$$

2.3.5 Eigenvectores y eigenvalores

v es un eigenvector y λ es un eigenvalor de la matriz cuadrada (nxn) **A** si cumplen con la relación (2.11).

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{2.11}$$

2.3.6 Descomposición en valores singulares

Cualquier matriz mxn puede expresarse como en (2.12):

$$\mathbf{A} = UDV^T \tag{2.12}$$

donde U son los eigenvectores de $\mathbf{A}\mathbf{A}^T$.

V son los eigenvectores de $\mathbf{A}^T\mathbf{A}$.

D está definida por la ecuación (2.13).

$$D = \sqrt{diag(eig(\mathbf{A}\mathbf{A}^T))} \tag{2.13}$$

2.3.7 Determinante

El determinante es un escalar; cuando el determinante es diferente de cero, significa que existe una solución única para el sistema matricial, de lo contrario, el sistema es singular (no hay una solución única definida para el sistema). Sea **A** una matriz cuadrada, el determinante se define por la ecuación (2.14).

$$\det(\mathbf{A}) = \prod_i \lambda_i \tag{2.14}$$

donde λ_i es el i-ésimo eigenvalor de la matriz **A**.

2.3.8 Regresión lineal

La regresión lineal es un método para modelar matemáticamente la relación entre una variable dependiente Y, una o más variables dependientes X_j y un término aleatorio ε. Dicho modelo es descrito por la relación (2.15).

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \tag{2.15}$$

donde Y_t es la variable dependiente.

$X_1, X_2, ..., X_p$ son las variables independientes.

$\beta_0, \beta_1, \beta_2, ..., \beta_p$ son parámetros que miden la influencia de las variables independientes sobre las dependientes.

p es el número de parámetros independientes a tener en cuenta en la regresión.

2.3.9 Regresión no lineal

La regresión no lineal corresponde a un modelado basado en datos multidimensionales (x,y), descrito por la ecuación (2.16).

$$y = f(x, \theta) + \varepsilon \tag{2.16}$$

donde f es una función no lineal con respecto a los parámetros desconocidos θ.

ε es un término aleatorio de ajuste.

En este tipo de regresión se pretende obtener los valores de los parámetros asociados con la mejor curva de ajuste.

2.4 Análisis de la investigación existente

En la mayoría de las publicaciones relacionadas con la clasificación de cáncer de mama para el auxilio en el diseño de terapias, los clasificadores discriminan entre los subtipos basándose en mediciones de la expresión de genes individuales. Sin embargo, a este tipo de enfoque pueden ser asociadas las siguientes observaciones.

1. Gracias a estas investigaciones se han detectado algunos genes asociados directamente a cada subtipo de cáncer de mama (oncogenes), pero la actividad de los genes no es independiente sino que la actividad de un gen es afectada por otros genes. La clasificación basada en expresiones individuales no toma en cuenta estas interacciones.

2. El diseño de tratamientos (farmacéuticos o de otra índole) enfocados a la actividad de genes individuales es sumamente compleja. Por ahora, es mucho más práctico diseñar terapias que ataquen a un conjunto de genes asociados al cáncer.

3. Cuando se tienen bases de datos (matrices) de expresión genética, es común que el número de genes sea mucho mayor que el número de observaciones ($m << n$), arbitrariamente, las técnicas de clasificación convencionales están diseñadas para bases de datos donde el número de observaciones u objetos es mayor que el número de mediciones de las características de éstos ($m > n$) por lo que hay que recurrir a técnicas más complejas y herramientas de extracción de características. Ordenar a los genes en conjuntos permitiría involucrar a todos los genes en el proceso de clasificación y, a la vez reducir la dimensión de las mediciones de características, facilitando la aplicación de algoritmos convencionales.

Puede notarse que la clasificación basada en conjuntos de genes tiene algunas ventajas potenciales sobre la clasificación basada en mediciones genes individuales, pero, al realizar la clasificación con este enfoque diferente ¿cómo se verá afectado el resultado con respecto al enfoque tradicional? Ya se tienen avances importantes usando la división del cáncer de mama en los cuatro subtipos moleculares mencionados a priori, si se hace una clasificación no supervisada (*clustering*) basada en conjuntos de genes ¿se agruparán los casos en estos mismos cuatro subtipos? La necesidad de una comparación entre los dos enfoques es la base del planteamiento del problema central y del objetivo de esta investigación.

Capítulo 3

Desarrollo

3.1 Bases de datos iniciales

El punto de partida fueron tres bases de datos. Las primera correspondía a una base de 493 muestras de personas con alguno de los cuatro subtipos de cáncer de mama, cada una con la medición de la expresión de sus genes. La segunda también era una base de muestras por genes, solo que estas muestras correspondían a personas sin cáncer de mama [14]. La imagen mostrada en la figura 2.1 corresponde a un segmento de la primera matriz, donde se observan las muestras como renglones y los genes como columnas.

La tercera base de datos correspondía a una lista de 1322 pathways por genoma completo (pathways por genes), encontrando para cada pathway, un 1 en la intersección con los genes que estaban relacionados a éste y un 0 en las intersecciones con los genes con los que el pathway no tenía relación comprobada. Para visualizar mejor esta base, un segmento de la misma es mostrada en la figura 3.1.

	row.names	ACAA1	ACAA2	ACACA	ACACB	ACAD8	ACADL	ACADM	ACADS
1	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	1	1	0	0	1	0	1	1
2	REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	1	0	1	1	0	1	1	1

Figura 3.1 Extracto de la base de datos que indica (con 1s) cuáles genes están asociados a las funciones metabólicas celulares (*pathways* contra genes).

Partiendo de la figura 3.1 puede observarse que los genes ACCA1, ACAA2, ACAD8, ACADM y ACADS participan de forma importante para llevar a cabo la función descrita por el primer $pathway$ (degradación de valina, leucina e isoleucina), mientras que para el segundo pathway (metabolismo de lípidos y lipoproteínas) hay evidencia documentada de una fuerte relación con la expresión de los genes ACCA1, ACACA, ACACB, ACADL, ACADM y ACADS.

3.2 Curado y limpieza de las bases de datos

Para llevar a cabo el proceso planeado fue antes necesario dar un preprocesamiento a las bases de datos, por lo que se escribió un código que realizara las siguientes funciones:

- Eliminar muestras con información incompleta o incorrecta.
- Quitar genes cuyo nombre correspondía a claves de identificación que hacían difícil reconocer exactamente a qué gen se refería y por lo tanto, no era posible asociarlo a los pathways.
- Dejar solo los genes que aparecían en las tres bases de datos que se tenían (base de datos de enfermos de los cuatro subtipos, de sanos y pathways-genes), ya que no tenía sentido conservar genes relacionados a pathways en la tercera base de datos si no se tenía medición de la expresión de éstos en las primeras dos; análogamente, tampoco tenía sentido conservar aquellos genes que aparecían en las bases de expresión de sanos o enfermos si no tenían relación con alguno de los pathways de la base de pathways-genes.
- Una vez que las tres bases de datos tenían la información de los mismos genes, hacer una sola matriz de expresión que contenía las muestras tanto de enfermos como de sanos.

3.3 Base de datos de expresión

El punto de partida formal del proceso fue la base de datos obtenida del preprocesamiento descrito a priori, la cual consistía de la medición de la expresión de 6345 genes para 554 observaciones (personas), es decir, una matriz de dimensión $554x6345$. Las observaciones de esta base de datos ya habían sido clasificadas previamente dentro de los subtipos moleculares o como personas sanas. En la tabla 3.1 se muestra la distribución de las observaciones.

Tabla 3.1 Distribución de la base de datos de expresión genética.

Subtipo	**Observaciones**
Luminal A	172
Luminal B	141
Basal	123
HER2+	57
Sanos	61
Total	**554**

3.4 *Pathifier*

Como se mencionó antes, la intención del presente trabajo es analizar la clasificación de observaciones de cáncer de mama basándose en mediciones hechas sobre conjuntos de genes, en comparación con las basadas en genes individuales. Una primera cuestión fue cómo elegir estos conjuntos, en donde los genes contenidos en cada uno estuvieran relacionados entre sí.

Es evidente que cuando el cáncer de mama está presente, las funciones celulares (*pathways*) se ven afectadas, incluso se ha detectado un número limitado de funciones asociadas

inmediatamente a esta afección [15] y cuya alteración de cada una parece interactuar con las alteraciones de las otras funciones confiriendo así al cáncer su naturaleza compleja.

Estas funciones están controladas por los genes; para cada *pathway* hay un conjunto de genes que es asociado a su regulación. Hasta aquí, la primera cuestión está resuelta, es posible obtener un cierto número de *pathways* donde cada uno involucra a un conjunto de genes, pero ahora, la segunda cuestión sería ¿cómo obtener una medición cuantitativa para cada uno de estos *pathways*? Esta segunda cuestión fue resuelta con el algoritmo *Pathifier* [16]-[18], implementado en el lenguaje de programación R.

Pathifier es un algoritmo que permite cuantificar las alteraciones en los *pathways* de las observaciones de cáncer. A este algoritmo se le indican una serie de *pathways* y cuáles genes están asociados a cada uno de ellos. También es alimentado con las mediciones de expresión genética de observaciones de pacientes sanos. Al comparar la expresión de los genes en los *pathways* de los sanos con la expresión de los genes en los pathways de los enfermos, se encuentra una diferencia numérica. Esta diferencia sirve como medida de alteración para cada uno de los pathways.

3.5 *Consensus Clustering*

Habiendo obtenido una base de datos con observaciones de cáncer de mama y mediciones de sus *pathways*, el siguiente paso consistía en observar si la clasificación basada en mediciones de *pathways* era consistente con la clasificación basada en mediciones de expresión. Para esta finalidad, la mejor técnica es el *clustering*, que permitiría comprobar si ambos enfoques formaban eran consistentes, tanto en el número de *clusters* como en la composición de los mismos, sin forzar a las observaciones a ser encasilladas dentro de alguno de los subtipos ya conocidos, como sería en el caso de la clasificación supervisada.

La primer técnica seleccionada para abordar esta parte de la investigación fue *Consensus Clustering* [19], específicamente la herramienta desarrollada para R [20].

Consensus Clustering, es un método de *clustering* que permite evaluar el consenso de los *clusters* formados a través de un comportamiento iterativo, es decir, *Consensus Clustering* realiza una y otra vez el proceso de *clustering* hasta que las observaciones dejan de cambiar del *cluster* al que fueron asignados en los intentos previos hacia otro *cluster* diferente (alcanza cierta estabilidad). *Consensus Clustering* debe entenderse más como una herramienta que guía y asiste en el uso de alguno de los algoritmos de *clustering* que se incluyen dentro de ésta misma, y que además, provee de herramientas gráficas que ayudan a reconocer el número correcto de grupos o *clusters* en que se dividen las observaciones. A continuación se describe de manera general el algoritmo que sigue *Consensus Clustering*.

1. El algoritmo realiza la primera iteración de la técnica de clustering seleccionada para dividir al conjunto de muestras en dos grupos y crea una matriz de conectividad o de adyacencia. La matriz de adyacencia coloca a las muestras en linea horizontal y vertical y coloca un 1 para las muestras que fueron colocadas en un el mismo grupo y 0 para las que no se ubicaron en un mismo grupo. En la tabla 3.2 puede verse un ejemplo de matriz de adyacencia para nueve muestras, cuando se busca dividirlas en tres grupos; ésta indica que el primer grupo está compuesto por las muestras m_1, m_3 y m_5, en un segundo grupo fueron colocadas las muestras m_1, m_3 y m_5, y al tercer grupo fueron asignadas las muestras m_2, m_6 y m_8.

   Para cada iteración i se crea una matriz de adyacencia y al final, se crea una matriz μ (matriz de consenso) que es la suma normalizada de las i matrices de adyacencia generadas en el proceso. Esto se hace primero para dividir las muestras en dos grupos y se repite hasta el número máximo de grupos que el usuario desea generar.

