UNIVERSIDAD DE SALAMANCA

Knowledge and biomarkers extraction system by integrating heterogeneous information sources

Juan Ramos González

Director: Juan Francisco De Paz Santana

Tesis doctoral para la obtención del
grado de Doctor en Ingeniería Informática

Facultad de Ciencias
Departamento de Informática y Automática

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Junio de 2019

Autorización del director de Tesis.

El Dr. Juan Francisco De Paz Santana, profesor Titular de Universidad del Departamento de Informática y Automática de la Universidad de Salamanca

HACE CONSTAR:

que como director de la tesis doctoral de Juan Ramos González, con DNI 70913080D, autoriza a este a presentar la tesis doctoral *"Knowledge and biomarkers extraction system by integrating heterogeneous information sources"*.

Fdo: Juan Francisco De Paz Santana

En Salamanca, a 27 de Mayo de 2019

"A scholar has been defined as someone who knows more and more about less and less. Pursuing minutiae is often an effective strategy in academic research, since becoming an expert in some narrow niche is often a good way to publish and secure tenure. For the more intellectually ambitious, however, it is much more exciting to pursue theoretical ideas that are both important and novel. How can such creativity be achieved? It helps, of course, to be a genius (...) Perhaps it takes a genius to work in a well-trodden area and manage to come up with something totally novel, but for the rest of us there is an easier road to creativity. Instead of focusing narrowly on one academic field, a researcher can cast a broader intellectual net and make new connections by tying together ideas from different disciplines".

Paul Thagard: *Cognitive Science, 1996*

UNIVERSIDAD DE SALAMANCA

Abstract

Facultad de Ciencias
Departamento de Informática y Automática
Doctorado en Ingeniería Informática

por Juan Ramos González

Director: Juan Francisco De Paz Santana

Cancer constitutes a major health problem nowadays and therefore cancer diagnosis, treatment and characterization are a crucial scientific challenge. The underlying issue in this research field is related to our limited comprehension about the human cell. The cell dynamics are defined by very complex nets of reactions, compounds and biomolecules. In the last few years, with the appearance of new technologies for genomics and related areas, our capacity to measure biomolecular data has grown faster than the capacity to analyze and interpret such data.

Transcriptomics technologies allow us to study the complete set of transcripts of a cell produced by the genome, by measuring and quantifying RNA. This provides us with information on how the behavior of gene expression changes between various biological conditions when comparing different samples.

Due to the high dimensionality and huge amount of data extracted by these technologies, it is not possible to properly study relations and extract patterns from data without computational power. Consequently, the use of machine learning and data mining techniques in this context has greatly increased in recent years, giving birth to the discipline of Bioinformatics. Specifically, biomarker selection involves now a broad research area of machine learning applications to the health field, constituting a motivation to evaluate, criticize and improve a variety of algorithms. By searching new biomarkers, we can define and characterize tumors, contributing both to knowledge extraction, diagnosis and patient management.

Cancer involves a wide diversity of alterations and, as different patients do not respond in the same way to the same treatment, an accurate diagnosis is critical

to allow the physicians to properly choose a treatment and improve the lifespan of cancer patients.

More concretely, by designing systems capable of finding biomarkers and signatures with high classification potential, we can improve medical research and favor the use of new markers in clinic, as also classify automatically tumor tissues and increase our comprehension about the molecular processes occurring in different types of cancer.

However, even if machine learning algorithms provide a way to extract useful information from high dimensional datasets, their use is not exempt from issues, which have been observed, for example, when comparing results from different researchers. Measuring gene expression is a complex process involving many stages (in both biochips and RNAseq technologies). During this process many problems can affect the performance of different proposals, such as the redundancy in the data, small sample sizes, technological and human errors, methodological differences limiting the comparison of results and even natural variability, among others. Moreover, understanding the behavior of cancer is an especially complex problem as it constitutes a group of disorders with high molecular instability and variability, factors which also affect the data analysis, being very difficult to establish a classification. Consequently, it is a major need to develop new systems that are able to obtain real significant genes for classification and that can be applied in a bigger generality, without sacrificing accuracy when classifying previously unexplored datasets a of same problem.

In summary, there are many obstacles in the way of selecting biomarkers and characterizing tumor subtypes that need to be addressed, considering also the fact that most proposals until now have focused more on reaching high accuracy results rather than in validating and selecting features that are meaningful for the context.

The present thesis aims to deal with some of the main issues concerning applicability and meaningful selection of cancer related genes by designing and applying different data mining frameworks involving hierarchical clustering, case based reasoning, a boosting algorithm and different feature selection techniques.

Different needs have been explored and, consequently, various systems prioritizing the fulfillment of different objectives have been proposed.

In order to evaluate if the developed systems are able to select potential biomarkers, which are relevant from both the biological and the statistical point of view, they have been applied to different study cases involving gene expression measurements from tumor tissues. Such study cases also allow us to extract knowledge about two types of cancer: non-small cell lung cancer (NSCLC) and pancreatic ductal adenocarcinoma (PDAC). More concretely, three frameworks have been implemented and applied over datasets coming from different researchers and labs. Not only these frameworks have proven to select genes with high-classification potential, they have proven to do so without neglecting the knowledge extraction process.

Agradecimientos

En primer lugar, quiero agradecer a mi director, el Doctor Juan Francisco de Paz Santana, su guía y amabilidad a lo largo de estos años. Me ha apoyado siempre en las muchas situaciones vividas, malas y buenas, y me ha ayudado y tranquilizado en las dificultades surgidas a lo largo del desarrollo de esta tesis, motivándome además durante mi breve experiencia docente.

También quiero dar las gracias a mi colega y colaborador, el Doctor José Antonio Castellanos Garzón, un compañero inmejorable, del que he aprendido mucho. Siempre ha tenido tiempo para mí y ha padecido conmigo, conservando siempre el buen talante por difícil que fuese la situación.

Muchas gracias a Daniel López Sánchez y a Jorge Revuelta Herrero pues, además de haberme brindado su ayuda, hemos compartido juntos esta etapa desde principio a fin, dentro y fuera del trabajo.

Agradezco al Doctor Juan Manuel Corchado y a los compañeros que han trabajado conmigo en equipo. Quiero dar también las gracias al Doctor Miguel Rocha y a todos mis compañeros portugueses, que me han hecho sentir como en casa.

Agradezco a los miembros del tribunal su tiempo y su atención dedicados a la lectura y evaluación de este trabajo de investigación.

La Doctora Mónica García Benito y el Doctor José Julián Calvo Andrés han sido para mí un modelo a seguir, un ejemplo de lo que debe ser un buen docente, personas que disfrutan de enseñar y un recordatorio de honradez y humildad, cualidades nunca suficientemente presentes y por las que les estoy profundamente agradecido.

Le agradezco sinceramente a la Doctora Liviu Badea haber cedido de forma desinteresada los datos clínicos referentes a los pacientes del segundo caso de estudio que presenta esta tesis. Fue un gesto generoso y no tan habitual.

Quiero expresar también mi gratitud hacia todo el personal de la Escuela de Doctorado y la Agencia de Investigación de la Universidad de Salamanca, por su amabilidad, ayuda y guía a través de las llamas del averno burocrático. De igual forma tengo que agradecer a la Comisión académica del programa y, en especial, a la Doctora María N. Moreno García, por haber resuelto todas mis dudas y estar siempre dispuesta a ayudar.

A continuación, expreso mi agradecimiento a otras personas que, de forma no tan indirecta, me han ayudado durante este proceso.

Agradezco a Asun, mi heroína, su ejemplo inspirador. Tengo mucho que agradecer también a mi tía Marola, por todo su cariño y las horas pasadas al teléfono ayudándome

con ejercicios de química, cuando aún desconocía la importancia que adquiriría para mí. A mi profesora Ana le doy las gracias por la música y por el refugio que ha supuesto. También quiero expresar mi gratitud hacia Sergio, Bryan y Michael que, junto con mis padres, me motivaron para adentrarme en una fascinante tercera disciplina.

Siempre es preciso contar con alguien que aporte una visión desenfadada y nos ayude a relativizar, y por ello le doy las gracias a mi buen amigo Bender. También se las doy a mis amigas Bea y María por su constante buen humor y por no dejar de tomarme el pelo, y con ellas al pequeño Uni. A mi amigo Javier le agradezco su disposición y apoyo, tanto en el trabajo como en el ocio, y su calma inalterable, que ojalá algún día me contagie. Doy las gracias a mi amigo Rubén por su buen humor y su decisiva ayuda zoológica. Quiero expresar mi gratitud también hacia Lucía, por ayudarme con su arrolladora vitalidad y por los buenos momentos vividos juntos. A mis amigas Ana y Eva, ingenio y sinceridad brutal en cada dosis, les doy las gracias por haberme apoyado en momentos difíciles y ofrecerme tantos buenos ratos.

Tengo mucho que agradecer a mi muy, muy querida tía Espe y a mis siempre divertidos primos Carmen y Pedro, con los que tan buenos momentos he pasado, y con ellos a mi querido tío Goyo, porque sin duda él hubiese disfrutado más que yo mismo de este paso. Es inevitable echarte de menos, tío.

Hay personas que lo dejan todo para ayudar a un amigo. Quiero darle por ello un cariñoso agradecimiento a mi buena amiga Salto, por ser la primera en llegar y la última en irse.

Han sido, sin duda, las no tan pequeñas cosas cotidianas las que más me han ayudado, por lo que quiero expresar aquí mi gratitud hacia mis amigos, mis hermanos Guillermo y Mario, por las largas y necesarias horas de disfrute y risas y lo que hemos creado alrededor de esa mesa de jardín. Agradezco a Guillermo todo su apoyo, todo el tiempo pasado con su inacabable ingenio y el buen ambiente que hemos tenido a lo largo de tantos años. Agradezco a Mario su ayuda incondicional y haber soportado más de lo exigible, así como el haberme ofrecido durante estos años una confianza y amistad absolutas. Quién lo hubiese dicho en aquellas escaleras.

No sabría por dónde empezar a agradecer a mi divertida, inteligente y humilde hermana Julia todo lo que me ha ayudado (tanto en cuestiones matemáticas como de todo tipo) durante este proceso e inspirado con su ejemplo, haciendo de mí un hermano maravillosamente mal acostumbrado.

Muchas más páginas de las presentes en este trabajo (con infinitesimales interlineado y fuente) necesitaría para agradecerles debidamente a mis padres, Susana y Tino. No podemos escapar de la genética y difícilmente de nuestra educación, y por ello hay quien se agobia al verse reflejado en sus padres. Yo tengo la suerte de no poder escapar de mi herencia.

Finalmente, quiero agradecer a la poco convencional personalidad que me apoya todos y cada uno de los días. Te doy las gracias, Alba, por haber puesto a mi disposición toda tu inteligencia, tu paciencia, tu bondad y, en definitiva, por la alegría que me inyectas cada día. Gracias por recordarme las cosas que me apasionan y que merecen la pena, y gracias por sacar lo mejor de mí. Tú subes la media.

Índice general

Índice de figuras

Índice de tablas

Abreviaturas

AI	**A**rtificial **I**ntelligence
AKR1B10	**A**ldo-**K**eto **R**eductase family **1** member **B10**
AR	**A**ction **R**esearch
BSW	**B**ackward **S**election **W**rapper
CAT	**C**orrelation **A**djusted **T**-scores
CB	**C**ase **B**ase
CBR	**C**ase **B**ased **R**easoning
DDA	**D**irection **D**ependence **A**nalysis
ECM	**E**volutionary **C**lustering **M**ethod
FSW	**F**orward **S**election **W**rapper
HDLSS	**H**igh **D**imensional **L**ow **S**ample **S**ize
ICA	**I**ndependent **C**omponent **A**nalysis
KDM6B	Lysine **D**e**M**ethylase **6B**
kNN	**k**-**N**earest **N**eighbours
KPCA	**K**ernel **P**rincipal **C**omponent **A**nalysis
KRT5	**K**e**R**a**T**in **5**
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
LDA	**L**inear **D**iscriminant **A**nalysis
LLE	**L**ocally **L**inear **E**mbedding
LUAD	**LU**ng **AD**enocarcinoma
LUSC	**LU**ng **S**quamous-cell **C**arcinoma
MCE	**M**ódulo de ***Clustering*** **E**volutivo
MCJ	**M**ódulo de ***Clustering*** **J**erárquico
MF	**M**ódulo de **F**ronteras

MFE	**M**ódulo de **F**iltrado **E**stadístico
MIC	**M**ódulo de **I**ntersección de ***Clustering***
MSC	**M**ódulo de **S**elección de ***Clustering***
NB	**N**aive **B**ayes
NSCLC	**N**on **S**mall **C**ell **L**ung **C**ancer
PCA	**P**rincipal **C**omponent **A**nalysis
PDAC	**P**ancreatic **D**uctal **A**deno**C**arcinoma
PKP1	**P**la**K**o**P**hilin **1**
RMA	**R**obust **M**ultichip **A**verage
SDA	**S**hrinkage **D**iscriminant **A**nalysis
SFTA2	**S**ur**f**ac**t**ant **A**ssociated protein **2**
SOX4	**S**RY (Sex determining Region Y) b**OX** 4
SPON1	**SPON**dine **1**
SVM	**S**upport **V**ector **M**achine
TRIM29	**TRI**partite **M**otif containing **29**

A mis padres, Susana y Tino.

Capítulo 1

Introducción

El diagnóstico y tratamiento del cáncer constituye en la actualidad una de las mayores áreas de investigación, debido a que supone un problema de salud extendido en todo el mundo e implica fenómenos de gran complejidad molecular. La predicción y el diagnóstico preciso ocupan un lugar importante dentro de esta área. En ambas tareas, la Minería de Datos, la Inteligencia Artificial y, más concretamente, el Aprendizaje Automático o *Machine Learning*, desempeñan un papel clave. Esto es debido a cómo el avance tecnológico en la medición y toma de datos y la propia complejidad biológica del cáncer han dado forma al principal reto de la medicina del cáncer y de muchas otras enfermedades hoy día: la necesidad de extraer información útil de conjuntos de datos de grandes dimensiones. Esta necesidad es conocida por todos en el ámbito de la investigación como impulsora de la Bioinformática, disciplina cuyo impacto en investigación no hace sino aumentar.

Como en cualquiera de sus muchas aplicaciones, el objetivo de las técnicas de Aprendizaje Automático en Biomedicina es producir un modelo que pueda ser utilizado para llevar a cabo tareas de clasificación, predicción y estimación, entre otras. Así, ya sea con fines de clasificación o caracterización o incluso tratamiento de enfermedades, el empleo de este tipo de técnicas en el estudio de patrones de expresión, genes diferencialmente expresados y biomarcadores, es fundamental en la investigación del cáncer y, a más largo plazo, en clínica.

Los biomarcadores son indicadores objetivos y cuantificables de un estado médico o biológico, ya sea una patología, el estado normal o el de la respuesta farmacológica a una terapia. Normalmente se trata de proteínas (o de los niveles en los que se manifiestan en

cada condición biológica). En la investigación del cáncer y otras patologías, el estudio de los genes que codifican dichas proteínas y sus niveles de expresión ha permitido el descubrimiento de muchos biomarcadores, así como la caracterización de distintos tipos de enfermedades (si bien eso no quiere decir que nivel de expresión equivalga a cantidad de proteína, como se especificará más adelante). La selección de nuevos biomarcadores mejora, en definitiva, la capacidad de diagnosticar diversas patologías así como de evaluar la eficacia de muchos fármacos, entre otras muchas aplicaciones biomédicas de gran utilidad.

El cáncer es un crecimiento anormal de células cuyos mecanismos de control de la división se hallan alterados. Se trata de un problema con base genética que puede originarse en distintos tejidos del cuerpo debido a una acumulación gradual de daño celular debida a factores diversos. El cáncer no es una sola enfermedad, sino un grupo de trastornos multigénicos, en el que las células presentan numerosos genes mutados y/o alterados, produciendo daños en diversos mecanismos de regulación y control que le permiten dividirse sin límite e incluso invadir otros tejidos y órganos.

Pese a ser una de las áreas de investigación que más atención recibe, si no la que más, la inconmensurable complejidad de la célula marca muchos retos aún sin resolver referentes a la comprensión de la misma, repercutiendo en todos los niveles superiores que atañen a la Biomedicina. A este respecto, la mayoría de problemas que rodean al tratamiento del cáncer se derivan en ultima instancia de que las células tumorales son humanas, muy similares en muchos aspectos a las células sanas. Diseñar una terapia específica que afecte exclusivamente al tumor sin dañar el tejido sano es, por tanto, uno de los retos fundamentales. Esto se une al hecho de que, dada la base genética del cáncer, los pacientes no reaccionan igual ante los tratamientos. Estos son algunos de los motivos por los que la búsqueda de biomarcadores y caracterización precisa de tumores es de vital importancia.

Ahora bien, no sólo el tratamiento constituye un problema: el diagnóstico eficaz y precoz, que ha demostrado a menudo ser la herramienta más eficaz para mejorar el pronóstico de muchas clases de cáncer, y quizás la que puede aportar más soluciones a más corto plazo, presenta también numerosas dificultades de diversa naturaleza.

Los avances en los últimos años de las nuevas tecnologías para la recogida de datos en el área de la Genómica y las otras disciplinas “ómicas”proporciona actualmente una

gran cantidad de información acerca de diversos aspectos de la célula, revolucionando la forma de investigar su biología. La recopilación de esta gran cantidad de datos ha supuesto un problema en la extracción de conocimiento. No obstante, esto permite la entrada de la Inteligencia Artificial y la Minería de Datos en esta área, conformando un enfoque interdisciplinar. Es aquí, por tanto, donde la Informática hace frente a problemas que las técnicas de análisis tradicionales no pueden manejar de forma eficiente, complementándolas y aportando una gran capacidad de procesamiento de datos. Esto permite a su vez extraer conocimiento de forma acorde al ritmo de medición de datos de las nuevas tecnologías.

Los sistemas de selección de genes en cáncer tienen varias aplicaciones, orientadas a apoyar la toma de decisiones médicas, el diagnóstico automático, la búsqueda de dianas terapéuticas para el diseño de nuevos fármacos, el establecimiento de nuevas clasificaciones, etc.

No obstante, en la actualidad también existen problemas que afectan a la selección de genes mediante datos transcriptómicos y a la clasificación automática de tumores. En primer lugar, la disposición de muestras es muy limitada y combinarlas para construir conjuntos de datos no siempre es posible, dadas ciertas diferencias metodológicas e incluso tecnológicas en la obtención de estos. Por otro lado, los propios criterios de selección son a veces muy específicos y los genes seleccionados pueden depender en gran medida de la técnica empleada. Además, pocas propuestas validan sus resultados en otros conjuntos de datos, lo que a menudo no permite conocer su capacidad real de generalización. En el diagnóstico automático, muchos trabajos dan prioridad a obtener elevadas tasas de acierto mediante tales propuestas específicas y ocasionalmente se pierde de vista el contexto biomédico. A estos y otros problemas relacionados se le suman otros de naturaleza biológica, como lo son la inestabilidad génica del cáncer y la variabilidad génica natural. Todo ello genera una gran complejidad que afecta a la clasificación y selección y se traduce en que la capacidad de generalización de muchos métodos y la significancia de sus resultados son limitadas, considerando además que no todos los métodos son igualmente apropiados.

Dadas las dificultades de origen biológico, muchos autores sostienen que la medicina personalizada es la línea más apropiada en el tratamiento del cáncer. Teniendo en cuenta que esto ocasiona una necesidad aún mayor en cuanto a capacidad de análisis,

clasificación y gestión de pacientes, entre otras, la búsqueda de soluciones basadas en el Aprendizaje Automático es igualmente imprescindible para la investigación en esta área. En definitiva, más allá de la implementación de herramientas de análisis diseñadas para ser usadas por profesionales del campo de la Biomedicina, el empleo de la creciente capacidad computacional al servicio de sistemas de análisis basados en Inteligencia Artificial permite adoptar un enfoque interdisciplinar a la hora de investigar y solucionar problemas médicos de toda clase (que resultan inabordables desde un enfoque exclusivamente biológico o químico).

Así pues, el uso de técnicas estadísticas y de aprendizaje automático, así como de herramientas informáticas, ocupa cada día un puesto más importante en la investigación biomédica y la necesidad de extraer conocimiento, procesar, interpretar y estructurar cantidades ingentes de datos han hecho de la Bioinformática una disciplina imprescindible e indivisible de la Biología Celular y la Medicina.

En esta tesis se presenta una propuesta de análisis de datos transcriptómicos orientada a la selección y descubrimiento de biomarcadores, así como a la clasificación de tejidos tumorales. Dicha propuesta trata de cubrir varias necesidades de la selección y predicción automáticas y abordar algunos de los problemas más habituales del área, considerando la significancia biológica de los resultados.

Este trabajo se enmarca, por tanto, en el área de la Inteligencia Artificial y la Mineria de Datos, pues estas ofrecen herramientas y algoritmos que se adaptan a la nueva problemática que define el ámbito de la investigación molecular y celular.

1.1. Hipótesis y objetivos

Tras el estudio de diversas técnicas de Aprendizaje Automático y de la problemática presente en el área de la detección de biomarcadores, así como de las diversas aproximaciones y técnicas que la Minería de Datos puede ofrecer en dicha área, junto con las carencias de un enfoque monodisciplinar en la resolución de problemas de selección de genes y clasificación de tumores, se formula la siguiente hipótesis:

- La complejidad en la selección de biomarcadores, el diagnóstico y clasificación de diversos tipos y subtipos de cáncer dada la variabilidad natural, la inestabilidad en los procesos de selección causados por diferencias metodológicas en la recogida y análisis de datos y la inestabilidad génica del propio cáncer, puede ser paliada mediante el diseño y aplicación de sistemas de extracción de información y análisis basados en Inteligencia Artificial y Minería de datos.

Alrededor de esta idea el trabajo presentado propone una propuesta para el análisis de datos transcriptómicos orientada a la selección y descubrimiento de biomarcadores para la detección de tumores, así como a la caracterización de tejidos cancerosos.

Para ver cumplida esta hipótesis, los obstáculos que definen la problemática del área, que serán expuestos con más detalle en el siguiente capítulo, nos permiten establecer una serie de objetivos necesarios para construir sistemas de selección y predicción eficaces.

A continuación se detallan, por tanto, los objetivos principales de esta tesis:

- Utilizar métodos que permitan descubrir biomarcadores complementarios, es decir, que no sean redundantes respecto a otros genes seleccionados a la hora de clasificar y que, juntos, aporten un elevado poder de clasificación.
- Diseñar metodologías capaces de realizar clasificaciones complejas entre clases de tumores con características moleculares comunes, más allá de la mera comparación de muestras tumorales y control.
- Minimizar el número de biomarcadores utilizados en diagnóstico maximizando la capacidad de clasificación, estableciendo grupos de genes reducidos y aptos para su estudio molecular en un laboratorio.
- Establecer metodologías de selección con capacidad de generalización y que sean robustas frente a la inestabilidad, priorizando la significancia biológica de los resultados en tamaños muestrales reducidos.
- Diseñar metodologías capaces de extraer conocimiento biológico en el contexto del cáncer, persiguiendo obtener información sobre la implicación de genes que puedan actuar como dianas terapéuticas y permitiendo el estudio de factores en relación con la expresión génica.

1.2. Metodología

Durante el desarrollo del presente trabajo de investigación, se ha seguido un protocolo de tipo *Action Research* (AR).

La metodología *Action Research* consiste en la recopilación sistemática de datos acerca de un sistema en relación a una meta o necesidad del mismo. El proceso de recopilación está motivado por una serie de hipótesis del investigador acerca de la naturaleza del sistema.

Los datos son suministrados al sistema como consecuencia de la observación de resultados y tiene lugar una "acción"que puede implicar la modificación de las hipótesis. De hecho, si bien algunos autores ven ciertas divisiones, otros lo ven como una aplicación del método científico a problemas específicos, llevada a cabo por un equipo de investigación. De acuerdo con este enfoque, en la actualidad se distinguen tres principales paradigmas de investigación: la aproximación empírica, la interpretativa y la crítica teórica [3].

