

UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

CLASIFICACIÓN DE QUIMIOSENSIBILIDAD TUMORAL EN MUESTRAS DEL
PROYECTO TCGA MEDIANTE LA INTEGRACIÓN DE DATOS GENÓMICOS

Tesis sometida a la consideración de la Comisión del Programa de Posgrado
en Ingeniería Eléctrica para optar al grado y título de Maestría Académica
en Ingeniería Eléctrica

JOSUÉ DAVID VARGAS AMADOR

Ciudad Universitaria Rodrigo Facio, Costa Rica

2019

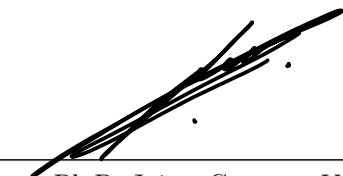
Dedicatoria

A mi familia y su memoria,
que sin saberlo son la guía
inmaterial de mis acciones.

Agradecimientos

Al profesor Francisco Siles por creer en mí y darme la oportunidad de pertenecer al *PRIS-Lab* y al profesor Rodrigo Mora por darme una idea de investigación que estoy seguro guiará el resto de mi vida académica.

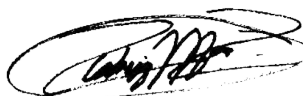
Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Ingeniería Eléctrica de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Ingeniería Eléctrica.



Ph.D. Jaime Cascante Vindas
Representante del Decano
Sistema de Estudios de Posgrado



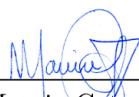
Dr. rer. nat. Francisco Siles Canales
Director de Tesis




Ph.D. Rodrigo Mora Rodríguez
Asesor



Dr. Juan Luis Crespo Mariño
Asesor



Ph.D. Marvin Coto Jiménez
Director
Programa de Posgrado en Ingeniería Eléctrica



Josué David Vargas Amador
Candidato

Resumen

La enfermedad del cáncer causa millones de muertes en todo el mundo y miles en Costa Rica, y el tratamiento del cáncer causa también una carga económica inmensa en el sistema de seguridad social porque las drogas contra la enfermedad son extremadamente caras. Han surgido en los últimos años múltiples estudios computacionales y estadísticos sobre el análisis de datos genómicos del cáncer, en especial los datos de número de copia de ADN y la expresión génica. Los datos de número de copia pueden ser empleados para describir un fenómeno llamado aneuploidía en el cual regiones enteras de los cromosomas desaparecen o se multiplican. La aneuploidía tiene implicaciones clínicas pues está relacionada a distintas consecuencias clínicas e incluso algunos investigadores piensan es el motor del crecimiento y la resistencia del cáncer. Dos conjuntos de datos genómicos del cáncer bien conocidos son el CCLE y el TCGA, donde el primero trata sobre estudios de líneas celulares y el segundo trata sobre estudios de muestras de cáncer. En este trabajo se plantea usar un método llamado GISTIC que permite encontrar las regiones con alteraciones cromosomales en un conjunto de muestras, donde se busca entrenar un clasificador de máquina de soporte vectorial y además generar reglas de asociación de datos para las regiones alteradas. El clasificador cumple con la expectativa de métrica esperada a pesar de la existencia de clases desbalanceadas. Un aporte realmente valioso del trabajo son las reglas de asociación de datos que permitirían determinar combinaciones de aberraciones vinculadas a la mortalidad y la resistencia a los tratamientos.

Abstract

The cancer disease causes millions of deaths worldwide and thousands in Costa Rica, and the treatment of cancer also causes a huge economic burden on the social security system because drugs against the disease are extremely expensive. In recent years there has been a great increase in computational and statistical studies on cancer genomic data, especially on DNA copy number and gene expression. The copy number data can be used to describe a phenomenon called aneuploidy in which entire regions of the chromosomes disappear or multiply. Aneuploidy has clinical implications because it's related to different clinical consequences and even some researchers think it is the engine of cancer growth and resistance. Two well-known genomic datasets of cancer are the CCLE and the TCGA, where the first deals with studies of cell lines and the second deals with studies of cancer samples. In this work we propose to use a method called GISTIC that allows us to find regions with chromosomal alterations in a set of samples, where we are looking to train a vectorial support machine classifier and also generate data association rules for regions with chromosomal alterations. The results of the classifier meet the expected metric despite the existence of unbalanced classes. A really valuable contribution of the work are the association rules that allow to determine aberrations linked to mortality and treatments' resistance.

Índice de Contenidos

| | |
|--|-----------|
| Resumen | v |
| Abstract | vi |
| Índice de Figuras | x |
| Índice de Cuadros | xi |
| 1 Introducción | 1 |
| 1.1. Justificación | 2 |
| 1.2. Propuesta | 4 |
| 1.3. Problema | 4 |
| 1.4. Hipótesis | 4 |
| 1.5. Objetivos | 4 |
| 2 Antecedentes | 6 |
| 2.1. Integración de datos genómicos | 7 |
| 2.2. Algoritmos de un solo tipo de datos | 10 |
| 3 Marco Teórico | 15 |
| 3.1. Cáncer y biología | 15 |
| 3.2. Reconocimiento de patrones | 25 |
| 4 Metodología | 29 |
| 4.1. Selección por medio de GISTIC | 30 |
| 4.2. Pruebas de clasificación | 32 |
| 4.3. Pruebas de reglas de asociación | 34 |

| | |
|--|-----------|
| 5 Resultados y Análisis | 36 |
| 5.1. Selección de GISTIC | 37 |
| 5.2. Tablas de la clasificación | 43 |
| 5.3. Reglas de asociación de datos | 45 |
| 6 Conclusiones y Recomendaciones | 58 |
| 6.1. Conclusiones | 58 |
| 6.2. Recomendaciones | 60 |
| 7 Bibliografía | 63 |
| Apéndice A Referencia de artículo | 68 |

Índice de Figuras

| | |
|--|----|
| 3.1. Flujo de trabajo reconocimiento de patrones | 26 |
| 3.2. Representación de un clasificador lineal de máquina de soporte vectorial | 28 |
| 4.1. Diagrama de flujo de trabajo de la metodología | 30 |
| 5.1. Regiones cromosómicas identificadas con amplificaciones de número de copias en el conjunto de datos CCLE. | 38 |
| 5.2. Regiones cromosómicas identificadas con eliminaciones de número de copias en el conjunto de datos del CCLE. | 38 |
| 5.3. Regiones cromosómicas identificadas con amplificaciones de número de copias en el subconjunto de cáncer de mama del TCGA. | 39 |
| 5.4. Regiones cromosómicas identificadas con eliminaciones de número de copia en el subconjunto de cáncer de mama del TCGA. | 40 |
| 5.5. Regiones cromosómicas identificadas con amplificaciones de número de copias en el conjunto de cáncer de mama del CCLE. | 41 |
| 5.6. Regiones cromosómicas identificadas con eliminaciones de número de copia en el conjunto de cáncer de mama del CCLE. | 41 |

Índice de Cuadros

| | |
|---|----|
| 4.1. Tabla de drogas con el número de líneas celulares etiquetadas como sensibles y resistentes. | 33 |
| 5.1. Tabla con el número de alteraciones en los cromosomas de las líneas celulares del CCLE de cáncer de seno | 42 |
| 5.2. Tabla con el número de alteraciones cromosómicas de las muestras del TCGA de cáncer de seno | 43 |
| 5.3. Tabla con los resultados de la clasificación de los datos usando validación cruzada | 44 |
| 5.4. Tabla con los resultados de la validación cruzada con los datos disponibles de sensibilidad a paclitaxel, topotecan y lapatinib | 45 |
| 5.5. Tabla con los resultados de emplear el clasificador entrenado con los datos del TCGA en los datos del CCLE | 45 |
| 5.6. Tabla con la métrica de soporte de las alteraciones cromosómicas de las muestras del TCGA de cáncer de seno | 47 |
| 5.7. Tabla con las reglas de asociación de las alteraciones cromosómicas de las muestras de TCGA de cáncer de seno | 47 |
| 5.8. Tabla con la métrica de soporte de las alteraciones cromosómicas de las muestras del TCGA de los pacientes muertos de cáncer de seno | 48 |
| 5.9. Tabla con las reglas de asociación de las alteraciones cromosómicas de las muestras de TCGA de los pacientes muertos de cáncer de seno | 49 |
| 5.10. Tabla con la métrica de soporte de las alteraciones cromosómicas de las muestras del TCGA de los pacientes vivos de cáncer de seno | 49 |
| 5.11. Tabla con las reglas de asociación de las alteraciones cromosómicas de las muestras de TCGA de los pacientes vivos de cáncer de seno | 50 |

| | |
|--|----|
| 5.12. Tabla con la métrica de soporte de las alteraciones en los cromosomas de las líneas celulares del CCLE de cáncer de seno | 50 |
| 5.13. Tabla de reglas de asociación de las alteraciones cromosómicas de las líneas celulares del CCLE de cáncer de seno | 51 |
| 5.14. Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a paclitaxel | 51 |
| 5.15. Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a paclitaxel | 52 |
| 5.16. Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a paclitaxel | 52 |
| 5.17. Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a paclitaxel | 53 |
| 5.18. Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a topotecan | 53 |
| 5.19. Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a topotecan | 54 |
| 5.20. Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a topotecan | 54 |
| 5.21. Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a topotecan | 55 |
| 5.22. Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a lapatinib | 55 |
| 5.23. Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a lapatinib | 56 |
| 5.24. Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a lapatinib | 56 |
| 5.25. Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a lapatinib | 57 |

Capítulo 1

Introducción

El cáncer alrededor del mundo es responsable de 8 millones de muertes con una incidencia de 14 millones. En el 2008, 12.7 millones de personas fueron diagnosticadas con la enfermedad, y en el mismo año, 24.6 millones de personas vivían con cáncer. Actualmente en Costa Rica, hay 8000 casos de cáncer diagnosticados cada año, y además 3500 personas mueren anualmente por la enfermedad. En el 2007, el cáncer fue responsable del 20 % de todas las muertes en Costa Rica, pero se espera que para el 2025 sea la causa del 50 %. La alta incidencia en el país supone una gran presión para todas las organizaciones relacionadas con la salud pública [Coto, Siles y Mora 2016].

El cáncer no es una sola enfermedad, sino más bien una familia de enfermedades que dependen de las peculiaridades de cada persona: género, edad, herencia, estilo de vida, el tejido afectado, entre muchas otras razones. Estas condiciones patológicas pueden ser tratadas terapéuticamente mediante quimioterapias. Un tratamiento poco efectivo induce más mutaciones haciendo las células malignas más resistentes a tratamientos posteriores.

Hoy día, existen ensayos clínicos de quimiosensibilidad que generan datos que relacionan el desarrollo celular del cáncer ante una concentración de una droga específica, donde a partir de estas pruebas en múltiples drogas se puede definir si el paciente es resistente o sensible a determinados tratamientos, y por ende convertirse en una ayuda en la decisión médica. Se han realizado múltiples ensayos clínicos, a nivel internacional, en el que se han recolectado información génica de líneas celulares [Barretina 2012, Costello et al. 2014], implantes en ratones [Bruna et al. 2016] y tumores de pacientes [Lee 2015], aunque solo algunos de estos proyectos han recolectado información de quimiosensibilidad. Algunos

de esos proyectos se llaman de **Cancer Cell Line Encyclopedia** (CCLE) [Barretina 2012] y **The Cancer Genome Atlas** (TCGA) [Lee 2015].

La idea central de almacenar los datos genómicos y clínicos de las muestras, es tener la capacidad de realizar experimentos in-silico (a partir de los datos genómicos almacenados en computadoras) que permitan encontrar relaciones entre los datos y así reconocer grupos de muestras con mejores respuestas a los tratamientos, propiciando un posible aumento de la sobrevivencia del paciente, sin la necesidad de realizar los ensayos clínicos de quimiosensibilidad.

Con el fin de enfrentar el problema mencionado anteriormente, en el PRIS-Lab se ha llevado a cabo investigaciones al respecto, tales como el proyecto Plataforma biocomputacional de análisis de datos genómicos para superar la resistencia a la terapia contra el cáncer [Coto, Siles y Mora 2016] y particularmente en J. Coto y Siles 2017, en donde se utilizaron modelos regresiones y agrupamientos para representar datos genómicos y de quimiosensibilidad.

1.1. Justificación

La terapia personalizada se refiere a todo el conjunto de tratamientos que recibe un paciente con base en el análisis de la información molecular y clínica para determinar cuál es la terapia más apropiada para las variantes de la enfermedad. A nivel mundial existen múltiples esfuerzos académicos que han tratado de reunir información de diversas formas (tumores de pacientes, cultivos celulares y pruebas en ratones), con el objetivo de realizar experimentos computacionales basados en diversas técnicas de reconocimiento de patrones para clasificar las muestras, encontrar nuevos subtipos de cáncer, identificar nuevos blancos terapéuticos para atacar con drogas, o determinar el mejor tratamiento para esas muestras de tumor basándose en las características genómicas.

En la Universidad de Costa Rica se ha venido trabajando la idea de una plataforma biocomputacional que permita construir modelos del comportamiento del cáncer con el objetivo de avanzar hacia la implementación de la terapia personalizada a largo plazo. Con dicha visión fue que se realizó un proyecto de investigación inscrito en el Centro de Investigaciones en Tecnologías de Información y la Comunicación con el nombre de Plataforma biocomputacional de análisis de datos genómicos para

superar la resistencia a la terapia contra el cáncer [Siles y Coto (2017)]. La motivación de tal proyecto fue inspirado en el potencial uso plataformas de secuenciación de nueva generación, donde en un futuro próximo sea posible contar con datos genómicos para cada paciente, lo cual podría permitir el diseño de terapias específicas para cada paciente acorde al perfil genómico de cada caso de cáncer.

En el 2017 se finalizó en el Laboratorio de Reconocimiento de Patrones y Sistemas Inteligentes (PRIS-Lab) el trabajo de Juan Carlos Coto, estudiante de la Maestría Académica en Ciencias de la Computación de la UCR, donde su trabajo final de graduación fue titulado **Modelado híbrido de datos genómicos de alta dimensionalidad para el reconocimiento de patrones en quimiosensibilidad del cáncer** [Coto, Siles y Mora 2016], dicho trabajo consistió en definir una plataforma para realizar experimentos computacionales con datos de la conocida base de datos **The Cancer Cell Line Encyclopedia** (CCLE) donde además de datos genómicos de líneas celulares se encuentran datos de la sensibilidad de las líneas celulares de cáncer ante distintas drogas. En la plataforma se incluyó la implementación de la base de datos de las líneas celulares así como el modelado de genes mediante regresiones del número de copia y la expresión.

Actualmente se sabe que la resistencia a las drogas está relacionada a los cambios en el cariotipo presentes en el cáncer, incluyendo la multiresistencia donde el cáncer ante una droga específica desarrolla resistencia contra otras drogas distintas simultáneamente [Duesberg et al. 2007]. La teoría de la especiación del cáncer de Duesberg propone que la aneuploidía (la ganancia o pérdida de cromosomas) en el cáncer es el cariotipo de una nueva especie, donde la misma inestabilidad de la aneuploidía es la que permite el desarrollo de la resistencia a las drogas [Nicholson y Cimini 2013].

En miles de cánceres se ha revelado la existencia de patrones en el cariotipo con aneuploidías que son encontradas recurrentemente en diferentes tipos de cáncer, así como otros patrones específicos en tumores individuales y en distintos tejidos de origen [Nicholson y Cimini 2013]. El tejido de origen es importante para determinar patrones en el cariotipo del cáncer, lo cual indica que la selección de dichos cariotipos específicos dependen de la fisiología específica de las células de diferentes tejidos. Por ende sería de gran valor encontrar patrones que estén relacionados a las líneas celulares con multiresistencia a las drogas.

1.2. Propuesta

Mediante un proceso de reconocimiento de patrones se buscará identificar regiones de genes con alteraciones (amplificaciones o eliminaciones de cromosomas) de líneas y muestras de cáncer en cada tejido. En dicho proceso primero se identificará las regiones con alteraciones estadísticamente significativas por medio del método GISTIC que además preprocesará los datos por regiones, y después se debe proceder a realizar un análisis de reglas de asociación con el fin de seleccionar las regiones cromosómicas con alteraciones más frecuentes haciendo una reducción de dimensionalidad.

Con la selección de alteraciones de interés para reducir la dimensionalidad se procederá a realizar una clasificación y una asociación de las líneas celulares del CCLE a partir de las alteraciones presentes. De forma parecida sucede con las alteraciones encontradas en las muestras del TCGA donde la idea es encontrar la asociación con la supervivencia y la vida de los pacientes. Al final se buscaría ver si existen alteraciones en común entre CCLE y TCGA.

1.3. Problema

¿Al entrenar un clasificador de máquina de soporte vectorial con las alteraciones cromosómicas del CCLE y el TCGA se puede determinar las clases resistentes y sensibles con una confiabilidad (precision score) de 0,7?

1.4. Hipótesis

Un clasificador de máquina de soporte vectorial entrenado con las alteraciones cromosómicas del CCLE y el TCGA determinará las clases resistentes y sensibles con una confiabilidad (precision score) de 0,7.

1.5. Objetivos

Objetivo General

Clasificar muestras del proyecto TCGA según la quimiosensibilidad tumoral mediante la integración intermedia de datos genómicos

Objetivos Específicos

1. Investigar bibliografía relacionada a la aplicación de algoritmos de reconocimiento de patrones en genómica del cáncer.
2. Desarrollar una plataforma computacional para el manejo de los datos y realización de experimentos.
3. Diseñar el algoritmo de clasificación basado en el número de copia y la posición de los genes de las líneas celulares.
4. Encontrar reglas de asociación de datos en las alteraciones cromosomales.
5. Implementar el algoritmo de reconocimiento de patrones en líneas celulares.
6. Validar los resultados del algoritmo mediante la información disponible del mismo proyecto TCGA.
7. Escribir un artículo científico para ser enviado a revisión a una conferencia o revista de relevancia en el área.

Capítulo 2

Antecedentes

El capítulo de Antecedentes está dividido según la taxonomía de artículos sobre algoritmos de reconocimiento de patrones de datos genómicos definido por el artículo de Ritchie et al. 2015 donde se explica que los algoritmos de datos genómicos pueden emplear un solo tipo de datos o más de un solo tipo de dato genómico. Los métodos que emplean más de un tipo de dato genómico se dividen en tres tipos principales: métodos basados en concatenación o tempranos, basados en integración o intermedios, y basados en modelos o tardíos.

Los algoritmos tempranos simplemente suponen que cada muestra está representada por un vector de genes por cada tipo de dato, entonces lo que sucede es que se concatenan los vectores de los tipos de datos y se aplica la clasificación o el agrupamiento. Los algoritmos intermedios lo que realizan es una transformación a los datos y crean una representación reducida de los datos originales y en dicha representación se aplican los algoritmos. Y por último los algoritmos tardíos que se basa en aplicar clasificación o agrupamiento, y después simplemente juntan los modelos mediante métodos de votación buscando coincidencias o correlaciones entre cada una de las clasificaciones realizadas.

También en el mismo artículo de Ritchie et al. 2015 se describe que los artículos pueden trabajar con un solo tipo de datos donde se sugiere dividirlos por su posible donde se sugiere separarlos por su funcionalidad: para clasificar subtipos moleculares, sensibilidad o tipos de tejidos.

2.1. Integración de datos genómicos

Integración temprana

Los algoritmos con integración temprana son mucho más fáciles de implementar, aunque no explotan las relaciones existentes entre los distintos tipos de datos. En el artículo de Curtis 2012 se presentó un análisis integral del número de copia y la expresión génica en un conjunto de aprendizaje y validación de 997 y 995 tumores primarios. Mediante un análisis no supervisado de los perfiles de ADN-ARN fueron revelados nuevos subgrupos con resultados clínicos distintos. Se usó clustering conjuntamente con el número de copia y la expresión génica. Los 100 genes cuyo número de copia está más asociado a su expresión génica, fueron elegidos como entradas de un análisis de agrupamiento integrativo basado en la inferencia del algoritmo de *Expectation Maximization*. El número de grupos fue determinado en el índice de Dunn. Los 10 grupos obtenidos fueron tipificados por aberraciones de número de copia bien definidos y dividieron varios de los subtipos intrínsecos en los cuales normalmente se agrupan los casos de cáncer de seno. En Taskesen, Babaei et al. 2015 mediante conocimiento previo de los subtipos moleculares que caracterizan la leucemia, realizan la clasificación por medio de un esquema de integración temprana.