Tabla 3.2 Ejemplo de matriz de adyacencia para nueve muestras.

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9
m_1	1	0	1	0	1	0	0	0	0
m_2	0	1	0	0	0	1	0	1	0
m_3	1	0	1	0	1	0	0	0	0
m_4	0	0	0	1	0	0	1	0	1
m_5	1	0	1	0	1	0	0	0	0
m_6	0	1	0	0	0	1	0	1	0
m_7	0	0	0	1	0	0	1	0	1
m_8	0	1	0	0	0	1	0	1	0
m_9	0	0	0	1	0	0	1	0	1

2. Una vez computada la matriz de consenso, ésta es reordenada, colocando de forma contigua a las muestras que resultaron pertenecer a un mismo grupo en al menos una iteración. En la tabla 3.3 se observa cómo quedaría la matriz de consenso reordenada, suponiendo aún una división en tres grupos y que éstos tuvieron la misma composición en todas las iteraciones (matriz de consenso perfecta compuesta de sólo unos y ceros).

   Si se reemplazan los valores numéricos de la matriz de consenso por intensidades de algún color (rojo, por ejemplo), donde 1 es la expresión más intensa y 0 es la más tenue (blanco), el resultado serían bloques de tonalidades. En una matriz de consenso perfecta, se distinguirían bloques de un solo color an un fondo blanco; por ejemplo, para una matriz de consenso de tres grupos, en donde éstos tuvieron la misma composición en todas las iteraciones (matriz de consenso perfecta) se obtendría un resultado como el que se muestra en la figura 3.2.

Tabla 3.3 Matriz de consenso μ reordenada.

	m_1	m_3	m_5	m_2	m_6	m_8	m_4	m_7	m_9
m_1	1	1	1	0	0	0	0	0	0
m_3	1	1	1	0	0	0	0	0	0
m_5	1	1	1	0	0	0	0	0	0
m_2	0	0	0	1	1	1	0	0	0
m_6	0	0	0	1	1	1	0	0	0
m_8	0	0	0	1	1	1	0	0	0
m_4	0	0	0	0	0	0	1	1	1
m_7	0	0	0	0	0	0	1	1	1
m_9	0	0	0	0	0	0	1	1	1

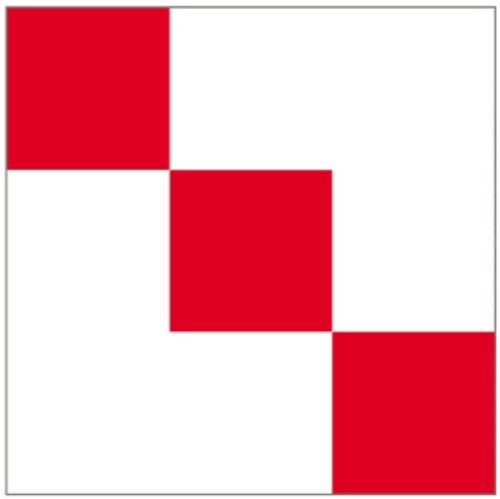


Figura 3.2 Matriz de tonalidades equivalente a la matriz de consenso de la tabla 3.3 [19].

La matriz de consenso visualizada con colores es la primer herramienta gráfica provista por *Consensus Clustering*.

3. Para cada *cluster* $k \in K$ el algoritmo define un parámetro llamado consenso de *cluster* $m(k)$; y para cada muestra $e_i \in D$ y cada cluster k se define el consenso de muestra $m_i(k)$. Los dos parámetros anteriores están descritos respectivamente por las ecuaciones (3.1) y (3.2).

$$m(k) = \frac{1}{N_k(N_k - 1)/2} \sum_{\substack{i,j \in I_k \\ i<j}} \mu(i,j) \tag{3.1}$$

$$m_i(k) = \frac{1}{N_k - 1\{e_i \in I_k\}} \sum_{\substack{j \in I_k \\ j \neq i}} \mu(i,j) \tag{3.2}$$

donde N_k es el número de elementos en el *cluster* k.

I_k es un indicador cuyo valor es 1 si las muestras involucradas están presentes en el *cluster* k[1].

$\{e_i \in I_k\}$ Es una función condicional que toma el valor de 1 si la condición dentro de los corchetes se cumple.

El parámetro *cluster* $m(k)$ es el consenso promedio entre los pares de muestras que pertenecen al cluster k. $m_i(k)$ mide el consenso promedio entre la muestra e_i y todas las demás muestras en el *cluster* k.

4. Se crea un histograma de los valores contenidos en la matriz de consenso, es decir, el algoritmo extrae todos los valores distintos de los elementos de la matriz y cuenta cuántos (densidad) elementos de la matriz tienen el mismo valor. Por ejemplo, para la matriz de la tabla 3.3, que tiene 81 elementos, el histograma mostraría que hay 24

[1] *Consensus Clustering* puede no tomar todas las muestras en todas las iteraciones, I_k sirve para tener control de este submuestreo.

elementos con valor 1 y 54 con valor de 0, y la gráfica de éste sería como la mostrada en la figura 3.3a. Es evidente que para el caso de una matriz de consenso perfecta, el histograma sólo mostrará conteos para valores de 0 y 1, mientras que para una matriz no perfecta, el histograma tendría más la forma de la figura 3.3b.

5. Se calcula la función de distribución acumulativa (CDF) del histograma de las matrices de consenso de las divisiones en los diferentes números de grupos. Esta función está definida por la ecuación (3.3) y su gráfica es de la forma mostrada en la figura 3.4.

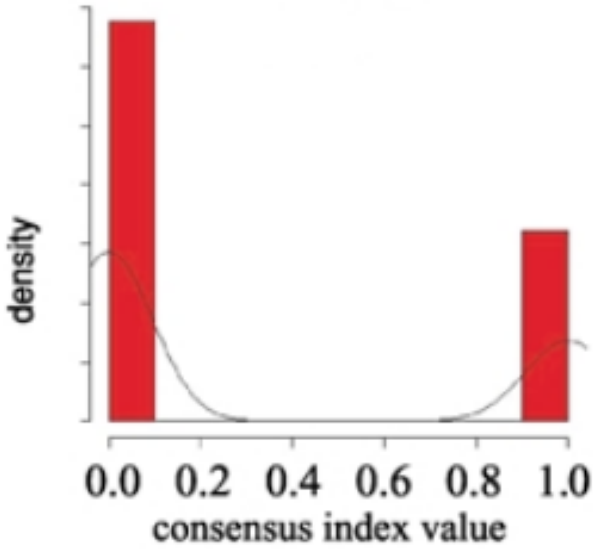
$$CDF(c) = \frac{\sum_{i<j} 1\{\mu(i,j)\} \leq c}{N(N-1)/2} \tag{3.3}$$

donde c son los valores de 0 a 1 en el histograma.
N es el número de muestras.

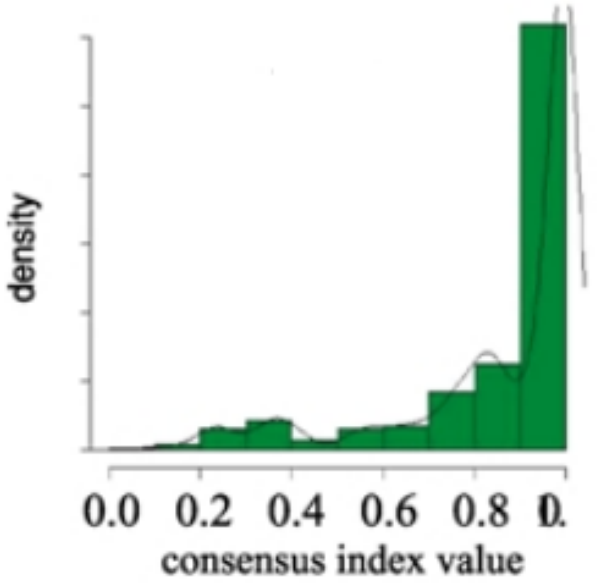
La gráfica de la CDF también supone una poderosa herramienta visual para determinar el número óptimo de *clusters*. Una forma inicial de interpretarla es medir la diferencia entre el valor de $CDF(0.8)$ y $CDF(0.2)$, una diferencia pequeña puede reflejar una buena estabilidad para la división en ese valor de k, aunque el valor más pequeño no siempre significa que esa curva corresponda al valor de k óptimo, también hay que analizar la forma de la curva. Además, la medición del área bajo la curva de las CDFs (*area under the (CDF) curve* AUC) de la gráfica para cada k significa otro parámetro de interés estadístico para hacer conclusiones.

6. Se calcula el AUC ($A(k)$) con la ecuación (3.4).

$$A(k) = \sum_{i=2}^{m} [x_i - x_{i-1}] CDF(x_i) \tag{3.4}$$

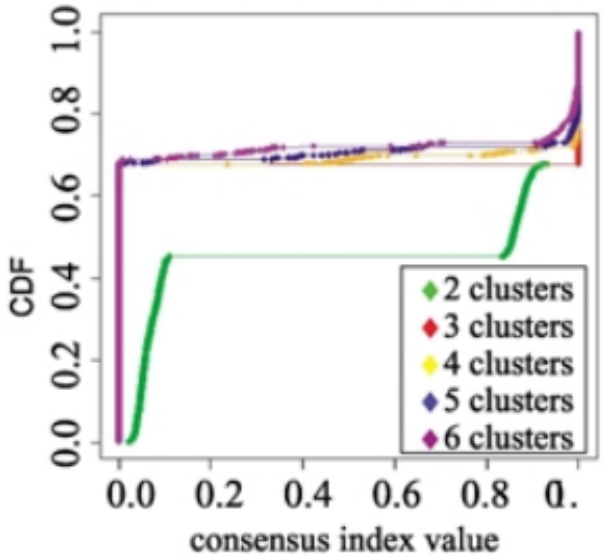


(a) Para matriz perfecta

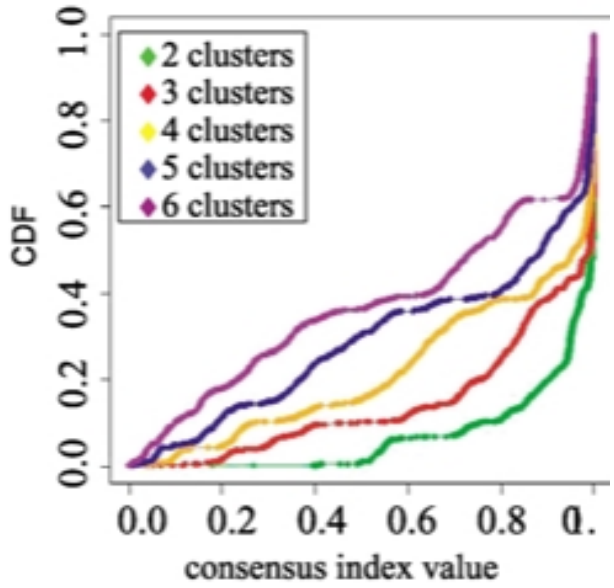


(b) Para matriz no perfecta

Figura 3.3 Histogramas para matrices de consenso [19].



(a) Para matriz perfecta ($k = 2$)



(b) Para matriz no perfecta

Figura 3.4 CDFs para histogramas. [19].