Más concretamente, en el proceso sistemático de *Action Research* de recopilar información y realizar periodos de análisis y validación, se pueden observar generalmente una serie de pasos que, pese a no ser inmutables, presentan una estructura central más o menos constante. De acuerdo con ella, en primer lugar se revisa la práctica investigadora específica, tras lo cual se identifica el aspecto que se desea mejorar. A continuación, se diseña la estrategia para llevar a cabo dicha mejora y se ejecuta. Los resultados generados se examinan y el plan se reajusta según las observaciones y descubrimientos realizados. Entonces se continúa con la "acción" y, posteriormente, esta acción modificada es evaluada. En esta clase de metodología, los pasos anteriores se repiten hasta alcanzar la meta del sistema [110].

Este modelo es, por tanto, cíclico y los procesos que implica se pueden resumir en las siguientes fases: análisis, descubrimiento de fenómenos, conceptualización, *planning*, ejecución, evaluación, y aprendizaje específico [3].

Así pues, en este caso podemos observar que, a lo largo del proceso y dadas las dificultades propias del área de estudio, el análisis del comportamiento de los sistemas

diseñados ha permitido retroalimentar la propuesta de acuerdo a los objetivos de selección y clasificación para mejorar su funcionamiento inicial y ajustarse al cumplimiento de los objetivos de investigación.

1.3. Organización

De acuerdo con la hipótesis establecida y a las observaciones realizadas durante el transcurso de la investigación, se expone a continuación la estructura del presente trabajo.

El Capítulo 2 detalla el contexto que envuelve la selección de genes y las dificultades investigadoras inherentes a esta área, así como las diferentes clases de aproximación al problema. En primer lugar, la Sección 2.1 expone el contexto referente a la necesidad médica de la selección de genes y clasificación mediante el uso de técnicas de Inteligencia Artificial. En la Sección 2.2, se explica con mayor detalle la problemática del campo que define los objetivos de esta tesis, especificando cuáles son las dificultades en el contexto de la recopilación y análisis de datos biomédicos para la selección de genes. La Sección 2.3 revisa las principales técnicas de selección de genes, describiendo los posibles enfoques que se han adoptado ante dicho proceso. De forma más concreta, se revisa en la Sección 2.4 el uso que se ha hecho hasta ahora de otras técnicas específicas utilizadas en la propuesta de esta tesis para la selección de biomarcadores y clasificación de tejidos tumorales. La Sección 2.5 expone brevemente algunas consideraciones acerca de la naturaleza de los datos que se han tenido en cuenta tanto en la interpretación como en el diseño de la propuesta. Finalmente, y como punto de interés tanto para contextualizar como para realizar un posterior análisis, la Sección 2.6 agrupa algunos de los biomarcadores seleccionados por otros autores y utilizados en clínica con respecto a los tipos de cáncer que se van a analizar como parte de los casos de estudio.

El Capítulo 3 contiene la propuesta, exponiendo en primer lugar el enfoque que se ha hecho del problema mediante el diseño de tres sistemas de análisis y especificando la base de la que se ha partido para abordarla, así como la organización interna y relación entre los diseños que componen dicha propuesta. A lo largo de tres secciones se explica la organización y funcionamiento de los diferentes *frameworks* de selección de genes que se

han diseñado para afrontar los diversos problemas expuestos en los capítulos anteriores, así como la motivación de diseñar tales sistemas y las observaciones realizadas durante la experimentación. Estas alimentan los diferentes enfoques desde los que se aborda cada diseño posterior. En la Sección 3.1 se presenta una aproximación basada en razonamiento basado en casos (CBR) con un módulo de selección de características integrado basado en *Gradient Boosting Regression Trees* (GBRT). En la subsiguiente Sección 3.2 se propone un *framework* de técnicas híbridas basado en la detección de puntos frontera en *clustering* jerárquico. Finalmente, la Sección 3.3 presenta un sistema de tipo *ensemble* con una etapa previa de filtrado y otra posterior basada en *wrappers* para la búsqueda de un *set* estable. En cada una de estas secciones se aplica la correspondiente metodología sobre un caso de estudio real que implica el uso de conjuntos de datos de expresión génica obtenidos a partir de tejidos tumorales; asimismo se exponen las observaciones inmediatas sobre los objetivos de cada sección.

El Capítulo 4 presenta las conclusiones finales, la discusión acerca del papel de la propuesta en el cumplimiento de los objetivos definidos en este capítulo. Se razona acerca de la idoneidad de los diferentes *frameworks* ante problemas específicos, los fenómenos observados durante su aplicación a los casos de estudio y de qué forma afectan al proceso de selección. Asimismo se exponen las capacidades de cada *framework*, así como sus fortalezas y debilidades.

Capítulo 2

Antecedentes

Dadas la hipótesis y objetivos establecidos, para comprender mejor las necesidades de la investigación en este capítulo se expone el contexto de la selección de genes, sus objetivos, los principales retos que conlleva en la actualidad y las técnicas empleadas, así como otros métodos de interés utilizados en la propuesta.

2.1. La selección de genes en el diagnóstico del cáncer: ¿por qué buscar biomarcadores?

Los conjuntos de datos de alta dimensionalidad definen hoy día el principal reto de la Bioinformática. La Minería de Datos y el Aprendizaje Automático ofrecen formas de hacer frente a esta dificultad permitiéndonos localizar y eliminar características redundantes o no relevantes para un problema específico, extrayendo así conocimiento útil a partir de la información contenida en un conjunto de datos [39]. En el ámbito de la medicina, estas disciplinas se aplican a la persecución de objetivos diversos, como el análisis de la importancia de parámetros clínicos, el uso de estos para el diagnóstico automático, la predicción de la progresión de enfermedades, la extracción de conocimiento médico, planificación de tratamientos y su gestión, etc [86].

La Inteligencia Artificial y la Minería de Datos en la selección de genes persiguen tanto mejorar el diagnóstico de enfermedades, apoyando la toma de decisiones, como localizar posibles dianas terapéuticas, además de obtener conocimiento sobre los mecanismos moleculares de la propia célula [96].

En la medicina del cáncer el diagnóstico preciso es clave en la eficacia del tratamiento [39, 86]. Hoy en día el diagnóstico del cáncer en clínica recae sobre técnicas moleculares y de observación histológica. Progresivamente se ha ido otorgando más importancia a la detección de alteraciones génicas clínicamente informativas. No obstante, existen muchos tumores cuya detección es primordialmente histológica, basándose en definitiva en la observación de una serie de características morfológicas microscópicas. Esto ocurre fundamentalmente en tumores pobremente caracterizados, dada una falta de marcadores que nos permitan distinguirlos y clasificarlos [151]. Así, si bien la Inteligencia Artificial también puede ayudarnos en el ámbito de las técnicas histológicas, apoyarse únicamente en ellas, pese al gran interés que suscitan en el área de la visión artificial, es un enfoque médicamente limitado.

La caracterización molecular de un tumor alcanza una elevadísima complejidad, si bien el uso de técnicas de Aprendizaje Automático en este ámbito otorga una mayor comprensión de la biología del mismo [86]. A este respecto, cabe destacar que el diagnóstico mediante técnicas histológicas se ve obstaculizado por varias dificultades derivadas de factores que no son capaces de contemplar, íntimamente ligados a la biología molecular de la célula [151]. Por ejemplo, el hecho de que tumores diferentes puedan compartir similitudes histológicas o de que haya variaciones celulares en tumores formados por un mosaico de células afectadas por distintas alteraciones son algunos de los problemas a los que se enfrentan las técnicas basadas en la observación histológica [151]. Así pues, precisamos de nuevos marcadores que nos permitan caracterizar y predecir aquellos tumores que representan un reto diagnóstico [132], ya que de esta forma se podría influir muy positivamente en la elección del tratamiento para los pacientes de cáncer, que es, junto con el diagnóstico en sí mismo, el objetivo principal de la clasificación [137, 147].

2.2. Los problemas de la caracterización de tumores y selección de genes.

Las dificultades que encontramos en el área de la selección de genes no son triviales y están relacionadas con nuestro limitado conocimiento de la dinámica de la célula, así como con determinadas circunstancias de realidad clínica e investigadora. De hecho,

podemos encontrar obstáculos tanto metodológicos, tecnológicos, humanos, etc., como de naturaleza biológica, algunos de los cuales se hallan fuertemente relacionados.

En el terreno de las dificultades metodológicas y técnicas, cabe destacar en primer lugar que, como se ha mencionado, cada muestra extraída de un paciente cuenta con una ingente cantidad de características. No obstante, al mismo tiempo, en el área médica el tamaño muestral de los conjuntos de datos es muy reducido [132], lo que supone una falta de información a otro nivel. Es decir, los conjuntos de datos contienen gran cantidad de información acerca de cada paciente, pero con frecuencia los pacientes son pocos para obtener conclusiones estadísticamente válidas. Estos *datasets* reciben el nombre de HDLSS o *High Dimension, Low-Sample Size data* [99, 156].

Además, a pesar de que la selección de genes aporte cierta información, muy a menudo las investigaciones y aproximaciones al respecto se centran exclusivamente en la clasificación, dificultando la extracción de conocimiento [7].

Por otra parte, también cabe mencionar que, normalmente las diferentes clases de una población se hallan desbalanceadas, lo cual puede repercutir en la capacidad de generalización de los clasificadores [7]. Por otra parte, los *datasets* de expresión génica suelen presentar un elevado número de características redundantes que dificultan la selección de subconjuntos de genes realmente relevantes para una condición concreta. De hecho, este es el problema central, junto con la dimensionalidad de los *datasets*, para encontrar grupos de marcadores realmente eficaces, lo que afecta también a la capacidad de predicción de muchas propuestas [159].

Respecto al trabajo técnico de laboratorio, debemos considerar que el diseño experimental, las diferencias entre reactivos utilizados e incluso la habilidad del personal técnico pueden dar lugar a resultados distintos y posibles interpretaciones erróneas [55]. Al mismo tiempo, también pueden producirse errores de medida debido a fallos tecnológicos que generan ruido en los datos.

En el caso de los *microarrays*, resulta extremadamente complicado llevar a cabo comparaciones entre diferentes plataformas, pues han sido construidas de acuerdo a estándares diversos, por lo que la composición y distribución de sondas no es la misma [55].

Además de los anteriores, existe otro problema fundamental que a menudo se sintetiza en una sola pregunta. Dado que la clasificación es en definitiva un artificio, ¿cuántas clases de células tumorales hay? [138]. No obstante, también debemos preguntarnos, ¿están clasificadas de acuerdo a un criterio apropiado? Para que la predicción automática de tipos tumorales que nos ocupa sea lo más fiable y exitosa posible, la clasificación de tumores debe reflejar los principales cambios moleculares existentes entre las células tumorales. El *National Cancer Institute* estadounidense establece alrededor de 200 tipos de cáncer, organizados por su localización [138]. La oncología se ha basado hasta ahora en esta clasificación deficiente que no favorece la adecuación de una terapia al paciente [62, 100]. No obstante, hoy en día muchos autores sugieren en diversos estudios específicos la existencia de gran cantidad de subtipos tumorales basándose en características moleculares más que clínicas, por lo que la subdivisión de clases tumorales tiende a aumentar [67]. Esto se debe fundamentalmente a que el número de muestras en los estudios crece de forma gradual, al tiempo que poco a poco se simplifica la integración de diversos datos "ómicos" proporcionados por los avances tecnológicos [138]. No obstante, no siempre las nuevas clasificaciones sugeridas son acertadas, pues los tamaños muestrales son aún bajos y a menudo se dan soluciones demasiado específicas a cada *dataset* [23]. Este último fenómeno conduce a malinterpretar en el conjunto de datos ciertas relaciones intrínsecas entre genes, que no se hallan realmente asociadas al cáncer [138].

Todo lo anterior se relaciona con el problema metodológico derivado de que una vasta cantidad de resultados experimentales dependen en gran medida del algoritmo de selección utilizado, lo que dificulta la aplicabilidad de muchas propuestas. Muchas técnicas presentan criterios de selección muy específicos, de tal forma que, cambiando la técnica empleada sobre un mismo *dataset*, los resultados pueden variar, a veces de forma significativa [1, 60].

Así pues, todos estos factores expuestos determinan un mismo fenómeno: la inestabilidad que afecta a la construcción del *dataset* y al propio proceso de selección, uno de los retos fundamentales a los que se enfrentan estos estudios [11].

Por otro lado, existen una serie de dificultades definidas por la propia biología de los tejidos tumorales, que constituye en definitiva el mayor desafío en la caracterización del cáncer. Damos nombre a los tumores principalmente por el tejido u órgano en el que se originan, agrupándolos en seis clases mayores [132]. No obstante, al ser una

enfermedad que afecta a gran cantidad de genes de forma diversa, el artificio teórico que constituye su clasificación a veces nos ayuda y otras nos limita desde un punto de vista investigador. Debemos considerar que el cáncer se origina por la sucesiva acumulación de daño en las células ligada íntimamente a los procesos de senescencia celular [70]. Pese a originarse a partir de un tejido concreto, las células cancerígenas, que tienen sus mecanismos de control de división y de regulación alterados, presentan una gran inestabilidad génica y cambian rápidamente acumulando gran cantidad de mutaciones a lo largo del tiempo. De esta forma, incluso dentro de un mismo tejido tumoral, diferentes zonas pueden tener distintas alteraciones génicas [109]; asimismo existen diferencias entre células metastásicas y del tumor original [24], ya que, de hecho, son algunas de estas mutaciones las que permiten a determinadas células migrar y favorecer la metástasis [95]. En resumen, la dificultad analítica que se manifiesta aquí es la inestabilidad génica y molecular del cáncer.

En esta misma línea, encontramos otro gran obstáculo de origen biológico para una selección significativa: la variabilidad natural. Diferentes individuos presentan pequeñas diferencias genéticas por las que sus niveles basales de expresión génica, así como de tantos otros factores biológicos, no son necesariamente iguales, aunque no estén afectados por ninguna patología.

Además, no solo las diferencias genéticas entre individuos pueden afectar a los niveles de expresión y a la existencia de distintas clases de células tumorales. De hecho, más allá de la identificación de mutaciones concretas, existen muchos factores que no son genéticos y añaden variabilidad a las células tumorales como, por ejemplo, la regulación epigenética, las jerarquías de diferenciación celular, el microambiente tumoral, etc [23]. Además, debemos ser conscientes de que la medición de los niveles de expresión no se realizan sobre una única célula y que, aunque solo podamos medirlos en un momento dado, no son constantes a lo largo del tiempo. En definitiva, estos fenómenos se traducen en que existe cierta estocástica en los niveles de expresión.

Debemos considerar, en primer lugar, que el cáncer es un grupo de enfermedades en el que están implicados muchos factores; después, que entre cada persona existen pequeñas diferencias, tanto genéticas como de otros tipos; y, finalmente, que las células de distintos tumores no contienen exactamente las mismas mutaciones ni se han producido necesariamente en el mismo orden. Tenemos que tener en cuenta además que, incluso

dentro de un mismo tumor, existe heterogeneidad, lo que supone, de hecho, un obstáculo incluso para la medicina personalizada [15]. Por estos motivos, resulta muy complejo definir grupos de tumores que, dentro de la variabilidad existente, mantengan una serie de características comunes.

Así pues, ambos fenómenos, inestabilidad génica y variabilidad natural, se unen a la inestabilidad del proceso de selección y constituyen el principal problema tanto de la extracción de conocimiento como del establecimiento de biomarcadores y propuestas con capacidad de generalización.

2.3. Las técnicas de selección de genes

La selección de características en el área de la extracción de biomarcadores se puede definir como el proceso de extraer grupos de genes cuyos valores de expresión son representativos de una condición médica y biológica específica [71, 87, 93].

Conocemos estos genes como "genes informativos" o "diferencialmente expresados" y constituyen la base para el empleo de clasificadores en el estudio del diagnóstico y pronóstico de enfermedades. Su análisis también atrae el interés de las compañías farmacéuticas, cuyos esfuerzos se centran en identificar proteínas que puedan ser la diana de ciertas drogas [123]. Aunque se hayan hecho considerables esfuerzos en el desarrollo de nuevas estrategias y métodos para descubrir genes informativos, aún hoy no existe una sola técnica capaz de solucionar todos las dificultades que rodean a este proceso.

En términos generales y dado el amplio abanico de posibilidades, las técnicas de selección de características se han clasificado en cinco grupos principales: filtros, *wrappers*, métodos embebidos, métodos híbridos y, más recientemente, se ha considerado también como una quinta aproximación los *ensembles* [93]. Puesto que los métodos híbridos y los *ensemble* hacen uso de las otras tres técnicas, no todos los autores los consideran como dos clases adicionales. Cada una de estas categorías demanda el uso de métodos distintos por lo que entre ellas se cuentan técnicas de aprendizaje tanto supervisado como no supervisado.

- Los **filtros** se hallan orientados a discriminar y descartar características en base a las propiedades intrínsecas de cada *dataset*, estimando una puntuación que refleje su relevancia con el fin de establecer un punto de corte para seleccionar las características con mejores puntuaciones asociadas [135].

- Los métodos de ***wrapper*** utilizan clasificadores para encontrar las características con mayor poder de discriminación mediante la minimización de una función de predicción de error [40, 164]. Estos métodos tienden a consumir mucho tiempo para su ejecución y sus resultados se hayan íntimamente ligados al tipo de clasificador que se utilice.

- Los métodos **embebidos** son similares a los *wrappers* en muchos aspectos, pero permiten que el método de aprendizaje que guía el proceso interactúe con la selección en sí, lo cual reduce el tiempo de ejecución requerido por los *wrappers* [66].

- Los métodos **híbridos** son aquellos que emplean de forma conjunta dos o más métodos de entre los anteriores, ya se basen o no en el mismo criterio. En su mayoría consisten en combinaciones de filtros y *wrappers*, que tratan de aprovechar las ventajas que ofrece cada uno [112].

- Algunos autores consideran los métodos ***ensemble*** como otro tipo de técnica de selección de características. Son relativamente recientes y se basan en recombinar resultados de otras técnicas de selección de características para obtener grupos más estables, ya que pequeñas variaciones en un *dataset* de entrenamiento pueden tener un efecto considerable en la selección de un método que se aplica individualmente [1, 113] Así, los métodos *ensemble* hacen frente a las dificultades relacionadas con la inestabilidad que presentan muchos de los métodos anteriores.

A continuación se exponen las características de estos métodos en mayor profundidad.

2.3.1. Filtros

Pese a ciertos problemas, como la dificultad para establecer un *threshold* para descartar características, una de las fortalezas de los filtros reside en que tratan de encontrar un subconjunto de genes relevantes sin depender de un modelo de clasificación. Los filtros

siguen dos tendencias alternativas que nos permiten dividirlos en grupos: aquellos que seleccionan las mejores características de acuerdo a un *ranking* (métodos de *ranking*) y otros que optimizan una función objetivo (métodos de búsqueda espacial) [115, 135]. Los primeros asignan una puntuación a cada característica mediante una función de *score* establecida de acuerdo al objetivo de filtrado [144]. Por su parte, los métodos de búsqueda espacial persiguen optimizar la función objetivo de tal forma que los subconjuntos de genes encontrados tengan la máxima relevancia y la mínima redundancia.

De acuerdo con esta distinción, se establece una taxonomía para los filtros según la cual los metodos de *ranking* pueden ser a su vez subdivididos en univariantes y bivariantes [94]. Los primeros pueden ser clasificados a su vez como paramétricos y no paramétricos, mientras que los métodos bivariantes pueden ser *greedy* (algoritmos voraces) o *all-pairs* (algoritmos duales) [20].

En general, los métodos de *ranking* tratan de seleccionar las características con mayor puntuación (*score*) discriminado el resto en una aproximación de cuatro etapas [144]:

- Seleccionar una función de *score* que asigna una puntuación a cada característica y ordenar el conjunto de datos en base a cada puntuación.
- Estimar la significancia estadística de las puntuaciones asignadas (p-valor).
- Seleccionar las características con mayor puntuación de acuerdo a las etapas anteriores.
- Validar el subconjunto de genes obtenido.

La dificultad principal de los métodos de *ranking* reside en el establecimiento de un *threshold* para descartar todas las características con una puntación inferior. No está claro cómo encontrar el *threshold* óptimo, ni existe una aproximación unificada para ello [146]. Pese a esta limitación, los filtros han sido frecuentemente utilizados en el análisis de datos de *microarrays*, pues se trata de una solución útil en extracción de conocimiento y clasificación, más rápida que los *wrappers* [98].

Por otro lado, los métodos de búsqueda espacial son considerados multivariantes [42]. A diferencia de los de *ranking*, estos optimizan la combinación de significancia y

redundancia para seleccionar conjuntos de genes siguiendo tres pasos [144], que consisten en:

- Construir una función objetivo para su optimización.
- Definir un algoritmo de búsqueda de genes mediante la función objetivo.
- Llevar a cabo una validación de la solución obtenida.

Este último punto es compartido tanto por los métodos de búsqueda espacial como por los de *ranking* y lógicamente es de gran importancia en cualquier proceso de selección de características. Los procesos de validación llevados a cabo sobre un subconjunto de genes informativos pueden estar orientados a la clasificación, en cuyo caso se evalúa de acuerdo a una tasa de acierto obtenida en un clasificador o si, por el contrario, el estudio persigue unicamente la caracterización mediante biomarcadores, entonces los genes se validan de forma independiente con respecto a la significancia estadística de sus puntuaciones asignadas [144].

Los filtros permiten que los algoritmos sean sencillos, realizando una selección o descarte secuencial, facilitando su diseño y también su comprensión por otros investigadores. Este es uno de los motivos (junto con la independencia de un modelo de clasificación) por el que suponen la aproximación a la selección de características más común [98]. Más concretamente, existen autores que afirman que los filtros consistentes en técnicas de *ranking* son la forma más apropiada para realizar selección de características en conjuntos de datos de alta dimensionalidad [56]. De acuerdo a Guyon [58] y Lazar *et al.* [93] el enfoque de los filtros podría favorecer la identificación de genes candidatos para ser utilizados como dianas terapéuticas.

2.3.2. *Wrappers*

Los *wrappers* realizan una búsqueda de posibles subconjuntos de características en el espacio y estos se evalúan mediante un modelo de clasificación específico. Para determinar todos los posibles subconjuntos, se utiliza un algoritmo de búsqueda que "envuelve" al clasificador.

Estos métodos ofrecen normalmente mejores resultados de clasificación, aunque existe un mayor riesgo de sobreajuste (cualidades que se explican fácilmente dado que su criterio de selección depende de un clasificador) [131].

Podemos dividir los *wrappers* en dos clases: determinísticos y *randomizados*. Los primeros presentan un menor riesgo de sobreajustarse a los datos de entrenamiento que en el caso de los *randomizados*, mientras que estos últimos son menos propensos a alcanzar un mínimo local. Los *wrappers randomizados* son en su mayoría algoritmos evolutivos y de enfriamiento simulado (o *simulated annealing*) [66].

Los *wrappers* son dependientes de los clasificadores que utilizan, lo que desemboca en una baja capacidad de generalización. Presentan además una menor eficiencia computacional, requiriendo tiempos de ejecución mayores [146]. Por otro lado, estos métodos interactúan con un clasificador y su consiguiente capacidad de clasificación hace que sean ampliamente utilizados, a menudo combinándose con filtros en *frameworks* híbridos. Otro aspecto positivo es que los *wrappers*, a diferencia de los filtros, consideran y permiten descubrir relaciones de dependencia entre las características. Las distintas aproximaciones basadas en máquinas de soporte vectorial son los *wrappers* de uso más común en el campo de la selección de genes [56].