El estudio de Wang, Fang y Chen 2016 es un análisis sobre la predicción de las respuestas de los pacientes con cáncer ante agentes terapéuticos. El artículo analiza otros enfoques que han trabajado primordialmente en alteraciones genómicas de líneas celulares tratadas con diversas drogas. A pesar de haber alcanzado resultados prometedores para ciertas drogas, estos enfoques no han incorporado información acerca de los compuestos e ignoran el hecho que las drogas funcional o estructuralmente parecidas pueden tener un efecto terapéutico similar.

En Goodspeed et al. 2016 se discute que las líneas celulares tienen aberraciones específicas distintas a los tumores. El número de copia de ADN y la expresión génica tienen mayor similaridad entre las líneas y los tumores que los datos de mutaciones. Los subtipos moleculares están representados también en las líneas celulares, aunque en muchos casos no todos los subtipos encontrados en las muestras de tumores están representados. Se encuentran marcadores similares en tumores de distintos tejidos.

Integración intermedia

En el artículo de Taskesen, Huisman et al. 2016 se propone una integración intermedia que emplea en los algoritmos distintos tipos de datos moleculares. Los algoritmos se corrieron en 4434 muestras del proyecto *TCGA* de 19 tipos de cáncer distintos, incluyendo la expresión génica y el número de copia de ADN. Mediante un algoritmo de reducción de dimensionalidad llamado t-SNE se creó un mapa de dimensión reducida de los datos en 2D, donde en dicho mapa se aplicó un análisis de agrupamiento que reveló 18 grupos de muestras. Este enfoque de integración intermedia es propuesto como una mejor alternativa, obteniendo mayor cantidad de grupos, pues les permite capturar las interacciones entre los distintos tipos de datos que representan distintas etapas de la expresión molecular, además que el algoritmo t-SNE agrupa los datos basándose en la similitud de las muestras pues minimiza la métrica de similitud llamada Kullback-Leiber [Taskesen, Huisman et al. 2016].

El estudio de Costello et al. 2014 se centró en predecir la respuesta a las drogas basado en conjuntos de datos de perfiles genómicos, epigenómicos y proteómicos. Se analizaron en total 44 algoritmos para la predicción de quimiosensibilidad. El algoritmo con mejor desempeño en la predicción tomó en cuenta las relaciones no lineales de los datos y a su vez incorporó información de las vías moleculares¹. Se encontró que la expresión génica proporcionó el mayor poder predictivo de los conjuntos de datos, sin embargo el desempeño mejora al aumentar los conjuntos de datos independientes.

El algoritmo con mejor desempeño fue el desarrollado por un equipo de la Universidad de Helsinki [Costello et al. 2014] los cuales desarrollaron un método de aprendizaje de máquina que incorporó múltiples conjuntos de datos mediante una regresión no lineal con el objetivo de aprender a predecir múltiples sensibilidades de drogas simultáneamente. El algoritmo es llamado *Bayesian multitask multiple kernel learning* (MKL), el cual está basado en 4 principios: regresión mediante funciones de kernel, aprendizaje de múltiples vistas, aprendizaje de múltiples tareas, e inferencia bayesiana. En el aprendizaje de múltiples vistas, datos de entrada (tipos de datos) heterogéneos son integrados en un solo modelo. Durante aprendizaje de múltiples tareas se comparte la información entre las drogas simultáneamente, lo cual modela las sensibilidades a lo largo de todas las drogas.

En el artículo A. Zhang et al. 2015 se presenta un modelo de dos capas en el cual se usa redes de

¹Las vías moleculares son las relaciones que existen entre las moléculas como proteínas, ARN y ADN, las cuales se pueden modelar matemáticamente como grafos.

similaridad de las drogas y de las líneas celulares usando un modelo de pesos. A partir del modelo se obtuvo un coeficiente de correlación de Pearson mayor al obtenido al usar modelos elásticos. Además se pudo predecir la respuesta para la mayoría de drogas, donde se logró predecir que algunas líneas celulares serían más sensitivas en las que no tenían una mutación en el gen BRAF. El modelo de dos capas modela la similaridad entre las líneas celulares representadas por su expresión génica y entre drogas en su estructura química 2D. Se emplean los algoritmos de reconocimiento de patrones Máquinas de Soporte Vectorial (SVM) y Árboles Aleatorios (RF). La doble capa no es un grafo bipartito pues no tiene todos los datos de respuesta a las drogas. Aquí se menciona que el parámetro usado para representar quimiosensibilidad es el parámetro llamado IC50 que representa la actividad celular en el 50 % de la concentración de la dosis de cada droga.

Integración tardía

La integración tardía requiere modelos complejos que permitan aplicar algoritmos por separado, y requiere de mayor complejidad de elaboración y no se asegura que se aprovechen completamente las relaciones entre los tipos de datos. En el artículo Hoadley et al. 2014 se emplea un algoritmo llamado *Cluster Of Clusters Assignments* que lo que hace es aplicar un algoritmo de agrupamiento a cada uno de los conjuntos de datos y después busca las coincidencias que pueden existir entre los grupos y así forma nuevos grupos mediante un esquema de integración tardía simple.

En el estudio de Cho 2011 se explica que los meduloblastomas son tumores heterogéneos que colectivamente representan el tumor más maligno en niños. Se identificaron 6 subgrupos moleculares, cada uno con una combinación única de aberraciones cromosomales que influyen la expresión génica. Para el análisis de clustering de los datos de expresión génica se empleó el método Non-negative Matrix Factorization, y por medio de un coeficiente de correlación se determinó el número de grupos. Después se realiza un análisis de correlación con el número de copia de ADN a los subgrupos ya detectados, con el fin de encontrar posibles aberraciones de importancia.

Otro artículo es Daemen et al. 2013 donde se usan máquinas de soporte vectorial (SVM) y Árboles Aleatorios (RF) para identificar características moleculares asociadas a las respuestas de 70 líneas celulares ante 90 drogas. Los conjuntos de datos analizados incluyen mediciones de número de co-

pias, mutaciones, expresión génica, metilación y expresión proteica. SVM es más predictivo para 35 compuestos mientras RF lo es para otros 55. La combinación de datos moleculares llevan a mejores predicciones en cada conjunto de datos por separado.

2.2. Algoritmos de un solo tipo de datos

Dentro de los artículos relacionados a los algoritmos de un solo tipo de datos, se pueden dividir en tres grupos importantes: caracterización de acuerdo al tipo de cáncer, caracterización de subtipos de cáncer y caracterización según quimiosensibilidad.

Los artículos más abundantes son los encargados de clasificar los subtipos donde tratan de estratificar pacientes de un cáncer específico. Los que siguen en cantidad de artículos son los que clasifican según tejido, los cuales tratan de caracterizar los distintos tipos de cáncer para analizar sus diferencias. Los últimos en abundancia están asociados con la quimiosensibilidad determinan la efectividad de las drogas aplicadas a los tejidos.

Subtipos moleculares

En Qiu, Bi y Song 2017 se investiga acerca de cáncer de pulmón, el cual es dividido en dos subtipos principales: adenocarcinoma y carcinoma escamoso, donde ambos tienen tratamientos distintos. Ellos entrenan un clasificador **Naive Bayes** con el número de copia del conjunto de datos del TCGA para luego clasificar muestras de otros conjuntos de datos, donde obtienen un 84 % de precisión. Depuydt et al. (2018) Depuydt et al. 2018 emplea un clasificador de **random forest** en conjunto con regresiones para separar las muestras de los pacientes de neuroblastomas de acuerdo a la sobrevivencia.

En Beroukhim 2007 se desarrolló un algoritmo llamado GISTIC diseñado con el fin de hallar regiones de los cromosomas con aberraciones estadísticamente significativas cuyos resultados sean consistentes con los resultados de otros estudios acerca de variaciones cromosómicas que ocurren en gliomas (cáncer cerebral). En este importante artículo, son analizados 141 muestras de glioma con el fin de comparar los resultados de otros estudios. GISTIC consiste de un valor G que es proporcional al total de la magnitud de las aberraciones en cada posición y luego aleatoriamente se cambian las posiciones aplicando un test de permutaciones con el fin de determinar que una alteración con un valor G no ha

sucedido aleatoriamente.

En Etemadmoghadam 2009 usando el método GISTIC para encontrar regiones con amplificaciones y eliminaciones entre los grupos de muestras con carcinomas de ovario resistentes y sensibles a las terapias convencionales basadas en carboplatino. En este artículo Ruiz 2017 por medio de las alteraciones del número de copia conocidas en estudios previos fue posible clasificar correctamente un 87% de las muestras en las tres categorías de gliomas. En Cimino 2016 GISTIC es empleado para identificar alteraciones junto con Análisis de Componentes Principales para visualizar los datos, que se agrupan en los tres subtipos donde cada uno tiene alteraciones características. En Zaman 2013 se clasifica el cáncer de seno empleando los datos de número de copia y de expresión empleando agrupamiento jerárquico.

En el artículo de Bruna et al. 2016, se señala que la estratificación molecular es el primer paso hacia la medicina de precisión para cáncer. Se han realizado estudios acerca de la taxonomía genómica del cáncer de seno, pero para capturar apropiadamente la heterogeneidad intratumor se necesita información proveniente de los subtipos tumorales y la evolución genómica del cáncer. En la búsqueda de una solución a este problema, se ha encontrado que una posible solución es la caracterización molecular de las xenotransplantaciones² y sus líneas celulares derivadas a corto plazo pues retienen la heterogeneidad y pueden ser usadas en ensayos clínicos. Por medio de un método basado en la pendiente de la curva de respuesta de la dosis se clasificaron los patrones de sensibilidad de las drogas en ocho grupos, donde compuestos con mecanismos de acción similares se agruparon juntos.

Existe otro conjunto de algoritmos que no se han empleado en el área específica de la quimiosensibilidad pero se han empleado para identificar subtipos de cáncer en distintos de tejidos usando un solo conjunto de datos, en este caso específico expresión génica. Por ejemplo el artículo de Collison et al. 2013, Liu, X. Zhang y S. Zhang 2014, Bailey et al. 2016] y Hughey y Butte 2015.

En el artículo de Collison et al. 2013 se describe que el cáncer de páncreas tiene 3 subtipos que presentan evidencia de resultados clínicos y respuestas terapéuticas diferentes: clásico, cuasi-mesenquimal y exocrino. Se aplicó Factorización de Matrices No-negativas (NMF) para agrupar e identificar subtipos por medio de datos de expresión génica. Luego se obtuvo un marcador de 62 genes, por medio de un análisis estadístico respecto al cambio de expresión relativo a la desviación estándar. La estratificación

²Xenotransplantaciones son muestras de cáncer de personas que se implantan en ratones para usarlos en ensayos clínicos.

transcriptómica (expresión génica) en subtipos proporcionó información de pronósticos para anotaciones clínicas.

En el artículo publicado por Liu, X. Zhang y S. Zhang 2014, las aberraciones en el genoma del cáncer observadas en investigación básica y clínica ha sido empleadas para categorizar pacientes en un esfuerzo para mejorar las decisiones clínicas y desarrollar tratamientos más efectivos. El cáncer de seno no debe ser visto como una sola enfermedad, por el contrario es una enfermedad heterogénea que consiste de diferentes subtipos en el nivel molecular y clínico, con diferentes implicaciones terapéuticas. Diferentes subtipos de cáncer tienen diferentes respuestas a los tratamientos. Dada la heterogeneidad entre pacientes, se busca encontrar tales distintos grupos. Se adoptaron los 5 grupos tumorales encontrados con el método PAM50 con el conjunto de datos definidos en el METABRIC. Después se realizó a cada uno de estos grupos un análisis de supervivencia, donde cada subtipo muestra pronóstico distinto.

En el artículo de Bailey et al. 2016] se observa el análisis integral de 456 muestras de cáncer de páncreas identificaron 32 genes mutados recurrentemente que se encuentran 10 vías moleculares. El análisis de expresión definido para los 4 subtipos: escamoso, progenitor pancreático, inmunogénico y diferenciado aberrantemente (ADEX), los cuales presentan características patológicas bien definidas. Se empleó clustering no supervisado en los datos de expresión génica mediante el algoritmo Non-negative Matrix Factorization, y se usó el coeficiente de correlación para determinar el número de grupos significativos en el espacio de muestras.

En el estudio de Hughey y Butte 2015 se desarrolló un método para realizar análisis de datos genómicos usando un algoritmo de regresión llamado *red elástica*, el cual proporciona un enfoque poderoso y versátil para la clasificación y la regresión. Usando 629 muestras a partir de los 5 conjuntos de datos, entrenaron un clasificador para distinguir entre 4 subtipos de cáncer de pulmón.

Quimiosensibilidad

En Carter et al. 2017 se entrenó una máquina de soporte vectorial con el objetivo de clasificar las muestras de pacientes con cáncer de pulmón en dos categorías: refractario (después de la primera quimioterapia, el paciente tiene una recaída dentro del primer tratamiento) o sensible (recaída después

de los primeros tres meses de tratamiento con etopósido o carboplatino). Se alcanza un 83 % de clasificación correcta usando validación cruzada de 10 iteraciones. Fueron detectadas 2281 posiciones con alteraciones el número de copia de los cromosomas. En Ni y Zhuo 2013 por medio de la observación de muestras de pacientes con cáncer de pulmón se observa que los cambios en el número de copia son eventos clave en la metástasis. Las variaciones en el número de copias permanecen constantes, aunque el tratamiento de la quimioterapia continúa avanzando. Al emplear clustering jerárquico diferentes subtipos de cáncer de pulmón tienen diferentes patrones de alteraciones.

Tipos de tejido

En N. Zhang et al. 2016 se emplea extracción de características para reducir la dimensionalidad de los datos de número de copia del TCGA donde se hace un ranking de los 200 genes más relevantes para la clasificación. El clasificador es un árbol de decisión usando un sistema de votación dividido en segmentos. Los datos del TCGA se emplearon para clasificar a cuál de 6 posibles tejidos pertenecían los datos, alcanzando un precisión del 0.75. En Xu et al. 2018 se emplea una máquina de soporte vectorial de kernel lineal basado en el número de copia para clasificar si la muestra es de tejido saludable o pertenece a cáncer de colon en etapa de desarrollo temprana. Se alcanzó una sensibilidad del 91.8 %. El clasificador está basado en las variaciones del número de brazos en cada cromosoma. Emplean un test estadístico Z para cada gen y con base en dicho resultado se entrena y aplica el clasificador.

En Sanchez-Garcia, Villagrasa y Matsui 2014 se usa un nuevo algoritmo de identificación de regiones con alteraciones en el número de copias llamado ISAR, con el fin de usar dichas regiones en conjunto con la información clínica para definir genes importantes en el crecimiento del cáncer. Ellos emplean un **Modelo de Mezclas Gaussianas** para seleccionar los genes más relevantes basados en la información conjunta de las alteraciones en el número de copia de las muestras en el TCGA junto con los genes que presentaban alteraciones en la expresión de las líneas celulares del CCLE.

En el artículo de Cheung et al. 2011 se analizan las líneas celulares con cáncer donde se encuentran alteraciones por medio de un método estadístico llamado pesos de evidencia que toma en consideración los números de copia y la expresión de los genes. En Arakawa et al. 2017 se observa que las alteraciones en el número de copias están relacionadas a distintas etapas del desarrollo del cáncer gástrico. Se em-

plea el test estadístico chi-cuadrado con el objetivo de encontrar regiones con alteraciones particulares en los cromosomas. En Yuan et al. 2012 se compara múltiples métodos para identificar alteraciones en el número de copia donde GAIA y GISTIC sobresalen por su identificación de blancos moleculares conocidos.

Capítulo 3

Marco Teórico

3.1. Cáncer y biología

Cáncer

El cáncer es una colección de enfermedades relacionadas, en donde algunas de las células del cuerpo comienzan a dividirse sin detenerse y propagarse a los tejidos circundantes. El cáncer puede comenzar casi en cualquier parte del cuerpo humano, que está formado por millones de células. En condiciones normales, las células humanas crecen y se dividen para formar nuevas células a medida que el cuerpo las necesita. Cuando las células envejecen o se dañan, mueren y nuevas células toman su lugar (NCI 2015).

Cuando el cáncer se desarrolla, sin embargo, este proceso ordenado se rompe, pues a medida que las células se vuelven más y más anormales, las células viejas o dañadas sobreviven cuando deberían morir y las nuevas células se forman cuando no son necesarias. Estas células adicionales pueden dividirse sin detenerse y pueden formar tumores (NCI 2015).

Las secuencias genómicas de las células individuales dentro del organismo están sujetas a alteraciones de su estructura mediante distintos mecanismos que a su vez modifican la información contenida en el genoma. Los genes mutados resultantes pueden desviar las células para adquirir nuevos fenotipos anormales. Dichos cambios pueden ser incompatibles con los roles normalmente asignados a la estructura y fisiología del organismo (Weinberg 2014).

Entre estos cambios inapropiados pueden estar las alteraciones en la proliferación celular, lo que

puede llevar a la aparición de grandes poblaciones de células que ya no siguen las reglas que rigen la construcción y el mantenimiento normal de los tejidos (Weinberg 2014).

Cuando se representan de esta manera, las células que forman un tumor son el resultado de un desarrollo normal que ha salido mal. A pesar de los mecanismos que tiene el organismo para prevenir su aparición, las células cancerosas de alguna manera logran desarrollarse y crecer. Las células normales están cuidadosamente programadas para colaborar unas con otras en la construcción de los diversos tejidos que hacen posible la supervivencia. Pero las células cancerosas se desarrollan siguiendo una sola consideración, la cual es hacer más copias de sí mismas (Weinberg 2014).

Los tumores cancerosos son malignos, lo que significa que pueden diseminarse o invadir los tejidos cercanos. Además, a medida que estos tumores crecen, algunas células cancerosas pueden desprenderse y viajar a lugares del cuerpo a través de la sangre o el sistema linfático y formar nuevos tumores lejos del tumor original (NCI 2015).

Las células cancerosas se diferencian de las células normales de muchas maneras que les permiten crecer fuera de control y convertirse en invasoras. En especial las células cancerosas son menos especializadas que las células normales. Es decir, mientras que las células normales maduran en tipos celulares muy distintos con funciones específicas, las células cancerosas no lo hacen. Esta es una de las razones por las que, a diferencia de las células normales, las células cancerosas continúan dividiéndose sin detenerse (NCI 2015).

Además, las células cancerosas pueden ignorar las señales que normalmente le dicen a las células que dejen de dividirse o que comiencen un proceso conocido como muerte celular programada, o apoptosis, que el cuerpo usa para deshacerse de las células innecesarias (NCI 2015).

Aneuploidía

El cáncer es un proceso evolutivo en el que las células han adquirido mutaciones que confieren rasgos fenotípicos beneficiosos, como la resistencia a la muerte celular, donde se expanden y superan las células vecinas. Las células cancerosas son notoriamente conocidas por alteraciones genómicas y

por su alta variabilidad (Duesberg et al. 2007).

La aneuploidía la cual puede ser definida como la existencia de un número desequilibrado de cromosomas, es una característica del cáncer presente en el 90% de los tumores sólidos. Aunque es una de las alteraciones descritas más antiguas del cáncer, y aunque los esfuerzos genómicos han permitido la caracterización precisa del cariotipo en los pacientes con cáncer, el papel de la aneuploidía en el surgimiento del cáncer es desconocido (Duesberg et al. 2007).

La aneuploidía es una alteración que afecta el número de copia de ADN de cromosomas enteros y de brazos cromosómicos. En cáncer, la aneuploidía afecta la mayor parte del genoma, más que cualquier otra alteración genética. Las alteraciones más frecuentes ocurren en 30% de los tumores sólidos. Los subtipos del cáncer están usualmente caracterizados por patrones específicos de alteraciones en los brazos cromosomales (Duesberg et al. 2007).