7. Para cada k se calcula el cambio en el área bajo la curva CDF ($\Delta(k)$). Si $k = 2$ se usa la ecuación (3.5), para $k > 2$, $\Delta(k)$ está definido por la ecuación (3.6)[2]. Si colocamos cada resultado con respecto a su k en una gráfica, lo que obtenemos es algo como lo que se muestra en la figura 3.5.

$$\Delta(k) = A(k) \tag{3.5}$$

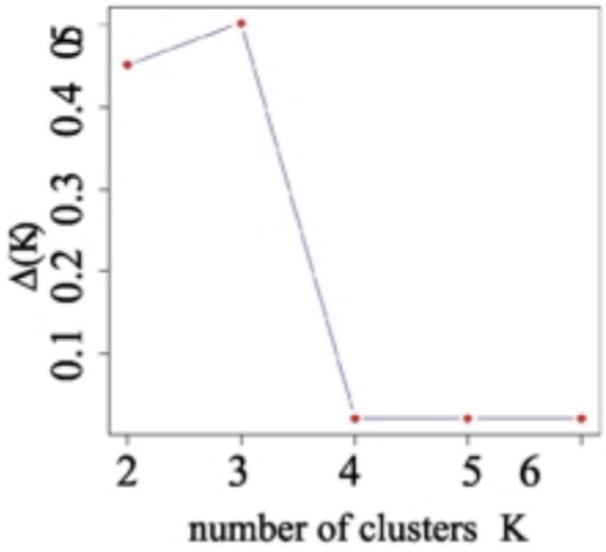
$$\Delta(k) = \frac{A(k+1) - A(k)}{A(k)} \tag{3.6}$$

Como la diferencia de la entre la $CDF(k)$ y $CDF(k+1)$ (y por lo tanto entre $\Delta(k)$ y $\Delta(k+1)$) se debe a cambios entre la composición de los $clusters$ al cambiar el número de clases, un cambio pequeño de una k a otra en la gráfica de $\Delta(k)$ significa que la composición de los grupos ha alcanzado cierta estabilidad, y sugiere que el número óptimo de $clusters$ probablemente ha sido encontrado.

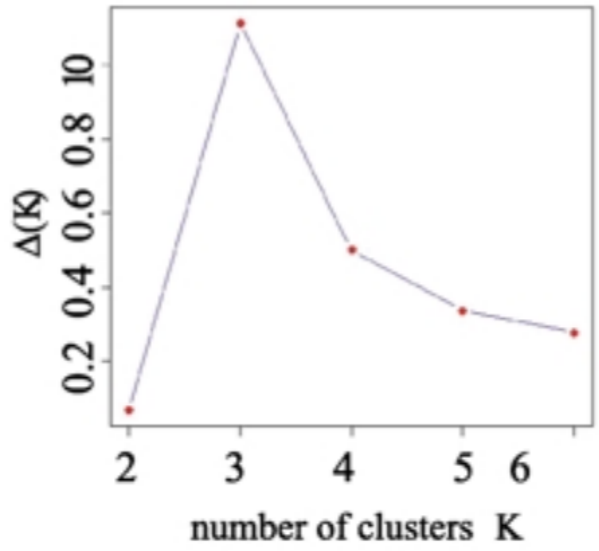
3.6 *K-means*

Después de una primera etapa de *clustering* , se decidió repetir el experimento con otra técnica, con la finalidad de comparar y validar los resultados; recordando además que *Consensus Clustering* es particularmente útil para conocer el número adecuado de *clusters* (k), ahora era necesaria una técnica para profundizar en el análisis de la composición de estos *clusters*. la herramienta elegida para llevar a cabo lo anterior fue *K-means* [21]. Como las demás técnicas de *clustering*, *K-means* distribuye un n observaciones en k *clusters*, donde las observaciones en un cierto *cluster* comparten características.

[2]Se usa una ecuación distinta para $k = 2$ ya que $\Delta(k)$ es una función que mide el cambio entre el resultado de $A(k)$ y los resultados de las ks subsecuentes, y dado que no hay análisis para $k = 1$, $\Delta(2)$ es el valor inicial.



(a) Para matriz perfecta (k=2)



(b) Para matriz no perfecta

Figura 3.5 $\Delta(k)$ para histogramas. [19].

Esta técnica puede ser clasificada como semi-supervisada, ya que no tiene una etapa de entrenamiento, pero se le indica el número de grupos k (definido en la etapa anterior con *Consensus Clustering*) en que debe dividir a las observaciones.

K-means realiza el proceso de *clustering* siguiendo estos pasos:

1. El algoritmo coloca puntos llamados centroides en un espacio n-dimensional, donde n es el número de observaciones, que en este caso son los 6345 genes para la matriz de expresión, o bien los 1322 *pathways* para la matriz de alteración. Cada punto representa un *cluster*. Los centroides pueden ser promedios de las mediciones de las observaciones o alguna otra medición estadística, incluso pueden ser colocados arbitrariamente.

2. Cada observación es colocada en el espacio y es asignada al centroide que resulta estar más cerca.

3. Se calculan los centros de cada *cluster* en base a las características de las observaciones que quedaron en cada *cluster* y este centro se convierte en el nuevo centroide.

4. Como los centroides fueron movidos, es posible que las observaciones estén ahora más cerca de un *cluster* diferente, al cual son reasignadas.

5. Cuando todas las observaciones son reasignadas, se vuelven a calcular a centroides.

6. El proceso de reasignación de las observaciones se repite hasta que la posición de los centroides ya no cambia significativamente.

Cabe mencionar que el procedimiento descrito arriba para K-means es la secuencia general en la que operan todas las técnicas de *clustering*, incluyendo a todas aquellas que son usadas por *Consensus Clustering*. Lo que varía entre las técnicas, es la medición estadística

con la que son calculados los centroides y la forma en que se mide la distancia entre estos últimos y las observaciones.

3.7 Diagrama de bloques

La figura 3.6 muestra el flujo de actividades seguido para la investigación en forma de diagrama de bloques.

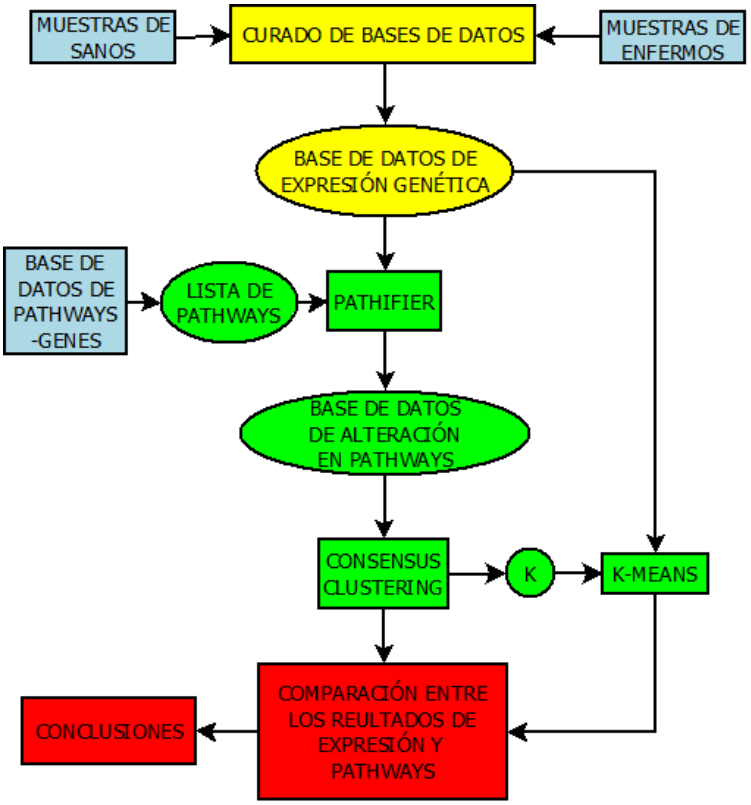


Figura 3.6 Diagrama de bloques de la investigación.

Capítulo 4

Pruebas

4.1 Parámetros finales seleccionados para los algoritmos empleados

4.1.1 *Pathifier*

La función principal para el uso del paquete de *Pathifier* en R es:

```
quantify_pathways_deregulation(data, allgenes, syms, pathwaynames,
normals, attempts)
```

data es la matriz numérica nxm donde n es el número de genes y m es el número de observaciones. En este caso $n = 6345$ genes y $m = 554$ observaciones.

allgenes es una lista que contiene los nombres (como caracteres) de los n genes.

syms es una lista de longitud igual al número de *pathways* que serán medidos, en este caso se usaron 1322 *pathways*. Cada elemento de esta lista contiene a su vez una sublista de los nombres de los genes (tal como aparece en *allgenes*) que están asociados a ese *pathway* particular.

pathwaynames es la lista que contiene los nombres de los 1322 pathways (el orden debe corresponder con el de *syms*).

normals es un vector booleano de la longitud de las observaciones que sirve como etiqueta, 0 para indicar que la observación corresponde a cáncer de mama, 1 para sanos. Recuérdese que en nuestra matriz se incluyen observaciones de 61 sanos.

attempts es el número de iteraciones. se cambió el valor preestablecido[1] de 100 a 50.

Pathifier entrega una medida de la alteración para cada *pathway* por cada observación, de esta forma es posible obtener una matriz de 554 observaciones x 1322 pathways a partir de la matriz inicial de 554 observaciones x 6345 genes.

4.1.2 *Consensus Clustering*

La función principal del paquete de Consensus Clustering es:

```
ConsensusClusterPlus(d, maxK, reps, clusterAlg=‘‘hc’’,
distance=‘‘pearson’’)
```

d es la matriz sobre la que será aplicado el *clustering*. En este caso, la matriz de expresión o bien, la de *pathways*.

maxK es el número máximo de *clusters* que debe formar el algoritmo. Es necesario aclarar que no se indica el número en que debe dividir a las observaciones, si no que el algoritmo repite la operación para $k = 2$ hasta $maxK$, para finalmente proveer las herramientas que ayudan a decidir cuál fue el mejor valor de k. En las pruebas (después de varios intentos previos) se decidió fijar $maxK = 10$.

reps es el número de iteraciones, en estas pruebas se eligió $reps = 5000$.

clusterAlg es el algoritmo de clustering en que *Consensus Clustering* basa su funcionamiento. El algoritmo establecido fue *Hierarchical Clustering (hc)*. El *clustering hc*(*clustering* jerárquico) funciona de la siguiente manera. Del universo de muestras, separa las que comparten características de resto, creando así las primera división con $k = 2$, del grupo de las muestras

[1]Los argumentos no enlistados se dejaron con los valores preestablecidos.

restantes toma las que son más parecidas ahora para crear un nuevo grupo ($k = 3$), y así sucesivamente conforme se pide aumentar el valor de k[22].

distance refiere al tipo de distancia que debe calcularse entre los centroides y las observaciones. Aquí se selecciono la distancia *pearson*.

4.1.3 *K-means*

En cuanto a K-means, R tiene una función para desarrollar el clustering con esta técnica, ésta es:

```
kmeans(x, centers, iter.max=10)
```

x, es la matriz, que nuevamente, en este caso es la matriz de expresión o de *pathways*.

centers es el número de *clusters*. Este parámetro fue seleccionado en base a los resultados entregados por *Consensus Clustering*.

iter.max es el número de iteraciones máximo permitido en caso de que no se llegue a una estabilidad en el posicionamiento de los centroides. Fue seleccionado un $iter.max = 5000$.

Capítulo 5

Resultados

5.1 Matriz de alteración en *pathways*

Gracias a *Pathifier* se obtuvo la matriz de 554 observaciones con la medición de la alteración en 1322 *pathways*. Ya se dijo antes que este algoritmo calcula esta alteración gracias a la comparación con las medidas de muestras de personas sanas. Para asimilar la medición de la alteración hecho por *Pathifier*, en la figura 5.1 se muestra el rango de mediciones (*score*) para las 61 observaciones de sanos incluidos en nuestra matriz y el rango de mediciones para las 493 observaciones restantes que corresponden a tumores de cáncer de mama (de los 4 subtipos moleculares) pertenecientes a uno de los 1322 pathways (regulación del ciclo de mitosis celular). La diferencia es notable. Un experimento extra probó que la distinción entre observaciones de sanos y enfermos es bastante sencilla si se emplea una técnica de clasificación sobre la matriz de alteraciones en *pathways* obtenida.

5.2 Número de *clusters* k

Tratando de encontrar el número de *clusters* contenido tanto en la matriz de expresión (para corroborar que corresponde con la bibliografía) como en la matriz de *pathways* (para verificar si corresponde con la matriz de expresión), se obtuvieron los siguientes resultados.