2.3.3. Métodos embebidos

Los métodos embebidos, al igual que los *wrappers*, utilizan un modelo de clasificación como criterio para seleccionar características pero, a diferencia de aquellos, el proceso de búsqueda de un espacio óptimo está integrado en el propio clasificador o, mejor dicho, se utiliza parte del proceso de aprendizaje del algoritmo para realizar la selección [56, 131]. Los métodos embebidos interactúan, por tanto, con el modelo de clasificación. Por estos motivos la selección depende en gran medida de dicho modelo (siendo normalmente el criterio empleado válido unicamente para ese método), pero al mismo tiempo son menos costosos computacionalmente que los *wrappers* y también consideran las relaciones entre características [45, 66]. Aunque los métodos embebidos no están libres del riesgo de sobreajustarse a los datos de entrenamiento, este es menor que en el caso de los *wrappers*, siendo además menos costosos computacionalmente [131].

Algunos de los procedimientos más habituales dentro de este tipo, empleados para la selección de genes, son, por ejemplo, las técnicas de regularización y los algoritmos de poda (o *pruning based*). Las primeras utilizan funciones objetivo para minimizar los errores de ajuste y eliminar aquellas características con coeficientes de regresión cercanos a 0 [103] mientras que en las segundas se utilizan todas las características en el entrenamiento para construir el modelo de clasificación y se eliminan de forma recursiva aquellas con menor coeficiente de correlación [56].

2.3.4. Métodos híbridos

La idea detrás de los métodos híbridos es utilizar diferentes criterios de selección en distintas etapas del proceso, combinando varias técnicas para mejorar la eficiencia de la selección y aumentar la capacidad de clasificación. Así pues, tratan de aunar las fortalezas de ambos métodos para dar lugar a un *framework* más eficaz [112], manteniendo cierta capacidad de generalización y clasificación simultáneamente. No es de extrañar, por tanto, la diversidad de *frameworks* híbridos existente en el área, normalmente ejecutando el filtro en primer lugar [58].

2.3.5. *Ensembles*

Los métodos *ensemble*, a diferencia de los híbridos, aplican los diversos criterios durante una misma etapa. Su objetivo es encontrar diferentes subgrupos de características utilizando distintas aproximaciones y posteriormente combinarlos para formar un único subconjunto [131]. Las ventajas principales de esta clase de enfoque son que la selección suele ser más robusta en los conjuntos de datos de alta dimensionalidad. Además, nos proporciona una visión de la importancia de las características seleccionadas, pues podemos ver qué características son extraídas por cada método y si cada una ha sido extraída por más de uno [1].

Podemos distinguir tres grupos de *ensembles*. Existe un primer grupo que considera la diversidad de los datos, y consiste en utilizar el mismo método de selección en diferentes subconjuntos de un *dataset*. En segundo lugar hay *ensembles* que atienden a la

diversidad funcional, aplicando diferentes técnicas sobre el mismo conjunto de datos. El tercer y último grupo es una combinación de ambos, que contempla tanto la variabilidad en los datos como el uso de distintas técnicas [11].

2.4. Otras técnicas de importancia utilizadas en la propuesta

Dado que también otras técnicas, no orientadas por sí solas a la selección de características, cumplen un importante papel en los procesos de clasificación y selección y, específicamente en el desarrollo del presente trabajo, se contextualizan brevemente a continuación en el área de la selección de genes a partir de datos transcriptómicos.

2.4.1. El *clustering* jerárquico

El análisis de *clustering* persigue establecer una división de un conjunto de datos en varios grupos de acuerdo a características específicas, de tal forma que los datos pertenecientes a un mismo grupo compartan mayor similitud entre sí que con aquellos pertenecientes a otros grupos [74]. En los últimos años, se han utilizado muchos algoritmos de *clustering* en el análisis de expresión génica, demostrando un gran potencial en la búsqueda de grupos de genes relevantes [74]. Las técnicas de *clustering* han resultado también de gran utilidad en el descubrimiento de funciones de genes, procesos de regulación y subtipos celulares.

Una de las particularidades del *clustering* de datos de expresión génica es que su aplicación puede resultar interesante tanto sobre los genes como sobre las muestras [74]. Normalmente, el uso de *clustering* para seleccionar genes relevantes pretende maximizar la diversidad, y sigue el siguiente orden: primero, se escoge una distancia para representar el espacio de características; después estas se agrupan utilizando el correspondiente algoritmo de *clustering* y, finalmente, se seleccionan las características más representativas de cada *cluster* [22]. No obstante, se debe tener en cuenta que el uso de diferentes distancias puede generar diversos *clusterings*, por lo que estas técnicas tampoco se hallan libres de sesgo para agrupar los datos [79].

Gran parte del éxito de estas técnicas en estudios de expresión génica es que permiten realizar una interpretación biológica de los resultados. [79]. A este respecto, los métodos de *clustering* jerárquico son los más frecuentemente utilizados en análisis y visualización de datos de expresión génica. [142]. En ellos, los datos van agrupándose o dividiéndose sucesivamente a lo largo de diferentes niveles siguiendo un jerarquía. Los algoritmos de *clustering* jerárquico utilizan una matriz de disimilitud como entrada; y pueden ser aglomerativos o divisivos, dependiendo de si el dendrograma se construye partiendo de una observación y realizando agrupaciones hasta formar el conjunto total de los datos o, por el contrario, dividiendo grupos hasta alcanzar el nivel de cada observación individual [32] . Así, se generan árboles reflejando diferentes niveles de *clustering*. Su aplicación en el campo de la biología se extendió inicialmente en el área de la filogenética [46].

Uno de los motivos de su éxito en el terreno que nos ocupa es que el *clustering* normalmente es exploratorio, por lo que en el análisis de expresión génica no se tiene a menudo conocimiento acerca de qué k escoger en los métodos particionales, existiendo la posibilidad de cometer errores durante la búsqueda [79]. No obstante, antes de interpretar el *clustering* desde el punto de vista biológico, cabe preguntarse si los resultados son reproducibles o si el nivel del árbol escogido para la selección de genes es apropiado; ambas cuestiones constituyen las dificultades principales en la aplicación del *clustering* jerárquico a este campo [158]. Además, desde el punto de vista biológico, también cabe preguntarse si existe alguna característica común que comparten los datos de un mismo *cluster*. No obstante, el *clustering* jerárquico ofrece, pese a las dificultades de escoger nivel, la posibilidad de barajar información que nos perderíamos en otros métodos de *clustering* que fijan los parámetros para generar un solo *clustering* a su medida. Por tanto, el *clustering* jerárquico nos proporciona ciertas ventajas de cara a la interpretabilidad de los resultados como, por ejemplo, permitirnos ordenar los genes en el *dataset*, puesto que, durante su ejecución, estos se irán disponiendo en el árbol en función de su similitud a lo largo de la jerarquía establecida. De esta manera, podemos ver diferentes grados de similitud entre genes. Además, la tabla que representa los datos ordenados puede mostrarse gráficamente, ofreciendo la posibilidad a los biólogos de estudiar las relaciones entre genes [46].

A este respecto, uno de los objetivos del *clustering* en expresión génica es observar la correlación de la expresión de diversos genes. En este sentido, algunos autores se basan fundamentalmente en el análisis visual para agrupar los genes [34, 36]. Sin embargo,

estos métodos presentan una alta subjetividad y su aplicación se torna más compleja según aumenta el número de muestras.

Eisen [46], por su parte, propone un método de *clustering* jerárquico de media de grupos (*pairwise average-linkage*) para poner de manifiesto la gran utilidad de las técnicas jerárquicas en la realización de interpretaciones biológicas. A partir de su influyente trabajo, se generaliza el uso del *clustering* jerárquico basado en disimilitud en el área y, con ello, el surgimiento de una línea de investigación para mejorar y evaluar dicho método en la selección de genes.

Waddell y Kishino [145] sugieren posteriormente el uso de una correlación parcial en lugar de la correlación estándar, sosteniendo que de esta forma la relaciones encontradas entre genes son teóricamente más sólidas. Kerr y Churchill [80] utilizan un modelo ANOVA (*ANalysis of VAriance*) y remuestrado basado en residuos para validar la calidad de diferentes algoritmos de *clustering*. Chen *et al.* [31] también evalúan y comparan varios métodos de *clustering* y mapas autoorganizados o *self organizing maps* (SOM), apoyándose en propiedades físicas de los *clusters* resultantes, tales como la homogeneidad y la separación. Por su lado, Datta *et al.* [38] introducen una nueva matriz de disimilitud basada en mínimos cuadrados parciales y utilizan datos de levaduras para evaluar diferentes algoritmos de *clustering* de acuerdo a varias medidas de validación, concluyendo que Diana, un método jerárquico divisivo, es el más consistente.

Posteriormente, muchos trabajos se han centrado en proporcionar procedimientos que nos permitan conocer hasta que punto un *clustering* está realmente representando una estructura de relaciones real, dadas la baja muestra y la alta dimensionalidad o, dicho de otro modo, valorar la significancia estadística del *clustering*. En este sentido, Suzuki *et al.* [139] desarrollan un paquete en el lenguaje R, basándose en técnicas de *bootstrapping* originalmente propuestas y utilizadas para evaluar la significancia en árboles filogenéticos [44, 82]. En el caso específico del *clustering* jerárquico, esta tarea es especialmente compleja por la gran cantidad de *tests* requeridos para abordar las muchas divisiones en la estructura de árbol que lo caracteriza. Liu *et al.* [99] proponen un método Monte Carlo para este mismo propósito orientado a k-*means*, pero no a algoritmos de *clustering* jerárquico. Maitra *et al.* [107], por su parte, adoptan también una aproximación basada en *bootstrap* que, pese a considerar cualquier número de *clusters* en un *dataset*, sigue sin abordar el problema jerárquico, ni tampoco el empleo

de *datasets* HDLSS. Huang *et al.* [68] desarrollan un método integrado en un paquete de R; su fundamento es calcular la significancia de dividir un *dataset* en dos *clusters*, y que puede ser aplicada iterativamente cuando hay más de dos *clusters*, contemplando así la posibilidad de que exista cualquier número de grupos, y aplicándose en este caso a *datasets* HDLSS. No obstante, sigue centrándose en k-means, un método particional. Recientemente, Kimes y Liu [82] proponen un nuevo método de Monte Carlo que sí está orientado a la evaluación de la significancia del *clustering* jerárquico, resolviendo el problema de la estructura del árbol mediante la aplicación un test de forma secuencial, que emplea datos de expresión génica de pacientes de cáncer.

Existe, en definitiva, una amplia gama de técnicas a nuestra disposición, aunque aún no contamos con guías de criterios generales acerca de la idoneidad de cada algoritmo en diferentes casos de estudio. En cualquier caso, el *clustering* jerárquico ha permitido gran cantidad de descubrimientos en el área bioinformática y su uso ofrece una visión estructurada del conjunto de datos [68].

2.4.2. Razonamiento basado en casos

El razonamiento basado en casos o CBR (*Case Based Reasoning*) es una aproximación para la resolución de problemas de predicción que trasciende el conocimiento general y hace uso de soluciones especificas obtenidas en situaciones pasadas similares. Así, esta técnica lleva a cabo un aprendizaje incremental, reteniendo cada nuevo caso en una base de casos para aplicarla sobre problemas futuros. En este aspecto reside la principal diferencia entre CBR y otras aproximaciones de Inteligencia Artifical; y esto se traduce en que cada vez es más utilizado en investigación [2].

Estos sistemas se han aplicado a la resolución de problemas diagnósticos en medicina [30, 118]. En esta área, la clasificación es decisiva para el diagnóstico y tratamiento, y la alta dimensionalidad y el elevado nivel de ruido en los datos afectan considerablemente al funcionamiento de muchos algoritmos [157]. Pese a ello, existen pocas aproximaciones de CBR al ámbito de la clasificación de datos transcriptómicos, ya que a menudo no se realizan clasificaciones complejas o se tiende a evaluar las propuestas en un mismo *dataset* de tamaño reducido [157].

En los *datasets* de expresión génica, la extracción de conocimiento es aún un reto debido al reducido tamaño de los conjuntos de datos de entrenamiento disponibles en las bases de datos [99]. Esta clase de conjuntos de datos presenta un gran número de atributos que, como se ha destacado, es mucho mayor que el número de muestras disponibles para el entrenamiento de un algoritmo, por lo que es necesario reducir el número de atributos almacenados en los casos que contiene la base de casos [9]. Por este motivo la selección de características ha recibido más atención en los últimos años como una forma de mejorar los clasificadores basados en CBR [6].

Aunque el uso de CBR en esta área es aún reducido, algunas propuestas han tratado de mejorar la clasificación mediante este procedimiento. A este respecto, Arshadi y Jurisica [9] incluyen técnicas de selección de características en un sistema CBR como un método para gestionar conjuntos de datos de alta dimensionalidad y lo evalúan sobre un *dataset* de cáncer de pulmón, logrando un incremento en la tasa de acierto desde 60 % a 70 % cuando lo comparan con un sistema similar sin una etapa de selección de características. Motivados por esta y otras aproximaciones, como la de Huang *et al.* [69], Anaiise *et al.* [6] combinan con éxito distintos métodos de selección de características con el pesado de las mismas, mejorando el proceso de recuperación de un sistema CBR y alcanzando elevadas tasas de acierto en clasificación de clases tumorales.

La lógica difusa también ha sido aplicada al ámbito del razonamiento basado en casos para el estudio de perfiles de expresión génica. Fernández *et al.* [48] ha propuesto un método que implica una representación difusa de los niveles de expresión génica de cada caso, apoyada en la discretización de los datos de expresión en un reducido número de funciones de pertenencia difusas. El procedimiento ha demostrado ser capaz de generalizar todas las muestras disminuyendo el número de genes necesario para la clasificación.

Finalmente, cabe decir que, pese al reducido uso que se ha hecho del razonamiento basado en casos en la selección de genes en cáncer, los diversos trabajos en este y otros campos hacen de esta técnica una vía interesante de cara a ser aplicada a la clasificación y caracterización de tejidos tumorales.

2.5. Aclaraciones acerca de los datos transcriptómicos

De acuerdo con el dogma central de la biología, reducido a lo esencial, partes del DNA (ácido desoxirribonucleico) del núcleo celular se desenrollan para que se formen cadenas complementarias de RNA (ácido ribonucleico) mensajero, que contiene la información para la formación de una proteína (transcripción). Estos transcritos de RNA salen del núcleo y la información contenida en ellos codifica la formación de las correspondientes proteínas (traducción) [134]. Se denomina, por tanto, "transcriptoma", al conjunto de todas las moléculas de RNA presentes en la célula en un momento dado.

Los datos de expresión génica empleados en selección de genes se obtienen mediante el uso de tecnologías como los *microarrays* (tanto chips de cDNA como *oligochips*) y el RNAseq. Aunque de forma distinta, ambas nos proporcionan una medida de la cantidad de un transcrito (o fragmento de RNA) que porta la información del gen a partir del cual se ha formado. Así pues, normalmente un conjunto de datos de expresión génica contiene los niveles de expresión pertenecientes a los genes que constituyen el genoma de la especie de estudio, medidos para una serie de muestras de tejido distintas, por regla general pertenecientes a distintos individuos. Estos datos, junto con determinadas características clínicas, son los empleados por la propuesta de este trabajo para realizar tareas de selección y clasificación en cáncer.

A este respecto, la tecnología de secuenciación nos ofrece una mayor precisión a la hora de medir niveles de expresión, más notable aún cuando estos son reducidos, si bien es cierto que requiere de un procesamiento más complejo y, al tratarse de una tecnología más joven, existen menos algoritmos específicos, lo que también se traduce en que todavía hay una menor (aunque creciente) disponibilidad de datos públicos de expresión obtenidos mediante RNAseq [162].

Los perfiles de expresión resultan de gran utilidad en diversos campos, fundamentalmente en el diagnóstico y en el desarrollo de tratamientos para distintas enfermedades [1]. No obstante, para no faltar a la realidad de la célula y en aras de la corrección científica, es preciso matizar que, aunque a menudo denominamos "valores de expresión" a aquellos obtenidos a través de RNAseq o *microarrays* de expresión, estos representan en realidad medidas de la cantidad de RNA de un gen. Esto quiere decir que no necesariamente una

gran cantidad de mRNA genera una gran cantidad de la proteína correspondiente; existen además procesos post-traduccionales que apenas alcanzamos a comprender. A pesar de que a menudo se asume una relación entre ambas biomoléculas, hay tanto estudios en los que se observa una correlación entre cantidad de mRNA y proteína como otros en los que no [121]. A este respecto, sería ideal poder medir la cantidad de todas las proteínas de la célula simultáneamente y en tiempo real, pero en la actualidad carecemos de una tecnología capaz de llevar a cabo esta tarea. En resumen, se pretende aquí resaltar que, como en otras disciplinas "ómicas", pueden asumirse ciertos riesgos teóricos derivados fundamentalmente del desconocimiento acerca la relación entre la cantidad de transcritos (de mRNA) y la de las correspondientes proteínas, así como de la dificultad para medir e integrar datos de diversos procesos celulares y observar su comportamiento a lo largo del tiempo.

Sin embargo, es importante resaltar que sí se observan cambios en el transcriptoma que parecen obedecer a distintos procesos y condiciones, lo cual nos permite estudiar la implicación de esos genes en la producción de proteína y comprobar si son buenos marcadores de una condición o clase de tejido.

2.6. La selección de biomarcadores en el adenocarcinoma ductal pancreático y en el cáncer de páncreas de células no pequeñas

La selección de biomarcadores constituye un campo muy activo, considerando que contribuye a mejorar el diagnóstico y proporciona conocimiento acerca de los procesos moleculares de los tejidos tumorales. No obstante, el cáncer constituye un grupo de patologías heterogéneo y complejo [86] que, como se ha expuesto, no resulta fácil clasificar.

Puesto que los casos de estudio del presente trabajo engloban el adenocarcinoma ductal pacreático (PDAC o *Pancreatic Ductal AdenoCarcinoma*) y la clasificación de subclases de cáncer de pulmón de células no pequeñas (NSCLC o *Non-Small Cell Lung Cancer*), se especifican a continuación algunos de los marcadores relevantes empleados y propuestos en el área.

En los últimos años se han hecho grandes esfuerzos por caracterizar las principales alteraciones génicas presentes en PDAC. Entre los muchos genes alterados (a menudo oncogenes y supresores tumorales) podemos destacar RAS, AKT, CDKN2A y TP53, DPC4 (SMAD4), así como genes que codifican para proteínas de importantes rutas de señalización como Sonic, Wnt y Notch, y que normalmente están afectados por mutaciones puntuales o pérdidas alélicas [65, 85]. Normalmente, los tumores pancreáticos presentan varias de estas mutaciones, de entre las cuales la alteración de los genes KRAS y CDKN2A son las más frecuentes [83]. Aunque existen clasificaciones de cara a sus propiedades histopatológicas [13], más recientemente se ha sugerido la existencia de cuatro subtipos principales de PDAC de acuerdo con características moleculares, denominados "estable", "localmente reorganizado", "diseminado" e "inestable" [19]. De acuerdo a la discusión desarrollada en esta tesis, es interesante destacar que, si bien las alteraciones mencionadas pueden actuar como un grupo de marcadores de cáncer pancreático, no son buenos marcadores para una subdivisión de clases pues se ha observado que su expresión es estable entre los diferentes subtipos.

Actualmente, el único marcador ampliamente utilizado en clínica para la detección de cáncer de páncreas es el CA19-9 (*carbohydrate antigen 19-9*). En el caso del PDAC se han propuesto gran cantidad de marcadores a través de diversas aproximaciones, tanto inmunohistoquímicas como bioinformáticas. No obstante, pese a los esfuerzos en investigación y las grandes inversiones económicas en el descubrimiento de nuevos marcadores mas efectivos que CA19-9, aún no se ha validado ningún candidato más efectivo [140]. La utilidad de este marcador es menor en los casos en los que hay obstrucción biliar, situación en la que se aconseja utilizar CEA (*carcinoembryonic antigen*). CA125 (*cancer antygen 125*) también ha sido obtenido en numerosos estudios y según el perfil molecular observado puede ser mas apropiado que CA19-9, si bien su uso no está igualmente extendido en clínica [101].

Novotny *et al.* [117] defienden el empleo de la piruvato kinasa M2-PK (*M2-pyruvate kinase*) como un marcador potencial de PDAC y se han realizado varios trabajos investigando su eficacia [14, 75]. Además, se han propuesto otras proteínas como CEMIP (*cell migration-inducing hyaluronan binding protein*) y se han estudiado marcadores más específicos, por ejemplo C4BPA (*C4b-binding protein alpha-chain*), como candidato para para distinguir el PDAC de la pancreatitis; e IGFBP2 y 3 (*insulin-like growth factor-binding protein*) se ha empleado para discriminar PDAC en fases tempranas, ofreciendo

además buenos resultados cuando se combinan con CA19-9 [90]. Asimismo también se ha estudiado el prometedor papel de varios RNA no codificantes en la detección de PDAC, aunque aún su validación se encuentra en fases tempranas [63].

En el caso del cáncer de pulmón de células no pequeñas, dado el problema diagnóstico que supone, existen muchos trabajos centrados en el estudio de las diferencias entre adenocarcinoma de pulmón y cáncer de pulmón de células escamosas, más que en la mera detección del cáncer. Sin embargo, en este caso la mayoría de autores se centra en el uso de técnicas inmunohistoquímicas para la validación de los biomarcadores, habiendo pocas aproximaciones bioinformáticas a este problema. Genes como DSC3, KRT5, KRT6 para el cancer de pulmón de células escamosas y ttf-1 y napsina-A para el adenocarcinoma pulmonar, entre otros, son algunos de los marcadores principales reportados por diferentes autores [8, 133, 141, 161] como poseedores de alto potencial para la subclasificación del cáncer de pulmón de células no pequeñas. No obstante, en trabajos más recientes, han surgido nuevas propuestas como MLPH, TMC5, SFTA3 [161] o ST6GALNAC1 y SPATS2 [141] para su uso como biomarcadores en este contexto. Recientemente, Takamochi *et al.* [141] ha presentado el primer análisis CAGE (*Cap Analysis of Gene Expresion*) orientado a tumores primarios de las clases de cáncer de pulmón de células no pequeñas, centrándose en el análisis de *clustering* y validando el conocimiento extraído en un *dataset* diferente. Alcanzan en dicha tarea una elevada tasa de acierto partiendo de técnicas inmunohistoquímicas, pero el conjunto de evaluación es tan reducido que realmente es necesario llevar a cabo más experimentos para conocer su eficacia.

En definitiva, se ha propuesto un elevado número de proteínas para su uso como biomarcadores, proceso durante el cual han aumentado los conocimientos acerca de los procesos celulares que tienen lugar en el cáncer. Lamentablemente, dar el paso desde *sets* de genes propuestos a biomarcadores aptos para clínica es un problema interdisciplinar de gran complejidad y a menudo poco exitoso [122]. Por tanto, la dificultad para validar los marcadores candidatos que han sido sugeridos por distintas metodologías, así como la propia diversidad de estas y su difícil automatización, constituyen un gran obstáculo hoy en día incluso para aquellas aproximaciones que ofrecen marcadores de gran potencial.

2.6.1. Conclusión del estudio de antecedentes

Se ha visto que las aproximaciones bioinformáticas basadas en la Inteligencia Artificial pueden ofrecer solución a muchos de los problemas definidos por el contexto biomédico irresolubles desde otros puntos de vista, tanto en la selección de posibles marcadores como en el diagnóstico, clasificación y predicción de enfermedades. No obstante, generan conocimiento a un ritmo muy superior a la capacidad de validación de muchas técnicas químicas y se precisa, por tanto, de una apropiada comunicación entre distintas disciplinas. Desafortunadamente, es común la existencia de trabajos que presentan nuevos biomarcadores, de los cuales algunos no son correctamente validados, mientras que otros, independientemente de su potencial en diagnóstico, no tienen la oportunidad de someterse a la experimentación requerida para su evaluación [28]. Por otro lado, las propuestas realizadas en selección distan mucho de ofrecer una solución única y no existe un algoritmo óptimo para lidiar con todos los problemas que este tipo de investigación implica, por lo que resulta complicado, aunque necesario, establecer métodos que aborden el máximo número de contingencias posibles durante el proceso de selección.