Las aneuploidías surgen de errores persistentes en la segregación de los cromosomas durante la división celular en un proceso conocido como la inestabilidad cromosómica. La inestabilidad cromosómica es un contribuyente principal a la variabilidad y heterogeneidad genética en el cáncer y es un determinante importante en el diagnóstico y la resistencia de los tratamientos (Duesberg et al. 2006).

Teoría cariotípica del cáncer

Muchas células con cáncer adquieren durante los tratamientos resistencia a muchas drogas o incluso antes del tratamiento ya son resistentes a estas, contrario a lo que sucede a las células normales que permanecen sensibles a las drogas durante períodos prolongados de tiempo (Duesberg et al. 2007).

A pesar de que ha habido investigación en el tema durante las últimas décadas, todavía está en debate cómo las células generan fenotipos con una alta resistencia ante muchas drogas más rápidamente que lo predicho por las mutaciones, y cómo pueden las células con cáncer generar dicha resistencia con los mismos genes que han sido heredados de las células normales. La búsqueda de las causas de la resistencia a las drogas se ha concentrado en las mutaciones genéticas y epigenéticas pero puede ser que sea necesario proponer nuevas posibilidades como algunos autores lo han propuesto con la teoría

cariotípica del cáncer (Duesberg et al. 2007).

Las teorías genéticas convencionales han fallado para explicar por qué el cáncer no es encontrado en recién nacidos y no es hereditario, es cromosomal y fenotípicamente inestable, tiene aneuploidías específicas al cáncer y fenotipos no selectivos como la resistencia a múltiples drogas, metástasis y la inmortalidad (Duesberg et al. 2006).

El desbalance de miles de genes corrompe el grupo de proteínas que segregan, sintetizan y reparan cromosomas (Duesberg et al. 2007). La aneuploidía es una fuente estable de variaciones cariotípicas y fenotípicas, pues la selección de aneuploidías estimula la evolución y la progresiones subsecuentes y malignas de las células con cáncer; esto hace que las células cancerosas sean como especies nuevas y distintas con cariotipos inestables, en vez de solo células mutadas. Las aneuploidías específicas del cáncer generan fenotipos complejos y malignos a través de las variaciones anormales de miles de genes.

CCLE

La traducción sistemática de los datos genómicos del cáncer al conocimiento de la biología del tumor y las posibilidades terapéuticas sigue siendo un desafío. Dichos esfuerzos deberían recibir una gran ayuda de sistemas de modelos preclínicos robustos que reflejen la diversidad genómica de los cánceres humanos y para los cuales se dispone de anotaciones genéticas y farmacológicas detalladas (Barretina 2012).

La Enciclopedia de Líneas Celulares de Cáncer (CCLE, por sus siglas en inglés) es un esfuerzo por compilar la expresión génica, el número de copias cromosómicas y los datos de secuenciación de más de 1000 líneas celulares de cáncer humano. Adicionalmente se tiene los perfiles farmacológicos de 24 fármacos anticancerosos en 479 de las líneas celulares, esta colección se generó de esta forma con el fin de identificar de factores genéticos y de predicción basados en la expresión génica y la sensibilidad farmacológica (Barretina 2012).

En el caso de cáncer de seno en el CCLE, solamente se tienen a disposición 59 líneas celulares de las cuales solamente de 29 se tiene acceso a información de la sensibilidad a las drogas.

El principal parámetro a considerar en el estudio del CCLE es el IC50 que significa el 50% de la concentración inhibitoria de la droga. Dicho valor se emplea como medida farmacocinética para indicar la sensibilidad o la resistencia de una línea celular ante una droga, donde de hecho también es el parámetro empleado en la pruebas clínicas de sensibilidad realizadas a pacientes de cáncer para determinar el mejor tratamiento.

La principal fuente bibliográfica de referencia de este conjunto de datos se encuentra en el artículo descrito por el Broad Institute en el artículo de (Barretina 2012) donde se describe la metodología con la que se realizó el estudio y la recolección de datos.

TCGA

El proyecto TCGA (The Cancer Genome Atlas) es una colaboración entre el Instituto Nacional del Cáncer en Estados Unidos (NCI, por sus siglas en inglés) y el Instituto de Investigación en el Genoma Humano, donde como un atlas se ha compilado la información genómica a lo largo de 33 tipos de cáncer. El conjunto entero de datos del TCGA contiene la información de 11 mil muestras de pacientes, lo cual ha contribuido enormemente a la investigación en ciencia básica a través de una gran red de investigaciones relacionados a dicha compilación de información. Cada uno de los tipos de cáncer es causado por células con un crecimiento descontrolado debido a errores en el genoma, por dicha razón, estudios como el TCGA buscan la recopilación masiva de información para consistentemente analizar los mecanismos del surgimiento del cáncer así como mejorar los métodos de diagnóstico y tratamiento de la enfermedad (Lee 2015).

Además tiene datos de expresión de ARN y miARN, números de copia de ADN y datos de secuenciación. El TCGA tiene disponible datos clínicos de los pacientes (aunque no disponible para cada caso, a veces hay ciertos datos incompletos). Los datos clínicos incluyen datos de género, grupo racial, supervivencia, el estado acerca de la supervivencia del paciente, y el estado sobre la desaparición del cáncer en el paciente.

En el caso del cáncer de seno del TCGA se encuentran 1058 muestras disponibles con la infor-

mación disponible del estado binario de supervivencia del paciente así como la supervivencia del paciente mismo.

El TCGA forma parte de un repositorio de datos genómicos del cáncer llamado *Genomic Data Commons* llevado a cabo por parte del Instituto del Cáncer Estadounidense (NCI, por sus siglas en inglés) con el objetivo de agrupar varias iniciativas, siendo las principales TCGA y TARGET, donde este último es una iniciativa para almacenar datos sobre cáncer infantil ya que suele tener variaciones importantes en relación al cáncer en adultos. TCGA posee datos de 33 tipos distintos de cáncer según tejido.

La página de *Genomic Data Commons* tiene una parte dedicada a la exploración y visualización de los datos, inclusive también aquí se pueden descargar manualmente todos los datos, muestra por muestra: <https://portal.gdc.cancer.gov/exploration>. Luego también tiene una parte dedicada propiamente a descargar los datos de los distintos conjuntos de muestras: <https://portal.gdc.cancer.gov/repository>. Para realizar la descarga se seleccionan las categorías de datos deseados del repositorio gráficamente por medio de las opciones del panel lateral izquierdo y se descarga un archivo llamado Manifest en conjunto con una herramienta llamada gdc-client (<https://gdc.cancer.gov/access-data/gdc-data-transfer-tool>). En general tiene muchas otras funcionalidades la página, hablan más acerca de los datos y también existe un API para comunicarse con los datos.

A pesar de lo anterior, la mejor forma para descargar los datos es por medio de una herramienta de consola llamada *gdctools* (<https://github.com/broadinstitute/gdctools>). Por medio de esta herramienta se pueden descargar los datos del repositorio de Genomic Data Commons. El GitHub de *gdctools* es <https://github.com/broadinstitute/gdctools>.

También hay que recordar que se encuentra la página *cBioPortal*, la cual contiene un repositorio con muchos trabajos de datos genómicos, dentro de los cuales se encuentran los de Genomic Data Commons, junto con otros proyectos de muchos otros centros de investigación http://www.cbioportal.org/data_sets.jsp. Ahí puede ver los datos clínicos de los pacientes para que se dé una idea del formato. Los archivos clínicos ya descargados vienen en formato XML.

FireBrowse (<http://firebrowse.org>) es una página que reúne los resultados de múltiples experimentos llevados a cabo con el TCGA incluyendo pruebas de alteraciones del número de copias de ADN. Es una página llena de visualizaciones sobre los distintos tipos y conjuntos de datos del TCGA. La página pertenece al reconocido instituto *Memorial Sloan Kettering Cancer Center* de Nueva York.

GISTIC

El método de Identificación de Objetivos Genómicos Significativos en Cáncer (GISTIC, por sus siglas en inglés) es el encargado de la identificación de alteraciones significativas en el número de copia en el genoma. Este método funciona mediante dos etapas (Beroukhim 2007). En la primer etapa se identifican las posiciones de las aberraciones cromosómicas en múltiples tumores y se calcula una puntuación estadística llamada G, la cual es proporcional a la magnitud total de cada una de las aberraciones en cada posición. En la segunda etapa al permutar las posiciones en cada tumor, GISTIC determina la frecuencia con la que se determina el valor G asignando un valor de probabilidad q mediante la aplicación de unas pruebas estadísticas llamadas FDR (*False Discovery Rate*). Se define un valor de umbralización de modo que las alteraciones encontradas no ocurren solo por casualidad, pues se consideran únicamente las alteraciones superiores a este umbral.

GISTIC se puede dividir en cuatro etapas que abordan 4 problemas distintos para identificar aberraciones cromosómicas (Beroukhim 2007):

- Caracterización de las aberraciones cromosómicas en cada tumor: Las aberraciones en cada uno de los tumores deben ser mapeadas con precisión.
- Diferenciación entre aberraciones relevantes y no relevantes: deben identificarse las aberraciones que se elevan por encima de las aberraciones aleatorias.
- Identificación de las regiones con picos de mayor probabilidad de contener genes objetivo: Para cada aberración relevante debe identificarse la posición en el genoma con mayor probabilidad de contener los genes específicos causantes de las alteraciones.
- Clasificación de los tumores en función de sus aberraciones: Los tumores deben clasificarse en cuanto a si son aberrantes (con amplificaciones o eliminaciones) en las posiciones del genoma predichos, de modo que se puedan estudiar los efectos de esas aberraciones.

La etapa 2 contiene las 2 características centrales del algoritmo: que puntúa cada marcador genómico de acuerdo con una medida integrada de la amplitud de los cambios en el número de copias, y que evalúa la significancia estadística de cada puntuación en comparación con los resultados esperados de la tasa de aberración promedio.

En esta etapa, se mapean las aberraciones cromosómicas en cada tumor. Aquí, el objetivo es maximizar la precisión con la que se identifican estas aberraciones. Las regiones cromosómicas con intensidades de señal altas se designan como amplificaciones y las regiones con intensidades de señal bajas se designan como eliminaciones.

Etapa 1

Existen varios métodos para reducir los efectos del ruido aleatorio en los conjuntos de datos de número de copia, mediante la identificación de regiones con variación de número de copia y el promedio de las intensidades de señal para todos los marcadores dentro de dichas regiones. Los ejemplos incluyen algoritmos de segmentación tales como el análisis de ganancia y pérdida de ADN (GLAD, por sus siglas en inglés). GLAD se utiliza debido a su alta sensibilidad para identificar cambios en el número de copias. Sin embargo, este alto nivel de sensibilidad a veces lleva a GLAD a informar cambios de número de copia inexistentes en segmentos muy pequeños (menos de cuatro marcadores).

En los conjuntos de datos de baja calidad, las variaciones de intensidad de la señal debidas a los cambios en el número de copias están ocultas por el ruido. Por lo tanto, se identifican conjuntos de datos de alta calidad que tienen picos separados, correspondientes a diferentes números de copia, en histogramas de los datos de intensidad de señal. Las muestras de baja calidad, particularmente aquellas con una extensa contaminación con ADN normal, generan una señal insuficiente para distinguir picos separados y se descartan.

Etapa 2

Esta etapa contiene las dos características principales de GISTIC. Primero, se califica cada marcador genómico como una región afectada por aberraciones relevantes mediante un puntaje denominado

G. Se tratan por separado las amplificaciones y las eliminaciones, lo que permite la posibilidad de que una región pueda señalarse como amplificada y eliminada simultáneamente. En los casos de amplificaciones y eliminaciones, se asume que tanto la frecuencia como la amplitud promedio de estos eventos se indican independientemente. Por lo tanto, se usa un puntaje integrado de la prevalencia los cambios en el número de copias por la amplitud promedio (transformada a \log_2).

En segundo lugar, comparamos estas puntuaciones G con la distribución de las puntuaciones esperadas si solo se observaran aberraciones aleatorias. Dicha distribución se puede determinar al restaurar el genoma después de permutar las ubicaciones de los marcadores dentro de cada muestra. La comparación de las puntuaciones reales con las generadas por el modelo de aberraciones aleatorias permite calcular la significancia estadística de cada puntuación G, representada por los valores q de la Tasa de Descubrimiento Falso (False Discovery Rate, en inglés) lo que representa la probabilidad de que los datos observados no hayan sido generados por casualidad.

Las regiones del genoma que son demasiado frecuentes o altamente aberrantes para ser explicadas solo por casualidad son seleccionadas como susceptibles de albergar aberraciones relevantes.

Etapa 3

En esta etapa, GISTIC identifica las ubicaciones de los genes más relevantes para las aberraciones en los cromosomas; esto partiendo del hecho que estos genes se pueden encontrar más frecuentemente en las aberraciones con mayor puntuación, donde pueden haber múltiples genes en una región alterada y pueden actuar a lo largo de una región cromosomal entera.

Etapa 4

Para determinar los efectos de las aberraciones relevantes identificadas en la etapa 2, se debe clasificar los tumores en cuanto a si tienen estas aberraciones. Debido a que es más probable que las regiones pico contengan los objetivos genómicos de estas aberraciones, GISTIC clasifica primero cada tumor según su estado de número de copia en las regiones pico. Para aberraciones amplias, que pueden estar afectando específicamente una gran región del genoma, GISTIC clasifica a cada tumor en cuanto a si es aberrante o no en la mayor parte de la longitud de la región.

Parámetros de GISTIC

Los parámetros de configuración de GISTIC tienen valores por defecto en caso que estos no sean proporcionados. Los cuatro parámetros más importantes son: `t_amp`, `t_del`, `q_vthresh` y `conf_level`.

`t_amp` y `t_del`, son los valores de umbral para que un número de copia pueda ser considerado como una amplificación o una eliminación, respectivamente. Ambos valores por defecto son 0.1 y el rango de valores sugeridos va de 0.1 a 0.3.

`qv_thresh` es el valor de significancia estadística empleado para escoger las alteraciones con mayor recurrencia. El valor por defecto es 0.25.

`conf_level` es el nivel de confianza con el que se puede definir los límites de una alteración. El valor por defecto es 0.75, aunque es recomendable tener valores superiores o iguales a 0.9 para asegurar una mayor confianza de la región local de la alteración.

GISTIC emplea como entrada principal, el archivo de segmentación, el cual está encargado de indicar los segmentos del genoma con alteraciones en el número de copia.

También GISTIC usa en su configuración un archivo llamado archivo de marcadores encargado de indicar las posiciones capturadas en los experimentos de **microarreglos** originales. Como tal archivo usualmente no se tiene, hay dos opciones: se toma como referencia el archivo de segmentación para generar las posiciones de los marcadores, o no se emplea dicho archivo pero la última versión del método GISTIC genera un archivo de seudomarcadores que simplemente define marcadores cada cierto número de bases.

Medicina personalizada

Los objetivos principales en la medicina contra el cáncer se encuentran en la prevención, detección y tratamiento. Se busca la compilación de información acerca de las alteraciones en el genoma del cáncer junto con conocimiento detallado de los estados epigenéticos y transcriptómicos de los genomas con el objetivo final de definir alternativas de detección y tratamiento de forma estratificada para los

diversos grupos de pacientes basándose en las características genéticas de caso particular. Pero falta mucho por resolver para poder emplear con fines clínicos dicha información recolectada pues se necesita entender los mecanismos del desarrollo del cáncer (Chin, Andersen y Futreal 2011).

El uso de herramientas de alta capacidad permiten caracterizar el ácido nucleico con el fin de analizar las alteraciones en el genoma del cáncer. La complejidad de las alteraciones genómicas necesita de métodos que permitan identificar la función de los genes a partir de los datos recolectados en múltiples esfuerzos. El desarrollo futuro de nuevas estrategias de tratamiento contra la enfermedad dependen de la interpretación detallada de las consecuencias de las alteraciones genómicas (Chin, Andersen y Futreal 2011).

3.2. Reconocimiento de patrones

Etapas

En el esquema clásico de reconocimiento de patrones, los métodos dividen su funcionamiento en distintas etapas: preprocesamiento, selección o extracción de características, clasificación y validación. Todas estas etapas pueden tener retroalimentación entre todas ellas, de modo que por ejemplo la validación afecte la selección de características y la clasificación (Duda, Stork y Hart 2000).

La etapa de preprocesamiento se encarga de limpiar los datos del ruido y los demás efectos indeseados que distorsionan las mediciones de los valores reales, un ejemplo de esto es el filtrado del ruido de las imágenes, o la eliminación de los elementos con valores vacíos o que se salen de los rangos de una tabla de datos. La etapa de selección de características es la encargada de decidir cuáles atributos son valiosos y cuáles no, como sucede en la selección de genes para mejorar la clasificación o el agrupamiento. La etapa de clasificación es la aplicación de algún método encargado de determinar la clase de un elemento basado en los atributos definidos. La etapa de validación está dedicada a determinar si los resultados de la clasificación están acordes a las métricas esperadas, sino hay que corregir las etapas anteriores con el fin de obtener resultados más óptimos (Duda, Stork y Hart 2000).

El flujo de tradicional de trabajo de los algoritmos de reconocimiento de patrones se ilustra en

la imagen 3.1 en donde se pueden observar que puede haber retroalimentación entre todas las partes que lo conforman. En el caso de este trabajo la parte de selección de características se lleva a cabo en la selección de regiones cromosómicas con el algoritmo de GISTIC mientras que los algoritmos de clasificación y la validación cruzada complementan el flujo de trabajo.

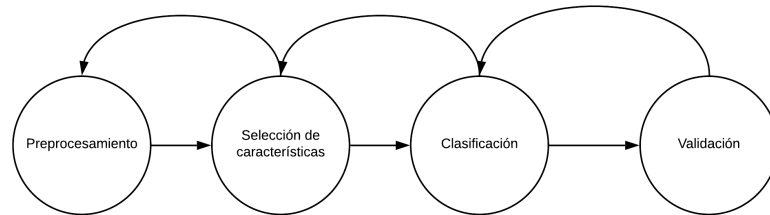


Figura 3.1: Flujo de trabajo reconocimiento de patrones

Reglas de asociación

Los elementos de un conjunto de datos están representados por un vector de características o atributos. Si dichos atributos son discretos entonces puede establecerse la frecuencia con la cual cada atributo aparece en el conjunto total de los elementos (esta métrica se le llama soporte) y podría incluso establecerse la relación entre los atributos y observar la frecuencia ocurren con la cual dichas relaciones dentro del conjunto de datos (esta métrica se le llama confianza) Witten y Frank 2005.

Estas frecuencias pueden encontrarse empleando el algoritmo Apriori, el cual emplea un enfoque de abajo hacia arriba, donde los subconjuntos más pequeños y más frecuentes son probados contra los datos. Se buscan las relaciones que cumplan la especificación de la métrica de cobertura (coverage) y confianza (confidence) que se solicita al inicio al ejecutarse el algoritmo Witten y Frank 2005.

El algoritmo Apriori, en términos generales, genera una estructura de datos, en la que se buscan una a una las relaciones y se eliminan aquellas que no cumplan el valor de la métrica solicitada Witten y Frank 2005. Primero se generan los conjuntos de elementos con la frecuencia mínima indicada (cobertura) y luego por cada elemento se determinan las reglas (relaciones entre atributos) considerando el valor mínimo de frecuencia de aparición mutua (confianza). Entonces al inicio se generan los conjuntos de elementos que cumplan la cobertura y luego se sigue con los conjuntos de dos elementos que cumplan la cobertura, y así sucesivamente podría seguir en adelante con los conjuntos de más elementos. Luego

después de generada esa lista se debe verificar el valor de confianza de cada combinación (subconjunto) de elementos y eliminar de la lista los que no cumplen dicho requerimiento mínimo de confianza. El algoritmo descrito se llama *search breadth first*.