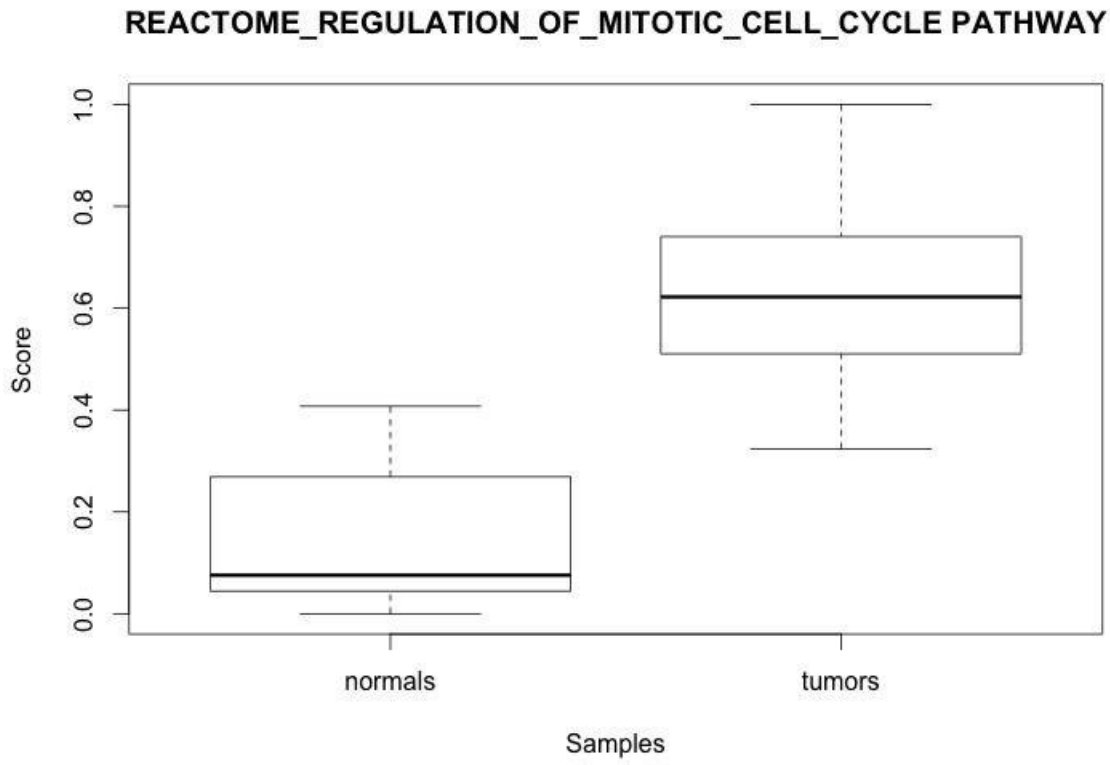
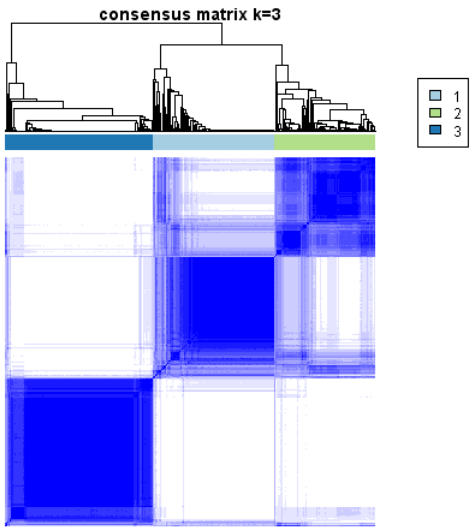
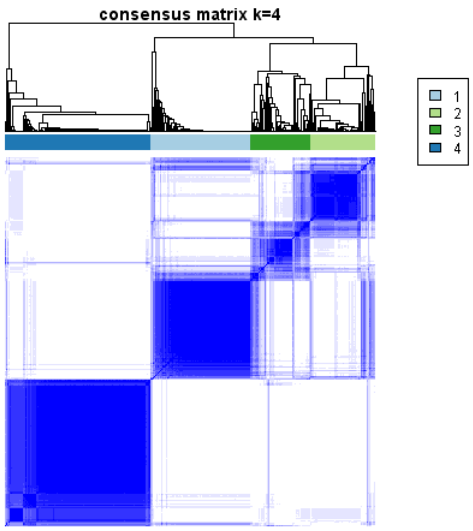


Figura 5.1 Mediciones de alteración en *pathway* de regulación de mitosis celular en sanos (*normals*) y enfermos de cáncer (*tumors*).

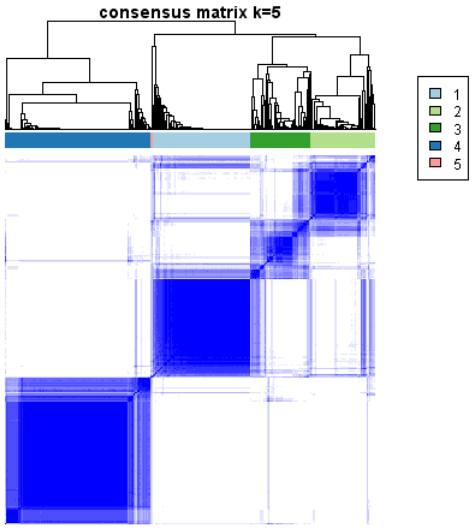
Recordando que la base de las herramientas gráficas y estadísticas de *Consensus Clustering* es la matriz de consenso, en la figura 5.2 se muestran dichas matrices, resultantes de ingresar las muestras correspondientes a la expresión genética de los enfermos. Las muestras de sanos fueron omitidas ya que su distinción del resto de las muestras era evidente, además de que el interés principal de la investigación es la distinción entre los subtipos de cáncer de mama. Sabiendo que el número de *clusters* real es de cuatro, se muestran las matrices de consenso para $k = 3$, $k = 4$, $k = 5$ y $k = 6$.



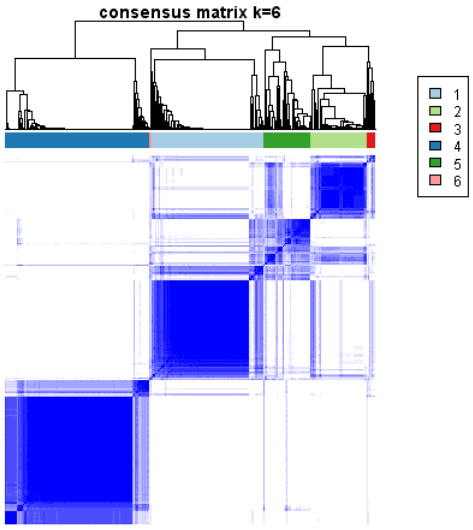
(a) $k = 3$



(b) $k = 4$



(c) $k = 5$



(d) $k = 6$

Figura 5.2 Matrices de consenso obtenidas para las muestras de expresión genética de los cuatro subtipos moleculares para la división en diferentes valores de k.

Para poder realizar una comparación inicial, en la figura 5.3 se muestran las matrices de consenso para $k = 3$, $k = 4$, $k = 5$ y $k = 6$ obtenidas al aplicar *Consensus Clustering* sobre la base de datos de alteración en pathways de las muestras de enfermos.

Después de obtener las matrices de consenso, se obtuvieron las gráficas de la CDF tanto para la matriz de expresión como para la de *pathways*. Los resultados se muestran en la figura 5.4.

Posteriormente, se calcularon y obtuvieron las gráficas para los valores de $\Delta(k)$ para ambas bases de datos, desde $k = 2$ hasta $k = 6$. Los resultados pueden observarse en la figura 5.5.

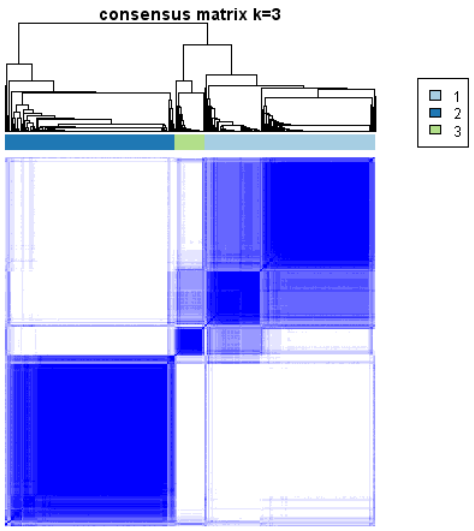
Como se ha mencionado antes, *Consensus Clustering* está enfocado principalmente a encontrar k sin embargo, a partir de las matrices de consenso para los diferentes valores de k, es posible construir una gráfica en la que se puede analizar en una sola imagen la evolución de la composición de los *clusters* a través de diferentes niveles de partición. Dichas gráficas resultantes para la matriz de expresión y pathways de muestran en la figura 5.6.

Finalmente, los valores de consenso de *cluster* obtenidos a partir de la ecuación (3.1) fueron trasladados a una gráfica de barras para observar la consistencia de los grupos creados con los diferentes valores de k. Está gráfica fue construida para ambas matrices y se muestra en la figura 5.7[1].

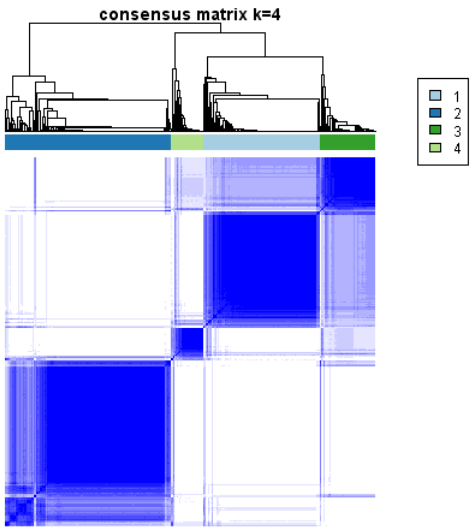
5.3 Composición de los *clusters* obtenidos con *consensus Clustering*

Se observa que, al menos el número de clusters corresponde al número de subtipos moleculares descritos en la bibliografía sin embargo, aún falta revisar si la composición de los

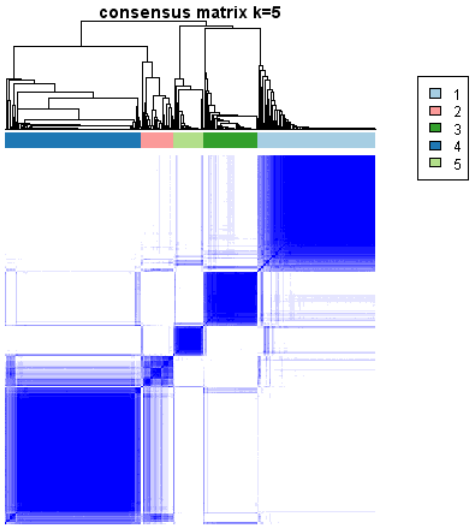
[1]Hay que recordar que las gráficas de las figuras 5.2-5.7 son los resultados tomando en cuenta solo las muestras de enfermos, sin tomar en cuenta las de sanos.



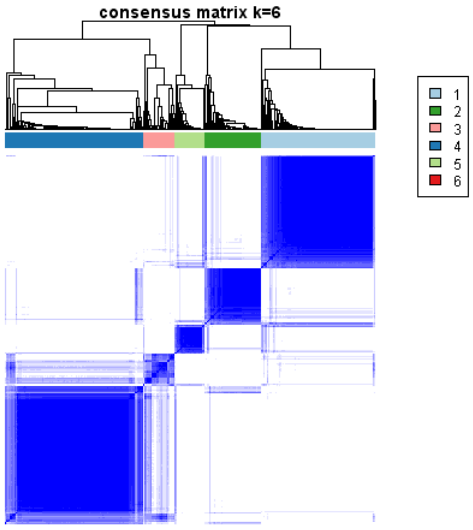
(a) $k = 3$



(b) $k = 4$



(c) $k = 5$



(d) $k = 6$

Figura 5.3 Matrices de consenso obtenidas para las muestras de alteración en pathways de los cuatro subtipos moleculares para la división en diferentes valores de k.

clusters también corresponde a los grupos de observaciones de los subtipos, con la ventaja de que se sabe a qué subtipo corresponden realmente las observaciones empleadas. *Consensus Clustering* también entrega la composición de los *clusters*, y estos resultados son mostrados a continuación.