Capítulo 3

Selección de biomarcadores para el diagnóstico del cáncer y la extracción de conocimiento biológico: afrontando las dificultades del análisis de datos en el área biomédica desde la perspectiva del aprendizaje automático

Dados los problemas inherentes al área biomédica expuestos en los dos capítulos anteriores, se ha perseguido desarrollar sistemas de selección que tratan, en definitiva, de poseer el mayor número de características deseables según las necesidades propias del contexto; es decir, sistemas tales que pretenden alcanzar el mayor grado posible de cumplimiento de todos los objetivos mencionados. Ahora bien, aunque los objetivos están relacionados entre sí, en realidad partimos de diferentes enfoques cuando se aborda el proceso de selección, si consideramos además el amplio abanico de criterios que utilizan los diversos algoritmos. Una vez estudiados distintos planteamientos y, teniendo en cuenta los objetivos y dificultades marcados por el contexto médico, se juzga pertinente diseñar e implementar varios *frameworks*, que priorizan aspectos diversos pese a compartir metas comunes. En este sentido, la aplicación de distintas metodologías ha permitido una mejor comprensión del problema, observando su evolución respecto al cumplimiento de la hipótesis, comparándolas entre sí y evaluando si en efecto cada una desempeña la función para la que fue diseñada o, por el contrario, se adecua mejor a otros objetivos de selección. Si bien varios objetivos se hallan íntimamente relacionados, en un primer momento se adopta un planteamiento más orientado al diagnóstico y clasificación mientras que otro prioriza la flexibilidad y la extracción de conocimiento.

Tras una revisión bibliográfica y el estudio de distintas técnicas orientadas a enfocar los problemas tratados anteriormente, se ha observado que algunos de los fenómenos más notables (y que precisan de atención con mayor urgencia) en el área de la selección y descubrimiento de biomarcadores son la baja uniformidad en los resultados de grupos de genes para clasificación y la falta de validaciones en datos externos. Este problema, junto con el uso de conjuntos muestrales de reducido tamaño, es la causa de la baja capacidad de generalización de muchas propuestas. Como consecuencia, se opta en primer lugar por enfocar el problema del diagnóstico desde el punto de vista de una selección y clasificación precisas, pero fundamentalmente aplicables, generando resultados manejables y susceptibles de su estudio en un laboratorio. Así, se toma este como el punto de partida alrededor del cual comenzar a construir la propuesta. Se inicia entonces el diseño de un *framework* que pretende apoyar y ofrecer resultados evaluables mediante métodos moleculares pero que, al mismo tiempo, necesita adaptarse al contexto de dichos métodos de trabajo para tener éxito en su tarea. Por consiguiente, es necesario proporcionar un método de selección eficaz que obtenga grupos de genes muy reducidos en el marco de una metodología apropiada para hacer frente a los bajos tamaños muestrales. Después de esta etapa inicial, y con la motivación de tratar la extracción de conocimiento, se implementa un segundo *framework* centrado en el análisis, tratando de salir al paso de diferentes problemas de selección. En este caso cobra una importancia especial la idea de utilizar técnicas que, mientras llevan a cabo la selección y descubrimiento de biomarcadores, ofrezcan al investigador en cuestión más información sobre la organización del *dataset* sobre el que trabaja.

Así pues, analizadas una vía con mayor orientación práctica y diagnóstica y otra que se centra en la exploración del conjunto de datos, permitiendo una mayor interacción del usuario y priorizando la extracción de conocimiento, surge la necesidad de proponer también una metodología que concilie en cierta medida estos conceptos. Se llega así a un tercer *framework* que, si bien se centra en la clasificación, trata de aportar cierta robustez frente al (gran) problema de la inestabilidad (tanto génica como del proceso de selección) considerando ciertos sesgos metodológicos que, más allá de la clasificación, pueden entorpecer la validez biológica de los resultados.

De esta manera se generan diferentes aproximaciones que abordan el proceso de selección de genes, la extracción de conocimiento derivada de este y el diagnóstico desde diferentes ángulos, tratando de proporcionar marcadores para la predicción y estudio de

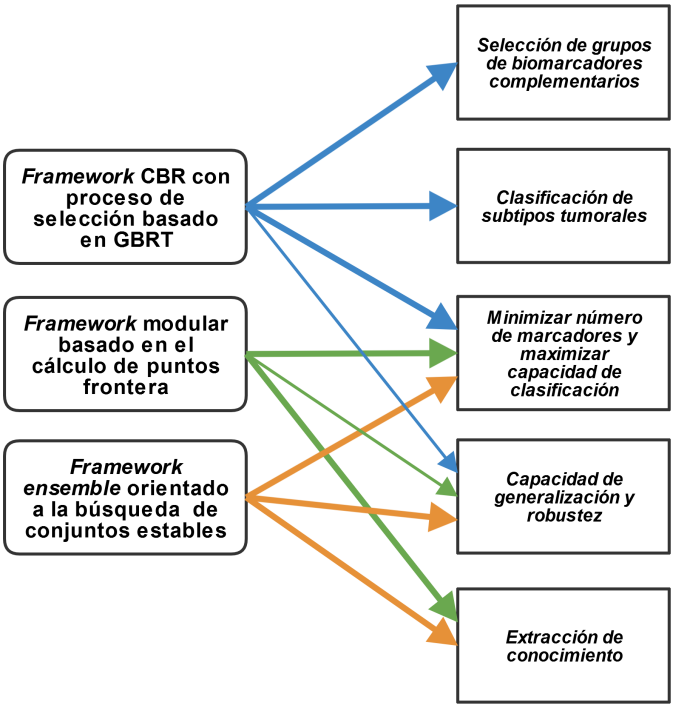

FIGURA 3.1: Diagrama indicando los principales objetivos iniciales de cada *framework* propuesto

tipos de cáncer de difícil detección y/o caracterización. En la Figura 3.1 se muestra un diagrama que refleja la orientación principal de cada uno de los *frameworks* implementados, si bien, como se ha dicho, todos los objetivos han sido tenidos en cuenta durante su diseño.

Este capítulo incluye tres *frameworks* de selección de genes a partir de datos transcriptómicos, evaluados a través de sendos casos de estudio que se centran en el cáncer de pulmón de células no pequeñas (NSCLC) y en el adenocarcinoma ductal pancreático (PDAC), patologías ambas cuya detección en el ámbito clínico es complicada, así como lo es su caracterización molecular. Estas cuestiones las he abordado en los trabajos [126], [125] y [27], que realicé conjuntamente con los autores que figuran en la bibliografía. En cada una de las siguientes secciones se presenta el correspondiente *framework* de selección, junto con la evaluación del mismo en un caso de estudio de interés y los resultados de tal aplicación.

3.1. *Framework* CBR para la predicción y caracterización de subtipos de cáncer mediante la integración de un módulo de selección de características basado en *gradient boosting*: estudio sobre muestras de cáncer de pulmón

La clasificación y el diagnóstico automáticos constituyen un área de creciente interés en la investigación oncológica. Normalmente, la toma de decisiones en este campo está basada en métodos moleculares y de observación de características histológicas. Estos últimos tienen un gran peso en el diagnóstico pero, dependiendo del tipo de tumor, proporcionar una clasificación afinada puede resultar tremendamente difícil. Esto quiere decir que, si bien existen diversos marcadores y métodos de selección para el diagnóstico de cáncer y su clasificación, el problema se torna más complejo cuando se pretenden discriminar subtipos moleculares, especialmente en el caso de los tumores que presentan células muy poco diferenciadas. Por tanto, existe una necesidad de crear sistemas que permitan encontrar grupos de biomarcadores más específicos que aseguren una mejor caracterización de diferentes subtipos moleculares de cáncer. Se trata, en suma, de encontrar firmas propias de diferentes clases de tumores, tarea que se complica una vez alcanzado determinado nivel de profundidad, dado que muchos cánceres comparten gran parte de las alteraciones génicas importantes.

Esta sección propone un sistema de predicción que se aplica sobre un caso de estudio para la discriminación de subtipos de cáncer de pulmón, en particular sobre dos subtipos de tumores de pulmón de células no pequeñas: el adenocarcinoma de pulmón (LUAD) y el carcinoma de pulmón de células escamosas (LUSC). El diagnóstico de esta clase de cáncer es complicado, ya que se ve afectado por los problemas mencionados, en especial por la baja diferenciación de muchos tumores de células escamosas en el momento del diagnóstico [120]. Hoy día aún no hay un consenso claro sobre los biomarcadores de elección para la discriminación de estas dos subclases.

Si bien, como se ha explicado anteriormente, las propuestas de análisis en este capítulo tratan de satisfacer todos los objetivos generales expuestos en el Capítulo 1 (al menos en cierto grado), de entre ellos, los objetivos prioritarios de esta propuesta son:

- Minimizar el número de biomarcadores utilizados en diagnóstico maximizando la capacidad de clasificación, estableciendo grupos de genes reducidos y aptos para su estudio molecular en un laboratorio.
- Diseñar metodologías capaces de realizar clasificaciones complejas entre clases de tumores con características moleculares comunes, más allá de la mera comparación de muestras tumorales y control.
- Utilizar métodos que permitan descubrir biomarcadores complementarios, es decir, que no sean redundantes respecto a otros genes seleccionados a la hora de clasificar y que, juntos, aporten un elevado poder de clasificación.

En cuanto al objetivo siguiente, se aborda fundamentalmente la idea de alcanzar un sistema con capacidad de generalización frente al problema de los bajos tamaños de muestra:

- Establecer metodologías de selección con capacidad de generalización y que sean robustas frente a la inestabilidad, priorizando la significancia biológica de los resultados en tamaños muestrales reducidos.

En cuanto al contexto médico de este problema, cabe destacar que el cáncer de pulmón es la primera causa de muerte por cáncer en todo el mundo [136]. Como ocurre en la mayoría de síndromes proliferativos, comprende un grupo heterogéneo de patologías sobre las que tenemos un conocimiento limitado pues, aunque presenten ciertas características comunes que nos permitan establecer categorías, la variabilidad génica natural y la gran cantidad de genes implicados se traduce en que no existen dos cánceres iguales. El cáncer de pulmón de células no pequeñas es además el más prevalente entre ellos, representando cerca del 90 % de todos los canceres de pulmón [29, 53, 114]. En el proceso diagnóstico, la capacidad para distinguir los adenocarcinomas de los carcinomas de células escamosas es crítica, y esto repercute en el tratamiento [161], pues su comportamiento ante este es diferente [165]. Sin embargo, existe cierta división e incertidumbre acerca de qué marcadores deberían ser utilizados en clínica [127]. Se han propuesto diferentes *sets* de marcadores, cuya eficacia varía en función de con qué otros estén combinados. En este sentido, como se ha mencionado anteriormente, nos encontramos la dificultad de que existen notables diferencias metodológicas en diversas investigaciones [161] que,

junto con la reducida cantidad de pacientes que componen la muestra y la complejidad e inestabilidad molecular del propio cáncer, se traducen en una baja uniformidad en los biomarcadores seleccionados para investigación y diagnóstico entre los investigadores.

Actualmente se destinan muchos esfuerzos a entender los procesos moleculares que rigen las células cancerígenas de tejidos pulmonares, pues son fundamentales para el desarrollo de nuevos tratamientos y para el diagnóstico. Desafortunadamente, la emisión de un diagnóstico precoz, a menudo el arma más eficaz contra un tumor, no es posible en la mayoría de los casos. Alrededor de un 40 % de los pacientes diagnosticados de algún tipo de cáncer de pulmón presentan tumores en estadio IV [160]. Esto no debe ser sino un aliciente para profundizar en nuestro conocimiento sobre el funcionamiento molecular de este tipo de patologías, así como en la búsqueda de nuevos grupos de marcadores y sistemas de diagnóstico más precisos.

Este capítulo presenta un sistema CBR que integra un proceso de selección de características que emplea *Gradient Boosting Regression Trees* (GBRT). El razonamiento basado en casos se ha utilizado en investigación médica debido a su capacidad de aprendizaje y de adaptación a diversos problemas de diagnóstico mediante el uso de información obtenida de experiencias previas [76], así como por su fácil aplicación [118].

Mediante la presente propuesta se persigue proporcionar un diagnóstico preciso de los subtipos de cáncer de pulmón de células no pequeñas. Para seleccionar los genes que nos permitan realizar la discriminación de clases, se debe realizar un proceso de selección de características. En este caso, se propone un sistema basado en GBRT, que no ha sido utilizado previamente en el contexto del razonamiento basado en casos hasta donde se ha podido comprobar. Como se expone en el presente capítulo, este algoritmo, combinado con algunas técnicas de preprocesado y filtrado, puede emplearse para identificar un pequeño conjunto de genes que nos permita distinguir los mencionados subtipos de cáncer. Por tanto, el objetivo fundamental de este *framework* reside en obtener un grupo reducido de genes significativos que sirvan de criterio para un sistema de clasificación que aprenda y se adapte a la entrada de nuevos datos de forma automática.

Las propiedades del razonamiento basado en casos lo hacen especialmente adecuado para este tipo de problema, teniendo en cuenta la existencia de una cantidad de muestra reducida o de bases de datos pequeñas que son actualizadas con el tiempo, así como la aparición de nuevas bases de datos que podrían ser combinadas. En este sentido, no

se trata exclusivamente de la capacidad para recuperar casos anteriores, sino también de otras cualidades del paradigma CBR en general y de esta propuesta en particular como, por ejemplo, permitir la corrección de diagnóstico en los casos analizados y contemplar dicha corrección para el aprendizaje, la capacidad para maximizar la precisión con tamaños muestrales bajos, etc.

Como se ha expuesto con anterioridad, la variabilidad natural en la expresión génica y las diferencias metodológicas para obtener los conjuntos de datos constituyen una dificultad de cara a la capacidad de generalización de muchas propuestas. Es decir, el potencial de predicción de la mayoría de sistemas de clasificación disminuye drásticamente cuando se evalúa sobre un nuevo *dataset*, lo que se agrava notablemente a causa de los bajos tamaños muestrales. Por tanto, y como se analizará en el apartado de resultados experimentales, el uso de un enfoque de CBR y su idoneidad para hacer frente a este problema queda justificado por los resultados obtenidos, ya que la capacidad de clasificación de la propuesta aumenta con la llegada de nuevos casos, mejorando en la medida que lo permite el ritmo de adquisición de datos. El sistema propuesto presenta también otras características importantes en cualquier método de selección, como su robustez frente al ruido y características no informativas. Además, en esta clase de problemas, para facilitar la aplicabilidad del sistema y la interpretación biológica, es importante utilizar métodos que no transformen el *dataset* combinando los datos referentes a diferentes genes. Asimismo, se ha procurado que el modelo de selección permita una sencilla adaptación o reimplementación para adaptarse, de ser necesario, a un sistema de trabajo de procesamiento en paralelo.

3.1.1. Estructura y metodología del sistema CBR

Esta sección presenta el método de selección de características y el sistema CBR en su conjunto. En primer lugar, se expone la organización del *framework*, cuya estructura general puede observarse en la Figura 3.2. Posteriormente, se explica en los siguientes apartados cada una de las etapas del proceso.

Este sistema se ha construido de tal forma que realiza dos etapas de extracción de características de forma sucesiva durante el proceso de entrenamiento. En la primera,

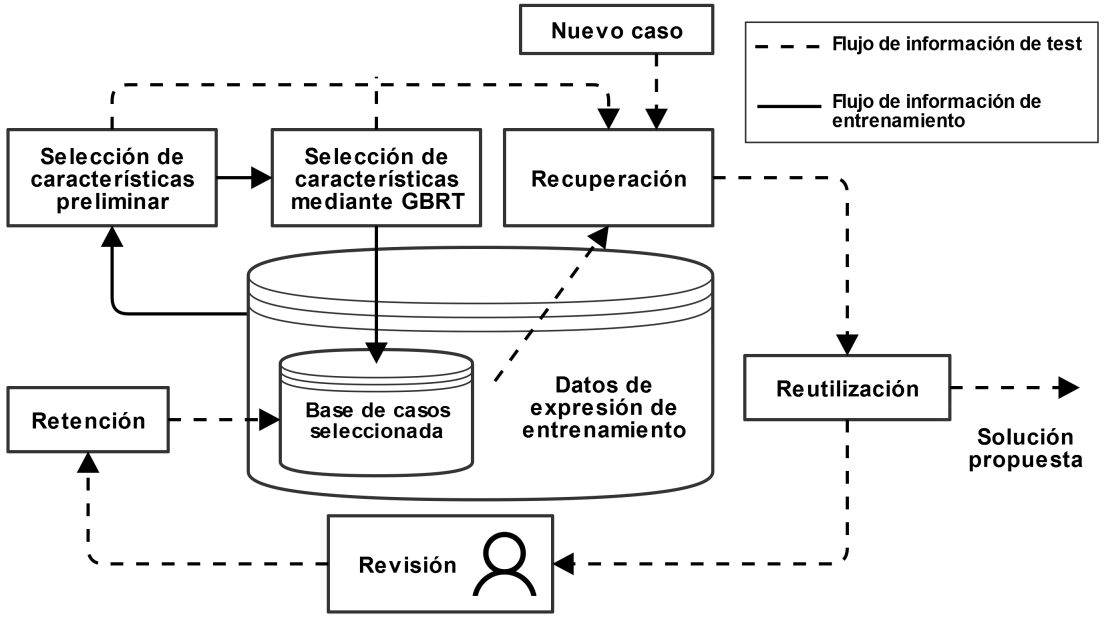

FIGURA 3.2: Esquema general del sistema CBR propuesto.

se descartan todos aquellos genes que no presentan una correlación con uno de los grupos, o que no varían su expresión de forma estadísticamente significativa de acuerdo a la pertenencia a una u otra clase de tumor (de entre las estudiadas en cada caso). Acto seguido, se aplica un segundo proceso de selección basado en GBRT encargado de seleccionar un *set* muy reducido de genes con sus respectivos valores de expresión, que permitan alcanzar y mantener una alta precisión durante la clasificación. Las etapas de selección almacenan los nombres de los genes seleccionados indicándonos, cuando llega un nuevo caso, qué genes se consideran no informativos para la predicción y deben ser descartados antes de que se inicie la fase de recuperación. Una vez finalizado el proceso de selección, la base de casos del CBR se construye con los niveles de expresión de los genes seleccionados. Para recuperar los casos durante la fase de recuperación se utilizará un método simple basado en distancia, una variante pesada del algoritmo *k-nearest neighbours* (kNN). Así, el sistema es capaz de aprender de los casos revisados y almacenados. Cabe destacar que, como se observa en la Figura 3.2, los flujos de información durante las etapas de entrenamiento y evaluación están representados por separado. En el caso de estudio, la eficacia de esta propuesta se ha comparado con otras técnicas de selección de genes en datos de expresión génica.

A continuación se describen los distintos procesos del *framework*.

3.1.1.1. Preprocesamiento y selección preliminar

En esta primera etapa, el objetivo es reducir la dimensionalidad del conjunto de datos y seleccionar genes que presentan una relación con la condición de estudio o, dicho de otro modo, descartar aquellos no relacionados con la diferencia entre los dos subtipos de cáncer estudiados. En primer lugar, se normaliza el *dataset*. Este punto puede variar en función de la tecnología utilizada para obtener los niveles de expresión génica. Puesto que en el caso de estudio que nos ocupa los datos provenían de biochips, se empleó el algoritmo RMA (*Robust Multichip Average*) [72]. Para cumplir con el objetivo mencionado, se aplicó a continuación la prueba no paramétrica Mann-Whitney. Este test ha sido ampliamente utilizado en la bibliografía para filtrar genes diferencialmente expresados [47] y se emplea aquí como punto de partida para realizar un filtrado inicial, cuando aún el *dataset* cuenta con un gran número de características. Se establece una hipótesis nula que sostiene que las muestras de dos grupos de estudio provienen de la misma población (presentando una distribución similar), mientras que la alternativa sostiene que provienen de diferentes poblaciones. Se utiliza para esta prueba un p-valor de corte de 0,05. Por tanto, cuando el p-valor es menor rechazamos la hipótesis nula, obteniendo genes cuyos valores de expresión se explican de forma significativa por la pertenencia a dos poblaciones distintas. Cabe decir, que antes de rechazar la hipotesis nula, el p-valor de corte escogido se somete a un proceso de ajuste mediante el método Benjamini & Hochberg, para disminuir las probabilidades de cometer errores tipo I (falsos positivos).

Algorithm 1 Selección de características preliminar

Require: dataset $T = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}$ of microarray samples $x_i \in \mathbb{R}^d$ and their corresponding class labels $y_i \in [0, 1]$.

1: $pvalues \leftarrow MannWhitney(T)$
2: **for** $i := 1$ **to** n **do**
3: $\quad x'^{(i)} \leftarrow [x_j^{(i)}]_{pvalues_j \leq 0{,}05}$
4: **end for**
5: $vars \leftarrow Var(\{x'^{(i)}, i = 1, \cdots, n\})$
6: $Th \leftarrow \frac{1}{2}(max(vars) - min(vars)) + min(vars)$
7: **for** $i := 1$ **to** n **do**
8: $\quad x'^{(i)} \leftarrow [x_j'^{(i)}]_{variances_j \geq Th}$
9: **end for**
10: **return** $\{(x'^{(i)}, y^{(i)}), i = 1, \cdots, n\}$

Ensure: The dataset T is returned preserving only the selected genes.

En este punto ya tenemos un conjunto de genes filtrados de acuerdo al objetivo de selección. Sin embargo, esta cantidad es aún muy elevada teniendo en cuenta que tratamos de maximizar la capacidad predictiva de la propuesta empleando el menor número de características posible. Así pues, se seleccionan los genes procurando no sólo que sus valores de expresión presenten una relación con una de las clases de estudio sino que además, desde un punto de vista clínico, nos interesa que estos valores presenten grandes variaciones. Por ello, y para reducir más la cantidad de genes, se calcula la varianza de los niveles de expresión de los genes seleccionados tras aplicar el test de Mann Whitney. De esta forma buscamos obtener genes que presenten grandes cambios de expresión (quizás fenotípicamente notables o puede que más fácilmente medibles), cuya repercusión en los procesos moleculares sea más fácilmente diferenciable desde un punto de vista experimental. Para filtrar los genes en este punto se establece un *threshold* a partir del cual se descartarán aquellos con valores de varianza inferiores. Se utiliza para dicho propósito el punto medio de todo el rango de valores que toman las varianzas de los genes. En definitiva, aquellos genes que presenten una varianza mayor que el valor de corte son escogidos para la segunda fase, que implica un proceso de selección más refinado y el método de selección central de la propuesta. Este proceso de selección preliminar se formaliza en el Algoritmo 1. Al final de estas dos subfases, contamos, por tanto, con un conjunto de genes significativos cuyos valores de expresión tienen una varianza elevada de acuerdo a la muestra disponible.

3.1.1.2. Selección de características con *Gradient Boosting regression Trees*

La etapa de selección preliminar descarta una elevada cantidad de genes, principalmente aquellos cuyos niveles de expresión no presentan una relación con las condiciones de estudio. Sin embargo, esta cantidad sigue siendo excesiva (especialmente para poder realizar cualquier comprobación experimental) y redundante; además se halla lejos de una pequeña firma de biomarcadores apropiada para la predicción. Por este motivo es necesario un filtrado más específico, que reduzca al máximo el número de genes necesarios para la clasificación y favorezca la aplicabilidad del método. La segunda etapa de selección de características llevada a cabo por el sistema CBR se basa en el algoritmo *Gradient Boosting Regression Trees* (GBRT) [52]. Este algoritmo, más comúnmente

empleado en tareas de clasificación que de regresión, se ha utilizado en este caso y por primera vez en el contexto de sistemas CBR como método de selección de características. La idea principal de GBRT consiste en construir, a partir de modelos de predicción débiles, un modelo de predicción más robusto como un *ensemble* de los primeros, constituyendo un "comité" con más poder de clasificación [50]. La salida del modelo de GBRT se calcula, por tanto, mediante el sumatorio pesado de M modelos de predicción débiles:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \tag{3.1}$$

donde $h_i(x)$ es la salida del i-ésimo modelo de predicción débil y γ_i es una constante que controla la contribución de cada modelo a la salida general conocida como *step length*. Es preciso señalar que, a diferencia de otros métodos de *ensemble*, en el caso de GBRT, los modelos (en este caso árboles de regresión), no se entrenan para obtener una misma salida, sino que son construidos de forma secuencial para predecir el error del modelo anterior, con el objetivo de aumentar su precisión:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{3.2}$$

De esta forma la salida de un árbol pretende aproximar lo máximo posible la diferencia entre la predicción del árbol anterior y la salida deseada.