Máquina de soporte vectorial

Los modelos lineales simples se pueden usar para la clasificación en situaciones donde todos los atributos son numéricos. Su mayor desventaja es que solo pueden representar límites lineales entre clases, lo que los hace demasiado simples para muchas aplicaciones prácticas. Existen un tipo de algoritmos llamados máquinas de soporte vectorial los cuales utilizan modelos lineales para implementar límites de clases no lineales. Estos límites se definen al transformar la entrada utilizando un mapeo no lineal; en otras palabras, transforman el espacio del vector de características original en un nuevo espacio. Con una asignación no lineal, una línea recta en el nuevo espacio no se ve como antes en el espacio de características original Witten y Frank 2005.

Los algoritmos de soporte de máquina vectorial han tenido una gran variedad de aplicaciones dentro de los estudios moleculares relacionados al cáncer han demostrado tener los mejores resultados al realizar marcos de comparación entre distintos algoritmos disponibles Costello et al. 2014.

Las máquinas de soporte vectorial se basan en un algoritmo que encuentra un tipo especial de modelo lineal: el hiperplano de máximo margen, definiéndose un hiperplano como un modelo lineal Witten y Frank 2005. El hiperplano de máximo margen es el que proporciona la mayor separación entre las clases. En la figura 3.2 se muestra un ejemplo de un clasificador lineal de máquina de soporte vectorial, en el que las clases están representadas por círculos abiertos y rellenos, respectivamente. Se supone que las dos clases son linealmente separables. Entre todos los hiperplanos que separan las clases, el hiperplano de margen máximo es el que está lo más alejado de ambos conjuntos de puntos, donde se define la bisectriz perpendicular de la línea más corta que conecta los conjuntos de datos.

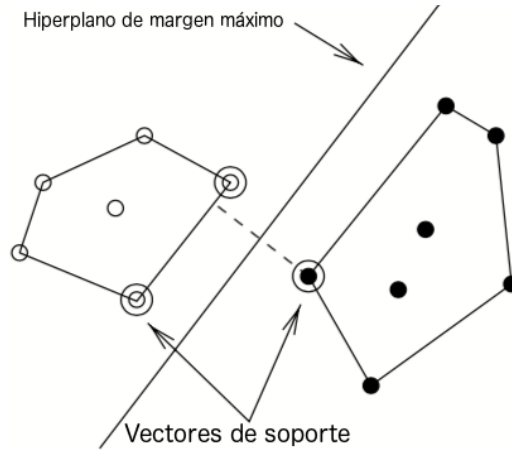


Figura 3.2: Representación de un clasificador lineal de máquina de soporte vectorial

Las instancias más cercanas al hiperplano de margen máximo, aquellas con una distancia mínima, se denominan vectores de soporte. Siempre hay al menos un vector de soporte para cada clase, y puede haber más de uno. Lo importante es que el conjunto de vectores de soporte define de manera única el hiperplano de máximo margen para el problema de aprendizaje. Dados los vectores de soporte para las dos clases, podemos construir fácilmente el hiperplano de máximo margen. Todas las demás instancias de entrenamiento son irrelevantes: se pueden eliminar sin cambiar la posición y la orientación del plano. La ecuación 3.1 muestra la estructura matemática del hiperplano de máximo margen en la cual γ_i es un valor relacionado a una clase específica (1 o -1), $\mathbf{a}(i)$ hace referencia a los vectores de soporte seleccionados en el entrenamiento y \mathbf{a} es una instancia en el entrenamiento. Los valores α_i y γ_i son parámetros que se obtienen durante el entrenamiento Witten y Frank 2005. La ecuación 3.2 muestra lo mismo que la ecuación 3.1, solo que se agrega que se calcula el valor de una función $k(\cdot)$ teniendo como argumento el producto punto del vector de soporte y el vector de entrenamiento, y así es como se hace el mapeo no lineal de la función.

$$x = b + \sum_{i \text{ is support vector}} \alpha_i \gamma_i \mathbf{a}(i) \cdot \mathbf{a} \quad (3.1)$$

$$x = b + \sum_{i \text{ is support vector}} \alpha_i \gamma_i k(\mathbf{a}(i) \cdot \mathbf{a}) \quad (3.2)$$

Capítulo 4

Metodología

El método GISTIC genera como salida las regiones cromosómicas con las alteraciones estadísticamente más significativas pero también proporciona el estimado de la mediana del número de copia para cada brazo cromosómico de las muestras. Con los valores de las medianas de los brazos cromosómicos se pueden caracterizar las muestras de cáncer de seno, de modo que se puede determinar cuáles muestras tienen cierto tipo de alteraciones (amplificaciones o eliminaciones).

En términos generales la metodología se puede separar en tres distintas partes funcionales:

- Selección de GISTIC.
- Pruebas de clasificación.
- Pruebas de reglas de asociación.

De primero se lleva a cabo la selección de regiones cromosómicas con alteraciones y luego se emplean dichas regiones en las pruebas de clasificación y en las pruebas de reglas de asociación. Después las pruebas de clasificación y reglas de asociación se aplican en los conjuntos de datos del CCLE y del TCGA. En el caso de las reglas de asociación se puede aplicar a los datos sin etiquetas (todos juntos) o con etiquetas (separados en clases). El diagrama de la figura 4.1 describe el trabajo llevado a cabo donde primero se preprocesa los datos con GISTIC para después llevar a cabo la clasificación y las reglas de asociación de datos para las regiones con alteraciones.

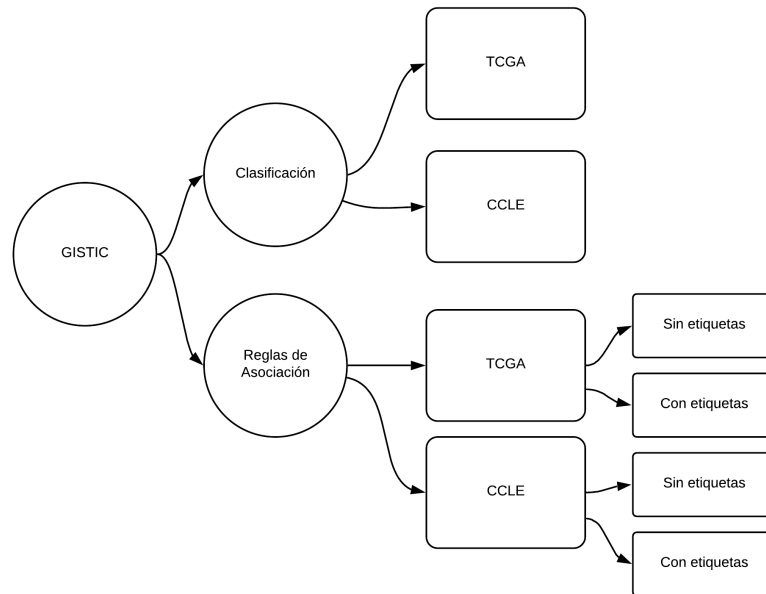


Figura 4.1: Diagrama de flujo de trabajo de la metodología

A los vectores de las alteraciones se les aplica un clasificador de modo que se separan las líneas según las etiquetas disponibles donde el espacio de muestras es el conjunto de brazos cromosómicos con alteraciones cuyos valores son continuos.

De esta manera las muestras pueden estar caracterizadas por las alteraciones cromosómicas si se consideran dichas alteraciones de forma discreta, al definir las como Amplificaciones o Eliminaciones, así se puede realizar un análisis de reglas de asociación de datos por medio del cual se puede determinar la relación entre pares y subconjuntos de alteraciones. La idea es encontrar las alteraciones más frecuentes en cada conjunto y subconjunto de datos, así como encontrar las asociaciones entre pares de alteraciones con el fin de determinar la confianza de que dicha relación ocurra en una línea o muestra específica.

4.1. Selección por medio de GISTIC

GISTIC permite seleccionar las regiones cromosómicas con alteraciones a lo largo de un conjunto de muestras. GISTIC tiene una configuración por defecto permite realizar una corrida sin configurar ningún parámetro, aunque se pueden cambiar por los rangos de valores sugeridos.

El valor del umbral de significancia estadística está definido por defecto como (0.25) aunque no tiene mucha influencia en el resultado porque los valores de significancia en las regiones de brazos cromosómicos con alteraciones suelen ser muy pequeños, cercanos a cero, aunque si se emplea un valor aún más pequeño como 0.05 puede ser que algunas alteraciones se filtren. Luego el valor de confianza descrito en el marco teórico se define como 0.95 aunque el valor por defecto es 0.75 con el objetivo de definir con completa certeza que las alteraciones encontradas si suceden.

En esta sección del trabajo se descargaron los archivos de segmentación con el número copia de ADN desde la página web del TCGA. En la descarga hay un archivo por cada muestra por lo que para lograr aplicar GISTIC se unificó todo en un mismo archivo y se aplicó GISTIC a los datos del proyecto de cáncer de seno.

La descarga de los archivos de número de copia de TCGA se llevó a cabo por medio de una herramienta llamada *gdctools* que funciona en la línea de comandos con la cual se puede decidir descargar cada uno de los tejidos disponibles así como cada uno de los tipos de datos genómicos disponibles. Los datos del CCLE ya se encontraban disponibles en el PRIS-Lab desde el trabajo de graduación del estudiante Juan Carlos Coto.

El método de GISTIC está implementado en Matlab y solamente corre en Ubuntu Linux, y genera la información del número de copia para cada uno de los genes así como indica cuáles regiones cromosómicas tienen una alteración. La ejecución de GISTIC para este experimento tarda en promedio entre una a dos horas, para procesar el algoritmo sobre el conjunto de datos de cáncer de seno del TCGA en una máquina virtual de VirtualBox que corre en una computadora MacBook con procesador i5.

Cuando se separaron los grupos para realizar las pruebas de GISTIC en el caso de TCGA se emplearon los datos de la vida clínica de los pacientes registrados para saber quiénes se encuentran vivos y cuáles se encuentran sin vida, y así se dividieron los grupos tanto como para aplicar GISTIC como para aplicar los clasificadores y las reglas de asociación.

Cuando se separaron los grupos para realizar las pruebas de GISTIC en el caso de CCLE se em-

plearon los datos de la resistencia a las drogas, en específico el parámetro IC50 donde si el valor de IC50 o sea la concentración cuando a la mitad de la actividad molecular es menor a la cuarta parte de la concentración total (en este caso la concentración total es 8uM).

GISTIC proporciona como salida varios archivos de texto junto algunas imágenes representativas de las amplificaciones y las eliminaciones encontradas. Entre los archivos de texto se encuentran un archivo `broad_significance_results.txt` que contiene las probabilidades de la prueba estadística de GISTIC que determina si hay una amplificación o una eliminación en un brazo cromosómico específico. Luego hay otro archivo importante que se llama `broad_value_by_arm.txt` en el cual se describe el valor de número de copia asignado a cada brazo cromosómico de cada una de las muestras disponibles.

GISTIC también solicita como entrada un archivo de marcadores con las posiciones de los puntos de interés del genoma pues esta información se emplea a la hora de mapear las regiones a un número de copia de ADN. Si no se tiene el archivo de marcadores hay dos opciones: no usar archivo de marcadores y GISTIC genera un archivo de seudomarcadores cada cierto número de bases en el genoma, o la otra opción es generar un archivo de marcadores a partir del archivo de segmentación de cada una de las muestras disponibles usando las posiciones del genoma que vienen en el archivo. La segunda opción es la usada para desarrollar estas pruebas en GISTIC.

4.2. Pruebas de clasificación

Se va a emplear las alteraciones estadísticamente más significativas como selección de características para luego aplicar un clasificador, donde entonces el espacio de entrada está definido por las alteraciones cromosómicas estadísticamente más significativas.

Para el TCGA se trata de evaluar un clasificador desde la perspectiva de la importancia de quién se toma la muestra, considerando si la muestra pertenece a un paciente vivo o uno fallecido y así se definen las etiquetas asignadas a cada muestra del conjunto de datos.

Para el CCLE se trata de evaluar un clasificador desde la perspectiva de la importancia de las líneas celulares tomando en consideración el criterio del IC50 para definir cuáles son sensibles o re-

sistentes, para ello primero se seleccionan los brazos cromosómicos estadísticamente más significativos que permitirían caracterizar las líneas celulares, en cada conjunto líneas con información de drogas disponibles. Se busca definir si cada línea celular es sensible o resistente a cada una de las drogas con más datos de muestras disponibles que en el caso del cáncer de seno son Topotecan, Lapatinib y Paclitaxel. Con la información anterior así se definirían las etiquetas de resistencia o sensibilidad de cada una de las drogas. En la tabla 4.1 se hace referencia a la cantidad de líneas celulares sensibles y resistentes disponibles a esas tres drogas, las demás drogas tienen muy pocas líneas o son todas resistentes ante esta.

| Droga | Sensible | Resistente |
|--------------|-----------------|-------------------|
| Paclitaxel | 21 | 8 |
| Lapatinib | 8 | 21 |
| Topotecan | 19 | 10 |

Cuadro 4.1: Tabla de drogas con el número de líneas celulares etiquetadas como sensibles y resistentes.

Los algoritmos empleados en la realización del trabajo están fundamentados en el trabajo teórico encontrado en las referencias bibliográficas de trabajos similares con datos genómicos donde se han aplicado las máquinas de soporte vectorial, donde en dichos artículos estos algoritmos son los que poseen los mejores resultados Costello et al. 2014. También otro tipo de algoritmo muy usado son los algoritmos de árboles (*random forests*), con un gran uso dentro de la literatura asociada.

Las implementaciones de los algoritmos se encuentran en el lenguaje de programación Python, en una importante librería llamada scikit-learn. Para sintonizar los algoritmos se emplea un método llamado GridSearchCV de la librería scikit-learn donde se realiza un barrido de los parámetros para obtener el mejor valor posible según una métrica específica, que en este caso se va a emplear el criterio de precisión de cada algoritmo de clasificación. De los valores de los parámetros a encontrar, en específico en la Máquina de Soporte Vectorial, se define el valor de la constante C de Costo de Error así como también se determina si usar un kernel lineal o un kernel gaussiano. En el caso del kernel gaussiano se tiene la necesidad de obtener un parámetro adicional llamado parámetro Gama, relacionado a la distribución gaussiana.

Tanto en el CCLE como en el TCGA se va a aplicar una validación cruzada de 10 divisiones, en la que se va a separar los datos con etiquetas disponibles en 10 grupos distribuidos aleatoriamente y se va

a analizar el comportamiento del clasificador. Esto se va a llevar a cabo por medio de los métodos de validación cruzada de *scikit-learn*, donde durante 10 ocasiones se van a dejar 9 grupos de los 10 para entrenar y 1 grupo diferente para probar el resultado del modelo. Con este método se puede encontrar el valor promedio de la métrica deseada de las pruebas. Por medio de estos métodos se realizaría una comparación entre los algoritmos de clasificación a partir de los mismos datos disponibles. Al final el resultado es saber cuál algoritmo proporciona mejores resultados.

Se entrenó un clasificador con los datos del TCGA para predecir los resultados del CCLE. Para realizar esta prueba se entrenó el clasificador con los datos del TCGA pero solo considerando las alteraciones que se encuentran en el CCLE, además de casi todas esas alteraciones se encuentran como alteraciones en TCGA, se procedió a predecir los valores de las etiquetas definidas en las líneas celulares por medio de los datos de las drogas.

4.3. Pruebas de reglas de asociación

Tomando en consideración las estimaciones del número de copia en cada brazo cromosómico se puede determinar las alteraciones que tiene cada muestra (TCGA) o línea (CCLE) con el objetivo de describir la frecuencia de ocurrencia y relación entre las distintas alteraciones. La idea es generar una lista con las alteraciones de cada muestra de modo que se pueda asociar dichas alteraciones con la sobrevida de los pacientes.

Las reglas de asociación de datos al ser una alternativa a los algoritmos de clasificación permiten definir relaciones difíciles de visualizar y que podrían describir de forma acertada las alteraciones de muestras con cáncer. Esto se puede lograr al aplicar una implementación del Algoritmo Apriori que permite encontrar las reglas de asociación de datos para las muestras de cáncer descritas por sus alteraciones cromosómicas.

Los resultados de las reglas de asociación de datos son obtenidos mediante la implementación del Algoritmo Apriori en una librería llamada *mlxtend* de Python. Las reglas de asociación de datos se obtienen de las regiones con alteraciones cromosómicas obtenidos como salida de la selección de cromosomas de GISTIC.

La idea es encontrar grupos delimitados de alteraciones que permitan establecer posibles relaciones entre los datos mediante las métricas de soporte y confianza. Los valores de soporte y confianza son definidos siempre al inicio donde el mínimo de confianza es 0.7 (o sea que la alteración se encuentre al menos en un 70% de las muestras).

Lo mismo para el valor de confianza entre pares o grupos de alteraciones donde se elige un valor de 0.7 como el valor mínimo de confianza. Los valores de dichos parámetros deseados se pueden disminuir de 0.7 con el fin de encontrar conjuntos no vacíos de solución. De esta sección los resultados serían tablas con las alteraciones de cada brazo cromosómico así como el número de soporte y con el valor de confianza superior o igual al mínimo establecido. En esta sección sería posible observar posibles coincidencias en las alteraciones entre el CCLE y el TCGA.

Al realizar un análisis de reglas de asociación de las alteraciones cromosómicas de cada uno de los posibles grupos (vivos o fallecidos) dado que con TCGA se tiene información acerca de los pacientes, sobre cuáles han logrado sobrevivir y cuales no, se va a generar tablas que resuman dichas alteraciones, que se pueden ir corroborando conforme exista nueva información disponible de nuevos pacientes.

Capítulo 5

Resultados y Análisis

La estructura de secciones del capítulo de Resultados sigue la estructura de etapas funcionales descritas en la *Metodología*:

- Pruebas de selección de GISTIC.
- Pruebas de clasificación.
- Pruebas de reglas de asociación.

La primera etapa concierne fundamentalmente a las pruebas realizadas mediante el programa GISTIC disponible en un archivo ejecutable listo para correr junto a los archivos de configuración de parámetros. El archivo debe correrse desde una computadora con Ubuntu Linux, el programa está escrito en una combinación de Matlab y C, debido a esto los desarrolladores compilaron el código mediante una herramienta de Matlab que permite crear un ejecutable exclusivo para la arquitectura sobre la que se compila el programa, en este caso Ubuntu Linux.

La segunda etapa versa sobre los resultados de los métodos de clasificación aplicados al espacio de características definidos por los valores de número de copia de ADN de las regiones con alteraciones cromosómicas obtenidos en la primer sección de *Pruebas de Selección de GISTIC*.

La tercer etapa muestra los resultados de las reglas de asociación de datos obtenidas por medio de la implementación en Python del Algoritmo Apriori en una librería llamada *mlxtextend*. Las reglas de asociación de datos emplean como entrada las regiones con alteraciones cromosómicas obtenidas en

la sección anterior de *Pruebas de Selección de GISTIC*.

5.1. Selección de GISTIC

Los resultados de alteraciones cromosómicas determinadas mediante GISTIC en TCGA, usando el archivo de marcadores, son bastante similares a los resultados publicados virtualmente en la plataforma web *FireBrowser* del *Broad Institute* (29 alteraciones en la corrida local contra 32 detectadas en los resultados publicados en la página) a pesar de que ahí se emplea el archivo de marcadores con el que se construyeron los archivos de segmentación. También GISTIC se aplicó a TCGA en una corrida donde no se usó el archivo de marcadores obteniendo resultados bastante similares en la detección de alteraciones de regiones cromosómicas (29 alteraciones con marcadores contra 31 sin marcadores). Aunque si se define un umbral de significancia estadística de 0.01 todos los casos tendrían el mismo número de alteraciones, lo cual es importante a la hora de considerar futuros experimentos, para definir el valor del umbral de la significancia.

Las figuras de la presente sección describen el trabajo llevado a cabo al ejecutar el método GISTIC en los conjuntos de datos del CCLE y el TCGA. En dichas figuras se describe en el eje vertical Y las posiciones en el genoma (según cada cromosoma) y en el eje horizontal X vienen los valores de la prueba estadística acumulativa Z empleada para encontrar las alteraciones.