5.3.1 Para el *clustering* sobre matriz de expresión

En la tabla 5.1 se observa la composición de los *clusters* obtenidos aplicar *Consensus Clustering* sobre la matriz de expresión.

Tabla 5.1 Tabla que muestra la cantidad de observaciones de cada subtipo en cada *cluster* obtenido por *Consensus Clustering* de la matriz de expresión.

Cluster \ Subtipo de las obs.	Luminal A	Luminal B	Basal	HER2+	Normal
1	44	50	0	5	0
2	5	31	0	8	0
3	28	34	0	0	0
4	94	7	1	5	61
5	1	19	122	39	0
TOTAL:	172	141	123	57	61

La tabla 5.2 refleja los resultados de la tabla 5.1, pero en porcentajes.

Dado que las muestras sanas pueden diferenciarse con mayor facilidad de las muestras de enfermos y pensando en que pudieran representar ruido al momento de realizar el clustering, fueron removidas y se repitió el *clustering*. Los resultados para $k = 4$ se observan en las tablas 5.3-5.4.

Tabla 5.2 Tabla que muestra el porcentaje de observaciones de cada subtipo en cada *cluster* obtenido por *Consensus Clustering* de la matriz de expresión.

Subtipo de las obs. / *Cluster*	Luminal A	Luminal B	Basal	HER2+	Normal
1	25.58	35.46	0	8.77	0
2	2.91	21.99	0	14.04	0
3	16.28	24.11	0	0	0
4	54.65	4.96	0.81	8.77	100
5	0.58	13.48	99.19	68.42	0

Tabla 5.3 Tabla que muestra la cantidad de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *Consensus Clustering* de la matriz de expresión.

Subtipo de las obs. / *Cluster*	Luminal A	Luminal B	Basal	HER2+
1	105	9	0	4
2	24	63	0	3
3	1	21	123	46
4	42	48	0	4
TOTAL:	172	141	123	57

5.3.2 Para el *clustering* sobre matriz de *pathways*

Siguiendo el procedimiento descrito anteriormente para la matriz de expresión, los resultados correspondientes a la matriz de *pathways* de observan en las tablas 5.5-5.8

Tabla 5.4 Tabla que muestra el porcentaje de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *Consensus Clustering* de la matriz de expresión.

Cluster \ Subtipo de las obs.	Luminal A	Luminal B	Basal	HER2+
1	61.05	6.38	0	7.02
2	13.95	44.68	0	5.26
3	0.58	14.89	100	70.70
4	24.42	34.04	0	7.02

Tabla 5.5 Tabla que muestra la cantidad de observaciones de cada subtipo en cada *cluster* obtenido por *Consensus Clustering* de la matriz de *pathways*.

Cluster \ Subtipo de las obs.	Luminal A	Luminal B	Basal	HER2+	Normal
1	122	6	0	6	0
2	14	49	0	0	0
3	5	49	123	45	61
4	0	0	0	1	0
5	31	37	0	5	0
TOTAL:	172	141	123	57	61

5.4 Validación de la composición de *clusters*

Dado que *Consensus Clustering* fue empleado principalmente para verificar el número óptimo de k, posteriormente se empleó *K-means* para comparar los profundizar en la composición de los clusters y comparar y validar en relación a los resultados obtenidos por *Consensus Clustering*.

Tabla 5.6 Tabla que muestra el porcentaje de observaciones de cada subtipo en cada *cluster* obtenido por *Consensus Clustering* de la matriz de *pathways*.

Cluster \ Subtipo de las obs.	Luminal A	Luminal B	Basal	HER2+	Normal
1	70.93	4.26	0	10.53	0
2	8.14	34.75	0	0	0
3	2.91	34.75	100	78.95	100
4	0	0	0	1.75	0
5	18.02	26.24	0	8.77	0

Tabla 5.7 Tabla que muestra la cantidad de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *Consensus Clustering* de la matriz de *pathways*.

Cluster \ Subtipo de las obs.	Luminal A	Luminal B	Basal	HER2+
1	122	6	0	7
2	14	49	0	0
3	5	49	123	45
4	31	37	0	5
TOTAL:	172	141	123	57

5.4.1 Para el *clustering* sobre matriz de expresión

En las tablas 5.9-5.10 se observan los resultados de la composición de los *clusters* obtenidos para la matriz de expresión usando *K-means*.

Tabla 5.8 Tabla que muestra el porcentaje de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *Consensus Clustering* de la matriz de *pathways*.

Subtipo de las obs. / *Cluster*	Luminal A	Luminal B	Basal	HER2+
1	70.93	4.26	0	12.28
2	8.14	34.75	0	0
3	2.91	34.75	100	78.95
4	18.02	26.24	0	0

Tabla 5.9 Tabla que muestra la cantidad de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *K-means* de la matriz de expresión.

Subtipo de las obs. / *Cluster*	Luminal A	Luminal B	Basal	HER2+
1	120	7	0	11
2	32	57	0	9
3	0	1	123	30
4	20	81	0	7
TOTAL:	172	141	123	57

5.4.2 Para el *clustering* sobre matriz de expresión

Repitiendo el proceso anterior con *K-means*, pero esta vez sobre la matriz de *pathways* se obtuvieron los resultados mostrados en las tablas 5.11-5.12.

5.5 Visualización de los clusters formados con *K-means*

Como se mencionó antes, *K-means* proyecta todos los *clusters* con cada una de sus muestras sobre un espacio n-dimensional; para observar la distribución de los grupos después

Tabla 5.10 Tabla que muestra el porcentaje de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *K-means* de la matriz de expresión.

Subtipo de las obs. / *Cluster*	Luminal A	Luminal B	Basal	HER2+
1	69.77	5.00	0	19.3
2	18.60	36.88	0	15.79
3	0	0.71	100	52.63
4	11.63	57.45	0	12.28

Tabla 5.11 Tabla que muestra la cantidad de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *K-means* de la matriz de *pathways*.

Subtipo de las obs. / *Cluster*	Luminal A	Luminal B	Basal	HER2+
1	27	50	37	12
2	0	14	15	6
3	145	33	1	9
4	0	44	70	30
TOTAL:	172	141	123	57

de aplicar esta técnica, se tomaron solo 2 componentes principales (proyección en espacio bidimensional) de las muestras ya clasificadas para poder realizar una gráfica.

5.5.1 Visualización de los *clusters* obtenidos con *K-means* para la matriz de expresión

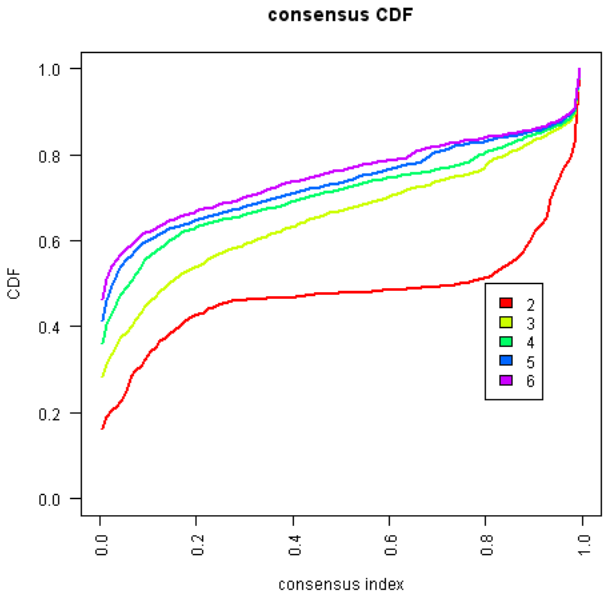
La figura 5.8 muestra la distribución de los *clusters* en 2 dimensiones cuando se aplicó *K-means* a la matriz de expresión.

Tabla 5.12 Tabla que muestra el porcentaje de observaciones de cada subtipo (sin tomar en cuenta sanos) en cada *cluster* obtenido por *K-means* de la matriz de *pathways*.

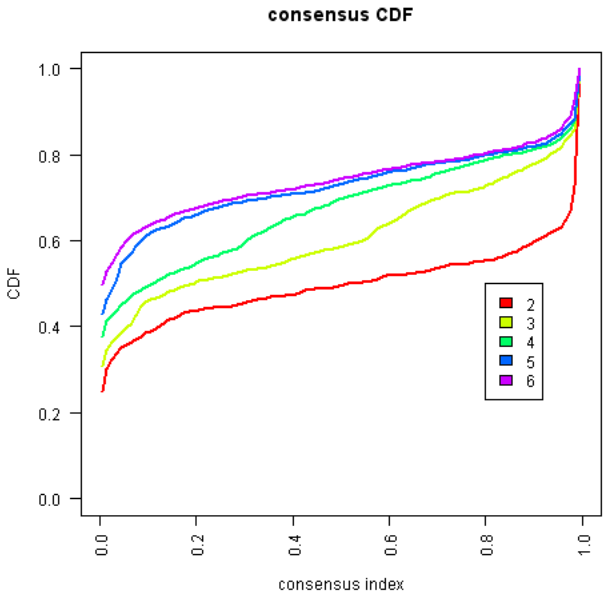
Subtipo de las obs. / *Cluster*	Luminal A	Luminal B	Basal	HER2+
1	15.70	35.46	30.08	21.05
2	0	9.93	12.20	10.53
3	84.3	23.4	0.81	17.79
4	0	31.21	56.91	52.63

5.5.2 Visualización de los *clusters* obtenidos con *K-means* para la matriz de *pathways*

La figura 5.9 muestra la distribución de los *clusters* en 2 dimensiones cuando se aplicó *K-means* a la matriz de *pathways*.

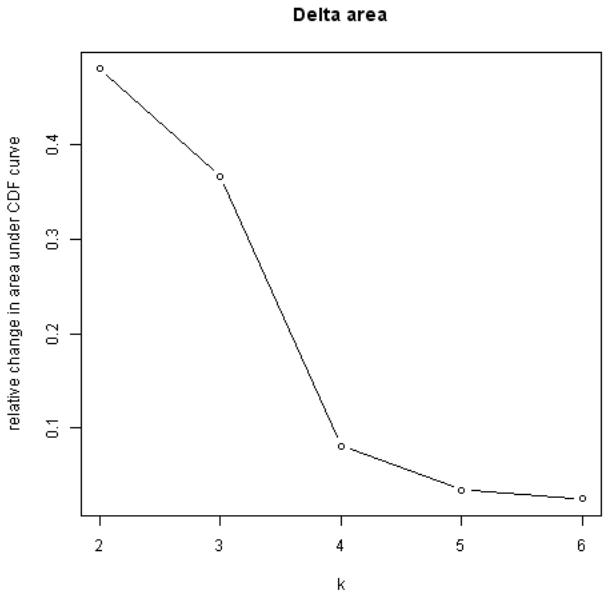


(a) CDF para expresión

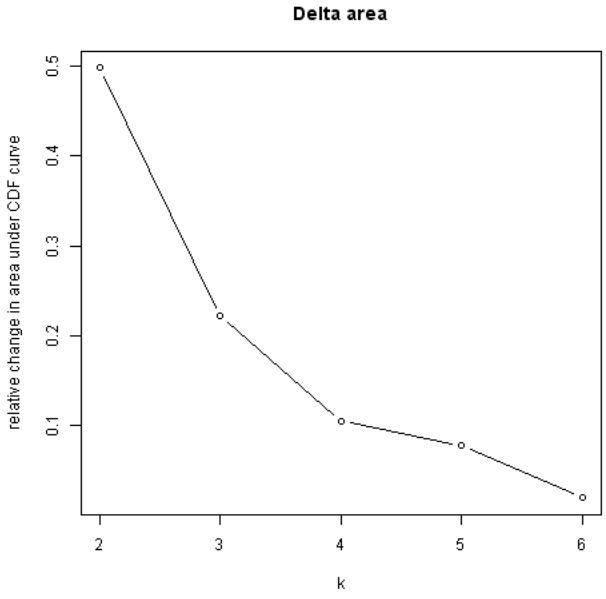


(b) CDF para *Pathways*

Figura 5.4 Gráficas de las CDF para expresión y *pathways* para k desde 2 hasta 6.