Ahora bien, puesto que en este caso no se va a utilizar con un fin de clasificación, sino para un proceso de selección de características relevantes, se analizarán los árboles resultantes, una vez construido el modelo, con el objetivo de determinar la importancia relativa de cada característica. Este proceso depende del criterio de división empleado durante la construcción de los árboles. En este caso usamos el error medio cuadrático o MSE (*Mean Squared Error*), uno de los criterios más ampliamente utilizados en la bibliografía. El MSE de un nodo t en un árbol de regresión dado se define como una función de la salida deseada para cada muestra en ese nodo:

$$MSE(t) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.3}$$

donde $\{y_1, y_2, \cdots y_n\}$ son las salidas deseadas para las n muestras en el nodo t e $\bar{y}$ es su valor medio.

Sólo en el primer modelo de predicción las salidas deseadas son las etiquetas de clase de cada muestra, mientras que los subsiguientes tratan de predecir el error del modelo anterior. En un nodo t que presente un buen criterio de división, los nodos hijos tendrán un MSE más bajo que dicho nodo. Podemos medir la eficacia de la división en un nodo dado como la reducción que se observa en el MSE de la siguiente forma:

$$\begin{aligned} MSEdecrease(t) = {} & samples(t) \cdot MSE(t) \\ & - samples(leftChild(t)) \cdot MSE(leftChild(t)) \\ & - samples(rightChild(t)) \cdot MSE(rightChild(t)) \end{aligned} \tag{3.4}$$

En este punto cabe resaltar que el MSE de cada nodo hijo se ha pesado en función del número de muestras en ese nodo. Al establecer una medida de calidad para la división ya podemos estimar la importancia de una característica en un modelo de predicción débil. Para hacerlo, se emplea una fórmula modificada de la propuesta por Friedman [52], adaptada para funcionar con el MSE, de tal forma que podríamos expresar formalmente la importancia de cada característica j en el árbol h_m:

$$I_j(h_m) = \frac{1}{n} \sum_t MSEdecrease(t) \cdot 1(v_t = j) \tag{3.5}$$

donde el sumatorio se realiza sobre los nodos terminales del árbol h_m, v es la variable de división evaluada en el nodo t, y n es el número de muestras en la raíz del nodo h_m. Para facilitar la comprensión de este punto, tomemos como ejemplo el primer árbol generado al construir el modelo sobre nuestro caso de estudio (Figura 3.3), expuesto más adelante. Según este árbol, aplicado sobre un *dataset* que consta (como se verá más adelante) de 58 muestras, solo hay dos características o genes con una importancia mayor que cero, representadas por los niveles de expresión de CLCA2 y AKR1B10. De acuerdo con la ecuación anterior, la importancia de estos dos genes sería $I(h_0)_{CLCA2} =$ 0,1659 e $I(h_0)_{AKR1B10} = 0{,}0198$. Ahora bien, para poder obtener la importancia de las características en todo el conjunto de árboles que componen el modelo, se calcula la media de los valores de importancia de cada característica obtenida en cada modelo:

$$\bar{I_j} = \frac{1}{M} \sum_{m=1}^{M} I_j(h_m) \tag{3.6}$$

Posteriormente las importancias relativas se normalizan de tal forma que sumen 1 y se

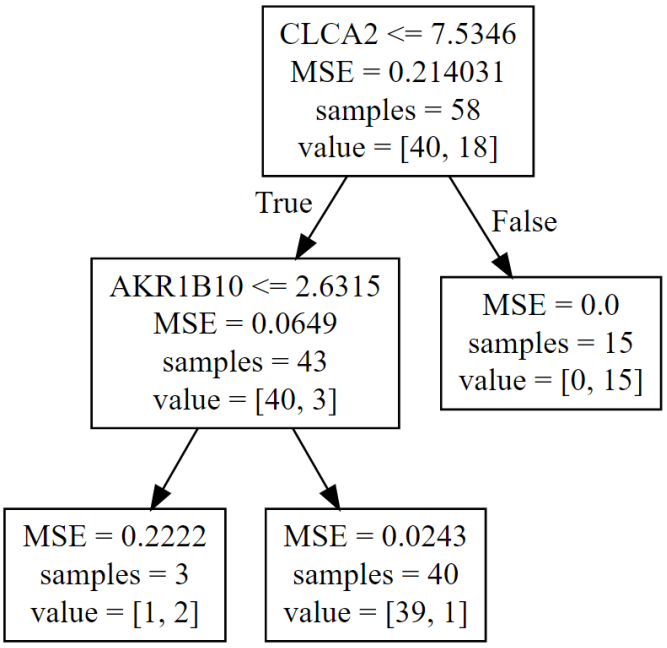

FIGURA 3.3: Ejemplo de árbol simple de profundidad dos. En cada nodo figuran la distribución de las muestras y el MSE.

disponen en un vector $\vec{I}$ con una longitud igual al número de características original. Se emplea así este vector de importancias de las características para refinar la representación de casos anteriormente descrita. Es en este punto donde, de acuerdo con la información obtenida, se debe decidir cuántas características conservar. Para ello, se siguen dos vías:

- En una de ellas seleccionamos todas aquellas características cuya importancia no fuese cero. Cabe resaltar que las características son pesadas y seleccionadas, por tanto, según su importancia. En este caso la representación final $x\prime$ de una muestra x con características $[x_j]_{j=1}^{d}$ sería:

$$x\prime = [\vec{I}_j x_j]_{\vec{I}_j \neq 0} \tag{3.7}$$

- La otra opción consiste en seleccionar manualmente las primeras k-ésimas características de acuerdo a cada necesidad, donde la representación final $x\prime$ de una muestra x se define por:

$$x\prime = [\vec{I}_j x_j]_{\vec{I}_j \geq th} \tag{3.8}$$

donde th es un *threshold* adaptativo igual a los k-ésimos valores más altos en $\vec{I}$. En función del valor de k en cada caso, este procedimiento es susceptible de reducir la tasa de acierto del sistema, o de no minimizar la redundancia. Sin embargo,

en el caso de estudio se proporciona una comparación entre ambos métodos y se explora empíricamente la variación de la tasa de acierto ante el cambio de número de características seleccionadas. Además, desde un punto de vista práctico puede ocurrir que todas las características con importancia no cero constituyan un número demasiado elevado para realizar procedimientos diagnósticos y o experimentales en un laboratorio.

3.1.1.3. Recuperación y reutilización

En esta fase tiene lugar la construcción de la base de casos, completado ya el proceso de selección de características. La base de casos está compuesta de un conjunto de n pares $CB = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, de tal forma que x_i es la representación del caso de una muestra y e y_i su solución correspondiente. En aquellos casos de clasificación binarios, como el que nos ocupa, cada solución y_i es una etiqueta de clase con valor 0 ó 1. En cuanto a la recuperación de casos, se utiliza una aproximación basada en distancia.

El proceso transcurriría de modo que, cuando llega una muestra nueva no etiquetada a la base, esta es preprocesada por las etapas de selección de características para obtener una representación de caso. Acto seguido, se definen los k vecinos más cercanos a dicho caso en la base de casos. Aunque en la base de casos inicialmente hay un número reducido de estos, esta etapa presenta problemas de escalabilidad según vayan presentándose nuevos casos. Encontrar los k vecinos más cercanos de x en la base de casos requiere un tiempo de $\mathcal{O}(ndk)$, donde d es el número de características que componen el caso y n es el número de casos almacenados en la base (CB). Para afrontar este problema se han propuesto diferentes técnicas, de las cuales en esta ocasión se emplea un árbol KD [128]. Este es un método de particionado de espacio que nos permite garantizar un tiempo de consulta sublineal mientras aumenta el número de casos en la base de casos. La fase de reutilización comienza cuando se han calculado los vecinos de la muestra sin etiquetar. Es en este punto cuando se utilizan las distancias de los k vecinos más cercanos para calcular lo que llamamos vector no normalizado de pesos [61]:

$$\vec{w} = [\frac{1}{||x - x_j||_2}, j = 1, 2, \cdots, k] \tag{3.9}$$

donde $||x - x_j||_2$ es la distancia euclídea entre el caso sin etiquetar y su j-ésimo vecino más cercano. Así, la probabilidad de que x pertenezca a la clase con etiqueta l es:

$$P(y = l \mid x, CB) = \frac{\sum_{i=1}^{k} \vec{w}_j \cdot 1(y_j = l)}{\sum_{i=1}^{k} \vec{w}_j} \tag{3.10}$$

Convenientemente:

$$\sum_{l} P(y = l \mid x, CB) = 1, \tag{3.11}$$

donde la suma se realiza sobre el *set* de todas sus posibles etiquetas de clase. Al final de la fase de reutilización, la etiqueta de una muestra x sin clasificar se predice para ser la que tenga mayor probabilidad:

$$y = \arg\max_{l} \; P(y = l \mid x, CB) \tag{3.12}$$

3.1.1.4. Revisión y retención

En esta fase reside la capacidad de aprendizaje del sistema en sí. Como se ha explicado, cuando una nueva muestra sin etiquetar es presentada al sistema, es sometida a los procesos de selección de características, entrenados antes de que se construyese la base de casos. Posteriormente, la fase de recuperación y reutilización se encarga de realizar una predicción etiquetando la muestra como una de las posibles clases de estudio. No obstante, para que el sistema mejore su eficacia, es preciso proporcionarle información acerca de cuáles de sus predicciones son correctas o incorrectas. Para ello hay que utilizar técnicas de diagnóstico alternativas. En el caso de estudio descrito más adelante se trata principalmente de técnicas de imagen y moleculares. Entre las primeras se cuentan la tomografía computerizada, la resonancia magnética y la broncoscopia [124, 129], y en cuanto a las moleculares, se emplean principalmente la citología de esputo y la biopsia, acompañadas del estudio de marcadores específicos y factores de crecimiento para caracterizar el tumor de forma más precisa[129]. Para emitir un diagnóstico, a menudo es necesario realizar más de una prueba a cada paciente y, pese a ser costoso, es lo que permite al sistema progresar para que alcance una mayor tasa de acierto por sí solo en el futuro. Por tanto, esta fase es especialmente importante, sobre todo al principio, hasta obtener una base de casos representativa. Se trata de una inversión para que el sistema alcance su máximo potencial obteniendo una alta tasa de acierto con la expresión de

los biomarcadores seleccionados. Toda esta información permite a un experto emitir un diagnóstico. De esta forma, una vez finalizada la revisión, el caso se almacena en la base de casos junto con la solución correspondiente revisada.

3.1.2. Caso de estudio

En el caso de estudio que nos ocupa, se comprueba la eficacia de este sistema en el diagnóstico y clasificación de dos subclases de cáncer de pulmón de células no pequeñas: el adenocarcinoma de pulmón y el carcinoma de células escamosas. Con ello se evalúa también la calidad del proceso de selección de características. Puesto que este sistema utiliza la información proporcionada por un conjunto de biomarcadores como criterio de clasificación, dicha etapa de selección resulta especialmente importante para realizar una predicción fiable. Para este trabajo, se tomaron dos conjuntos de datos referentes a adenocarcinoma de pulmón y carcinoma de células escamosas. Los dos conjuntos de datos consisten en perfiles de expresión génica de tejidos provenientes de pacientes de cáncer y que han sido obtenidos mediante el uso de *micorarrays* de expresión. Ambos *datasets* se han generado a partir del mismo modelo de chip, el U133 2.0 de Affymetrix; han sido normalizados mediante el algoritmo RMA, y ambos están disponibles de forma pública en el repositorio de NCBI (*National Center of Biotechnological Information*). En uno de los dos *datasets* también se presentaba un pequeño grupo de muestras de tejido de carcinomas de pulmón de células grandes, que fue retirado del conjunto para poder realizar el análisis, puesto que no era necesario para el propósito de este trabajo. Con el objetivo de evaluar de forma científicamente rigurosa el funcionamiento de la propuesta y, concretamente, su capacidad de generalización (así como la de otros algoritmos de trabajos relacionados), se utilizó uno de los conjuntos de datos para el entrenamiento y otro para la evaluación únicamente. El conjunto de entrenamiento cuenta con 58 muestras de tejido de distintos pacientes, siendo 40 de ellas de adenocarcinoma y las otras 18 de cáncer de células escamosas [88]. El otro conjunto de datos, destinado a la evaluación, contiene 172 muestras, 66 de las cuales pertenecen a pacientes con carcinoma de células escamosas y las restantes 106 a adenocarcinomas [21].

3.1.2.1. Resultados experimentales

El objetivo de las comprobaciones experimentales que se detallan a continuación no es sólo validar la eficacia del sistema para clasificación, sino también constatar su habilidad para aprender y mejorar su tasa de acierto cuando se presentan nuevos casos no vistos anteriormente. A este respecto, entre los experimentos de evaluación llevados a cabo podríamos distinguir tres grupos. Un primer grupo está destinado a medir la eficacia y valorar la efectividad de las etapas de selección de genes junto con la de la organización CBR. El segundo está orientado a comparar exclusivamente la eficacia del proceso de selección con otros métodos. El tercer grupo compara otras técnicas de selección y clasificación empleadas con datos transcriptómicos en la bibliografía con nuestro sistema CBR en su conjunto.

Como se ha mencionado, se empleó el *dataset* de 58 muestras [88] para entrenar el sistema y el de 172 para evaluar la precisión en clasificación [21]. Con este fin no solo se ha medido el porcentaje de acierto, sino también la precisión, el *recall* y la Kappa de Cohen [10]. Se ha tomado,, a efectos de clasificación el carcinoma de células escamosas como la clase positiva. Se ha procedido de esta forma debido a que el cáncer de células escamosas se halla menos representado en la medida de tasa de acierto por su menor ocurrencia en los *datasets* de entrenamiento y evaluación. Además, este desbalanceo entre clases en estos y otros conjuntos de datos explorados de estas mismas patologías es fácilmente explicable por la menor incidencia del cáncer de células escamosas respecto al adenocarcinoma en la población mundial. Para poder valorar el progreso del sistema CBR, se simuló el proceso de revisión y retención mediante la posterior incorporación de 40 casos del *dataset* de evaluación con sus correspondientes etiquetas de clase a la base de casos. De esta forma se medía la tasa de acierto del sistema en las muestras restantes, pudiéndose observar así el impacto del CBR en la eficacia global de la propuesta. En la primera fase experimental este mismo procedimiento se llevó a cabo empleando diferentes combinaciones de los procesos de selección propuestos, inhabilitándolos en los casos pertinentes para comprender su contribución al proceso.

En la Tabla 3.1 se muestran los resultados obtenidos para diferentes combinaciones de dichos métodos, habilitando e inhabilitando distintas etapas para observar su efecto en los resultados finales. Estos resultados contemplan no sólo el entrenamiento de la base

de casos inicial sino también el efecto de la posterior incorporación de las 40 muestras retenidas.

Selección preliminar	**Selección secundaria**	**Clasificador**	**Base de casos inicial**			**40 casos retenidos**		
			Acc.	κ	Prec/rec	Acc.	κ	Prec/rec
Var & MW (53 carac.)	-	CBR (wkNN)	94.1 %	0.875	89.3/95.1 %	95.45 %	0.900	87.7/100.0 %
Var & MW (53 carac.)	GBRT (non-zero) 24 carac	CBR (wkNN)	97.0 %	0.938	95.4/96.9 %	97.7 %	0.950	93.8/100.0 %
Var & MW (53 carac.)	GBRT (top-5) 5 carac.	CBR (wkNN)	96.5 %	0.926	95.4/95.4 %	97.7 %	0.950	93.8/100.0 %
- (54,613 carac.)	-	CBR (wkNN)	95.5 %	0.901	93.9/93.9 %	96.2 %	0.919	97.9/92.1 %
- (54,613 carac.)	GBRT (non-zero) 20 carac	CBR (wkNN)	91.2 %	0.817	90.0/86.9 %	92.4 %	0.837	91.6/88.0 %
- (54,613 carac.)	GBRT (top-5) 5 carac	CBR (wkNN)	90.6 %	0.807	93.9/83.7 %	90.9 %	0.807	91.6/84.6 %

Tabla 3.1: Resultados experimentales del sistema CBR utilizando diferentes combinaciones de los métodos empleados durante la selección de características. También se incluyen los resultados referentes a la capacidad de aprendizaje del sistema después de atravesar las etapas de revisión correspondientes a la llegada de nuevas muestras.

Se puede ver que la combinación de un filtro basado en las técnicas estadísticas empleadas junto con GBRT (es decir, la versión del proceso de selección más completa) ofrece los mejores resultados tanto cuando se parte de la base de casos inicial como cuando se incluyen los retenidos, como era esperable en el momento del diseño del proceso de selección. También vemos que, incluso considerando solo las cinco características más relevantes (en este caso nuestros marcadores), el sistema sigue siendo capaz de alcanzar una elevada tasa de acierto. Esto resulta de especial interés puesto que un bajo número de marcadores aumenta la aplicabilidad del método que los utiliza, considerando que, desde el punto de vista del trabajo de laboratorio, una elevada cantidad de genes no sería viable para determinadas comprobaciones experimentales como las inmunohistoquímicas.

A continuación se lleva a cabo una segunda fase experimental destinada a comparar la eficacia de la etapa de selección de características con la de otros métodos recurrentes en la bibliografía, pero manteniendo el marco de un sistema CBR. Es decir, solo se sustituyeron las etapas de selección, lo cual también nos permite valorar, simultánea pero independientemente, por un lado el proceso de selección de características y por otro la idoneidad del sistema CBR para este problema. En la Tabla 3.2 se muestran los resultados de estos experimentos.

Método de selección	Clasificador	Base de casos inicial			40 casos retenidos		
		Acc.	κ	Prec/rec	Acc.	κ	Prec/rec
Boruta [92] 27 carac.	wkNN	96.5 %	0.925	92.4/98.3 %	96.2 %	0.917	93.7/95.7 %
Spikeslab 54 carac.	wkNN	87.2 %	0.711	66.6/100.0 %	95.4 %	0.902	95.8/92.0 %
SDA [4] 24 carac.	wkNN	96.5 %	0.926	95.4/95.4 %	93.9 %	0.867	89.5/93.4 %
Alpha-Investing [163] 24 carac.	wkNN	88.95 %	0.774	95.4/79.7 %	93.1 %	0.849	85.4/95.3 %
RFE [59] 24 carac.	wkNN	96.5 %	0.925	93.9/96.8 %	95.4 %	0.900	91.6/95.6 %

Tabla 3.2: Resultados experimentales obtenidos por el CBR valiéndose de otros métodos de selección de la bibliografía. Se incluyen igualmente los resultados referentes a la capacidad de aprendizaje del sistema después de atravesar las etapas de revisión correspondientes a la llegada de nuevas muestras.

Después tiene lugar una tercera fase experimental destinada a comparar los siguientes algoritmos, comúnmente utilizados en selección de genes en la bibliografía [41, 66], con este sistema completo: PCA (*Principal Component Analysis*), KPCA (*Kernel Principal Component Analysis*), ICA (*Independent Component Analysis*), LLE (*Locally Linear Embedding*) y *Random Forest*. En cuanto a la clasificación, se evalúan kNN (*k-Nearest Neighbours*), NB (*Naive Bayes*) y SVM (*Support Vector Machine*) [18]. Para que los resultados de tasa de acierto fueran comparables, los datos de entrenamiento y test fueron exactamente los mismos que los empleados para validar esta propuesta. Para parametrizar correctamente los distintos algoritmos, se ejecuta un proceso de selección de modelos usando una validación cruzada (*k-fold cross validation*) sobre el conjunto de entrenamiento en la que $k = 10$. En la Tabla 3.3 se muestran los resultados de esta fase, incluyendo precisión y *recall*.

3.1.2.2. Genes seleccionados

Finalmente obtenemos un conjunto de genes seleccionados por nuestro sistema CBR. La etapa de selección mediante GBRT extraía 24 genes con importancia no cero, como se puede ver en la Figura 3.4. Teniendo en cuenta que se pretende reducir al máximo la cantidad de genes empleada para la predicción de clases y el estudio experimental previo, se ha considerado que la cantidad de genes puede ser reducida a 5 sin sufrir una pérdida significativa de la capacidad de clasificación. En este caso no se sacrifica, de

M. Ext.	Clasificador	Acc.	κ	Prec/rec
PCA (k = 24)	SVM(lineal) C=0.001	94.7 %	0.887	89.3/96.7 %
PCA (k = 24)	NaiveBayes	95.9 %	0.913	92.4/96.8 %
PCA (k = 24)	kNN (k=3)	95.3 %	0.901	92.4/95.3 %
FastICA (k = 24)	SVM(lineal) C=0.5	92.44 %	0.835	81.8/98.1 %
FastICA (k = 24)	NaiveBayes	95.3 %	0.901	93.9/93.9 %
FastICA (k = 24)	kNN (k=4)	92.4 %	0.835	83.3/96.4 %
LLE (k = 24)	SVM(lineal) C=0.001	94.7 %	0.888	90.9/95.2 %
LLE (k = 24)	NaiveBayes	91.2 %	0.820	95.4/84.0 %
LLE (k = 24)	kNN (k=3)	94.7 %	0.890	96.9/90.1 %
RF (300) 774 carac	SVM(lineal) C=0.001	94.7 %	0.890	93.9/92.5 %
RF (300) 765 carac	kNN (k=3)	94.7 %	0.889	92.4/93.8 %

Tabla 3.3: Resultados reflejando la eficacia de métodos alternativos de extracción de característicos comunes en la bibliografía.

hecho, tasa de acierto. Analizamos esos genes y contrastamos la información obtenida con las evidencias biológicas y otros trabajos relativos al cáncer de pulmón de células no pequeñas. Con esto se pretende averiguar si esos marcadores se han empleado de forma conjunta antes o si son coherentes con otros estudios, pese a no haber utilizado técnicas de laboratorio. En la Figura 3.5 se muestran los niveles de expresión de los marcadores seleccionados y en la Figura 3.6 se puede observar su distribución entre los dos subtipos de estudio.