Selección de imágenes

En las figuras 5.1 y 5.2 se puede apreciar todas las amplificaciones y las eliminaciones de todas las líneas celulares del CCLE, esa es la razón por la que se ve tan poblado cada uno de estos gráficos llenos de diversas alteraciones concernientes a distintos tipos de cáncer. En dichas imágenes el eje vertical representa la posición relativa de las alteraciones en cada cromosoma y el eje horizontal describe la puntuación Z acumulativa relacionada con la frecuencia de ocurrencia de amplificaciones estadísticamente significativas en todo el conjunto de datos del CCLE.

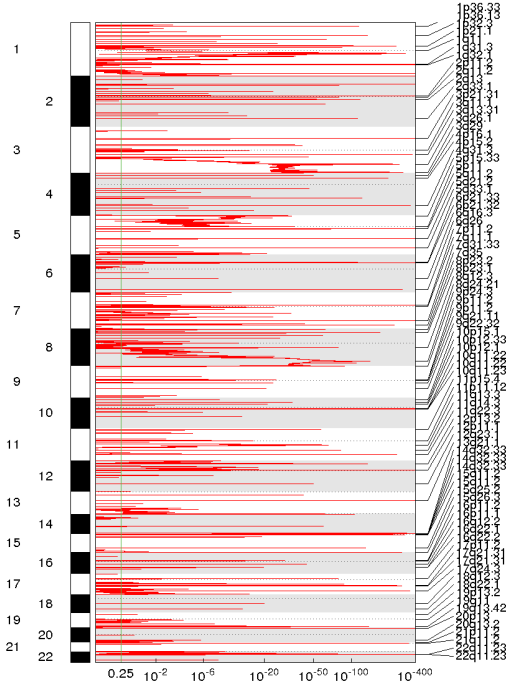


Figura 5.1: Regiones cromosómicas identificadas con amplificaciones de número de copias en el conjunto de datos CCLE.

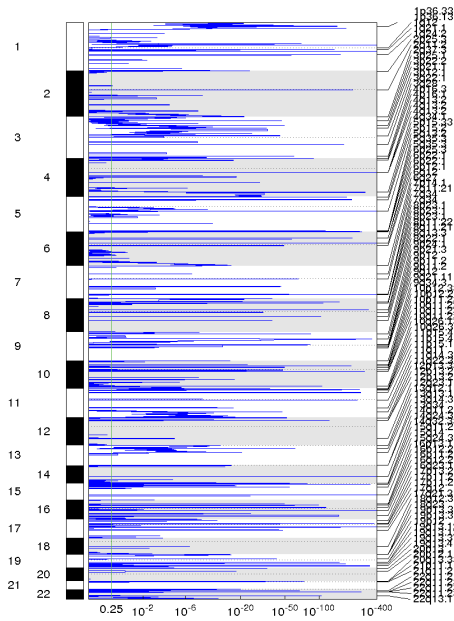


Figura 5.2: Regiones cromosómicas identificadas con eliminaciones de número de copias en el conjunto de datos del CCLE.

En las figuras 5.3 y 5.4 se pueden observar las alteraciones presentes en el conjunto de datos de cáncer de seno del TCGA, en el cual de manera gráfica se puede observar tanto alteraciones en regiones cromosómicas de brazos enteros así como alteraciones puntuales de regiones pequeñas.

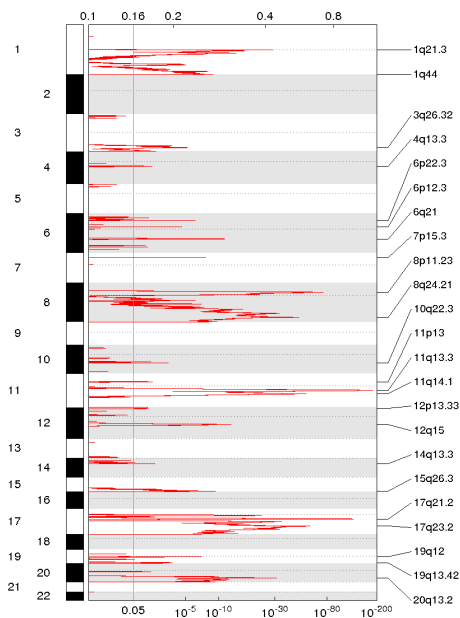


Figura 5.3: Regiones cromosómicas identificadas con amplificaciones de número de copias en el subconjunto de cáncer de mama del TCGA.

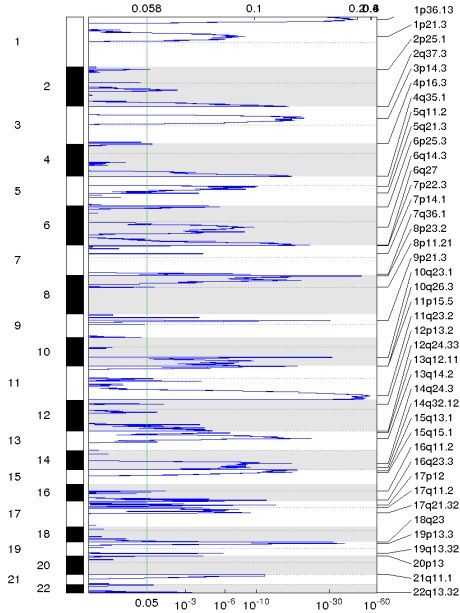


Figura 5.4: Regiones cromosómicas identificadas con eliminaciones de número de copia en el subconjunto de cáncer de mama del TCGA.

Las figuras 5.5 y 5.6 muestran las alteraciones presentes en el conjunto de datos de cáncer de seno del CCLE. Aquí se puede observar visualmente que hay menos alteraciones que en el conjunto de datos de cáncer de seno del CCLE, debido posiblemente a la mayor cantidad de muestras respecto a la cantidad de líneas celulares junto con una menor variedad de alteraciones.

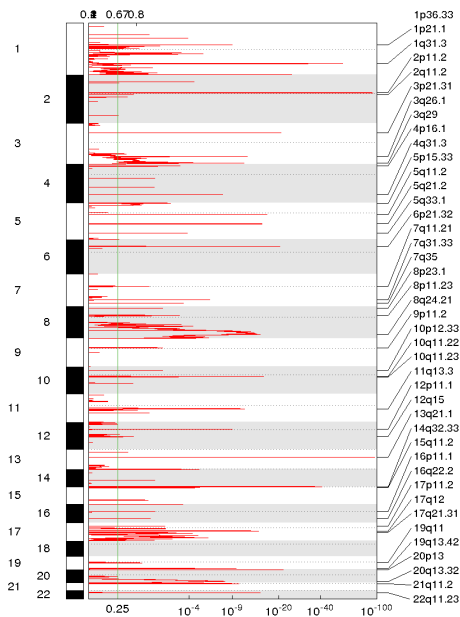


Figura 5.5: Regiones cromosómicas identificadas con amplificaciones de número de copias en el conjunto de cáncer de mama del CCLE.

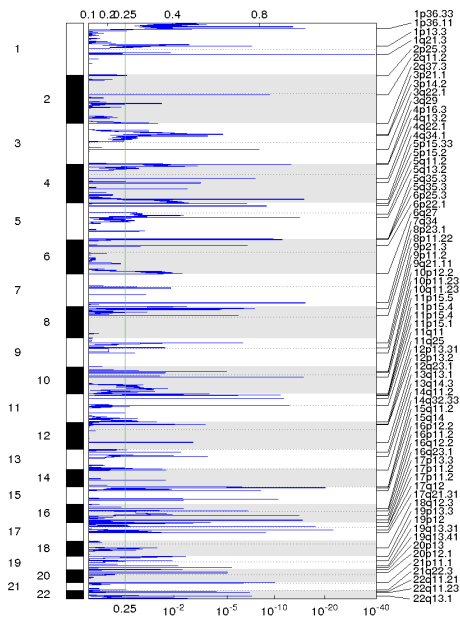


Figura 5.6: Regiones cromosómicas identificadas con eliminaciones de número de copia en el conjunto de cáncer de mama del CCLE.

Tablas de GISTIC

En esta subsección se presentan las tablas que resumen las principales alteraciones cromosómicas encontradas mediante GISTIC tanto en el TCGA como en el CCLE. Las imágenes son de gran valor ilustrativo pero difícilmente logran capturar de manera simple las principales alteraciones cromosómicas por ende es fundamental resumir estas relaciones en una tabla.

En la tabla 5.1 se tienen las alteraciones estadísticamente más significativas de las líneas celulares del CCLE para cáncer de seno. La columna **Brazo** hace referencia al brazo cromosómico respectivo, las columnas **Amplificaciones** y **Eliminaciones** se refieren respectivamente al número de ese tipo alteraciones presentes en dicho brazo cromosómico.

| Brazo | Amplificaciones | Eliminaciones |
|-------|-----------------|---------------|
| 7q | 25 | 12 |
| 8q | 25 | 20 |
| 20q | 39 | 6 |
| 8p | 6 | 37 |
| 11q | 12 | 22 |
| 13q | 7 | 30 |
| 15q | 6 | 32 |
| 17p | 8 | 32 |
| 17q | 12 | 23 |
| 18q | 9 | 33 |
| 9p | 9 | 32 |
| 21p | 13 | 37 |
| 22q | 11 | 28 |

Cuadro 5.1: Tabla con el número de alteraciones en los cromosomas de las líneas celulares del CCLE de cáncer de seno

En la tabla 5.2 se tiene las alteraciones estadísticamente significativas de las muestras del TCGA para cáncer de seno. La columna **Brazo** hace referencia al brazo cromosómico respectivo, las columnas **Amplificaciones** y **Eliminaciones** se refieren respectivamente al número de ese tipo alteraciones presentes en dicho brazo cromosómico. Nueve brazos presentan amplificaciones y diecinueve presentan eliminaciones, siendo el brazo 8p el que coincide en ambos casos, siendo estadísticamente significativo tanto como amplificación como eliminación. Las alteraciones que coinciden tanto en el TCGA como en el CCLE son las que se encuentran en las regiones 8p, 8q, 9p, 11q, 13q, 15q, 17p, 17q, 20q y 22q.

| Brazo | Amplificaciones | Eliminaciones |
|-------|-----------------|---------------|
| 1p | 232 | 233 |
| 1q | 746 | 34 |
| 4p | 71 | 385 |
| 4q | 69 | 348 |
| 5p | 321 | 135 |
| 6q | 136 | 312 |
| 7p | 314 | 122 |
| 8p | 295 | 389 |
| 8q | 579 | 77 |
| 9p | 144 | 339 |
| 9q | 120 | 314 |
| 10q | 104 | 266 |
| 11p | 124 | 304 |
| 11q | 106 | 393 |
| 13q | 91 | 454 |
| 14q | 135 | 302 |
| 15q | 80 | 336 |
| 16p | 416 | 171 |
| 16q | 131 | 602 |
| 17p | 68 | 607 |
| 17q | 224 | 260 |
| 18p | 155 | 332 |
| 18q | 135 | 345 |
| 20p | 390 | 108 |
| 20q | 455 | 50 |
| 22q | 84 | 492 |

Cuadro 5.2: Tabla con el número de alteraciones cromosómicas de las muestras del TCGA de cáncer de seno

5.2. Tablas de la clasificación

En esta sección se muestran las tablas con los resultados de distintos clasificadores aplicados al conjunto de datos de cáncer de seno del TCGA y del CCLE. En el caso del TCGA los clasificadores son entrenados con los datos de vida de los pacientes. Por medio de la técnica de validación de cruzada se puede obtener los valores promedio de la métrica **precisión**. Las diferencias de las métricas obtenidas en cada clasificador son bastante pequeñas como se puede apreciar en la tabla 5.3. En este caso se empleó una validación cruzada dividida en 10 grupos aleatorios (un grupo dedicado para entrenamiento y nueve dedicados para validación). Los parámetros de la máquina de soporte vectorial fueron seleccionados empleando un método llamado **GridSearch** del software **scikit-learn** que explora los parámetros posibles del clasificador y define los mejores parámetros de acuerdo a los resultados de una métrica que en este caso fue **precisión**. En este caso se empleó una máquina de soporte vectorial

de kernel radial (SVM rbf) con parámetros gama de 0.01 y la constante de penalización C de 0.01. También se empleó con el mismo resultado una máquina de soporte vectorial de kernel lineal con una constante de penalización C de 1 obteniendo el mismo resultado que el caso anterior. Los parámetros fueron encontrados empleando el método de **GridSearch**.

También se emplearon otros tres clasificadores: un *Linear Discriminant Analysis* (LDA), un *Random Forest* (RF) y un Árbol de Decisión (DT). En todos los casos se alcanza un valor casi idéntico de la métrica de precisión. El RF se emplea con el parámetro de profundidad de 80 encontrado mediante el método **GridSearch**.

| SVM rbf | SVM linear | LDA | RF | DT |
|----------|------------|----------|----------|----------|
| 0.904851 | 0.904851 | 0.902946 | 0.904851 | 0.906785 |

Cuadro 5.3: Tabla con los resultados de la clasificación de los datos usando validación cruzada

En el caso de los datos de cáncer del CCLE se entrenan los clasificadores con los datos de sensibilidad a las drogas disponibles, con lo que se puede dividir las líneas celulares en dos categorías. Aquí se da un problema que sucede también en el TCGA, pues existe un gran desbalance de clases, o sea existen considerablemente más elementos de una clase que otra lo cual produce resultados poco satisfactorios a la hora de emplear el clasificador con otros conjuntos de datos.

En el caso del paclitaxel el mejor resultado lo obtuvo (ver tabla 5.4) un clasificador de máquina de soporte vectorial con los parámetros seleccionados mediante **GridSearch** con parámetros gamma de 0.01 y la constante de penalización C de 25. También se empleó un clasificador de máquina de soporte vectorial lineal con constante C igual a 1 y un *Linear Discriminant Analysis* (LDA). Los clasificadores de *Random Forests* y el Árbol de Decisión no pudieron emplearse adecuadamente por los problemas de desbalance de clases que hay por la naturaleza intrínseca de los datos.

De manera análoga sucedió con el caso para la droga Topotecan (ver tabla 5.4) donde el mejor resultado se obtiene con el clasificador de máquina de soporte vectorial de kernel radial con los parámetros gamma de 0.01 y constante C de 50, y también con el caso de Lapatinib (ver tabla 5.4) en la cual el mejor resultado se obtuvo con la máquina de soporte vectorial con gamma de 0.01 y C de 0.01.

| Droga | SVM lineal | SVM radial | LDA |
|--------------|-------------------|-------------------|------------|
| Paclitxel | 0.4666 | 0.5555 | 0.4666 |
| Topotecan | 0.4444 | 0.6111 | 0.3555 |
| Lapatinib | 0.6619 | 0.7259 | 0.5222 |

Cuadro 5.4: Tabla con los resultados de la validación cruzada con los datos disponibles de sensibilidad a paclitaxel, topotecan y lapatinib

Se entrenó un clasificador con los datos del TCGA para predecir los resultados del CCLE. Para realizar esta prueba se entrenó el clasificador con los datos del TCGA pero solo considerando las alteraciones que se encuentran en el CCLE, además de casi todas esas alteraciones se encuentran como alteraciones en TCGA, se procedió a predecir los valores de las etiquetas definidas en las líneas celulares por medio de los datos de las drogas.

Los datos de las métricas muestran valores bajos donde el clasificador de máquina de soporte vectorial entrenados con el TCGA no es capaz de predecir los valores de las etiquetas del CCLE con un precision mayor a 0.7 para ninguna de las drogas con las cuales se probó dicho clasificador. En el caso de Lapatinib fue la droga donde se obtuvo las mejores métricas como ocurre en el caso de la validación cruzada para el CCLE.

Vale la pena recordar que las líneas celulares son modelos simples de células humanas de los cuales no se puede obtener los mismos resultados aunque a pesar de todo se obtienen resultados diferentes a cero, por lo cual puede ser que exista cierto potencial para definir el impacto de las drogas en líneas celulares.

| Droga | Paclitaxel | Topotecan | Lapatinib |
|------------------|-------------------|------------------|------------------|
| Precision | 0.166 | 0.277 | 0.666 |
| Recall | 0.375 | 0.5 | 0.571 |
| Accuracy | 0.31 | 0.374 | 0.482 |

Cuadro 5.5: Tabla con los resultados de emplear el clasificador entrenado con los datos del TCGA en los datos del CCLE

5.3. Reglas de asociación de datos

Al obtener los resultados de las pruebas de GISTIC se puede realizar un análisis de reglas de asociación de datos con las principales alteraciones presentes en los distintos brazos cromosomales de modo que se pueden crear pares de relaciones cromosómicas que suceden bajo ciertas métricas mínimas, con cierta frecuencia (soporte) y cierta confianza mínimas.

Vale la pena resaltar que las alteraciones cromosómicas pueden ser definidas cada una como si fuera una amplificación o como una eliminación, eso para aplicar las reglas de asociación sobre datos categorizados. GISTIC genera un archivo de salida que se llama `broad_value_by_arms.txt` en el cual se encuentra una tabla de datos separados por espacios con las alteraciones.

Las reglas de asociación de datos se podrían aplicar a los datos del TCGA y al CCLE, de forma general, es decir sin separar los conjuntos de datos, o separados específicamente según los datos de vida los pacientes o de sensibilidad en las líneas, por lo cual hay una subsección para cada uno de estos casos.

En las tablas con los valores de soporte (como la tabla 5.6) en la columna de **Alteraciones** se indica simultáneamente el brazo cromosómico y el tipo de alteración pues la palabra **del** indica una eliminación y la palabra **amp** indica una amplificación. En la columna **Soporte** vienen los valores de dicha métrica.

En las tablas con las reglas de asociación las columnas de **Antecedente** y de **Consecuente** se muestran los pares de eventos para los cuales ante una alteración antecedente ocurre simultáneamente una determinada alteración consecuente. En estas tablas las columnas **Soporte** (frecuencia de la ocurrencia) y de **Confianza** vienen los valores de dichas métricas.