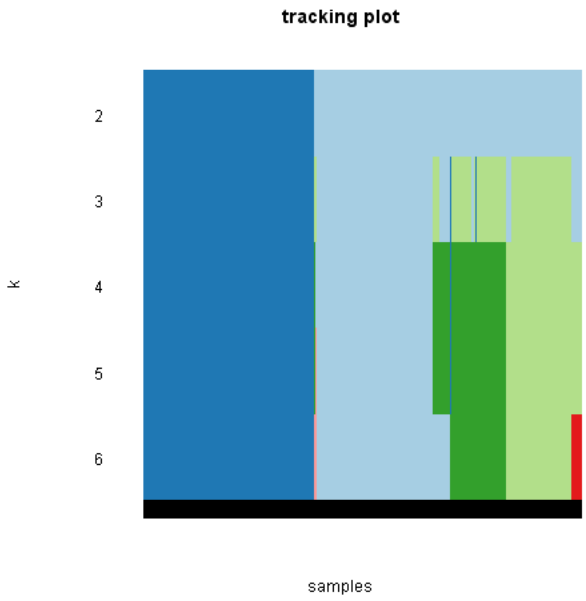


(a) $\Delta(k)$ para expresión

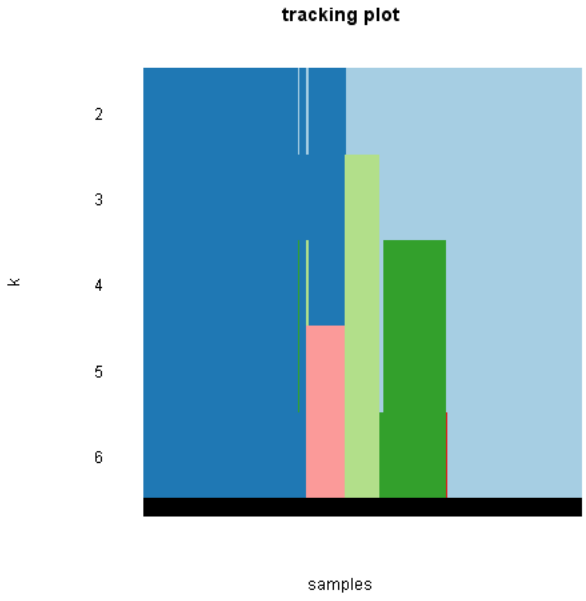


(b) $\Delta(k)$ para *Pathways*

Figura 5.5 Gráficas de las $\Delta(k)$ para expresión y *pathways* para k desde 2 hasta 6.

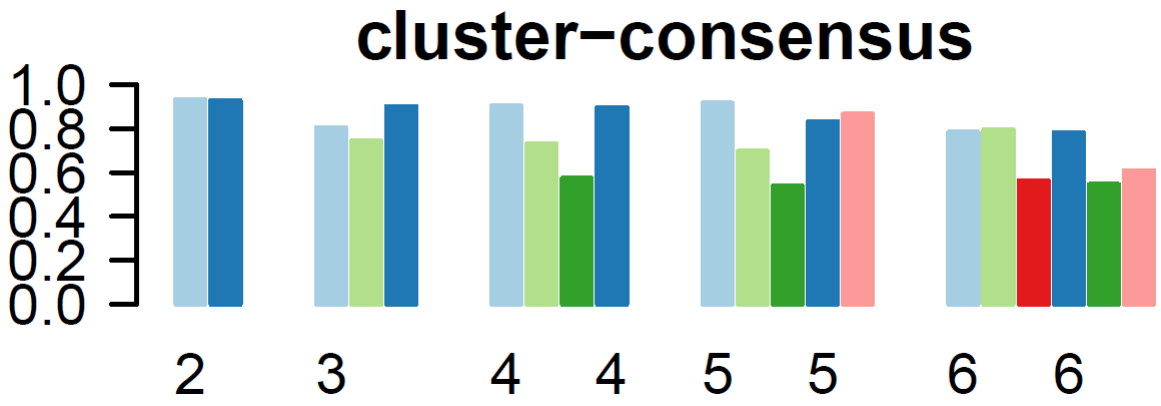


(a) Evolución de la partición para expresión

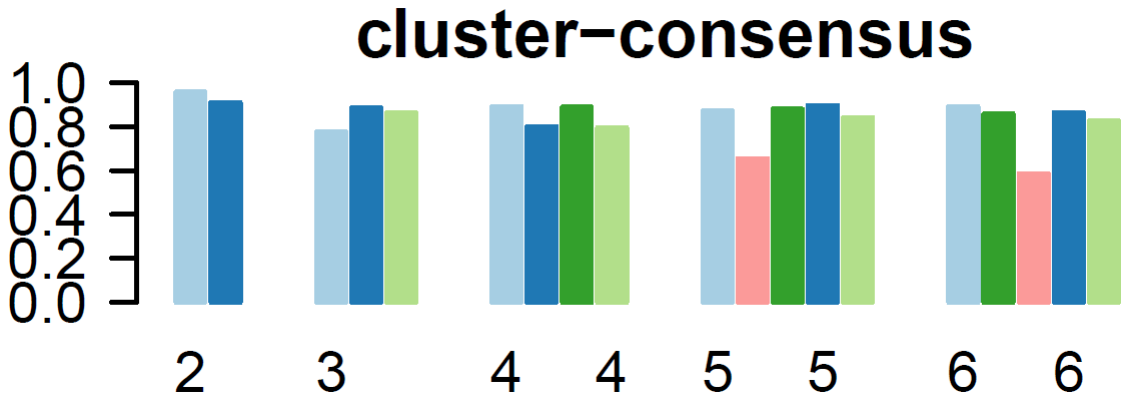


(b) Evolución de la partición para *Pathways*

Figura 5.6 Evolución de la división de los *clusters* a través de distintos valores de k.



(a) Para expresión



(b) Para *Pathways*

Figura 5.7 Consenso de los *clusters* formados con distintos valores de k.

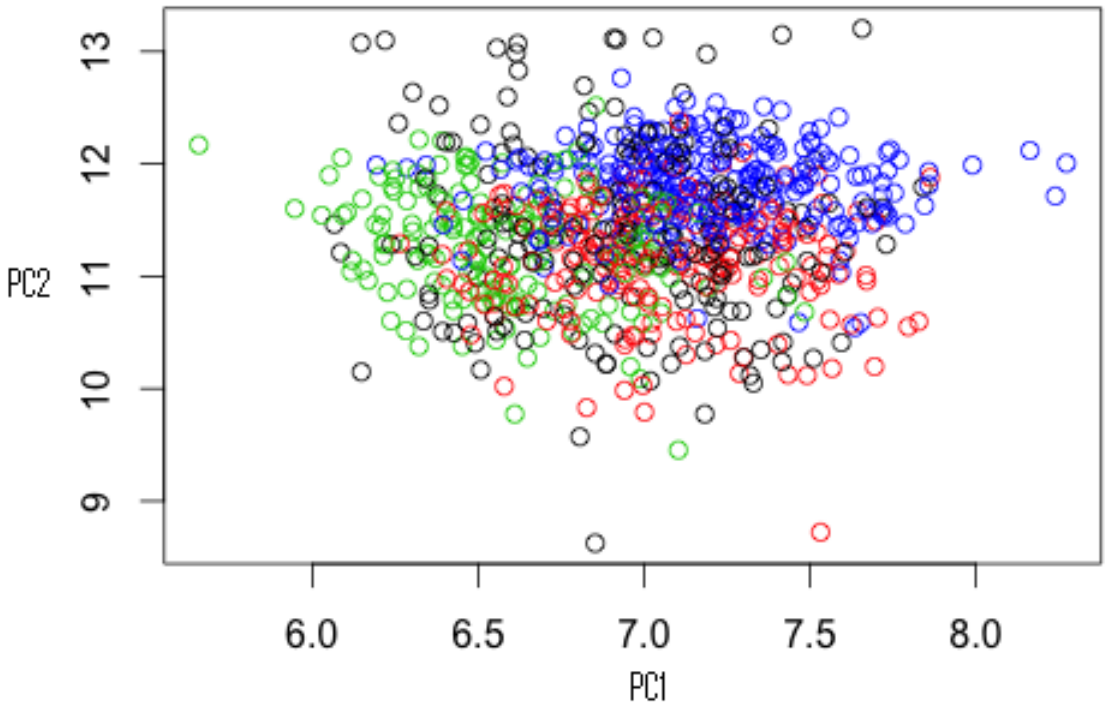


Figura 5.8 *Clusters* formados al aplicar *K-means* sobre la matriz de expresión con $k = 4$ (dos componentes principales).

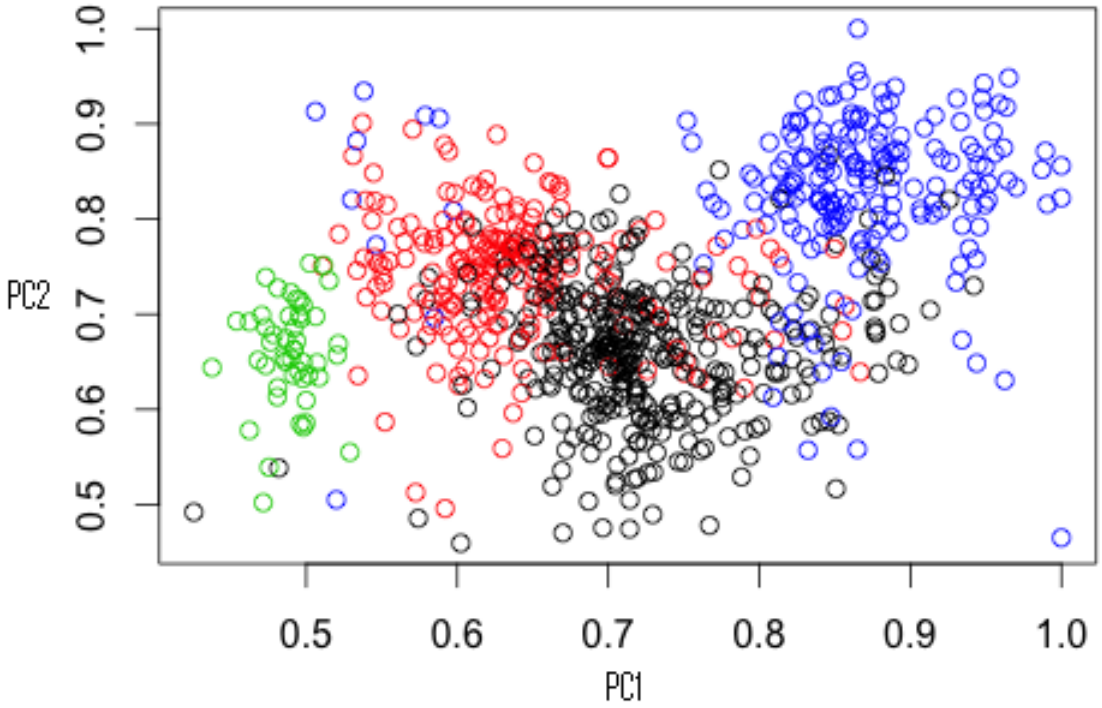


Figura 5.9 *Clusters* formados al aplicar *K-means* sobre la matriz de *pathways* con $k = 4$ (dos componentes principales).