Cuando se toman las cinco características con mayor importancia, el conjunto resultante de la selección está compuesto por los genes TRIM29, KRT5, SFTA2, PKP1 y AKR1B10. Finalmente, y como se ha mencionado, se recomienda usar este grupo frente a aquel resultante de utilizar todas las características con importancia no cero (24 en total). Cabe mencionar que KRT5 y TRIM29 han sido empleados como marcadores para la discriminación de LUAD y LUSC en numerosos estudios. Por su parte, PKP1 y AKR1B10 se han utilizado con menor frecuencia, aunque sí se ha comprobado la existencia de una expresión diferencial significativa. Por su parte, SFTA2 no ha sido

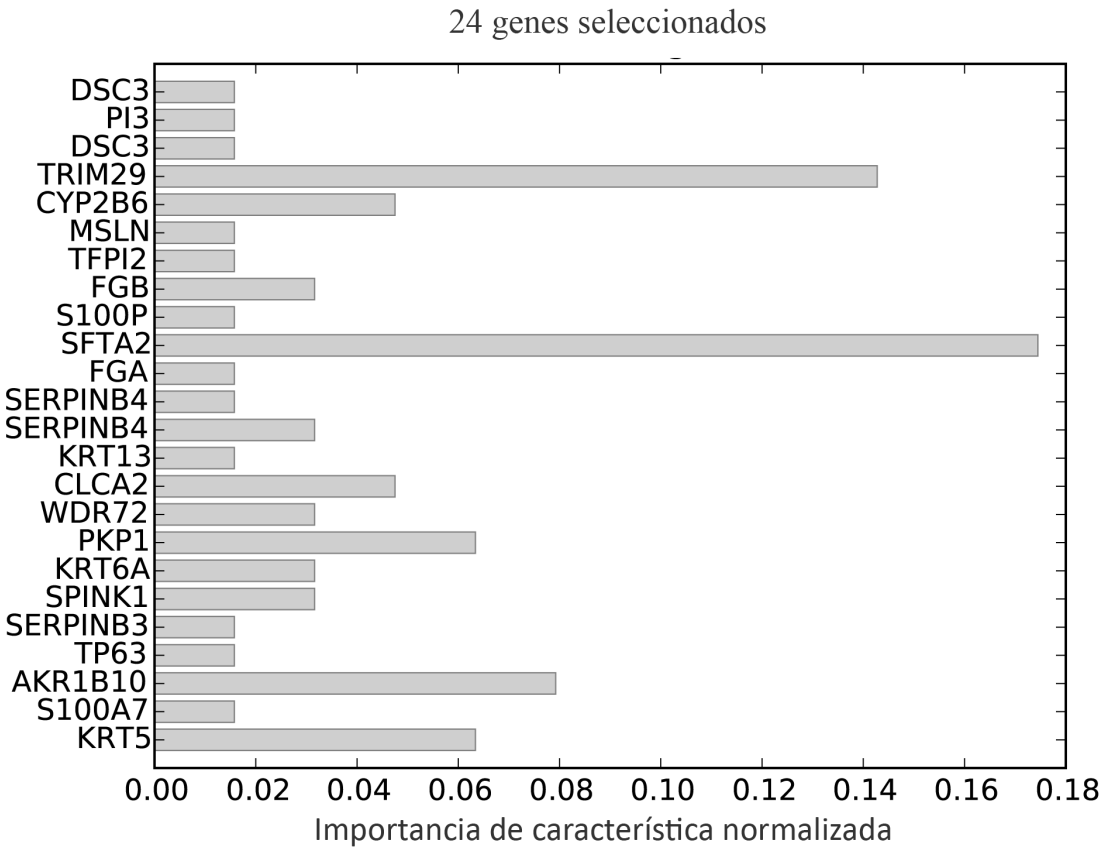

FIGURA 3.4: 24 genes con importancia no cero de acuerdo con el *ensemble* de árboles GBRT. Los genes se hallan ordenados en orden decreciente de *score* en la fase de selección preliminar.

previamente propuesto como marcador en este contexto, pero en el presente trabajo se sugiere su uso, tras comprobar que puede alcanzar un gran potencial para el diagnóstico. En cualquier caso, es la combinación de estos genes, más que su poder individual, lo que determina la eficacia del sistema.

- **TRIM29**. La familia de proteínas TRIM interviene en una serie de procesos de gran importancia para el organismo, como lo son la migración celular, la proliferación, los procesos de apoptosis y de diferenciación [154]. No es de extrañar, por tanto, que se haya comprobado la existencia de una desregulación de TRIM29 en diversos tipos de cáncer, normalmente relacionada con la proliferación y la progresión tumoral [97]. Resulta interesante que, dependiendo del contexto celular, TRIM29 puede ejercer un papel oncogénico o de supresión tumoral [97]. En cuanto al potencial de TRIM29 como marcador, cabe destacar que Zhou *et al.* [165] han

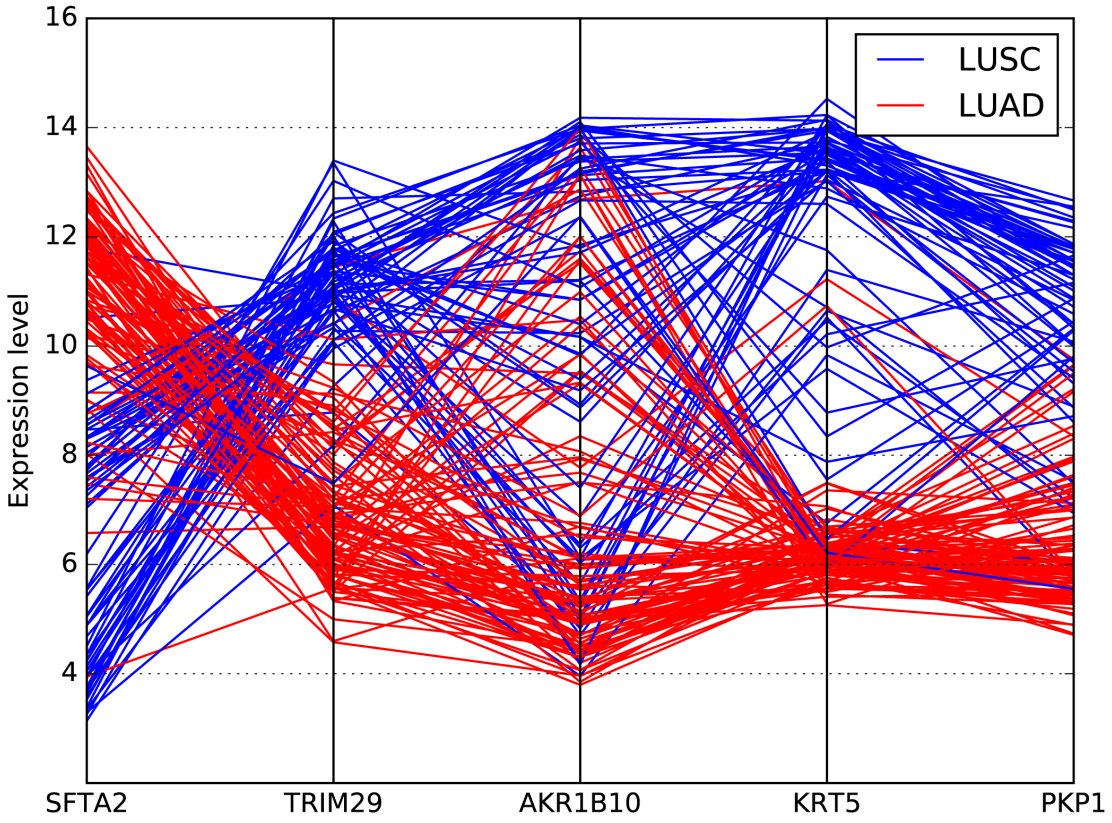

FIGURA 3.5: Gráfico de coordenadas paralelas reflejando los niveles de expresión de los 5 genes más relevantes en el conjunto de evaluación según el método de selección propuesto en este capítulo. Este conjunto contiene 172 muestras (66 de LUSC, 106 de LUAD).

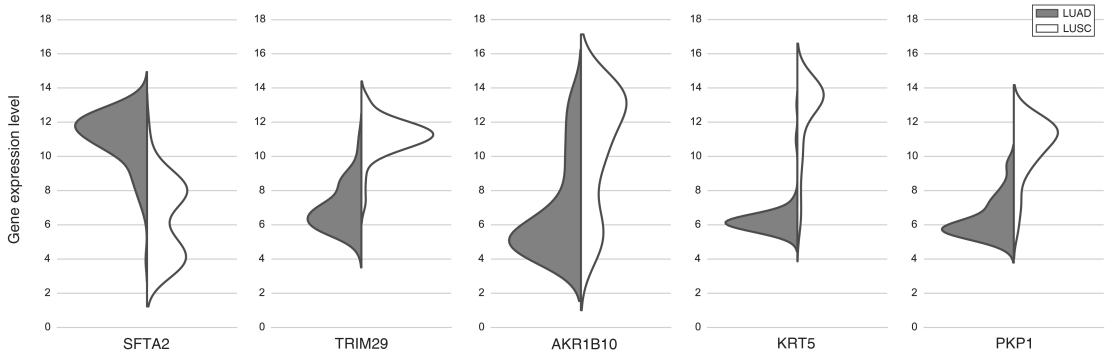

FIGURA 3.6: *Violin plots* mostrando la distribución de los 5 genes más relevantes de acuerdo al método propuesto. Los genes están ordenados de izquierda a derecha en orden decreciente de importancia.

observado una expresión diferencial significativa entre adenocarcinomas y carcinomas de células escamosas pobremente diferenciados (una de las situaciones más complejas para el diagnóstico tradicional). Dada esta información, recomiendan el uso de TRIM29 con KRT5, también seleccionado por nuestro sistema, para la

discriminación de ambos subtipos. Como se puede apreciar en la Figura 3.5 y coincidiendo con otros estudios, TRIM29 se presenta sobreexpresado en el cáncer de células escamosas [133].

- **KRT5.** La proteína codificada por este gen está implicada un proceso biológico conocido como transición epitelial mesenquimal (EMT), por el cual una célula pierde su capacidad de adhesión y su polaridad [35]. Este fenómeno se produce en diversas situaciones, como en el desarrollo embrionario. En el caso del cáncer es responsable de favorecer la capacidad de muchas células cancerígenas para migrar y con ello favorecer la metástasis del tumor. KRT5 es un marcador frecuentemente estudiado y ampliamente reconocido para la discriminación de subtipos de cánceres de pulmón de células no pequeñas. En general, las keratinas presentan patrones de expresión heterogéneos en NSCLC y su sobreexpresión se asocia generalmente al cáncer de células escamosas [133]. El hecho de que sea seleccionado por el sistema propuesto refuerza su papel como biomarcador al tiempo que habla en favor de la metodología empleada.

- **SFTA2.** En condiciones normales, SFTA2 es un gen que se encuentra altamente expresado en el pulmón [111]. Se trata de una proteína que, como su propio nombre indica, se asocia al surfactante pulmonar. Hay pruebas de su implicación en el adenocarcinoma pero, hasta donde se ha podido comprobar, no se había propuesto previamente como marcador, y no hay muchos estudios que hagan referencia a su papel en dicha patología[111]. Zhan *et al.* [161] seleccionan un grupo de genes sobreexpresados en adenocarcinoma de pulmón, entre los que figura este gen. No obstante, en las aproximaciones inmunohistoquímicas no lo utilizan porque no se ha conseguido obtener hasta la fecha un anticuerpo primario válido para la proteína correspondiente. Esto quiere decir que no se podía marcar la proteína y medirla a partir de muestras tisulares. Sin embargo, es destacable el hecho de que SFTA3, una proteína de la misma familia, sí se ha utilizado como marcador en algunos estudios [161]. A este respecto, debemos considerar que los genes que codifican las proteínas asociadas a surfactante son marcadores de un linaje celular AT2 (*Alveolar Type 2 lineage*), que comparte muchas características moleculares con las células de adenocarcinoma [81]. Esto sugiere que las células tumorales de muchos adenocarcinomas podían tener relación con este linaje celular. De hecho, tras consultar la literatura al respecto, se ha observado que recientemente, Xu *et*

al. [155] han concluido que las células AT2 son las predominantes iniciadoras de un tipo de adenocarcinomas inducidos por la proteína K-Rasg12d [155]. Por este motivo, sería interesante estudiar más en profundidad el comportamiento de la expresión de SFTA2 en el adenocarcinoma. Posteriormente a la realización y publicación de este análisis, Xiao *et al.* [153] han detectado y propuesto nuevamente SFTA2 como gen alterado con alto potencial para su uso como marcador en este mismo problema, reforzando así los resultados del presente estudio.

- **AKR1B10.** Aunque este gen se ha encontrado sobreexpresado en diferentes tumores pulmonares de células no pequeñas, dicha alteración es significativamente más común en el caso del cáncer de células escamosas [77]. Muy a menudo se ha relacionado este gen con los carcinomas de células escamosas asociados al tabaquismo [54]. Este gen, según podemos observar en la Figura 3.5, aparenta ser el menos fiable del grupo para la discriminación de las dos subclases del estudio. Sus niveles de expresión, aunque parecen indicar cierta tendencia, nos permiten ver que hay una superposición significativa de los niveles de expresión entre las clases. Es decir, que existen pacientes de adenocarcinoma que presentan el gen sobreexpresado, mientras que en otros con cáncer de células escamosas se halla subexpresado. Esto puede, en efecto, sugerir la existencia de subgrupos dentro de cada clase, por lo que sería interesante disponer de información acerca de los hábitos de tabaquismo de los pacientes, entre otros datos. No obstante, y si bien no es una característica con gran potencial predictivo por sí solo, e incluso pudiera parecer que su utilidad es limitada respecto a otros marcadores, se ha podido comprobar que es la combinación con otros genes del conjunto seleccionado lo que hace de AKR1B10 un gen útil en este grupo. Este gen está aportando una información para la clasificación que se complementa con la de otros marcadores y no es redundante. La selección de este tipo de características, como se explicará más adelante, ocurre gracias al empleo de GBRT en la etapa de selección.

- **PKP1.** La placofilina 1 o PKP1 es una proteína desmosomal. Pese que no ha acaparado la atención de muchos estudios, sí es cierto que cada vez son más las investigaciones referentes al papel de las proteínas de unión y se conoce su influencia en la proliferación celular y en la metástasis [130], por lo que no resulta extraño que pueda desempeñar un papel importante como marcador. De hecho, Sánchez *et al.* [133] y Kuner *et al.* [89] han observado una sobreexpresión de este

gen en carcinomas de células escamosas, aunque no se haya ido más lejos en la investigación de su potencial como biomarcador.

En definitiva, mediante el uso de estos genes, el sistema propuesto es capaz de predecir el subtipo tumoral de los pacientes de cáncer de pulmón de células no pequeñas alcanzando una elevada tasa de acierto, aprendiendo de su experiencia y haciendo frente a algunas de las dificultades más habituales de la investigación de esta área.

3.1.3. Discusión y conclusiones

El sistema presentado en esta investigación es un *framework* CBR que utiliza datos transcriptómicos para realizar clasificación de subtipos de cáncer. Consta de dos etapas de selección de características que persiguen reducir el número de genes considerados por el sistema para predecir la clase a la que pertenece cada muestra. La propuesta se ha evaluado sobre datos de dos clases de cáncer de pulmón de células no pequeñas.

En relación con el primer objetivo específico de la propuesta, que reside en lidiar con el problema de los bajos tamaños muestrales del área, se ha comprobado la capacidad del *framework* para afrontar este problema. La propia naturaleza del paradigma CBR hace de este un sistema adecuado para ello. Incrementar de forma gradual una base de casos no sólo permite mejorar la predicción con la llegada de nuevas muestras disponibles, sino que es bastante aproximado a la situación real: un pequeño conjunto de muestras de pacientes que va aumentando con el tiempo. Los resultados de la primera y segunda fase experimental verifican que el sistema CBR es capaz de aprender y mejorar su tasa de acierto con la incorporación de nuevos casos a la base de casos sin necesidad de volver a ejecutar la fase de selección de características cada vez.

En cuanto al objetivo de minimizar el número de marcadores maximizando la precisión, cabe mencionar que una de las características más destacables de esta propuesta es que los genes no se seleccionan individualmente, sino por su relevancia combinada. Esto quiere decir que, aunque un gen no sea relevante por si mismo, su combinación con otros puede aumentar la tasa de acierto del sistema. Esto es posible gracias a la etapa de selección realizada por GBRT, que, al descubrir interacciones no lineales entre características [52], permite seleccionar un grupo de genes que funcionen bien juntos para

la predicción de clases. Así pues, aunque a menudo se haya reducido a eso, no se trata solo de seleccionar genes con un gran potencial de clasificación. Si bien encontrar dichos genes es importante, para alcanzar la máxima tasa de acierto posible es probable que algunos de los genes seleccionados no se escogiesen como marcadores en otro contexto, o de forma individual. En este sentido, se ha alcanzado una selección respetuosa con el concepto de "grupo" para clasificación.

3.2. Extracción de conocimiento en selección mediante un *framework* híbrido modular basado en el cálculo de puntos frontera.

Desafortunadamente, dentro del área de la selección de genes, la pugna por presentar tasas de acierto competitivas a menudo tiene como consecuencia que la extracción de conocimiento se torna en realidad colateral a la clasificación. La especificidad de un sistema automático orientado exclusivamente a solucionar problemas de clasificación concretos ofrece ciertas ventajas, al tiempo que puede resultar más opaco desde el punto de vista de la extracción de conocimiento. Aunque la propia selección genere conocimiento, resulta interesante proponer métodos que aumenten nuestra comprensión, al menos, del conjunto de datos. A este respecto, esta sección propone un nuevo *framework* híbrido de selección de genes informativos que emplea técnicas de diversa naturaleza partiendo de una fuerte base en el *clustering*. Este *framework* persigue proporcionar mayor flexibilidad para establecer diferentes objetivos de selección y permitir estudios de distinto tipo, que aumenten nuestra comprensión del tumor en cuestión. Presenta además un algoritmo de cálculo de fronteras de *clustering*, que constituye una parte fundamental del criterio de selección. En los últimos años se ha incrementado el uso de técnicas híbridas en el campo de la selección de genes. En este sentido, existen muchas propuestas centradas en la clasificacion, mostrando gran variedad de resultados. Así, como se expone el el Capítulo 2, realmente la selección de genes en cáncer asume, como es comprensible dada la limitación tecnológica, varios riesgos teóricos. Si bien algunos son difícilmente salvables, al menos hasta que progrese nuestra capacidad de analizar proteínas y de hacer mediciones en un contexto temporal, es necesario plantear metodologías que nos

proporcionen resultados a partir de los que se puedan emitir interpretaciones biológicas. De hecho, la mera selección de genes en cáncer permite extraer conocimiento, aunque este se halle fundamentalmente orientado a la categorización de tipos de cáncer, a su diagnóstico y también, en cierta medida, al tratamiento. Sin embargo, se trata, en definitiva, de caracterizar un tejido tumoral. La selección a menudo genera conocimiento, que puede ser utilizado en el ámbito de la farmacogenómica y el diagnóstico en un plazo relativamente reducido, pero, desde el punto de vista de la comprensión molecular del cáncer, un gen de forma individual aporta poca información, pues son en realidad las rutas metabólicas y de señalización las que rigen la dinámica celular. Como consecuencia, en los últimos años ha aumentado el interés investigador en encontrar formas de integrar diferentes datos "ómicos". Dadas las necesidades expuestas y, en relación a los objetivos establecidos, esta sección aborda fundamentalmente las siguientes tareas de entre los objetivos principales:

- Diseñar metodologías capaces de extraer conocimiento biológico en el contexto del cáncer, persiguiendo obtener información sobre la implicación de genes que puedan actuar como dianas terapéuticas y permitiendo el estudio de factores en relación con la expresión génica.
- Minimizar el número de biomarcadores utilizados en diagnóstico maximizando la capacidad de clasificación, estableciendo grupos de genes reducidos y aptos para su estudio molecular en un laboratorio.
- Establecer metodologías de selección con capacidad de generalización y que sean robustas frente a la inestabilidad, priorizando la significancia biológica de los resultados en tamaños muestrales reducidos.

En cuanto a este último objetivo, se persigue fundamentalmente la obtención de resultados que sean biológicamente significativos. Así, a diferencia del *framework* anterior, en la presente sección se aborda el problema marcado por la hipótesis de investigación desde la perspectiva de dotar a un *framework* de análisis de cierta flexibilidad, no sólo para permitir el desarrollo de diferentes estudios más allá de la mera detección de tejido tumoral, sino para alcanzar una comprensión más profunda del *dataset*, priorizando la extracción de conocimiento frente al diagnóstico. Pese a ello, durante el análisis del caso

de estudio se realiza también un experimento de clasificación para evaluar los subconjuntos de marcadores seleccionados en la detección de tumores. Para validar esta propuesta, y a la vez investigar acerca de la posibilidad de usar el cálculo de fronteras como método de extracción de conocimiento, se ha centrado el estudio en analizar tejidos de adenocarcinoma ductal pancreático (PDAC). De hecho, tanto esta como la próxima sección centran sus casos de estudio en tejido pancreático, motivadas por la gran complejidad que envuelve este tipo de cánceres.

PDAC es uno de los cánceres más agresivos que, al igual que en el caso del cáncer de pulmón, se caracteriza en el ámbito médico por un diagnóstico tardío, dada la falta de síntomas tempranos [12, 37]. Se trata de la cuarta causa de muerte por cáncer en el mundo y se prevé que en el año 2020 se sitúe en segundo lugar [143]. La supervivencia media es de menos de 6 meses y la supervivencia a 5 años vista es inferior al 5 % [64]. De hecho, entre un 60 y un 70 % de los pacientes ya presentan metástasis cuando se les diagnostica. Estas terribles cifras han impulsado gran cantidad de estudios centrados en el microambiente tumoral que han incrementado notablemente el conocimiento del PDAC en los últimos años. Aun así, todavía estamos lejos de desarrollar tratamientos eficaces. Por tanto, la caracterización precisa de PDAC es de gran importancia en el ámbito médico. Su diagnóstico tardío está íntimamente ligado a la presencia de un estroma desmoplásico de gran volumen que presenta una relación muy dinámica con el tumor, lo que reduce de forma considerable la eficacia de cualquier tratamiento[17]. De hecho, se ha observado que, comúnmente, el estroma actúa como una barrera protectora del tumor. Al igual que otros cánceres se trata de un problema multifactorial que, como es habitual, tiene cierta relación con la edad y la senescencia celular. No obstante, en el PDAC esta relación tiene aparentemente más peso que otros factores. De hecho, el 85 % de los pacientes diagnosticados tienen más de 65 años, y una edad media de 73 [84].

Mediante la aplicación del *framework* híbrido a este caso específico se pretende no sólo comprobar su eficacia y la del papel de genes frontera en la selección de genes, sino también averiguar si existe una influencia de la edad en los niveles de expresión de determinados genes relacionados con el cáncer, y ver con ello si supone algún cambio en la gravedad de la enfermedad. Se pretende descubrir, en definitiva, si existen genes relacionados con PDAC que presenten al mismo tiempo una relación con el factor edad y cambios relevantes de expresión durante la progresión del tumor, averiguando así además si esto pudiese tener repercusiones clínicas o de severidad del cáncer. Este capítulo,

por tanto, plantea una estrategia orientada a la extracción de conocimiento basada en diversas técnicas que realizan *clustering* jerárquico u operan sobre él, constituyendo este el núcleo del *framework*. El uso de este tipo de *clustering* como elemento central aporta además una mayor flexibilidad en la toma de decisiones de análisis y permite comprender mejor la estructura del conjunto de datos. Este sistema se ha implementado en módulos que ejecutan uno o varios algoritmos según su objetivo, los datos de entrada y la fase del análisis, centrándose así en proporcionar flexibilidad. Gracias a ello, es posible implementar más métodos de selección que actúen sobre los resultados de *clustering*. De hecho, se presentan dos vías de análisis alternativas y los resultados de selección de cada una de ellas.

3.2.1. Estructura y metodología del *framework* híbrido basado en *clustering* jerárquico

Este *framework* híbrido pretende agrupar técnicas diversas de minería de datos para permitir al investigador reconocer relaciones y patrones, teniendo en cuenta los objetivos de esta sección, orientados a la extracción de conocimiento y al estudio de variables concretas más que al diagnóstico. No obstante, se mostrará que también es capaz de realizar esta tarea satisfactoriamente. Así pues, se emplean técnicas estadísticas, evolutivas, de *clustering* e incluso de representación visual. En este sentido y en comparación con la propuesta anterior, el uso de estas técnicas ya manifiesta una finalidad diferente, más exploratoria, dada su flexibilidad, menos centrada en el contexto médico del estudio diagnóstico, pero que permite perseguir diferentes objetivos de investigación específicos. En las siguientes subsecciones se expone el funcionamiento y propósito de los diferentes módulos de forma secuencial, de acuerdo con la organización del *framework*, como se puede ver en la Figura 3.7: un diagrama que conecta todos los procesos llevados a cabo por el mismo.