Reglas de asociación de datos de TCGA generales

En la tabla 5.6 se pueden observar la frecuencia con la que aparecen las alteraciones en las muestras de cáncer de seno del TCGA. Las alteraciones estadísticamente más significativas son aquellas donde el valor de la prueba estadística Z de GISTIC es mayor al umbral mínimo definido en este caso como 0.05. Las alteraciones estadísticamente más significativas que hay en común con el CCLE son las amplificaciones en 8q y 20q, y las eliminaciones en 8p, 9p, 11q, 13q, 15q, 17p, 17q y 22q.

| Alteraciones | Soporte |
|--------------|---------|
| 11q del | 0.361 |
| 13q del | 0.417 |
| 15q del | 0.308 |
| 16p amp | 0.382 |
| 16q del | 0.552 |
| 17p del | 0.627 |
| 18p del | 0.305 |
| 18q del | 0.317 |
| 1q amp | 0.684 |
| 20p amp | 0.358 |
| 20q amp | 0.417 |
| 22q del | 0.451 |
| 4p del | 0.353 |
| 4q del | 0.319 |
| 8p del | 0.357 |
| 8q amp | 0.531 |
| 9p del | 0.311 |
| 1q amp | 0.395 |
| 8p del | 0.357 |
| 8q amp | 0.531 |
| 9p del | 0.311 |
| 1q amp | 0.684 |

Cuadro 5.6: Tabla con la métrica de soporte de las alteraciones cromosómicas de las muestras del TCGA de cáncer de seno

En la tabla 5.7 se tienen las reglas de asociación de las alteraciones cromosómicas presentes en las muestras de cáncer de seno de TCGA. Por ejemplo la amplificación de 20p ocurre en el 32.8% de todas las muestras, y si la amplificación 20p sucede, simultáneamente ocurre una amplificación en el 20q con una frecuencia del 91.8%.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 16q del | 1q amp | 0.395 | 0.716 |
| 20q amp | 1q amp | 0.311 | 0.745 |
| 20p amp | 20q amp | 0.328 | 0.918 |
| 20q amp | 20p amp | 0.328 | 0.787 |
| 20q amp | 8q amp | 0.300 | 0.719 |

Cuadro 5.7: Tabla con las reglas de asociación de las alteraciones cromosómicas de las muestras de TCGA de cáncer de seno

Reglas de asociación de datos de TCGA etiquetados

En la tabla 5.8 se pueden observar la frecuencia con la que aparecen las alteraciones en las muestras de pacientes con cáncer de seno reportados muertos en el conjunto de datos del TCGA. En este conjunto las alteraciones más frecuentes son las amplificaciones en el 8q (0.63) y en el 1q (0.59), así como la eliminación del 17p (0.62).

| Alteraciones | Soporte |
|--------------|---------|
| 13q del | 0.500 |
| 14q del | 0.400 |
| 15q del | 0.370 |
| 16p amp | 0.390 |
| 16q del | 0.470 |
| 17p del | 0.620 |
| 18p del | 0.430 |
| 18q del | 0.410 |
| 1q amp | 0.590 |
| 20p amp | 0.400 |
| 20q amp | 0.440 |
| 22q del | 0.490 |
| 4p del | 0.390 |
| 4q del | 0.360 |
| 8p del | 0.480 |
| 8q amp | 0.630 |
| 9p del | 0.350 |

Cuadro 5.8: Tabla con la métrica de soporte de las alteraciones cromosómicas de las muestras del TCGA de los pacientes muertos de cáncer de seno

En la tabla 5.9 se tienen las reglas de asociación de las alteraciones cromosómicas de las muestras de TCGA de los pacientes de cáncer de seno sin vida. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.7 con el fin de dar una mayor confiabilidad a la existencia de ambas relaciones.

| Antecedente | Consecuente | Suporte | Confianza |
|-------------|-------------|---------|-----------|
| 13q del | 17p del | 0.350 | 0.700 |
| 13q del | 1q amp | 0.350 | 0.700 |
| 18q del | 18p del | 0.360 | 0.878 |
| 18p del | 18q del | 0.360 | 0.837 |
| 20p amp | 20q amp | 0.360 | 0.900 |
| 20q amp | 20p amp | 0.360 | 0.818 |

Cuadro 5.9: Tabla con las reglas de asociación de las alteraciones cromosómicas de las muestras de TCGA de los pacientes muertos de cáncer de seno

En la tabla 5.10 se pueden observar la métrica de soporte de las alteraciones cromosómicas de las muestras del TCGA de los pacientes vivos de cáncer de seno. En este conjunto las alteraciones más frecuentes son las amplificaciones en el 1q (0.689) y en el 8q (0.523), así como la eliminación en el 17p (0.553) y en el 16q (0.55). Las alteraciones se filtran a partir del valor de 0.35 de soporte.

| Alteraciones | Soporte |
|--------------|---------|
| 11q del | 0.370 |
| 13q del | 0.406 |
| 16p amp | 0.380 |
| 16q del | 0.555 |
| 17p del | 0.553 |
| 1q amp | 0.689 |
| 20p amp | 0.353 |
| 20q amp | 0.415 |
| 22q del | 0.446 |
| 4p del | 0.352 |
| 8q amp | 0.523 |

Cuadro 5.10: Tabla con la métrica de soporte de las alteraciones cromosómicas de las muestras del TCGA de los pacientes vivos de cáncer de seno

En la tabla 5.11 se tiene las reglas de asociación de las alteraciones cromosómicas presentes en las muestras del TCGA de los pacientes vivos de cáncer de seno. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.5 ya que solo una de las relaciones logró alcanzar más de 0,7 de confianza.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 16q del | 1q amp | 0.405 | 0.729 |
| 1q amp | 16q del | 0.405 | 0.588 |
| 1q amp | 17p del | 0.369 | 0.536 |
| 17p del | 1q amp | 0.369 | 0.667 |
| 8q amp | 1q amp | 0.364 | 0.696 |
| 1q amp | 8q amp | 0.364 | 0.528 |

Cuadro 5.11: Tabla con las reglas de asociación de las alteraciones cromosómicas de las muestras de TCGA de los pacientes vivos de cáncer de seno

Reglas de asociación de datos de CCLE generales

En la tabla 5.12 se pueden observar los valores de la métrica de soporte de las alteraciones en las líneas celulares del CCLE de cáncer de seno.

| Alteraciones | Soporte |
|--------------|---------|
| 13q del | 0.508 |
| 15q del | 0.542 |
| 17p del | 0.542 |
| 18q del | 0.559 |
| 20q amp | 0.661 |
| 21p del | 0.627 |
| 22q del | 0.475 |
| 8p del | 0.627 |
| 8q amp | 0.424 |
| 9p del | 0.542 |

Cuadro 5.12: Tabla con la métrica de soporte de las alteraciones en los cromosomas de las líneas celulares del CCLE de cáncer de seno

En la tabla 5.13 se tienen las reglas de asociación de las alteraciones cromosómicas de las líneas celulares del CCLE de cáncer de seno. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.7 con el fin de dar confiabilidad a la existencia de ambas relaciones.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 13q del | 20q amp | 0.407 | 0.800 |
| 15q del | 20q amp | 0.407 | 0.750 |
| 15q del | 8p del | 0.407 | 0.750 |
| 17p del | 21p del | 0.407 | 0.750 |
| 17p del | 8p del | 0.424 | 0.781 |
| 18q del | 20q amp | 0.424 | 0.758 |
| 18q del | 21p del | 0.407 | 0.727 |
| 18q del | 8p del | 0.441 | 0.788 |
| 8p del | 18q del | 0.441 | 0.703 |
| 18q del | 9p del | 0.407 | 0.727 |
| 9p del | 18q del | 0.407 | 0.750 |
| 21p del | 20q amp | 0.441 | 0.703 |
| 8p del | 20q amp | 0.441 | 0.703 |
| 9p del | 20q amp | 0.441 | 0.812 |
| 21p del | 8p del | 0.458 | 0.730 |
| 8p del | 21p del | 0.458 | 0.730 |
| 8p del | 9p del | 0.441 | 0.703 |
| 9p del | 8p del | 0.441 | 0.812 |

Cuadro 5.13: Tabla de reglas de asociación de las alteraciones cromosómicas de las líneas celulares del CCLE de cáncer de seno

Reglas de asociación de datos de CCLE etiquetados

En la tabla 5.14 se pueden observar la frecuencia con la que aparecen las alteraciones en las líneas (con datos de quimiosensibilidad disponibles) que se pueden definir como resistentes a paclitaxel en el CCLE. En este conjunto las alteraciones más frecuentes son las amplificaciones en el 20q (0.619), así como la eliminación en el 17p (0.619) y en el 8p (0.714). Las alteraciones se filtran a partir del valor de 0.5 de soporte.

| Alteraciones | Soporte |
|--------------|---------|
| 15q del | 0.571 |
| 17p del | 0.619 |
| 18q del | 0.524 |
| 20q amp | 0.619 |
| 21p del | 0.762 |
| 22q del | 0.524 |
| 8p del | 0.714 |
| 9p del | 0.571 |

Cuadro 5.14: Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a paclitaxel

En la tabla 5.15 se tienen las reglas de asociación de las alteraciones cromosómicas presentes en

las líneas celulares de CCLE resistentes a paclitaxel. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.7 con el fin de tener más certeza de la existencia de ambas relaciones. La relación con mayor confianza está compuesta por la eliminación de 9p y la eliminación de 8p con un 0.917.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 17p del | 21p del | 0.524 | 0.846 |
| 20q amp | 21 pdel | 0.524 | 0.846 |
| 21p del | 8p del | 0.571 | 0.750 |
| 8p del | 21p del | 0.571 | 0.800 |
| 9p del | 8p del | 0.524 | 0.917 |
| 8p del | 9p del | 0.524 | 0.733 |

Cuadro 5.15: Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a paclitaxel

En la tabla 5.16 se pueden observar la frecuencia con la que aparecen las alteraciones en las líneas (con datos de quimiosensibilidad disponibles) que se pueden definir como sensibles a paclitaxel en el CCLE. En este conjunto las alteraciones más frecuentes son las amplificaciones en el 20q (1.0), así como la eliminación en el 9p (0.875). Las alteraciones se filtran a partir del valor de 0.6 de soporte.

| Alteraciones | Soporte |
|--------------|---------|
| 15q del | 0.625 |
| 17p del | 0.625 |
| 18q del | 0.875 |
| 20q amp | 1.000 |
| 8p del | 0.625 |
| 8q amp | 0.625 |
| 9p del | 0.875. |

Cuadro 5.16: Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a paclitaxel

En la tabla 5.17 se tienen las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a paclitaxel. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.7 con el fin de dar confiabilidad a la existencia de ambas relaciones. Hay varias relaciones con una confianza de 1.0, esto se debe a las pocas muestras sensibles que existen para este caso.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 15q del | 20q amp | 0.625 | 1.000 |
| 17p del | 20q amp | 0.625 | 1.000 |
| 18q del | 20q amp | 0.875 | 1.000 |
| 20q amp | 18q del | 0.875 | 0.875 |
| 18q del | 8p del | 0.625 | 0.714 |
| 8p del | 18q del | 0.625 | 1.000 |
| 18q del | 9p del | 0.750 | 0.857 |
| 9p del | 18q del | 0.750 | 0.857 |
| 8p del | 20q amp | 0.625 | 1.000 |
| 8q amp | 20q amp | 0.625 | 1.000 |
| 9p del | 20q amp | 0.875 | 1.000 |
| 20q amp | 9p del | 0.875 | 0.875 |

Cuadro 5.17: Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a paclitaxel

En la tabla 5.18 se pueden observar la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a topotecan. En este conjunto las alteraciones más frecuentes son las amplificaciones en el 20q (0.619), así como la eliminación en el 17p (0.619) y en el 8p (0.714). Las alteraciones se filtran a partir del valor de 0.5 de soporte.

| Alteración | Soporte |
|------------|---------|
| 15q del | 0.579 |
| 17p del | 0.632 |
| 17q del | 0.526 |
| 18q del | 0.684 |
| 20q amp | 0.789 |
| 21p del | 0.579 |
| 8p del | 0.684 |
| 8q del | 0.526 |
| 9p del | 0.737 |

Cuadro 5.18: Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a topotecan

En la tabla 5.19 se tienen las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a topotecan. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.7. En este caso la relación con mayor confianza se da entre la eliminación de 18q y la amplificación de 20q con un 0.923.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 18q del | 20q amp | 0.632 | 0.923 |
| 20q amp | 18q del | 0.632 | 0.800 |
| 18q del | 9p del | 0.579 | 0.846 |
| 9p del | 18q del | 0.579 | 0.786 |
| 9p del | 20q amp | 0.632 | 0.857 |
| 20q amp | 9p del | 0.632 | 0.800 |
| 9p del | 8p del | 0.579 | 0.786 |
| 8p del | 9p del | 0.579 | 0.846 |

Cuadro 5.19: Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a topotecan

En la tabla 5.20 se pueden observar la frecuencia con la que aparecen las alteraciones en las líneas (con datos de quimiosensibilidad disponibles) que se pueden definir como sensibles a topotecan en el CCLE. En este conjunto las alteraciones más frecuentes son las amplificaciones en el 20q (0.619), así como la eliminación en el 17p (0.619) y en el 8p (0.714). Las alteraciones se filtran a partir del valor de 0.5 de soporte.

| Alteración | Soporte |
|------------|---------|
| 13q del | 0.700 |
| 15q del | 0.600 |
| 17p del | 0.600 |
| 20q amp | 0.600 |
| 21p del | 0.800 |
| 8p del | 0.700 |
| 8q amp | 0.600 |

Cuadro 5.20: Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a topotecan

En la tabla 5.21 se tienen las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a topotecan. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.7. En este caso la relación con mayor confianza se da entre la eliminación de 17p y la eliminación de 21p con un 1.0.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 21p del | 17p del | 0.600 | 0.750 |
| 17p del | 21p del | 0.600 | 1.000 |
| 21p del | 8p del | 0.600 | 0.750 |
| 8p del | 21p del | 0.600 | 0.857 |

Cuadro 5.21: Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a topotecan

En la tabla 5.22 se pueden observar la frecuencia con la que aparecen las alteraciones en las líneas (con datos de quimiosensibilidad disponibles) que se pueden definir como resistentes a lapatinib en el CCLE. En este conjunto las alteraciones más frecuentes es la eliminación de 8p (0.875) y la amplificación de 20q (0.75). Las alteraciones se filtran a partir del valor de 0.5 de soporte.

| Alteración | Soporte |
|------------|---------|
| 13q del | 0.500 |
| 15q del | 0.500 |
| 17p del | 0.625 |
| 18q del | 0.750 |
| 20q amp | 0.750 |
| 21p del | 0.750 |
| 22q del | 0.500 |
| 8p del | 0.875 |
| 8q del | 0.625 |
| 9p del | 0.750 |

Cuadro 5.22: Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a lapatinib

En la tabla 5.23 se tienen las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a lapatinib. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0,7. La confianza es en casi todos los casos igual a 1,0 posiblemente debido a las pocas muestras sensibles disponibles.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 13q del | 17p del | 0.500 | 1.000 |
| 13q del | 21p del | 0.500 | 1.000 |
| 13q del | 8p del | 0.500 | 1.000 |
| 13q del | 9p del | 0.500 | 1.000 |
| 15q del | 8p del | 0.500 | 1.000 |
| 17p del | 9p del | 0.625 | 1.000 |
| 21p del | 8p del | 0.750 | 1.000 |
| 8p del | 21p del | 0.750 | 0.857 |
| 22q del | 8p del | 0.500 | 1.000 |
| 8q del | 8p del | 0.625 | 1.000 |

Cuadro 5.23: Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE resistentes a lapatinib

En la tabla 5.22 se pueden observar la frecuencia con la que aparecen las alteraciones en las líneas (con datos de quimiosensibilidad disponibles) que se pueden definir como resistentes a paclitaxel en el CCLE. En este conjunto las alteraciones más frecuentes son las amplificaciones en el 20q (0.714), así como la eliminación en el 17p (0.619) y en el 8p (0.619). Las alteraciones se filtran a partir del valor de 0.5 de soporte.

| Alteración | Soporte |
|------------|---------|
| 15q del | 0.619 |
| 17p del | 0.619 |
| 17q del | 0.524 |
| 18q del | 0.571 |
| 20q amp | 0.714 |
| 21p del | 0.619 |
| 8p del | 0.619 |
| 8q amp | 0.571 |
| 9p del | 0.619 |

Cuadro 5.24: Tabla con la métrica de soporte de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a lapatinib

En la tabla 5.25 se tienen las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a lapatinib. En esta tabla se han filtrado las relaciones entre alteraciones con el fin que estas tengan una confianza superior a 0.7. En este caso la relación con mayor confianza se da entre la eliminación de 18q y la amplificación de 20q con un 0.917.

| Antecedente | Consecuente | Soporte | Confianza |
|-------------|-------------|---------|-----------|
| 20q amp | 18q del | 0.524 | 0.733 |
| 18q del | 20q amp | 0.524 | 0.917 |
| 20q amp | 9p del | 0.524 | 0.733 |
| 9p del | 20q amp | 0.524 | 0.846 |

Cuadro 5.25: Tabla con las reglas de asociación de las alteraciones cromosómicas de las líneas celulares de CCLE sensibles a lapatinib

Capítulo 6

Conclusiones y Recomendaciones

6.1. Conclusiones

Existen alteraciones cromosómicas comunes entre las muestras de cáncer real y las líneas celulares lo cual arroja importantes indicios sobre la similaridad en ciertos patrones de alteraciones genómicas (8 alteraciones coincidentes, en el TCGA hay 29 y en el CCLE hay 10) esto posiblemente porque las líneas son un modelo simple del cáncer y no tienen la diversidad de alteraciones cromosómicas presentes en el cáncer real.

Los resultados de los clasificadores al aplicar el método de validación cruzada cumple con la meta planteada en los objetivos (acerca de obtener la métrica de precisión superior a 0,7) pero el desbalance de clases presentes tanto en el CCLE como en el TCGA, dificulta que se pueda crear un modelo generalizable al conjunto de datos sin etiquetas, aunque puedan existir métodos para aumentar sintéticamente el conjunto de datos, no siempre se pueden obtener resultados satisfactorios. Se probaron los modelos de clasificación sobre los datos sin etiquetas y todos las muestras se saturan a la clase con más muestras.

Además los modelos de clasificación obtenidos del TCGA no sirvieron para generalizar algún resultado en el CCLE ni viceversa tampoco, pues al aplicarse el clasificador en el otro conjunto de datos todas las muestras se saturan a la clase con más muestras. Aunque también posiblemente las muestras del TCGA y el CCLE tienen problemas asociados a la inseparabilidad del conjunto de datos y la insuficiente cantidad de muestras para crear un modelo más generalizable sobretodo en el caso del CCLE que es un conjunto de datos con información tan escasa (solo 29 muestras con algún tipo de etiqueta

sobre la información de resistencia mientras hay 59 líneas celulares en total).

Las reglas de asociación de datos en este trabajo han mostrado ser una alternativa a los algoritmos de clasificación pues proporcionan una descripción analítica de los datos más allá de simplemente un modelo predictivo donde es posible establecer relaciones entre las muestras caracterizadas por las alteraciones cromosómicas. Las reglas de asociación han encontrado muchos casos de pares de alteraciones con valores de métrica de confianza superiores a 0.7, donde dichas reglas podrían emplearse para predecir la resistencia o mortalidad latente de un paciente.

Con las reglas de asociación de datos se han podido obtener descripciones a partir de cada uno de los conjuntos de datos aunque las generalizaciones son complicadas de realizar entre los distintos conjuntos de datos porque existen diferentes alteraciones cromosómicas tanto en el TCGA como en el CCLE aunado al hecho que el TCGA tiene más muestras y la complejidad molecular es mayor. Los casos de los pares de alteraciones que tienen una frecuencia alta (reflejado en la métrica de soporte) y que tienen una confianza alta requieren de un análisis más a profundidad. Por ejemplo en el caso de los pacientes fallecidos del TCGA se presentan amplificaciones en los pares cromosómicos de los brazos cromosómicos 20p y 20q donde dichas alteraciones pueden ser valiosas para la comprensión de los fenómenos de resistencia a los tratamientos.

Tener descrito un conjunto de alteraciones cromosómicas tanto para el TCGA y el CCLE así como para cada uno de sus subgrupos (resistentes o sensibles) es vital como trabajo intermedio para caracterizar las vías moleculares vinculadas a cada uno de los casos, así como para definir mejores estrategias de diagnóstico como se realiza en ciertos tipos de cáncer actualmente.

6.2. Recomendaciones

Es necesario llevar a cabo un análisis conjunto de las alteraciones cromosómicas con la expresión génica para crear un modelo estadístico que identifique los patrones de expresión de los genes pertenecientes a las regiones con alteraciones en el número de copia de ADN. Se podrían emplear reglas de asociación de datos así como pruebas de análisis de correlación que permitan establecer relaciones que no se pueden percibir a simple vista y seleccionar los genes que pueden estar vinculados a redes de compensación y a la resistencia a tratamientos. La idea es realizar, en conjunto con la identificación de genes, un análisis de vías moleculares en la cual se van a agrupar los genes alterados según su utilidad molecular para en una última instancia se pueda sugerir marcadores moleculares sujetos a posibles investigaciones de tratamientos en donde se analicen las vías moleculares interesantes a intervenir.

Para trabajos futuros sería apropiado considerar no solamente las alteraciones de brazos cromosómicos sino también las alteraciones de regiones puntuales detectadas también por GISTIC y que podrían poseer una gran importancia en los análisis de expresión y de vías moleculares.

También otro detalle a consideración para el trabajo futuro es que en TCGA los pacientes que se encuentran vivos algunos son más recientes que otros y dado que 5 años es el período en el que a un paciente se le considera libre de recaída, podría existir una división en la que se considera los pacientes vivos en espera y los pacientes vivos que terminaron el período de riesgo de recaída. Otro punto de mejora para el futuro es analizar si es necesario considerar las reglas de asociación de datos para más de dos alteraciones cromosómicas simultáneamente, aunque su probabilidad de ocurrencia sea bastante menor.