Capítulo 6

Discusión y conclusión

6.1 Discusión de los resultados

La figura 5.1, referente a la medición de la alteración en *pathways* explica claramente porqué es tan fácil para el método computacional implementado el distinguir entre muestras de sanos y enfermos; por lo anterior, se repostan los resultados gráficos derivados de *Consensus Clustering* y *K-means* sólo para las muestras de enfermos ya que, como muestran las tablas, cuando se usaron las muestras de sanos el cluster de éstas estaba perfectamente definido además de que el objetivo central era clasificar las muestras de células cancerígenas.

En cuanto a las matrices de consenso, para aquellas que fueron obtenidas empleando la información de expresión genética (figura 5.2) es evidente la tendencia a $k = 4$, mientras que para las matrices obtenidas para la información de alteración en *pathways* (figura 5.3), a partir del resultado 5.3(b) el *cluster* más grande comienza a subdividirse, mostrando más una tendencia hacia $k = 5$.

En la figura 5.4 se puede ver también con cierta facilidad, basándose en la manera de interpretar estas curvas descrita a priori, un índice de valor óptimo para $k = 4$ en la curva de CDF para las muestras de expresión. En cuanto a los resultados para las muestras de *pathways*, pareciera nuevamente que el valor óptimo es de $k = 5$ sin embargo, como se

mencionó antes la pendiente más pequeña no siempre es indicio de haber encontrado el valor óptimo de k, la forma de la curva es igualmente importante. En este caso es conveniente observar qué la curva para $k = 5$ es prácticamente igual a la correspondiente a $k = 6$. Durante las pruebas se observó que para las curvas de CDF, cuando a partir de cierto valor de k las curvas comienzan a empalmarse, hay un indicio fuerte de haber sobrepasado el valor óptimo de k; como muestra, vale la pena observar la figura 3.4(a). Estas curvas corresponden a la clasificación de muestras bien diferenciadas en donde se sabe que $k = 2$, y en ellas se observa cómo a pesar de que todas las curvas tienen una diferencia muy baja entre $CDF(0.8)$ y $CDF(0.2)$, a partir de la curva para $k = 3$ es muy difícil seguir la trayectoria, pues las gráficas están muy juntas. Con esta evidencia, es más factible elegir un valor de $k = 4$ para las curvas de pathways.

Para la gráfica de $\Delta(k)$ (figura 5.5), observar que para los datos de expresión el cambio significativo comienza a detenerse a partir de $k = 4$; también se observa, y puede causar confusión el hecho de que entre $k = 5$ y $k = 6$ el cambio es menor que entre $k = 4$ y $k = 5$ sin embargo, hay que recordar que para esta gráfica es más importante el punto en donde el cambio disminuye significativamente que el segmento de cambio con la pendiente menor. Siguiendo el criterio anterior, la gráfica para pathways otorga más soporte para $k = 4$.

Las gráficas mostradas en la figura 5.6 son esencialmente otra forma de ver las matrices de consenso para diferentes valores de k en un solo gráfico. Siendo así, en la figura 5.6a puede confirmarse lo que se decía para las matrices de consenso obtenidas con datos de expresión (figura 5.2) respecto a un mejor resultado para $k = 4$. De forma similar, para la gráfica correspondiente a los datos de alteración en *pathways* se observa desde otro manera lo que se observó en la figura 5.3, donde no es fácil decidir por un valor mejor de $k = 4$ o $k = 5$. Aquí, nuevamente se observa que el primer grupo que se dividió del resto de las muestras (azul marino) comienza a subdividirse creando así la confusión. Recordando que el algoritmo seleccionado como base del funcionamiento de *Consensus Clustering* fue *hc*,

hay que mencionar que se observó que debido al funcionamiento intrínseco de esta técnica, cuando se pide al programa dividir a las muestras en valores mayores a una k conocida, éste comienza a buscar dividir el primer grupo que separó al realizar la división en $k = 2$, tal como se observa en este caso.

A pesar de lo que se observa en la figura 5.6b, en la gráfica de consenso (figura 5.7b) es evidente el bajo consenso (confianza) que muestra el nuevo grupo (rosa) formado a partir de la división del grupo azul marino; incluso el consenso de grupo original crece, lo que se explica si consideramos que las muestras que se separan hacia un nuevo grupo son generalmente las que más trabajo costó al algoritmo identificar como parte de el grupo inicial.

Es evidente que la decisión por un valor de k empleando Consensus Clustering no debe basarse en el análisis de las herramientas gráficas por separado. hacer eso es un error que puede llevar al usuario a tomar un valor errado de k. Para hacer un análisis válido es necesario evaluar todos los recursos gráficos y visuales como un conjunto y poner especial atención en la forma en que interactúan entre ellos.

6.2 Conclusión

Con base en los resultados obtenidos es posible afirmar que, al igual que las mediciones de expresión genética, las mediciones de alteración en *pathways* pueden ser empleadas para clasificar cáncer de mama como alguno de los cuatro subtipos de la división más empleada (luminal A, luminal B, basal o HER2+) utilizando aprendizaje automático y además, el uso de mediciones sobre *pathways* representa ciertas ventajas, mismas que se describen a continuación.

Al pasar del uso de información de expresión genética a alteración de *pathways* se reduce la dimensión de las mediciones con respecto al mismo número de muestras, lo que siempre supone una ventaja en el uso de técnicas de clasificación automática, y aún más cuando se

trata de información genética. Al mismo tiempo se siguen manteniendo involucrados los genes originales y una forma de interacción entre ellos. El provecho de esta ventaja se ve reflejado en los resultados obtenidos con *K-means* mostrados en la figuras 5.8 y 5.9; es evidente que esta técnica logra obtener una distancia de separación entre *clusters* mayor cuando se basa en las mediciones de alteración en *pathways* en comparación con la distancia que se obtiene al usar las mediciones de expresión, lo que representa un menor costo computacional para clasificar las muestras.

A partir de esta discriminación a partir de mediciones de alteración en *pathways* puede darse otro enfoque al análisis de las estrategias de combate al cáncer de mama y al diseño de tratamientos y terapias más accesibles y efectivas.

6.3 Perspectivas

Con el cumplimiento de los objetivos delimitados y conclusión del presente trabajo, surge la motivación para proyectos futuros, siendo los principales:

- Una vez probado que las muestras de cáncer pueden dividirse en los cuatro subtipos al usar mediciones de la alteración en *pathways*, queda pendiente encontrar un método que maximice el porcentaje de exactitud en la clasificación de estas muestras.

- Se ha mencionado que hay *pathways* identificados cuya alteración es representativa del cáncer en general. Partiendo de las bases de datos que se usaron en el presente trabajo y técnicas computacionales, parece asequible encontrar un grupo de *pathways* asociados directamente al cáncer de mama o incluso, a cada uno de los subtipos moleculares.

- Encontrar más herramientas gráficas y estadísticas que refuercen la confianza de los resultados obtenidos.

- Encontrar técnicas cuantitativas para analizar las gráficas obtenidas más allá de la inspección visual.

Apéndice
Algoritmos implementados en R

A.1 Curado de bases de datos

```
## Programa para curar las bases de datos de sanos, enfermos y genes-pathways.
## Se dejan sólo los genes comunes a las tres bases de datos.



setwd("C:/Documents and Settings/CINVESTAV/Escritorio/bases/")

#Lectura de las bases de datos almacenadas:
her<-as.matrix(read.table("EXPMATRIX_HER2_COLAPSED.txt",sep="", header=TRUE))
luma<-as.matrix(read.table("EXPMATRIX_LUMA_COLAPSED.txt",sep="", header=TRUE))
lumb<-as.matrix(read.table("EXPMATRIX_LUMB_COLAPSED.txt",sep="", header=TRUE))
basal<-as.matrix(read.table("EXPMATRIX_BASAL_COLAPSED.txt",sep="", header=TRUE))
sanos<-as.matrix(read.table("EXPMATRIX_SANOS_COLAPSED.txt",sep="", header=TRUE))
mpath<-as.matrix(read.table("mpath.txt",sep=","))

#Extracción de los genes que aparecen en las bases de enfermos:
genes_enfermos<-luma[,1]
luma<-luma[,-1]
```

```
lumb<-lumb[,-1]
her<-her[,-1]
basal<-basal[,-1]

#Extracción de los genes que aparecen en las bases de sanos:
genes_sanos<-sanos[,1]
sanos<-sanos[,-1]

#Agrupación de todos los subtipos en una sola base:
exp2<-cbind(luma, lumb, basal, her)
rownames(exp2)<-genes_enfermos #nombres de los renglones (genes) de enfermos

rownames(sanos)<-genes_sanos #nombres de los renglones (genes) de sanos

#Reacomodo de las bases como matrices:
A<-as.matrix(exp2)
B<-as.matrix(sanos)
C<-as.matrix(t(mpath))

#Extracción de los genes comunes a las tres bases de datos
c1<-intersect(rownames(A),rownames(B))
c2<-intersect(c1,rownames(C))

## Obtención y escritura de la base de datos de enfermos...
##...solo con genes comunes.
expresion<-A[rownames(A)==c2[1]]
```

```
for(i in 2:length(c2)){
  expresion<-rbind(expresion,A[rownames(A)==c2[i]])
}
rownames(expresion)<-c2
colnames(expresion)<-colnames(A)
write.table(expresion, file="expresion.txt", row.names=TRUE, col.names=TRUE)

## Obtención y escritura de la base de datos de sanos...
##...solo con genes comunes.
sanos<-B[rownames(B)==c2[1]]
for(i in 2:length(c2)){
  sanos<-rbind(sanos,B[rownames(B)==c2[i]])
}
rownames(sanos)<-c2
colnames(sanos)<-colnames(B)
write.table(sanos, file="sanos.txt", row.names=TRUE, col.names=TRUE)

## Obtención y escritura de la base de datos genes-pathways...
##...solo con genes comunes.
mpath<-C[rownames(C)==c2[1]]
for(i in 2:length(c2)){
  mpath<-rbind(mpath,C[rownames(C)==c2[i]])
}
rownames(mpath2)<-c2
colnames(mpath2)<-colnames(C)
write.table(mpath, file="mpath2.txt", row.names=TRUE, col.names=TRUE)
```

A.2 *Pathifier*

```
## Programa que emplea Pathifier para cuantificar la alteración...
##...en los pathways de los enfermos con respecto a los sanos.

setwd("C:/Documents and Settings/CINVESTAV/Escritorio/bases")

#Lectura de las bases de enfermos y sanos
data<-as.matrix(read.table("expresion.txt",sep=""))
sanos<-as.matrix(read.table("sanos.txt",sep=""))
data<-cbind(data,sanos) #sanos y enfermos en una sola base
#Etiquetas que indican cuáles son sanos (1) y cuáles enfermos (0)
normals<-vector("logical",length=554)
normals[494:554]<-TRUE #últimos 61 son sanos (1)

allgenes<-rownames(data) #nombres de los genes

#Lectura de base de genes-pathways:
pathways<-as.matrix(read.table("mpath2.txt", sep="",header=TRUE))

g_names<-rownames(pathways) #nombres de los pathways

#Lista de pathways y sus genes correspondientes
x<-cbind(g_names,pathways)
s<-x[,1:2]
v2<-as.matrix(s[s[,2]==1])
```

```
a<-length(v2)/2
v2<-v2[1:a]
v<-list(matrix(v2))
for(i in 2:(dim(x)[2])-1){
  s<-x[,c(1,i+1)]
  v2<-as.matrix(s[s[,2]==1])
  a<-length(v2)/2
  v2<-v2[1:a]
  v[i]<-list(matrix(v2))
}
gs<-v
pathwaynames<-colnames(pathways)
pathwaynames<-as.list(pathwaynames)

#Lista de la matriz de expresión y su información:
expresion<-list(data=data,allgenes=allgenes,normals=normals)
#Lista de la matriz de pathways y su información:
pathways<-list(gs=gs,pathwaynames=pathwaynames)

#Inclusión de la biblioteca "Pathifier":
source("http://bioconductor.org/biocLite.R")
biocLite("pathifier")
library(pathifier)

#Función principal de Pathifier:
PDS<-quantify_pathways_deregulation(expresion$data,
```

```
                                    expresion$allgenes, pathways$gs,
                                    pathways$pathwaynames,
                                    expresion$normals, attempts=50)

#Creación de la matriz que contiene la cuantificación...
# ...de la alteración en los pathways de enfermos
disreg<-PDS$scores[[1]]
for(i in 2:length(PDS$scores)){
  j<-PDS$scores[[i]]
  disreg<-rbind(disreg,j)
}

#Escritura de la matriz como base de datos:
write.table(disreg, file="deregulations.txt", row.names=TRUE, col.names=TRUE)
```