3.2.1.1. Módulo de filtrado estadístico (MFE)

Una vez se ha seleccionado un *dataset*, este es el primer módulo en ejecutarse, de acuerdo al diagrama de la Figura 3.7. Este módulo realiza un filtrado preliminar que permite obtener una cantidad de genes manejable por los siguientes módulos al tiempo

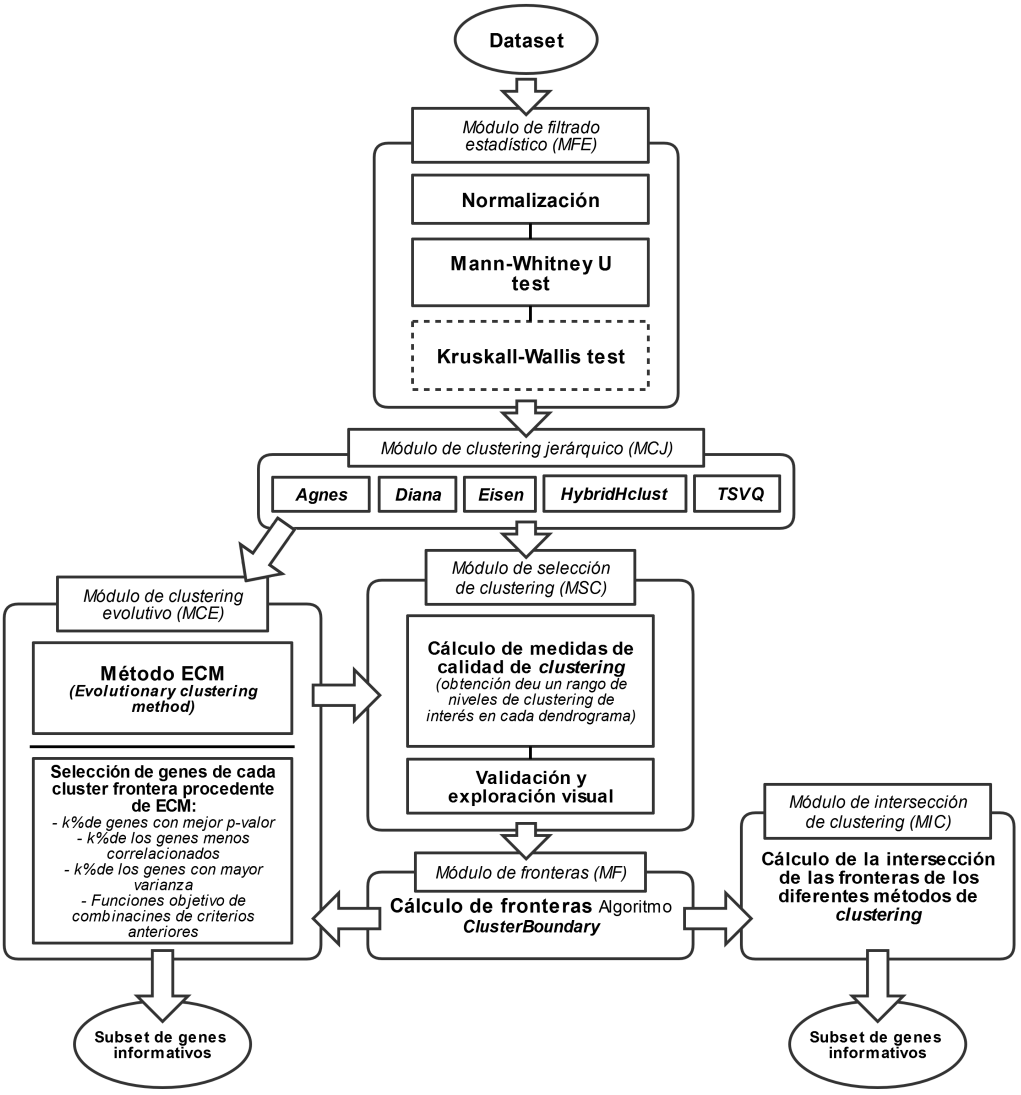

FIGURA 3.7: Diagrama representando el *framework* híbrido de selección basado en genes frontera.

que marca el objetivo principal de selección y determina los factores de estudio. En primer lugar, se normaliza el conjunto de datos, aunque el método utilizado puede variar en función del origen de estos. En el caso de estudio expuesto en esta sección se utiliza el algoritmo RMA. Se eliminan a continuación las sondas control y se aplica un algoritmo de tratamiento de datos faltantes o *missing data*. Una vez realizado este preprocesamiento, este módulo ejecuta varias pruebas para seleccionar genes estadísticamente relevantes de cara al caso de estudio en base a su significancia. Una vez completado el preprocesamiento del conjunto de datos, el primer test en ser aplicado es el test no paramétrico

de Mann-Whitney para obtener aquellos genes cuyos valores se explican por provenir de dos poblaciones diferentes (como se expuso en el capítulo anterior) [148], que en el caso de estudio descrito más adelante se corresponden con muestras de tejido control y tumorales de PDAC. De esta forma resultan genes cuyos niveles de expresión varían de forma significativa respondiendo a su pertenencia a una clase. El p-valor de corte inicial empleado para esta prueba es de 0,05, así que, para todos aquellos genes que presenten un p-valor inferior al obtenido tras el ajuste correspondiente, se rechaza la hipótesis nula y se seleccionan como genes con alta probabilidad de estar relacionados con la patología en cuestión. A continuación, dependiendo del objetivo del estudio, se puede aplicar un test de Kruskal Walis. Aunque se podría utilizar solo un método de filtrado en función de las necesidades, en este caso, este segundo test puede adquirir gran relevancia, pues está orientado al análisis de una variable adicional de estudio, lo que añade la posibilidad de una selección más específica. Mediante el empleo de Kruskal-Walis, una alternativa no paramétrica al ANOVA de varios factores, se introduce la idea de estudiar la relación con la patología de un factor de estudio adicional, más allá de la habitual expresión diferencial. Se trata, por tanto, de buscar dentro del conjunto que, tras la aplicación de Mann-Whitney, ya presenta una relación con la patología, cuáles de los genes que pasan el corte presentan además una asociación con otro factor (en el caso de estudio propuesto, la edad). Por lo tanto, los filtros estadísticos empleados en esta fase ya reflejan los criterios de búsqueda generales y alcanzan parcialmente el objetivo de seleccionar características de acuerdo al caso de estudio. Es decir, que los genes que han sido seleccionados tras aplicar los tests cumplen teóricamente los objetivos de haber sido escogidos por su relación con las variables del estudio. No obstante, desde un punto de vista práctico, este filtrado no es suficiente por un gran margen, pues no elimina la información redundante y el número de características sigue siendo muy elevado para nuestro objetivo. Este módulo establece, por tanto, un punto de partida fundamental al considerar los objetivos específicos de selección del estudio, pero requiere de un filtrado mucho más elaborado y restrictivo, capaz de afrontar las dificultades de la selección de biomarcadores y obtener un *set* reducido de genes relevantes.

3.2.1.2. Módulo de clustering jerárquico (MCJ)

Este módulo no realiza ningún filtrado por sí solo. En realidad su objetivo es dividir el conjunto de datos para trasladar el proceso de selección de genes desde un gran *dataset* a subconjuntos de menor tamaño. Si bien habitualmente se seleccionarían las características que pasen un *threshold* o las mejores k-ésimas de acuerdo con un índice de significancia o calidad aplicado sobre todo el conjunto, en este caso se busca encontrar genes relevantes, pero sin llevar a cabo esa selección hasta que no haya grupos establecidos por similitud, donde efectuar la criba por separado en cada conjunto. Este módulo aplica paralelamente diversas técnicas de *clustering* jerárquico para agrupar los genes en base a su expresión. Los algoritmos integrados en el *framework* son Agnes [78], Diana [104], Eisen [46], TSQV [152] y HybridHclust [33]. A continuación, se da una breve noción de los aspectos distintivos del funcionamiento de cada uno.

Estos algoritmos construyen una jerarquía de *clusterings* que podemos expresar mediante un dendrograma. En el caso de Agnes (*AGglomerative NESting*), en el primer nivel, cada observación o dato es un *cluster* por sí mismo y, a través de los siguientes niveles, los *clusters* se van fusionando de acuerdo a su similitud para dar lugar finalmente a un grupo que contiene todo el *dataset* [78]. En cada nivel, cada *cluster* se une al más cercano y constituyen uno solo en el siguiente nivel (*clustering* aglomerativo). Por su parte, Diana (*DIvise ANAlysis*) es en realidad un algoritmo muy similar, pues realiza esta misma tarea pero en el orden inverso: empezando desde un gran *cluster* que engloba todos los datos y dividiéndolo hasta alcanzar tantos *clusters* como observaciones (*clustering* divisivo) [78]. Eisen lleva a cabo un *clustering* jerárquico aglomerativo en el que cada *cluster* está representado por el vector medio de los datos contenidos en él [46]. Estos tres métodos se han utilizado con distancia euclídea. En cuanto a TSVQ (*Tree-Structured Vector Quantization*), este lleva a cabo un *clustering* divisivo, por el cual el conjunto de datos es subdividido de forma recursiva en 2 *clusters*, utilizando k-means siendo $k = 2$ para encontrar la subdivisión. Esta técnica constituye además la base de *Hybridhclust*, pues es un algoritmo en el que se aplica TSVQ, pero añadiendo el requisito de que los *mutual cluster* no pueden ser divididos [33]. Por *mutual cluster* se entiende aquel grupo de puntos cuya distancia máxima intragrupo es menor que la distancia al punto externo más cercano. Es, por tanto, común que los algoritmos de *clustering* jerárquico tiendan a preservar *clusters* mutuos, ya que no son divididos hasta que son aislados en

el dendrograma. Dentro de cada *cluster* mutuo, TSVQ se aplica para calcular un híbrido en sentido descendente en el cual se retiene una estructura de mutual *clusters*. Al estar basado en TSVQ, utiliza distancia euclídea cuadrada.

A continuación, en función de la vía de análisis escogida, se ejecuta el módulo de selección de *clusterings* o el método alternativo MCJE, cuyos resultados posteriormente también son sometidos al módulo de selección.

3.2.1.3. Módulo de selección y validación de *clusterings* (MSC)

El objetivo de este módulo consiste en seleccionar, a partir de los dendrogramas obtenidos del *clustering* jerárquico, aquellos niveles o rangos de niveles que presenten un *clustering* de mejor calidad. Con este fin, se calculan varios indicadores internos de cada *clustering* (es decir, de cada uno de los niveles de cada dendrograma). Las medidas utilizadas para evaluarlos son homogeneidad, ancho de silueta y separación. En definitiva, entendemos por un *clustering* de calidad aquel en el que se da la existencia de grupos compactos y bien diferenciados entre ellos. Una vez calculados estos indicadores, se seleccionan en cada dendrograma los tres niveles que presenten los mejores valores de cada una de las tres medidas. De estos tres, entre los dos mejores niveles se crea acto seguido un intervalo que comprende aquellos niveles que deberían ser analizados. La razón de ser de la selección de dicho intervalo es que el *clustering* de mejor calidad o más apropiado no tiene por qué ser el mejor en una sola de las medidas. Entonces, cada intervalo de niveles se explora para seleccionar el nivel de mayor calidad, descartando todos aquellos que presentan índices más bajos. Podría ocurrir que hubiera algún empate o dificultad para comparar dos *clusterings* cuyas medidas de calidad no bastan para decidir cuál escoger. En este caso, se recurre al análisis visual. Se construyen entonces visualizaciones que comprenden mapas de calor enlazados con dendrogramas y coordenadas paralelas para cada *clustering*, como se verá más adelante. Además, se representan en un *scatterplot* tridimensional las fronteras de cada *cluster*. De esta forma, en lugar de seleccionar aleatoriamente entre dos niveles, se pueden explorar dos *clusterings* para apoyar la toma de decisión, así como para conocer el número de *clusters* o cómo evoluciona cada *cluster* en el contexto del dendrograma. Al final de esta fase, se ha seleccionado un nivel de cada dendrograma, es decir, un único *clustering* resultante para cada método empleado.

3.2.1.4. Módulo de cálculo de fronteras (MF)

Este módulo se encarga de realizar una selección de características mediante la extracción de genes frontera de cada grupo en los *clusterings* resultantes de aplicar los anteriores métodos. Se obtiene un nuevo *clustering* en el cual se presenta cada uno de los *clusters* existentes pero incluyendo sólo los correspondientes genes frontera. Una frontera está formada por un subconjunto de puntos pertenecientes a la clausura topológica de un conjunto, pero no a su interior. Los puntos frontera definen cada *cluster* y, por tanto, a partir de ellos se puede sintetizar la información de cada *cluster* o discriminar todos los genes que hay en su interior. En consecuencia, los conjuntos de puntos frontera podrían agrupar genes que se hallasen a su vez diferencialmente expresados y que fuesen buenos candidatos en el proceso de selección de biomarcadores. En la medida en que estos genes poseen las características propias de pertenecer a la frontera, los conjuntos de genes que cumplan ambos criterios de selección pueden tener un impacto positivo en el proceso de extracción a la hora de lidiar con el problema de la redundancia. De esta forma se pretende evaluar el potencial de los genes frontera como biomarcadores, y ver su efecto en la clasificación. Para extraer dichos genes se emplea el algoritmo de puntos frontera en [26].

El algoritmo *ClusterBoundary* consta de cuatro fases. En primer lugar, se calculan de forma incremental a través de varias iteraciones los puntos extremos de un *cluster* (se trata de $2n$ puntos extremos, siendo n en número de dimensiones). A partir de estos, en la fase subsiguiente, se halla el centroide. En tercer lugar, se calculan los puntos medios de cada par de puntos extremos (a excepción de aquellos que pasan por el centroide) y se trazan los posibles radios desde el centroide hasta dichos puntos. Finalmente, se construye una bola (n dimensional) que contiene los puntos interiores a partir de uno de los posibles radios, considerándose genes frontera los que queden fuera de ella. De esta forma, en función del radio escogido de entre los disponibles, se puede ser más o menos restrictivo a la hora de obtener el conjunto de genes frontera. Para más información, [26] se puede encontrar un apoyo visual del funcionamiento del algoritmo. Se ha integrado PCA en este punto del proceso pues nos permite reducir la dimensionalidad del conjunto sin perder información, en este caso con el objetivo principal de poder representar gráficamente los puntos frontera y realizar cualquier tratamiento visual, para lo que se han escogido los tres primeros componentes. Una vez se ha calculado el *set* de

genes frontera, el *framework* se bifurca y la salida de este método puede ser analizada de forma alternativa por dos módulos: el de intersección de *clustering* (MIC) o el de *clustering* jerárquico evolutivo (MCJE).

3.2.1.5. Módulo de intersección de *clustering* (MIC)

Como se ha expuesto, se proponen dos vías alternativas para descubrir genes informativos. El método de intersección de *clustering* se encarga de calcular las intersecciones entre fronteras obtenidas a partir de distintos métodos de *clustering*. La idea consiste en que los genes que resultan de la intersección de diferentes métodos basados en distintos enfoques, son buenos candidatos para ser genes informativos, puesto que son relevantes para un amplio abanico de aquellos. Se ha desarrollado, por tanto, un algoritmo para este propósito, formalizado en Algoritmo 2.

Este algoritmo realiza la intersección de n conjuntos de genes frontera obtenidos en los distintos métodos de *clustering* alternativos, seleccionando aquellos genes que se repiten en varios procedimientos, es decir, que son un gen frontera de un *cluster* en diferentes métodos de *clustering*. Para ello se ejecutan diferentes niveles de intersección, en función de las coincidencias encontradas. Esto quiere decir que, en un primer nivel, se realizaría la intersección de todos los genes frontera de todos los métodos. El proceso se detiene si el *set* de genes frontera intersectados queda no-vacío al final de la comprobación de este nivel. De no ser así, se aplica la intersección al siguiente nivel, donde se calcula la unión de todas las posibles intersecciones formadas por $n-1$ fronteras de entrada (es decir, se disminuye la restricción de no haber coincidencias). De nuevo se comprueba si el *set* de genes está vacío. Si lo está, se pasa al nivel 3. Así, se repite la operación, tomando en este caso todas las posibles intersecciones formadas por $n-2$ fronteras de entrada.

El proceso se repite hasta que el *set* no este vacío o se alcance el nivel $n-1$. Sería extremadamente inusual que en el nivel de intersección $n-1$ no se encontrase ningún gen en común en las fronteras; no obstante, en caso de ocurrir, se unirían todas las fronteras para construir el *set* de salida, pues se considerarían igualmente importantes y en este punto no podrían ser filtradas en base a este criterio.

Algorithm 2 Algoritmo de intersección de fronteras

Input: A set of clustering boundaries $\mathfrak{I} = \{CB_1, CB_2, \cdots, CB_n\}$.
Output: IG, an informative gene set.

1: $\mathfrak{B} := \emptyset$
2: **for all** CB in $\mathfrak{I}$ **do** ▷ Computing the union of all clusters for each CB
3: Add($\mathfrak{B}$, $\bigcup_{i=1}^{|CB|} C_i$), where each C_i is a cluster boundary of CB; ▷ Converting clustering boundaries to sets and adding them to $\mathfrak{B}$
4: **end for**
5: $IG := \bigcap_{i=1}^{n} F_i,\ F_i \in \mathfrak{B}$ ▷ Computing intersection level 1
6: $l = 2$ ▷ Starting the loop with intersection level 2
7: **while** $IG \neq \emptyset\ \ \&\ \ l < n$ **do** ▷ Computing intersection level l
8: $IG := \bigcup_{i=1}^{\binom{n}{n-l+1}} \left(\bigcap_{j=1}^{n-l+1} F_{ij} \right),\ F_{ij} \in \mathfrak{B}$ ▷ Computing the union of all possible intersections with $n-l+1$ sets taken from $\mathfrak{B}$
9: $l = l + 1$
10: **end while**
11: **if** $IG = \emptyset$ **then** ▷ Computing the union of all boundary sets
12: $IG := \bigcup_{i=1}^{n} F_i,\ F_i \in \mathfrak{B}$ ▷ At this point all sets are disjoint so that, all genes are important
13: **end if**

3.2.1.6. Módulo de *clustering* jerárquico evolutivo (MCJE)

El siguiente método de extracción de genes informativos es ECM (*Evolutionary Clustering Method*), construido a partir del modelo evolutivo para *clustering* jerárquico EMHC en [25]. Dicho modelo ha sido reimplementado modificando previamente una serie de parámetros fijos dependientes del problema específico para lograr un método de *clustering* que se adapte al tipo de casos analizados por el *framework*. Como cualquier algoritmo genético, ECM precisa de una población inicial, que este caso está compuesta por una serie de dendrogramas obtenidos en la fase anterior. Esta etapa pretende refinar y mejorar el resultado del módulo de intersección de *clustering* partiendo de la idea de que los dendrogramas de salida de ECM heredan, alteran, recombinan y mejoran parte

del código genético de las soluciones de salida de otros métodos empleados para la selección (en este caso, los *clusters* de mayor calidad). Se espera, por tanto, que los genes localizados en la frontera de los *clusters* de las salidas tengan potencial como genes informativos. Los dendrogramas seleccionados por los métodos anteriores no desaparecen del *framework* sino que son conservados, al tratarse de un método alternativo, lo cual de hecho nos permite comparar resultados con la línea paralela de análisis. En el módulo se distinguen dos subfases; una, encargada de ejecutar el algoritmo sobre los dendrogramas proporcionado por la fase anterior, dando lugar a un dendrograma de salida. A continuación este dendrograma se somete al procesamiento de MSC y MF. Seguidamente tiene lugar la segunda subfase de este módulo, que consiste en seleccionar los genes frontera obtenidos tras la aplicación de MF al dendrograma de salida del método evolutivo.

Soluciones Pareto óptimas

Como se ha mencionado, el módulo ECM es modificado para ajustarse a la selección de genes informativos. Los cambios fundamentales atañen a la función de aptitud del dendrograma y sus *clusterings*. Aquí separamos los objetivos de la función de aptitud en [25] (separación y homogeneidad). Aparentemente esto generaría un problema, pues pretendemos seleccionar los mejores *clusterings* en función de dos características diferentes al mismo tiempo. Nos basamos entonces en el concepto de optimalidad de Pareto. Según este principio, existe un conjunto de soluciones óptimas o efectivas. Sin información adicional estas soluciones se consideran igualmente satisfactorias y se persigue obtener el mayor número posible de estas. Si la reasignación de recursos no puede mejorar el coste de una sin aumentar el de otra, entonces se considera que la solución es Pareto óptima. No obstante, el considerar todos los objetivos juntos hace que sea difícil encontrar una situación que satisfaga la condición de que un único vector represente la solución óptima para todos los objetivos. Formalmente, la definición de optimalidad de Pareto para tratar con un problema de máximos según Fonseca [49] sería: un vector de decisión $\vec{x}^*$ se denomina "Pareto óptimo" si y solo si no hay $\vec{x}$ que domine sobre $\vec{x}^*$ es decir, que no hay $\vec{x}$ que:

$$\forall i \in [1,k], f_i(\vec{x}) \leq f_i(\vec{x}^*) \text{ and } \exists i \in [1,k], \text{where } f_i(\vec{x}) < f_i(\vec{x}^*).$$

Una solución domina totalmente sobre otra si es mejor que ella en todos los objetivos.

En este sentido, la optimización de varios objetivos simultáneos pretende alcanzar un conjunto de soluciones sobre las que no domine ninguna otra. En vista de ello, introducimos la modificación en ECM para que sea un algoritmo evolutivo de Pareto, transformando las funciones de aptitud para los dendrogramas y para el *clustering*:

$$f_d(\mathfrak{G}) = \frac{1}{|\mathfrak{G}| - 1} \sum_{i=1}^{|\mathfrak{G}|-1} f_c(\mathfrak{C}_i), \tag{3.13}$$

donde $\mathfrak{G}$ es un dendrograma, $\mathfrak{C}_i$ es el *clustering* del nivel i en $\mathfrak{G}$ y f_c es la función objetivo recurrente para evaluar un *clustering* de $\mathfrak{G}$, formalmente:

$$f_c(\mathfrak{C}_{i+1}) = \frac{\mathcal{S}_1^*(\mathfrak{C}_{i+1})}{g - k + 1} - \frac{\mathcal{H}_1^*(\mathfrak{C}_{i+1})}{k - 1} + \operatorname{máx} \mathfrak{D}, \tag{3.14}$$

donde $S_1^*(\mathfrak{C}_{i+1})$ y $\mathcal{H}_1^*(\mathfrak{C}_{i+1})$ son la separación y homogeneidad del *clustering* $\mathfrak{C}_{i+1}$, respectivamente. $k = |\mathfrak{C}_i|$ y $g = \binom{k}{2}$ es el número de distancias entre los *clusters* de $\mathfrak{C}_{i+1}$ y $\operatorname{máx} \mathfrak{D}$ es la distancia máxima en la matrix de proximidad $\mathfrak{D}$ del *dataset* actual. El objetivo consiste, en definitiva, en obtener dendrogramas de alta calidad. La situación deseable en cualquier *clustering* es que la homogeneidad sea cuanto menor posible mientras la separación es alta. Estas funciones de aptitud pueden ser redefinidas como un vector de dos componentes objetivo que miden la separación y la homogeneidad por separado:

$$f_d^*(\mathfrak{G}) = \langle \mathcal{S}(\mathfrak{G}), \operatorname{máx} \mathfrak{D} - \mathcal{H}(\mathfrak{G}) \rangle, \tag{3.15}$$

donde $\mathcal{S}$ y $\mathcal{H}$ son medidas de separación y homogeneidad de los dendrogramas respectivamente. El objetivo es maximizar los dos componentes de f_d^*. Por su lado, definimos $\mathcal{S}$ y $\mathcal{H}$ en relación a los clusterings $\mathfrak{C}_i$ de $\mathfrak{G}$ como:

$$\mathcal{S}(\mathfrak{G}) = \frac{1}{|\mathfrak{G}| - 1} \sum_{i=1}^{|\mathfrak{G}|-1} \mathcal{S}_1^*(\mathfrak{C}_{i+1}), \tag{3.16}$$

$$\mathcal{H}(\mathfrak{G}) = \frac{1}{|\mathfrak{G}| - 1} \sum_{i=1}^{|\mathfrak{G}|-1} \mathcal{H}_1^*(\mathfrak{C}_{i+1}), \tag{3.17}$$

donde la función f_c^* para un *clustering* $\mathfrak{C}$ ha sido definida como un problema de maximización:

$$f_c^*(\mathfrak{C}) = \langle \mathcal{S}_1^*(\mathfrak{C}), \operatorname{máx} \mathfrak{D} - \mathcal{H}_1^*(\mathfrak{C}) \rangle. \tag{3.18}$$

Una vez completado este proceso y teniendo en cuenta que estas funciones no han sido

diseñadas para proporcionar más de un valor de aptitud, debemos definir un procedimiento para comparar las soluciones y llevar a cabo una selección. Se utiliza un método de selección por torneo [57], enfocado desde el criterio de comparación de Pappa *et al.* [119].