Las muestras de cáncer como se muestran en los resultados de las tablas de las reglas de asociación de datos tienen posibles patrones de alteraciones cromosómicas que se repiten tanto en los grupos de pacientes en condición con vida como los sin vida. Es importante resaltar que dichos patrones podrían hipotéticamente representar cariotipos que proporcionan la posibilidad de sobrevivir al cáncer que en otras condiciones sería inviable su existencia en los procesos de división celular.

En caso de que existan dichos cariotipos alternativos, estos podrían representar posibles estadios del cáncer, donde sería de gran interés estudiar a mayor profundidad cuáles de los genes de esas regiones

se encuentran sobreexpresados y subexpresados de modo que se pueda estudiar posibles mecanismos de la regulación de la expresión génica.

Al hacer el emparejamiento de cuáles genes tienen alteraciones en su número de copia con cuáles genes tienen variaciones en su expresión se puede analizar patrones de compensación asociados a ciertos grupos de genes. Luego valdría la pena agrupar esos genes por las funciones que representan.

Al hacer un análisis funcional de los genes se podría obtener cuáles son los mecanismos a los cuales dichos genes hacen referencia y se podría tener posibles hipótesis de cuáles son los caminos moleculares que facilitan a las células con cáncer existir a pesar de tener cariotipos en un principio completamente irregulares e inconsistentes con la vida misma.

Al agrupar funcionalmente los genes expresados diferencialmente se permitiría establecer un marco de posibilidades en el cual se pueda determinar cuáles son los genes que tienen un rol en la supervivencia del cáncer y en procesos como la muerte celular.

Eventualmente cuando se tengan estudios más a profundidad a partir de los datos, se podría llegar a hacer experimentos en laboratorio con el objetivo de corroborar los mecanismos de acción por los cuales el cáncer sobrevive considerando los genes antes seleccionados, al modificar de manera sintética los genes mediante una técnica de manipulación avanzada como CRISPR para desactivar ciertos genes.

El cáncer siempre se ha observado como un fenómeno desde la óptica de la genética pero quizás también deba observarse desde la óptica de la evolución donde existe en la naturaleza mecanismos semejantes donde algunos seres vivos evolucionan a través de la variación de su cariotipo.

Para finalizar vale la pena recalcar la idea que inspiró este trabajo: autores como Peter Duesberg sugieren con vehemencia que la aneuploidía sea vista como el motor del crecimiento y la supervivencia del cáncer y han propuesto una alternativa a la teoría mutacional del cáncer (Duesberg et al. 2006; Duesberg et al. 2007). Aunque es un fenómeno poco estudiado, en caso de ser cierta dicha afirmación, es fundamental tener descrito las relaciones cromosómicas con el objetivo final de seguir explorando los posibles mecanismos encargados de producir dichas aberraciones masivas que ocurren en nuestros

genomas.

Capítulo 7

Bibliografía

- Arakawa, N. et al. (2017). «Genome-wide analysis of DNA copy number alterations in early and advanced gastric cancers». En: *Molecular Carcinogenesis*, Wiley.
- Bailey, P. et al. (2016). «Genomic analyses identify molecular subtypes of pancreatic cancer». En: *Nature*.
- Barretina, J. (2012). «The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity». En: *Nature*.
- Beroukhi, R (2007). «Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma». En: *PNAS*.
- Bruna, A. et al. (2016). «A Biobank of Breast Cancer Explants with preserved intra-tumor heterogeneity to Screen Anticancer compounds». En: *Cell*.
- Carter, L. et al. (2017). «Molecular analysis of circulating tumor cells identifies distinct copy-number profiles in patients with chemosensitive and chemorefractory small-cell lung cancer». En: *Nature Medicine*.
- Cheung, H. et al. (2011). «Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer». En: *Proceedings of National Academy of Science (PNAS)*.
- Chin, L., J.N. Andersen y P.A. Futreal (2011). «Cancer genomics: from discovery science to personalized medicine.» En: *Nature Medicine* 17.3, pp. 297-303.
- Cho, YJ. (2011). «Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome». En: *Clinical Oncology*.

- Cimino, P.J. (2016). «Multidimensional scaling of diffuse gliomas: application to the 2016 World Health Classification system with prognostically relevant molecular subtype discovery». En: *Acta Neuropathologica Communications*.
- Collison, EA. et al. (2013). «Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy». En: *Nature Medicine*.
- Costello, J. et al. (2014). «A community effort to assess and improve drug sensitivity prediction algorithms». En: *Nature Biotechnology*.
- Coto, J. y F. Siles (2017). «High dimensional genomic data hybrid modeling to pattern recognition on cancer chemosensitivity». Tesis de maestría. Computer Science.
- Coto, J, F. Siles y R. Mora (2016). «A Biocomputational Platform as a First Step to Personalized Therapy in Costa Rica». En: *IEEE CONCAPAN BIP 2016*.
- Curtis, C. (2012). «The genomic and transcriptomic architecture of 2000 breast tumors reveals novel subgroups». En: *Nature*.
- Daemen, A. et al. (2013). «Modeling precision treatment of breast cancer». En: *Genome Biology*.
- Depuydt, P. et al. (2018). «Genomic Amplifications and Distal 6q Loss: Novel Markers for Poor Survival in High Risks Neuroblastoma Patients». En: *Journal National Cancer Institute*.
- Duda, H., DG. Stork y PE. Hart (2000). *Pattern Classification*. Wiley.
- Duesberg, P. et al. (2006). «Aneuploidy and Cancer: From Correlation to Causation.» En: *Infection and Inflammation: Impacts on Oncogenesis. Contributions to Microbiology*. 13, pp. 16-44.
- (2007). «Cancer drug resistance: the central role of the karyotype». En: *Drug Resistances Updates. Elsevier*. 10.51-58.
- Etemadmoghadam, D. (2009). «Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas». En: *Clinical Cancer Research*.
- Goodspeed, A. et al. (2016). «Tumor derived cell lines as molecular models of cancer pharmacogenomics se describe dicho algoritmo». En: *Molecular Cancer Research*.
- Hoadley, KA. et al. (2014). «Multiplatform analysis of 12 canncer types reveals molecular classification within and across tissues of origin». En: *Cell*.
- Hughey, JJ. y AJ. Butte (2015). «Robust meta analysis of gene expression using the elastic net». En: *Nucleic Acids Research*.
- Lee, H. (2015). «The Cancer Genome Atlas Explorer». En: *Genome Biology*.

- Liu, Z., X. Zhang y S. Zhang (2014). «Breast tumor subgroups reveal diverse clinical prognostic power». En: *Nature Scientific Reports*.
- NCI (2015). *What Is Cancer? by National Cancer Institute*. URL: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>.
- Ni, X. y M. Zhuo (2013). «Reproducible copy number variation patterns among circulating tumor cells of lung cancer patients». En: *PNAS*.
- Nicholson, J. y D. Cimini (2013). «Cancer Karyotypes: Survival of the Fittest.» En: *Frontiers in Oncology*.
- Qiu, Z., Z. Bi y K. Song (2017). «Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer». En: *Genes Chromosomes Cancer*.
- Ritchie, MD. et al. (2015). «Methods of integrating data to uncover genotype-phenotype interactions». En: *Nature Reviews Genetics*.
- Ruiz, V. (2017). «Molecular subtyping of tumors from patients with familial glioma». En: *Neuro-oncology, Oxford University Press*.
- Sanchez-Garcia, F., P. Villagrasa y J. Matsui (2014). «Integration of genomic data enables selective discovery of breast cancer drivers». En: *Cell*.
- Taskesen, E., S. Babaei et al. (2015). «Integration of gene expression and DNA-methylation profiles improves molecular subtype». En: *IAPR conference on Pattern Recognition in Bioinformatics*.
- Taskesen, E., S. Huisman et al. (2016). «Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics». En: *Nature Scientific Reports*.
- Wang, Y., J. Fang y S. Chen (2016). «Inferences of drugs responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties». En: *Nature Scientific Reports*.
- Weinberg, R. (2014). *The Biology of Cancer*. Garland Science, Taylor y Francis Group, LLC.
- Witten, I. y E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Second. Morgan Kaufmann, Elsevier.
- Xu, J. et al. (2018). «A Novel Method to Detect Early Colorectal Cancer Based on Chromosome Copy Number Variation in Plasma». En: *Cellular Physiology and Biochemistry*.
- Yuan, X. et al. (2012). «Comparative Analysis of Methods for Identifying Recurrent Copy Number Alterations in Cancer». En: *PLoS ONE*.
- Zaman, N. (2013). «Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-Specific Drug Targets». En: *Cell Reports*.

Zhang, A. et al. (2015). «Predicting anticancer drug responses using a dual-layer integrated cell line».

En: *Computational Biology*.

Zhang, N. et al. (2016). «Classification of cancers based on copy number variation landscapes». En:

Biochimica et Biophysica Acta, Elsevier.

Apéndices

Apéndice A

Referencia de artículo

Genome Copy Number Feature Selection Based on Chromosomal Regions Alterations and Chemosensitivity Subtypes

J. Vargas, Rodrigo Mora-Rodríguez and F. Siles

Pattern Recognition and Intelligent Systems Laboratory

Department of Electrical Engineering, School of Engineering

Universidad de Costa Rica, Sna José, Costa Rica

josue.vargasamador@ucr.ac.cr, rodrigo.mora@ucr.ac.cr, francisco.siles@ucr.ac.cr

Abstract—Cancer disease causes millions of deaths throughout the world and thousands in Costa Rica, and cancer treatment causes an immense economic burden on the social security system because drugs against the disease are extremely expensive. Through DNA sequencing techniques, the copy number of each gene could be found in order to describe how many times a gene is repeated within chromosomes. This type of data is of great importance since the cancer manifests a phenomenon called aneuploidy that consists of alterations in the number of chromosomes copies. Theories about the importance of aneuploidy in cancer supposed this phenomenon is a evolutionary engine that allows the disease to grow and resist changes produced by the organism and the treatments applied to these tissues. Studies have collected data on the number of copies in well-known databases such as CCLE and TCGA, which can be used to analyze the relationship between alterations in DNA and resistance to chemotherapies. In this study as a contribution, it was proposed to use statistical pattern recognition methods in order to identify chromosomal regions of DNA related to cancer chemosensitivity subtypes (resistant and sensitive subgroups) and then use these regions for CCLE cell lines labeling and TCGA samples classification.

Index Terms—cancer; chemotherapy; chromosomes; copy number; CCLE; TCGA

I. INTRODUCTION

Cancer around the world is responsible for 8 million deaths with an incidence of 14 million. In 2008, 12.7 million people were diagnosed with the disease, and in the same year, 24.6 million people were living with cancer. Currently in Costa Rica, there are 8,000 cases of cancer diagnosed each year, and in addition, 3,500 people die annually from the disease. In 2007, cancer was responsible for 20% of all deaths in Costa Rica, but it is expected around 2025 it will be the cause of 50%. The high incidence in the country is a great pressure for all organizations related to public health [1]. Cancer is not a single disease, but rather a family of diseases that depend on the peculiarities of each person: gender, age, inheritance, lifestyle, the affected tissue, among many other reasons. These pathological conditions can be treated therapeutically by chemotherapy. A less effective treatment induces more mutations making the malignant cells more resistant to subsequent treatments.

Nowadays, chemosensitivity clinical trials generate data that relate the cellular development of the cancer to a concentration of a specific drug, where from these tests in multiple drugs it can be defined if the patient is resistant or sensitive to certain treatments, and therefore become an aid in medical decision. Multiple clinical trials have been carried out, internationally, in which gene information has been collected from cell lines [2], [3], implants in mice [4] and patients' tumors [5], although only some of these projects have collected information on chemosensitivity. Some of those projects are called Cancer Cell Line Encyclopedia (CCLE) [2] and The Cancer Genome Atlas (TCGA) [5].

Central idea of storing genomic and clinical data is to have the capacity to perform in-silico experiments (from genomic data stored in computers) that allow to find relationships between the data and thus recognize groups of samples with better responses to treatments, promoting a possible increase in patient survival, without the need to perform clinical trials of chemosensitivity. It's now known that drug resistance is related to the changes in the karyotype present in cancer, including multiresistance where cancer before a specific drug develops resistance against other drugs simultaneously [6]. Duesberg's theory of cancer speciation proposes that aneuploidy (gain or loss of chromosomes) in cancer is the karyotype of new species, where same instability of aneuploidy is what allows development of drug resistance [7].

But a challenge that arises from past research [1] [8] is the reduction of the dimensionality of the problem faced by the pattern recognition algorithms where each gene represents a feature and since data is about thousands of genes per sample, there is a sampling problem so called "curse of dimensionality".

In this article the objective is to propose and employ a method of reducing the dimensionality problem of pattern recognition on genomic data by identifying regions of chromosomes alterations with their copy number (amplifications or eliminations of DNA number copies). The identification of such regions would allow to define sets of genes linked to both the groups of samples that present greater resistance and those that present greater sensitivity to drugs, in order to delimit the

necessary gene set or tune the methods of pattern recognition such as classification and clustering of cancer samples.

In the next section it will introduce a review of previous work described in the literature related to methods for the identification of regions with alterations in DNA number copies present in certain chromosomal regions of cell lines and cancer samples in order to then apply those genome alterations found to train pattern recognition methods which can be applied to find categories or subgroups with different phenotypes or clinical responses.

II. RELATED WORK

In addition to the algorithms of multiomic data analysis there are algorithms that only work with a type of omic data, in the case of this study given the importance of studying the effect of aneuploidy, DNA copy number data has been chosen. Within the algorithms of DNA copy number can be found three important groups of articles: classification according to type of cancer, classification of specific cancer subtypes and cancer classification by chemosensitivity.

Within these articles the most abundant are those that classify the cells or tissues in the subtypes of a specific cancer because this type of tests are very useful for the clinical decision of the patients where the best possible treatment can be determined as well as the possibility of survival.

The following articles in order of abundance allow to determine if cells or tissues belong to a type of cancer or even if they do not belong to any. It is very useful because it is related to samples whose exact origin is not known due to the type of collection or source from which the sample has been extracted.

Finally, the most scarce articles are those linked to the classification according to chemosensitivity, where there are very few studies because they require a clinical complexity associated with the performance of tests with conventional chemotherapy drugs.

A. Subtyping classification

Qiu et al. (2017) [9] investigate in their article about small cell lung cancer (SCLC), which is divided into two main subtypes: adenocarcinoma and squamous cell carcinoma, where both subtypes have different treatments. They use the data from the copy number of the TCGA database to train a "Naive Bayes" classifier to classify samples from other databases, where they obtained 84 percent accuracy in the classification. Depuydt et al. (2018) [10] used a random forest classifier in conjunction with regressions in order to separate the samples of neuroblastoma patients according to survival.

Beroukhi et al. (2007) [11] developed an algorithm called GISTIC designed to find regions in the chromosome with statistically significant alterations whose results are consistent with the results of other studies about known chromosomal variations in cancer such as glioma. In this article, 141 glioma samples are analyzed in order to compare with the results of other studies. GISTIC consists of giving a G value proportional to the total magnitude of the aberrations in

each position and then randomly changing positions using a permutation test in order to determine that a specific alteration with a G value did not happen by chance.

Etemadmoghadam et al. (2009) [12] using the GISTIC method find regions with amplifications and deletions of DNA copies with significant differences between the groups of samples with resistant ovarian carcinomas and those sensitive to conventional carboplatin-based chemotherapy. In this article [13] through the alterations of the number of copies known in previous studies it was possible to classify correctly 87 % of the samples of the three categories in which the gliomas are found. In [14] use GISTIC to find alterations and a technique called Oncoscape, based on PCA, is used to visualize and integrate the data. There are three clusters coinciding with glioma subtypes where each one has regions with matching alterations. Zaman et al 2013 [15] classify breast cancer into its main subtypes using data of copy number and expression using hierarchical clustering.

B. Chemosensitivity classification

Carter et al. (2017) [16] trained a radial kernel support vector machine with the aim to classifying samples of patients with small cell lung cancer (SCLC) in chemorefractories (after the first chemotherapy the patient has a relapse within the first treatment) or chemosensitive (relapse after the first three months of treatment with etoposide or carboplatin). They achieve an 83% correct classification using cross-validation of 10 iterations (10-fold). 2281 positions were detected on the chromosomes with variations in the number of copies between the chemorefractory and the chemosensitive ones. Ni et al. (2013) [17], through the experimental observations of samples from patients with lung cancer, it's observed that changes in the number of copies are key events in the metastasis and evolution of cancer. Variations in the number of copies remain constant although the treatment of chemotherapy continues to advance. The different lung cancer subtypes have different patterns of alterations as shown by hierarchical clustering.

C. Types classification

Zhang et al. (2016) [18] employ feature selection techniques to reduce the dimensionality of the TCGA copy number data where they rank the 200 most relevant genes for classification. The classifier is a decision tree using a voting system by dividing the data into different segments. TCGA data of 6 different types of cancer in order to detect which tissue belongs to each sample reaching a precision of 0.75. Xu et al. (2018) [19] employ a support vector machine with a linear kernel based on the copy number to classify whether the sample is healthy tissue or tissue with colon cancer at an early stage of development. A sensitivity of 91.8 % is reached. The classifier is based on the variations of the number arms of each chromosome. They used a statistical test Z to give a value to each gene and based on these values train and apply the classifier to the data.

Sanchez et al. (2014) [20] use a new algorithm to identify regions with alterations in the copy number called I\$AR in

order to use these regions together with information from clinical trials to define important genes in the growth of cancer. They use a bayesian mixture model to select the most relevant genes based on the joint information of copy number alterations on TCGA samples and genes with altered expression on CCLE cell lines. Cheung et al. (2011) [21] analyze cell lines with cancer where they find alterations by means of a statistical method called weight of evidence that takes into account the copy number and the expression of the genes. In [22] the alterations in the number of copies are related to the different stages of development (early and advanced) in gastric cancer. The chi-square statistical test is used to analyze the differences between samples with different tumor development in order to find regions with particular alterations in the chromosome. In [23] compares multiple methods to identify copy number alterations where GAIA and GISTIC algorithms stand out for their identification of molecular targets.

III. RESULTS

This article consists on a unsupervised analysis of DNA copy number data in cell lines and samples with cancer, where those alterations occur in some genes clustered at chromosomal locations. Those alterations could be divided on two types of alterations: amplifications (genes with extra copies) or deletions (genes with less copies).

An important point is that alterations typically occur in regions of contiguous genes. Thereby it's not only about DNA copy number data itself but more about specific positions within genome where these alterations occur, following these logical order we applied a statistical method called GISTIC that allows to identify statically significant regions with DNA alterations in cell lines separated by tissue and then divided in two groups by the threshold of IC50 chemosensitivity parameter that indicates sensitivity or resistance to certain drugs according to the standards in which drug tests are applied at clinical level.

The GISTIC algorithm is a statistical method that allows, in a cumulatively way, identify statistically significant alterations (amplifications or deletions) present throughout all the lines of the study. This method has different parameters to be tuned: the level of confidence, tolerance thresholds, maximum sampling space, among others. The idea is to present two sets of different results: the first seeks to compare alterations found in each tissue at both CCLE and TCGA datasets, and the second is to compare the alterations that exist between the most resistant and least resistant groups to the drugs in each tissue in the CCLE.

The GISTIC method receives as inputs, files with a data format called seg that describes the level of copy number in different locations of chromosomes. By means of a genomic annotation feature present in the same GISTIC implementation found on software website, it's possible map to each gene copy number from chromosomal regions data. In the images 1 and ?? you can see the result of applying GISTIC to the entire available data set of the CCLE. In an analogous way,

the method was applied in the subset of breast cancer data as can be seen in figure 2.

In order to carry out these experiments, the copy number dataset is used in the seg format of both the CCLE project and the TCGA project, in addition to the chemosensitivity data of the CCLE project database that had been implemented in a past research project. For the first part of the study, the GISTIC algorithm is applied with the default recommended conditions for the determination of the chromosomal regions with alterations in their copy number more frequently along the cell lines of the CCLE and the TCGA samples corresponding to tissues with breast cancer and ovaries as can be seen in the figures 2 and 3.