A.3 *Consensus Clustering*

```
##Programa que usa Consensus Clustering para realizar...
##...la clasificación de las muestras de enfermos.


setwd("C:/Documents and Settings/CINVESTAV/Escritorio/bases")

#disreg<-as.matrix(read.table("disregulations.txt",sep="",header=TRUE))
#disreg<-disreg[,-820:-880]
```

```
#disreg<-as.matrix(read.table("expresion.txt",sep="",header=TRUE))
#disreg2<-as.matrix(read.table("sanos.txt",sep="",header=TRUE))
#disreg<-cbind(disreg,disreg2) ##61 sanos

#PARA EXTRAER LOS YA CLASIFICADOS (SEGUROS)
#disreg<-as.matrix(read.table("disregulations.txt",sep="",header=TRUE))
#disreg<-disreg[,-820:-880]
#disreg2<-as.matrix(read.table("sanos.txt",sep="",header=TRUE))
#disreg<-cbind(disreg,disreg2) ##61 sanos
# disreg<-t(disreg)
# a<-disreg
# b<-as.matrix(read.table("tags.txt",header=TRUE,sep=""))
# comp<-matrix(nrow=length(b),ncol=dim(a)[2])
# for(i in 1:dim(a)[1]){
#   for(j in 1:length(b)){
#     if(rownames(b)[j]==rownames(a)[i]){
#       comp[j,]=a[i,]
#     }
#   }
# }
# disreg<-as.matrix(comp)
# rownames(disreg)<-rownames(b)
# disreg<-t(disreg)
#disreg<-as.matrix(disreg[,-494 -554])

#Lectura de la base de datos de expresión o pathways:
```

```
disreg<-as.matrix(read.table("expresion.txt",sep="",header=TRUE))

#disreg<-disreg[,-494:-554] #quitar muestras de sanos

#Preprocesamiento estadístico de la base de expresión o...
#...pathways con distacia de Pearson:
disreg<-sweep(disreg,1,apply(disreg,1,median))
disreg2<-as.dist(1-cor(disreg,method="pearson"))

#Inclusión de la biblioteca "ConsensusClusterPlus"
library(ConsensusClusterPlus)
#Función que realiza el algoritmo de Consensus Clustering:
results = ConsensusClusterPlus(disreg2,maxK=6,reps=5000,
                               pItem=0.8,pFeature=1,
                               clusterAlg="hc",title=getwd(),
                               distance="pearson",plot="png")
#ick almacena los resultados del clustering:
ick=calcICL(results,title=getwd())

k=4 #k es el número de clusters
a<-results[[k]][["consensusClass"]] #resultados para k

#etiquetas de las muestras:
t1<-as.matrix(rep("LumA",172))
t2<-as.matrix(rep("LumB",141))
t3<-as.matrix(rep("Basal",123))
```

```
t4<-as.matrix(rep("HER",57))
t<-rbind(t1,t2,t3,t4)
rownames(t)<-colnames(disreg)


#Identificar el subtipo de las muestras...
#... en cada cluster:
comp<-matrix(nrow=length(t),ncol=2)
for(i in 1:length(a)){
  for(j in 1:length(t)){
    if(rownames(t)[j]==names(a)[i]){
      comp[j,]=cbind(t[j],a[i])
    }
  }
}
comp<-as.matrix(comp)
rownames(comp)<-rownames(t)

#Creación de las tablas para mostrar resultados:
comp<-as.matrix(comp)
empt<-matrix(data=NA, nrow=k+1,ncol=4)
colnames(empt)<-c("LumA","LumB","Basal","HER")
rownames(empt)<-c(1:k,"Total:")

for(i in 1:k){ # K
  comp1<-comp[comp[,2]==i,]
```

```
  for(j in 1:4){ # col
    empt[i,j]<-sum(comp1[,1]==colnames(empt)[j])
  }
}
for(x in 1:4){
  empt[dim(empt)[1],x]<-sum(empt[-dim(empt)[1],x])
}
empt

#Tablas de resultados en porcentajes:
perc<-matrix(data=NA, nrow=k,ncol=4) ## k,5
colnames(perc)<-colnames(empt)
for(i in 1:k){ # ren
  for(j in 1:4){ # col=5
    perc[i,j]<-empt[i,j]/empt[dim(empt)[1],j]*100
  }
}
perc
```

A.4 *K-means*

```
## Programa que realiza la clasificación con K-means
##para la validación de resultados

setwd("C:/Documents and Settings/CINVESTAV/Escritorio/bases")
```

```
#Lectura de matrices de expresión o pathways:
disreg<-as.matrix(read.table("deregulations.txt",sep="",header=TRUE))
##disreg2<-as.matrix(read.table("sanos.txt",sep="",header=TRUE))
##disreg<-cbind(disreg,disreg2) ##61 sanos
disreg<-t(disreg)
##Para path-desreg:
#disreg<-as.matrix(read.table("disregulations.txt",sep="",header=TRUE))
disreg<-disreg[-494:-554,] #quitar sanos
#disreg<-t(disreg)

require(graphics) #necesario para gráfica
k=4 #Elegir número de clases
#Función que realiza K-means:
(cl <- kmeans(disreg, k, nstart = 4, iter.max = 5000))
plot(disreg, col = cl$cluster) #gráfica
#points(cl$centers, col = 1:4, pch = 10)

a<-cl$cluster #composición de los clusters

#Etiquetas de las muestras:
t1<-as.matrix(rep("LumA",172))
t2<-as.matrix(rep("LumB",141))
t3<-as.matrix(rep("Basal",123))
t4<-as.matrix(rep("HER",57))
b<-rbind(t1,t2,t3,t4)
rownames(b)<-rownames(disreg)
```

```
#Identificando a qué subtipo corresponden...
#...las muestras en cada cluster:
comp<-matrix(nrow=length(b),ncol=2)
for(i in 1:length(a)){
  for(j in 1:length(b)){
    if(rownames(b)[j]==names(a)[i]){

      comp[j,]=cbind(b[j],a[i])
    }
  }
}
#Creación de tablas de resultados:
comp<-as.matrix(comp)
rownames(comp)<-rownames(b)
comp<-comp[-494:-554,]
empt<-matrix(data=NA, nrow=k+1,ncol=4)
colnames(empt)<-c("lumA","lumB","basal","HER2+")
rownames(empt)<-c(1:k,"Total:")
for(i in 1:k){ # K
  comp1<-comp[comp[,2]==i,]
  for(j in 1:4){ # col
    empt[i,j]<-sum(comp1[,1]==colnames(empt)[j])
  }
}
for(x in 1:4){
  empt[dim(empt)[1],x]<-sum(empt[-dim(empt)[1],x])
```

```
}
empt
perc<-matrix(data=NA, nrow=k,ncol=4) ## k,5
colnames(perc)<-colnames(empt)

#Creación de tablas de resultados en porcentajes:
for(i in 1:k){ # ren
  for(j in 1:4){ # col=5
    perc[i,j]<-empt[i,j]/empt[dim(empt)[1],j]*100
  }
}
perc
```

Referencias

[1] World Health Organization, "Breast cancer: prevention and control," [en línea, actualización de 2014] http://www.who.int/cancer/detection/breastcancer/en/.

[2] Instituto Mexicano del Seguro Social, "Cáncer de mama," Salud en línea [en línea, actualización de noviembre de 2014] http://www.imss.gob.mx/salud-en-linea/cancer-mama.

[3] M Guedj, L Marisa, A de Reynies, et al., "A refined molecular taxonomy of breast cancer," Oncogenomics, vol. 31, pp. 1196-1206, 2012.

[4] Cartes d'Identité des Tumeurs research program, "citbcmst: a package to assign CIT breast cancer molecular subtypes from expression data," Ligue Nationale Contre le Cancer, [en línea, actualización de enero de enero de 2014] http://cit.ligue-cancer.net.

[5] Charles Perou, Sorlie T, Eisen MB, et al., "Molecular portraits of breast tumours," Nature, vol. 406, pp. 747-752, agosto de 2011.

[6] Aleix Prat & Charles Perou, "Deconstructing the molecular portraits of breast cancer," Molecular Oncology, vol. 5, pp. 5-23, noviembre de 2010.

[7] Fan Zhang, Tao Liu, Mu Wang, et al., "Dual-function biomarkers for detection of breast cancer and its cancer type: invasive versus non-invasive," IEEE International Conference on Bioinformatics and Biomedicine, pp. 345-348, 2013.

[8] Jorma de Ronde, Marc Jan, Esther Lips, et al., "Breast cancer subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform classifiers trained on all subtypes," Plos One, vol. 9, pp. 1-8, febrero de 2014.

[9] Markus List, Anne-Christin Hauschild, Qihua Tan, et al., "Classification of breast cancer subtypes by combining gene expression and DNA methylation data," Journal of Integrative Bioinformatics, vol. 11, 2014.

[10] Forough Firoozbakht, Iman Rezaeian, Lisa Porter, et al., "Breast cancer subtype identification using machine learning techniques," IEEE 4th International Conference on Computational Advances in Bio and Medical Sciences, junio de 2014.

[11] Pilar Eroles, Ana Bosch, J. Pérez-Fidalgo, et al., "Molecular biology in breast cancer: intrinsic subtypes and signaling pathways," Cancer Treatment Reviews, vol. 38, pp. 698-707, 2011.

[12] Ciro Donalek, "Supervised and unsupervised learning," Ay/Bi 199, abril de 2011, [en línea] http://www.astro.caltech.edu/~george/aybi199/Donalek_classif1.pdf.

[13] Kaare Brandt & Michael Syskind, "The matrix cookbook," [en línea, actualización de noviembre de 2012] http://matrixcookbook.com.

[14] K. Baca-López, E. Hernández Lemus & M. Mayorga, "'Information-theorical analysis of gene expression data to infer transcriptional interactions," 2. Gene expresion data analysis, Revista Mexicana de Física, vol. 55(6), pp. 457-461, diciembre de 2009.

[15] Douglas Hanahan & Robert A. Weinberg, "Hallmarks of cancer: the next generation," Cell, vol. 144, Elsevier Inc., pp. 646-668, marzo de 2011.

[16] Yotam Drier, Michal Sheffer & Eytan Domany "Pathway-based personalized analysis of cancer," PNAS, vol. 110, no. 16, pp. 6388-6393, abril de 2013.

[17] Yotam Drier, "Quantify deregulation of pathways in cancer," [en línea, actualización de abril de 2015] http://bioconductor.jp/packages/3.2/bioc/vignettes/pathifier/inst/doc/Overview.pdf.

[18] Bioconductor, "Package 'pathifier'," [en línea, actualización de febrero de 2015] http://www.bioconductor.org/packages/release/bioc/manuals/pathifier/man/pathifier.pdf.

[19] Stefano Monti, Pablo Tamayo, Jill Mesirov, et al., "Consensus Clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," Machine Learning, vol. 52, pp. 91-118, 2003.

[20] Matthew D. Wilkerson, "ConsensusClusterPlus (tutorial)," [en línea, actualización de abril de 2015] http://www.bioconductor.org/packages/release/bioc/vignettes/ConsensusClusterPlus/inst/doc/ConsensusClusterPlus.pdf.

[21] Ray Li, “Top 10 data mining algorithms in plain english,” [en línea, actualización de mayo de 2015] http://rayli.net/blog/data/top-10-data-mining-algorithms-in-plain-english/.

[22] Osama Abu Abbas, “Comparison between data clustering algorithms,” Computer Science Department, Yarmouk University, mayo de 2007.

==========