For the second part of the study GISTIC algorithm was applied again but only to the lines of CCLE separated as sensitive or resistant according to the parameter of chemosensitivity called IC50 responsible for describing the concentration of a drug when it's half of the cellular activity. This separation is based on the standard criteria of clinical chemosensitivity tests where a cell line is defined as sensitive if IC50 parameter is less than a quarter of the highest concentration, in this case 8.0 uM. In this exploratory study, cell lines were separated according to the following criteria: the number of times each cell line was resistant to each drug was counted, and the average number of times each cell line is resistant was defined as the separation criteria for the lines by below or in the average (in this study they are called for simplicity in terminology as sensitive) and lines above the average was defined as resistant in this study. This criterion was inspired by the fact that aneuploidy and alterations in the number of DNA copies are related to the mechanisms of resistance to drugs and the development of multiple and simultaneous resistance in cancer [6] [7]]. Figures 5 and 4 describe alterations found for chemosensitivity subtypes of both breast cancer and ovarian cancer. The average defined as separation threshold in this study is 18, considering all cell lines and all drugs available in CCLE dataset.

Table I
BREAST AND OVARY CANCER RELEVANT CHROMOSOMAL REGION ALTERATIONS PRESENT IN CCLE AND TCGA DATABASES

| | Breast | | Ovary | |
|------|----------------|-----------|----------------|-----------|
| | Amplifications | Deletions | Amplifications | Deletions |
| CCLE | 1p21.1 | 1p36.33 | 13q21.1 | 4p16.3 |
| | 2p11.2 | 1p36.11 | 5q11.1 | 11p13.3 |
| | 13q21.1 | 9p21.3 | 2q13 | 20p13 |
| | 14q32.33 | 22q13.1 | 1q31.3 | 22q13.1 |
| | 15q11.2 | | | 21p11.1 |
| TCGA | 1q 21.3 | 1p36.33 | 20p13 | 4p16.3 |
| | 8q24.21 | 1p36.13 | 20p11.1 | 11p11.12 |
| | 11q14.12 | 9p21.3 | 17q21.32 | 1p13.3 |
| | 20q13.2 | 22q13.1 | 5q11.2 | 20p13 |
| | | | 3q22.1 | 21p11.1 |

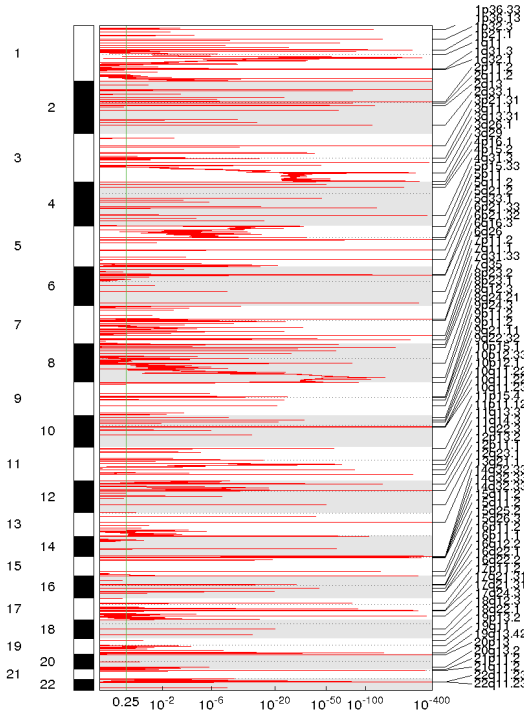


Figure 1. Identified chromosomal regions with copy number amplifications in CCLE dataset. In this image Y axis depicts relative position of alterations in every chromosome and X axis depicts cumulative G score related with frequency of statistically significant amplifications appearance throughout whole CCLE dataset.

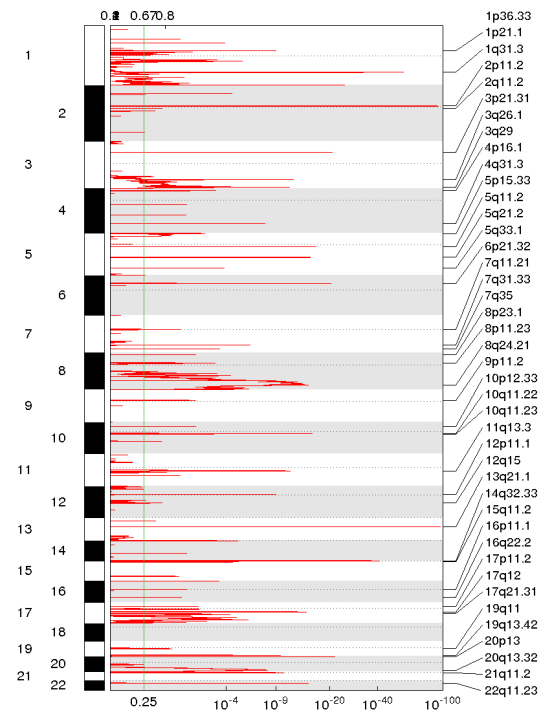


Figure 2. Identified chromosomal regions with copy number amplifications in CCLE breast cancer subset. In this image Y axis depicts relative position of alterations in every chromosome and X axis depicts cumulative G score related with frequency of statistically significant amplifications appearance throughout CCLE breast cancer subset.

Table II
RESISTANT AND SENSIBLE ALTERATIONS BREAST AND OVARY CANCER
IN CCLE DATASET

| | Breast | | Ovary | |
|-----------|----------------|-----------|----------------|-----------|
| | Amplifications | Deletions | Amplifications | Deletions |
| Resistant | 1q31.1 | 1p36.13 | 13q21.1 | 19p12 |
| | 4q31.3 | 7q34 | 3p21.21 | 3p14.2 |
| | 13q21.1 | 11q11 | 2q33 | 20p12 |
| | 20q13.31 | 18q12.21 | 13q21.1 | |
| Sensible | 11q13.2 | 1p36.33 | 3p21.3 | 3p26.3 |
| | 19q13.42 | 11p15.4 | 9p13.1 | 19p13.3 |
| | | 19q13.31 | 17q25.3 | |

IV. DISCUSSION

Normally, pattern recognition algorithms have been using predetermined gene lists or use heuristic techniques to choose such gene sets without considering the biological nature of DNA copy number data. Using an alternative technique of feature selection based on the biological cancer foundations could be of great benefit to not only decrease the dimensionality problem but also to carry out future studies on the relationship of DNA copy number alterations with the expression of genes as well as gene regulation networks.

The use of GISTIC is an important step in the development of the implementation of pattern recognition algorithms capable of identifying molecular mechanisms related to cancer

resistance on drug therapies, where alterations in the chromosomal regions are a key element in disease understanding. When alterations in the chromosomal regions of CCLE and TCGA datasets are compared both for breast cancer and for ovarian cancer, it can be clearly observed that there are differences in specific regions but also many regions are between positions very close to each other, which indicates how literature suggests that cell lines and cancer samples have common region alterations.

In the next phase of the experiment, the IC50 parameter was used, which allowed the separation of the cell lines into two groups based on the number of times each line is below the average number of times designated for the whole subset of lines of a specific tissue. As expected in both groups, the sensitive and resistant lines present different genomic alterations, as can be seen in the figure 2 and 3. In table II can be seen some of different alterations in chromosomal regions that exist among cell lines on both chemosensitivity subtypes.

Considering that the present study is an exploratory and general analysis about the resistance to multiple drugs, it is fundamental to follow up this study to make a specific analysis of resistance and sensitivity for each drug in all the cell lines with which it can be discussed from a biological perspective the action mechanisms of each drug and the

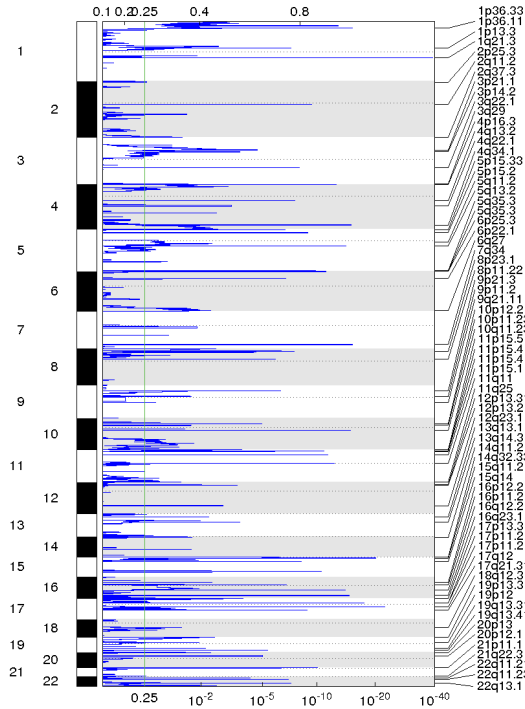


Figure 3. Identified chromosomal regions with copy number deletions in CCLE breast cancer subset. In this image Y axis depicts relative position of alterations in every chromosome and X axis depicts cumulative G score related with frequency of statistically significant deletions appearance throughout CCLE breast cancer subset.

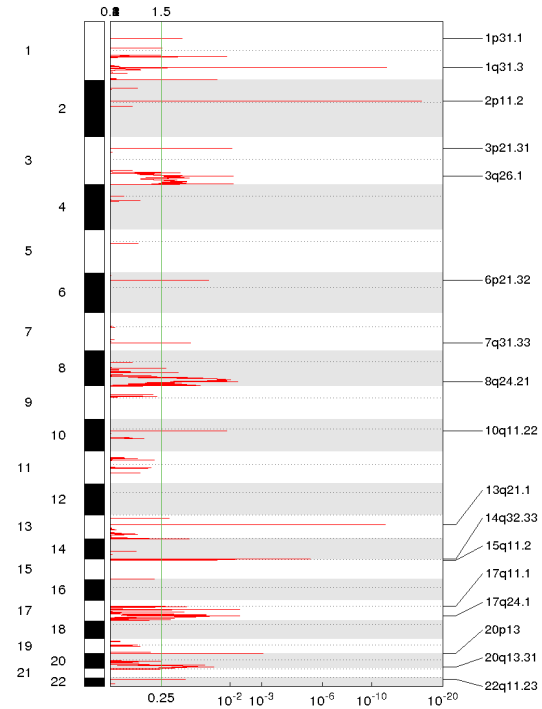


Figure 4. Identified chromosomal regions with copy number amplifications in CCLE resistant breast cancer subset. In this image Y axis depicts relative position of alterations in every chromosome and X axis depicts cumulative G score related with frequency of statistically significant amplifications appearance throughout CCLE resistant breast cancer subset.

possible relationship in chromosomal alterations among the different cell lines across the different available drugs.

In the following stages of this research, the accuracy expected by the classification algorithms for chemosensitivity described in the literature must be greater than 70 % [16] [17], but it should be considered that this study makes this classification between two data sets from different sources (data genomic values of CCLE cell lines and TCGA patient samples genomic data). In the original article where the data of the CCLE is published, it's described that there is a high correlation between the number of DNA copies of cancer samples and cell lines corresponding to samples, which certainly opens the possibility to study the relationship between cell lines and cancer samples [2].

V. CONCLUSIONS AND FUTURE WORK

The GISTIC method has made it possible to find breast and ovarian cancer tissues that have altered chromosomal regions in both the CCLE and TCGA databases. On the other hand, the cell lines defined as sensitive or resistant have some statistically significant alterations in each subgroup, which in theory it would allow to define genes of those altered regions as the space of input characteristics.

Using GISTIC as a feature selection method, defining each copy number gene as a feature, allows to reduce the

dimensionality problem in classification considerably where is taken into account the biological phenomenon of cancer aneuploidy, which can be used to discriminate cell lines and samples with the objective of identifying the mechanisms of resistance to drugs present in cancer and related to the same aneuploidy. In follow-up articles will be compared GISTIC with other methods for detecting chromosomal alterations such as GAIA [23].

It would be worth analyzing the relationship between the DNA copy number and RNA gene expression involved in these altered chromosomal regions as well as their correlation with the expression of other genes to study possible mechanisms of regulation, all of this in conjunction with a priori knowledge of molecular interactions known in each particular type of cancer.

As work is expected to continue, it would be essential to extract for each drug a list of all the genes contained in the areas of interest with alterations to perform two different data studies: the first a functional analysis of gene ontology (an analysis about up-regulated genes overrepresented in all cell lines and samples) and the second with previous knowledge compare the networks of gene interaction reported in literature for resistance or sensitivity to each available drug in each tissue to identify new possible interactions.

In a future work we will explore a classification based on

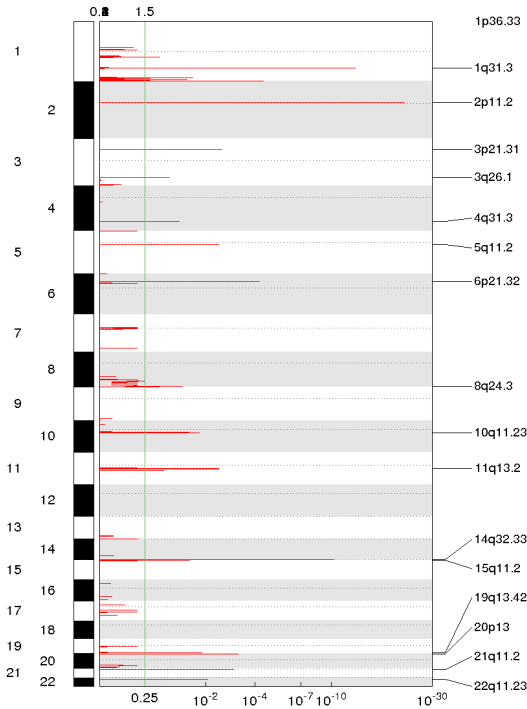


Figure 5. Identified chromosomal regions with copy number amplifications in CCLE sensitive breast cancer subset. In this image Y axis depicts relative position of alterations in every chromosome and X axis depicts cumulative G score related with frequency of statistically significant amplifications appearance throughout CCLE sensitive breast cancer subset.

selected gene sets related to altered chromosomal regions that are found through the GISTIC method in CCLE data set, in order to find phenotypic and clinical categories in TCGA dataset. According to methods described in state of the art, classification will be carried out by a support vector machine with radial kernel where its input space will be bounded by genes that belong to the subset of chromosomal regions with copy number alterations in a statistically significant way. Finally, we will seek to validate the effectiveness of the classification method when comparing results with respect to survival and rejection information of the patients' therapies.

ACKNOWLEDGMENTS

The present paper is part of the work to fulfilling the Master Program in Electrical Engineering, which has been supported by the Postgraduate Program in Electrical Engineering of the Postgraduate Studies System, UCR. Also, members of the PRIS-Lab and of the LabQT, both at UCR, have provided me with fruitful discussions that have help me to continue with this project.

REFERENCES

[1] Coto, J.C., Siles, F. and Mora, R., "A biocomputational platform as a first step to personalized therapy in Costa Rica," *IEEE CONCAPAN BIP 2016*, 2016.
 [2] J. Barretina, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, Marzo 2012.

[3] J. Costello, L. Heiser, J. Saez-Rodriguez, S. Kaski, J. Gray, and G. Stolovitzky, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, 2014.
 [4] A. Bruna, O. Rueda, W. Greenwood, and C. Garnett, M.J. Caldas, "A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds," *Cell*, 2016.
 [5] H. Lee, "The cancer genome atlas explorer," *Genome Biology*, 2015.
 [6] P. e. a. Duesberg, "Cancer drug resistance: the central role of the karyotype," *Drug Resistances Updates. Elsevier.*, 2007.
 [7] J. Nicholson and D. Cimini, "Cancer karyotypes: Survival of the fittest." *Frontiers in Oncology.*, 2013.
 [8] F. Siles and J. Coto, "High dimensional genomic data hybrid modeling to pattern recognition on cancer chemosensitivity," Master's thesis, Computer Science, 2017.
 [9] Z. Qiu, Z. Bi, and K. Song, "Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer," *Genes Chromosomes Cancer*, 2017.
 [10] P. Depuydt, V. Boeva, T. Hocking, R. Cannoodt, I. Ambros, P. Ambros, S. Asgharzadeh, E. Attiyeh, V. Combaret, R. Defferrari, M. Fischer, B. Hero, M. Hogarty, M. Irwin, J. Koster, S. Kreissman, R. Ladenstein, E. Lapouble, G. Laureys, W. London, K. Mazzocco, A. Nakagawara, R. Noguera, M. Ohira, J. Park, U. Pötschger, J. Theissen, G. Tonini, V.-C. D., L. Varesio, R. Versteeg, F. Speleman, J. Maris, G. Schleiermacher, and K. De Preter, "Genomic amplifications and distal 6q loss: Novel markers for poor survival in high risks neuroblastoma patients," *Journal National Cancer Institute*, 2018.
 [11] R. Beroukhir, "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma," *PNAS*, 2007.
 [12] D. Etemadmoghadam, "Integrated genome-wide dna copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas," *Clinical Cancer Research*, 2009.
 [13] V. Ruiz, "Molecular subtyping of tumors from patients with familial glioma," *Neuro-oncology, Oxford University Press*, 2017.
 [14] P. Cimino, "Multidimensional scaling of diffuse gliomas: application to the 2016 world health classification system with prognostically relevant molecular subtype discovery," *Acta Neuropathologica Communications*, 2016.
 [15] N. Zaman, "Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets," *Cell Reports*, 2013.
 [16] L. Carter, D. Rothwell, B. Mesquita, C. Smowton, H. Leong, F. Fernandez-Gutierrez, Y. Li, D. Burt, J. Antonello, C. Morrow, C. Hodgkinson, K. Morris, L. Priest, and M. Carter, "Molecular analysis of circulating tumor cells identifies distinct copy-number profiles in patients with chemosensitive and chemorefractory small-cell lung cancer," *Nature Medicine*, 2017.
 [17] N. X., "Reproducible copy number variation patterns among circulating tumor cells of lung cancer patients," *PNAS*, 2013.
 [18] C. of cancers based on copy number variation landscapes, "Zhang, n. and wang, m. and zhang, p. and huang, t." *Biochimica et Biophysica Acta, Elsevier*, 2016.
 [19] J. Xu, Q. Kang, X. Ma, Y. Pan, L. Yang, P. Jin, X. Wang, C. Li, X. Chen, C. Wu, S. Jiao, and J. Sheng, "A novel method to detect early colorectal cancer based on chromosome copy number variation in plasma," *Cellular Physiology and Biochemistry*, 2018.
 [20] S. G. F., "Integration of genomic data enables selective discovery of breast cancer drivers," *Cell*, 2014.
 [21] H. Cheung, G. Cowley, B. Weir, J. Boehm, S. Rusin, J. Scott, A. East, L. Ali, P. Lizotte, T. Wong, G. Jiang, J. Hsiao, C. Mermel, G. Getz, J. Barretina, S. Gopal, P. Tamayo, J. Gould, A. Tsherniak, N. Stransky, B. Luo, Y. Ren, R. Drapkin, S. Bhatia, J. Mesirov, L. Garraway, M. Meyerson, E. Lander, D. Root, and C. Hahn, "Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer," *Proceedings of National Academy of Science (PNAS)*, 2011.
 [22] N. Arakawa, T. Sugai, W. Habano, M. Eizuka, R. Sugimoto, A. R., Y. Toya, E. Yamamoto, K. Koeda, A. Sasaki, T. Matsumoto, and H. Suzuki, "Genome-wide analysis of dna copy number alterations in early and advanced gastric cancers," *Molecular Carcinogenesis, Wiley*, 2017.
 [23] X. Yuan, J. Zhang, S. Zhang, G. Yu, and Y. Wang, "Comparative analysis of methods for identifying recurrent copy number alterations in cancer," *PLoS ONE*, 2012.



Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Josué David Vargas Amador, con cédula de identidad 1-1540-0368, en mi condición de autor del TFG titulado Clasificación de quimiosensibilidad tumoral en muestras del proyecto TCGA mediante la integración de datos genómicos del cáncer

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. SI NO *

*En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Josué David Vargas Amador

Número de Carné: B16841 Número de cédula: 1-1540-0368

Correo Electrónico: josue.david.vargas@gmail.com

Fecha: 17/06/2020 Número de teléfono: 83136615

Nombre del Director (a) de Tesis o Tutor (a): Francisco Siles Canales

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.