En primer lugar, se seleccionan los individuos no dominados de la población, que pasan a ser parte del proceso de reproducción. Los individuos restantes son entonces seleccionados de acuerdo a la regla de torneo. Así, el individuo con mejor aptitud de la población restante pasa también de generación a generación y se somete a los procesos de cruzamiento y mutación. Respecto al método para comparar las soluciones durante el torneo, se presenta el problema de que los individuos no dominados entre sí son como tales no comparables, dado que el concepto de dominancia les impone un orden parcial (es decir, dentro de los no dominados no siempre se puede decir qué individuo es mejor). El método utilizado para solucionar esta incompatibilidad se basa en el planteamiento de Pappa *et al.* [119] para poder establecer un orden total capaz de comparar los individuos no dominados. Se propone, por tanto, un criterio que siga el principio de la dominancia de Pareto: dados dos individuos no dominados (en nuestro caso dendrogramas) Id_1 e Id_2, calculamos el número de individuos dominados en la población actual por Id_1 como $d_{1>}$ y los dominantes como $d_{1<}$. Lo mismo ocurre con Id_2 para determinar $d_{2>}$ y $d_{2<}$. Después, el individuo con un *score* más alto de entre $\{d_{1>} - d_{1<}, d_{2>} - d_{2<}\}$ es seleccionado. Si bien es muy infrecuente, si se diese el caso de que estas diferencias coincidieran se elegiría aleatoriamente. De esta forma, se completa el proceso de comparación de individuos afrontando la dificultad de evaluar a los individuos no dominados.

Operadores genéticos

A continuación, se especifican los dos operadores genéticos que utiliza ECM, el operador de mutación y el de sobrecruzamiento.

El operador de mutación consiste en una alteración que se aplica a un solo dendrograma explorando sus ramas; solo una parte del dendrograma será modificada mientras que otra permanecerá intacta. Teniendo en cuenta que, de hecho, un dendrograma es un tipo especial de árbol, el operador de mutación mueve un *cluster* de una rama del dendrograma a otra del mismo dendrograma. El operador de cruzamiento recombina la información relevante de dos individuos (aquí dendrogramas) para dar lugar a uno

solo que hereda el código genético de los primeros. Este operador, por tanto, realiza una búsqueda en anchura por el dendrograma. En primer lugar, selecciona de forma aleatoria el mismo nivel en dos dendrogramas distintos (o dendrogramas padre). En cada uno de ellos, que constituye un *clustering*, selecciona los mejores *clusters* para combinarlos y formar un nuevo *clustering* al que denominamos "semilla", utilizando la mitad de los mejores grupos seleccionados de cada *clustering*. Para concluir, se genera el dendrograma hijo aplicando el operador de mutación sobre el *clustering* semilla para obtener los niveles superiores. Para generar los niveles inferiores se aplica una estrategia divisiva sobre el *clustering* semilla de tal forma que, por cada nivel inferior, el *clustering* con menor homogeneidad es dividido en dos.

3.2.2. Caso de estudio

Esta sección incluye la aplicación del *framework* presentado a un caso de estudio de pacientes de cáncer pancreático y los resultados obtenidos, así como un pequeño análisis comparativo con otros métodos.

3.2.2.1. Resultados experimentales

El conjunto de datos empleado contiene los niveles de expresión 78 muestras de tejido pancreático extraídas de 36 pacientes, de cada uno de los cuales se extrae una muestra de tejido normal y otra tumoral. Los antecedentes clínicos y otros datos necesarios para el estudio fueron generosamente proporcionados por la doctora Liviu Badea [12]. La población consiste en un conjunto de muestras de tejido pancreático obtenidas por resección quirúrgica. Los datos de expresión fueron obtenidos mediante chips de Affymetrix, concretamente el modelo U133 2.0, que comprende 54675 sondas. Los datos fueron normalizados mediante el algoritmo RMA (*Robust Multichip Average*). Este estudio persigue seleccionar genes relevantes para el PDAC y conocer la implicación de la edad en el transcriptoma en pacientes con esta patología. Así, se pretende seleccionar genes de acuerdo a un factor de selección adicional y realizar un filtrado para extraer un *set* de marcadores que mantenga una relación con dicho factor.

Teniendo en cuenta que el *framework* está compuesto por varios módulos, entre los que hay dos vías alternativas de análisis, el proceso se va a exponer de forma secuencial, mostrando los resultados intermedios de cada etapa y detallando y comentando los resultados finales

Aplicando el módulo de filtrado estadístico

La función principal de este módulo es, en realidad, la de filtrar ruido y características claramente no informativas. Se aplicaron los tests estadísticos de forma secuencial de acuerdo al objetivo de selección establecido. Los datos fueron normalizados mediante RMA. En primer lugar, se aplicó Mann Whitney utilizando un p-valor de corte de 0.05. Aquí, desde el punto de vista biológico, se obtienen genes que se expresan de forma diferencial, es decir, cuyos valores se explican por la división entre tejido normal y tumoral. Ahora bien, como se parte de la idea de seleccionar no sólo genes implicados en PDAC, sino de buscar si existen variaciones de estos genes con la edad, se aplica a continuación la prueba Kruskal Walis (con el mismo corte de p-valor). La edad de los pacientes oscilaba entre 45 y 77 años. En este punto es preciso establecer intervalos de edad para poder aplicar el test. Se juzgó conveniente establecer 4 grupos, de tal forma que, dada la muestra disponible, los grupos no fuesen excesivamente pequeños ni demasiado pocos. Los grupos se distribuyeron primero de acuerdo a la proximidad de edad de los pacientes, intentando en segundo lugar que fuesen de un tamaño similar. Esta distribución se muestra en la Tabla 3.4.

De las 54675 sondas iniciales de los biochips en este caso de estudio resultaron 1299 al final de esta etapa. Aunque es aún un número elevado para un estudio más especifico, vemos que la inclusión de una variable adicional de estudio, en este caso la edad, ya desemboca en un mayor filtrado en la etapa estadística inicial. Se asume, por tanto, que los genes representados por las características seleccionadas al final de esta etapa presentan una alta probabilidad de tener una relación con el cáncer y con el factor edad. Este conjunto de genes sirve de entrada, en consecuencia, para el siguiente módulo.

Grupos	Grupo#1	Grupo#2	Grupo#3	Grupo#4
Intervalo de edad	$[45, 54]$	$[55, 61]$	$[63, 67]$	$[68, 77]$
Número de pacientes	9	9	8	10

Tabla 3.4: Subgrupos de edad establecidos para llevar a cabo el análisis de correlación de este factor con el nivel de expresión génica.

Aplicando el módulo de *clustering* jerárquico

El *dataset* resultante del módulo de filtrado estadístico constituye la entrada de este. Como se ha explicado, el conjunto de datos se somete a diferentes técnicas de *clustering* paralelas: Agnes, Diana, Eisen, Hybridhclust y TSVQ. Con todos los métodos se utiliza distancia euclídea (al cuadrado en los dos últimos) y en el caso de Agnes y Diana se utiliza la media como distancia inter-*cluster*. Al final de esta etapa se han construido 5 dendrogramas (1 por cada método empleado). La salida de esta etapa se dirige alternativamente a los módulos de selección de *clustering* y a ECM (que también pasará por SC).

Aplicando el módulo de selección de *clusterings*

Este módulo actúa tanto procesando la salida de MCJ directamente como interviniendo en el módulo ECM. En cualquier caso, su objetivo es apoyar la selección de un *clustering* del dendrograma. Según las medidas de calidad calculadas para los 5 dendrogramas de salida de la fase anterior, esta etapa establece un intervalo de niveles de interés para cada uno de ellos (como se ha mencionado, establecidos entre los dos niveles que presentan las medidas de calidad más altas). De esta forma se puede acotar la búsqueda entre aquellos niveles en el intervalo que van a alcanzar unos valores que indiquen *clusterings* de calidad y/o interesantes para nuestro caso. La Figura 3.8 contiene los dendrogramas y mapas de calor obtenidos mediante la ejecución de este módulo, especificando el número de *clusters* en los que se divide el conjunto en el nivel seleccionado de cada dendrograma.

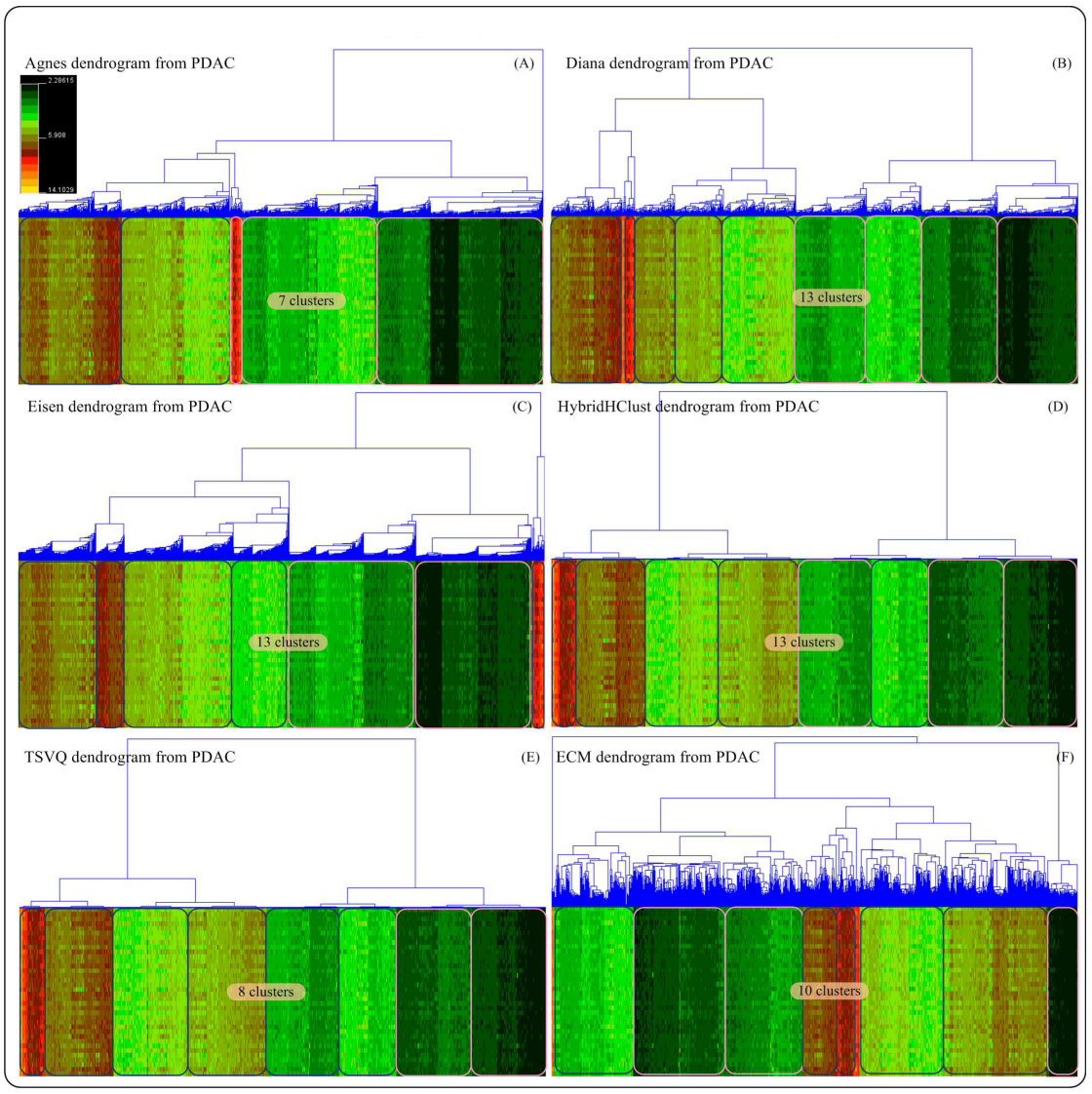

FIGURA 3.8: *Clusterings* seleccionados de cada dendrograma a partir de los distintos métodos jerárquicos aplicados sobre el conjunto de datos de PDAC

Aplicando el módulo de fronteras (MF)

A partir de los niveles seleccionados en el módulo MSC, MF extrae las fronteras correspondientes a cada uno creando nuevos *clusterings* que contienen exclusivamente puntos frontera. Además de los 5 *sets* de fronteras obtenidos de los métodos de *clustering* jerárquico, también se consigue un *set* proveniente del método ECM. Mientras que los primeros se dirigen al módulo MIC para realizar la intersección, el *set* de ECM se somete a la selección llevada a cabo por la segunda subfase del módulo de *clustering* evolutivo.

Ejecutando el módulo de *clustering* jerárquico evolutivo

Una vez aplicado el algoritmo evolutivo en sí y calculadas las fronteras de los *clusters*, la segunda subfase de MCJE trata de seleccionar finalmente un *set* de genes informativos. Con este propósito se tiene en cuenta un criterio basado en varianza y significancia. Se aplica entonces una función objetivo para seleccionar el 25 % de los genes de cada *cluster* según su *score*. La función formalmente se expresa como:

$$Score(g) := \alpha_1 \cdot significance(g) + \alpha_2 \cdot variance(g), \tag{3.19}$$

donde g es el gen y aα_1, α_2 son escalares que pueden ser definidos como $\alpha_1 = -1$ (ya que la significancia se expresa normalmente en un intervalo real $[0, 1]$) y $\alpha_2 = \frac{1}{maxvar}$. $maxvar$, por su lado, es la mayor varianza calculada en el conjunto de datos. Por tanto, cuanto mayor el valor de la función *score*, mayor es la relevancia del gen, lo que implica seleccionar genes con bajos valores de p-valor en la prueba de Mann-Whitney frente a elevados valores de varianza. Tras la aplicación de la función *score*, se obtuvieron 22 genes informativos. Posteriormente se estudiaron estos resultados sobre los gráficos de coordenadas paralelas. Para evitar cualquier saturación que dificulte la interpretación visual, en la representación gráfica se omiten los genes que mostraban un patrón de expresión similar a otros del grupo, manteniendo en dichos casos los de mayor *score*.

Ejecutando el módulo de intersección de *clustering*

En este módulo se procesan las salidas de MF, que consisten en fronteras de los grupos de diversos métodos de *clustering*. En este caso se encontraron 26 genes frontera tras aplicar el primer nivel de intersección del algoritmo, por lo que no fue necesario aplicar un segundo nivel, menos restrictivo. Obtenemos, por tanto, un grupo de genes con mayor significancia que supuestamente incluye genes informativos. Este número podría ser aún más reducido, como veremos tras observar el comportamiento de los genes que se muestran en las gráficas de coordenadas paralelas en la Figuras 3.9 y 3.10.

Grupos de genes seleccionados

A continuación se exponen los resultados de aplicar el *framework* híbrido al conjunto de datos de este caso de estudio de PDAC. Se han empleado alternativamente los dos métodos, MIF y ECM. Se pretende evaluar la implicación de MIF en el proceso, así como de los genes frontera. En las tablas 3.5 y 3.6 se muestran los *sets* de genes informativos obtenidos tras aplicar cada uno de los dos métodos alternativos, así como información acerca de si habían sido identificados en alguna base de datos o investigación en asociación al PDAC de acuerdo a la bibliografía disponible y a la base de datos PED (*Pancreatic Expression Dataset*) [108]. Se debe tener en cuenta que estos no son ni pretenden ser los marcadores de elección para el diagnóstico de PDAC, sino genes que han presentado asociación al PDAC a la vez que a la edad. Cabe destacar que ambos métodos seleccionan 10 genes comunes (resaltados en las tablas). Esto hace que consideremos dichos genes especialmente relevantes, por lo que podemos filtrar aún más el conjunto.

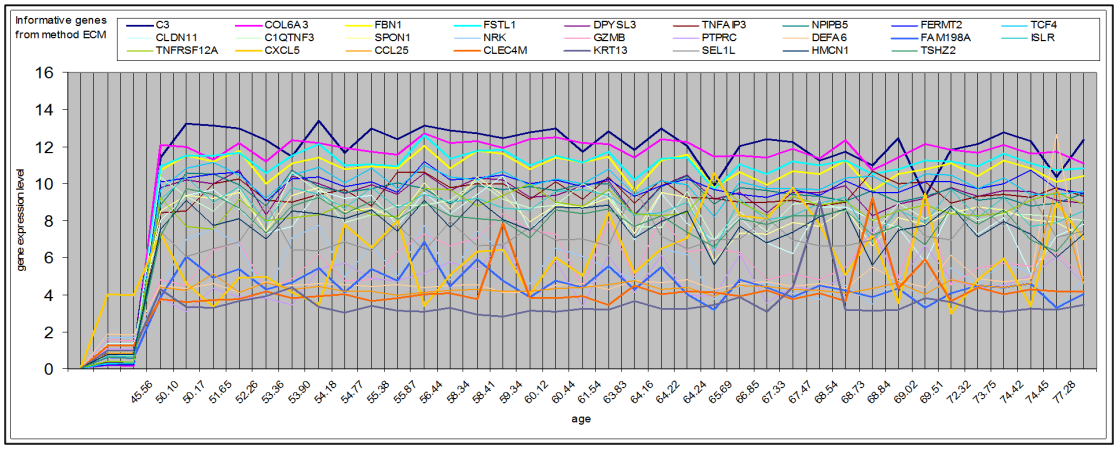

FIGURA 3.9: Niveles de expresión de los genes seleccionados mediante el módulo alternativo ECM ordenados según el factor edad.

3.2.2.2. Analizando la implicación del cálculo de fronteras

Aunque no es la finalidad del estudio, pues los genes seleccionados no se han buscado para ser marcadores de PDAC, sino como genes que varían con la edad y cuya expresión se halla alterada en esta clase de tumores, se han utilizado los genes seleccionados por MF para realizar una sencilla clasificación contemplando los tejidos no afectados y los

Identificador	Nombre del gen	Alteración registrada
C3	complement component 3	Sí
COL6A3	collagen type VI alpha 3	Sí
FBN1	fibrillin 1	Sí
FSTL1	follistatin-like 1	Sí
DPYSL3	dihydropyrimidinase like 3	Sí
TNFAIP3	TNF alpha induced protein 3	Sí
NPIPB5	nuclear pore complex interacting protein family, member B5	No
FERMT2	fermitin family member 2	Sí
TCF4	transcription factor 4	Sí
CLDN11	claudin 11	Sí
C1QTNF3	C1q and tumor necrosis factor related protein 3	Sí
SPON1	spondin 1	Sí
NRK	Nik related kinase	Sí
GZMB	granzyme B	Sí
PTPRC	protein tyrosine phosphatase, receptor type C	Sí
DEFA6	defensin alpha 6	Sí*
FAM198A	family with sequence similarity 198 member A	No
ISLR	immunoglobulin superfamily containing leucine-rich repeat	Sí
TNFRSF12A	tumor necrosis factor receptor superfamily member 12A	Sí
CXCL5	chemokine (C-X-C motif) ligand 5	Sí
CCL25	chemokine (C-C motif) ligand 25	Sí*
CLEC4M	C-type lectin domain family 4 member M	Sí*
KRT13	keratin 13	Sí
SEL1L	SEL1L ERAD E3 ligase adaptor subunit	Sí
HMCN1	hemicentin 1	Sí
TSHZ2	teashirt zinc finger homeobox 2	Sí

*Genes alterados previamente identificados en cáncer pancreático, sin especificar subtipo.

Tabla 3.5: 26 genes informativos obtenidos mediante el método ECM del *framework* propuesto en esta sección. Se muestran el identificador del gen, el nombre y si su alteración en cáncer pancreático ha sido registrada en estudios previos. Se resaltan además aquellos genes que también han sido seleccionados mediante la aplicación del módulo alternativo MIC.

tejidos tumorales en el *dataset*. Con ello se pretende ver la contribución de este módulo al *framework* y la relevancia o validez de los genes frontera como marcadores. Con este propósito se utiliza un clasificador basado en kNN (*k Nearest Neighbors*), teniendo en cuenta además que se trata de un *lazy model* o algoritmo perezoso que no necesita reconstruir el modelo de aprendizaje ante cambios en el *dataset* de entrenamiento, como ocurre en muchos clasificadores. Se retiró el módulo MF para ver la influencia que tenía sobre la clasificación, y se conectó la salida de MSC a las entrada de la segunda parte de MCJE. De esta forma se aplica una selección de genes sobre cada *cluster* de entrada antes de ejecutar el módulo de intersección de *clustering*, ya que no se han calculado genes frontera y es necesaria alguna otra clase de reducción. Se aplica entonces la función de *score* descrita anteriormente a cada *cluster* de cada uno de los *clusterings*. Los resultados

Identificador	Nombre del gen	Alteración registrada
NKIRAS1	NFKB inhibitor interacting Ras-like 1	No
TNFAIP3	TNF alpha induced protein 3	Sí
BICC1	BicC family RNA binding protein 1	Sí
SPON1	spondin 1	Sí
ENTPD1	ectonucleoside triphosphate diphosphohydrolase 1	Sí
CXCL5	chemokine (C-X-C motif) ligand 5	Sí
GZMB	granzyme B	Sí
PEG3	paternally expressed 3	Sí
PTPRC	protein tyrosine phosphatase, receptor type C	Sí
KRT13	keratin 13	Sí
SEL1L	SEL1L ERAD E3 ligase adaptor subunit	Sí
COPZ1	coatomer protein complex subunit zeta 1	Sí*
CLEC4M	C-type lectin domain family 4 member M	Sí*
C3	complement component 3	Sí
NRK	Nik related kinase	Sí
AFAP1-AS1	AFAP1 antisense RNA 1	No
GSTO2	glutathione S-transferase omega 2	Sí
GBA3	glucosidase, beta, acid 3	Sí*
PRSS35	protease, serine 35	Sí**
SAMSN1	SAM domain, SH3 domain and nuclear localization signals 1	Sí
MYLK3	myosin light chain kinase 3	No
RPL37A	ribosomal protein L37a	Sí

*Genes previamente identificados en cáncer de páncreas sin especificación acerca de su vinculación con un subtipo.
**Genes previamente identificados en otros subtipos de cáncer de páncreas distintos a PDAC.

Tabla 3.6: 22 genes informativos obtenidos mediante el módulo MIC de del *framework* propuesto en esta sección. Se muestran el identificador del gen, el nombre y si su alteración en cáncer pancreático ha sido registrada en estudios previos. se resaltan además aquellos genes que también han sido seleccionados mediante la aplicación del módulo alternativo ECM.

de los *clusterings* de entrada se intersecan en el método del módulo de intersección. Luego, en la segunda parte del MCJE, se mantiene la reducción de los genes escogidos en cada *cluster* de los *clusterings* de entrada, solo que en este caso el porcentaje de genes seleccionados en cada *cluster* es 12 %. En esta ocasión se ha restringido mucho más el porcentaje de selección respecto al que se utilizaba sobre los conjuntos de genes frontera, ya que un *cluster* tiene muchos más genes que un *cluster* frontera, y es necesario trabajar con un subconjunto pequeño de genes relevantes para poder comparar y poner a prueba ambas metodologías (con y sin módulo de cálculo de fronteras). Nótese que, en este caso, de cara a comparar la validez y efecto de usar puntos frontera, se están seleccionando aquellos genes mejores de acuerdo a la función de *score* de todo el *cluster* frente a los conjuntos de genes frontera. Una vez se ha adaptado el *framework* para funcionar sin el

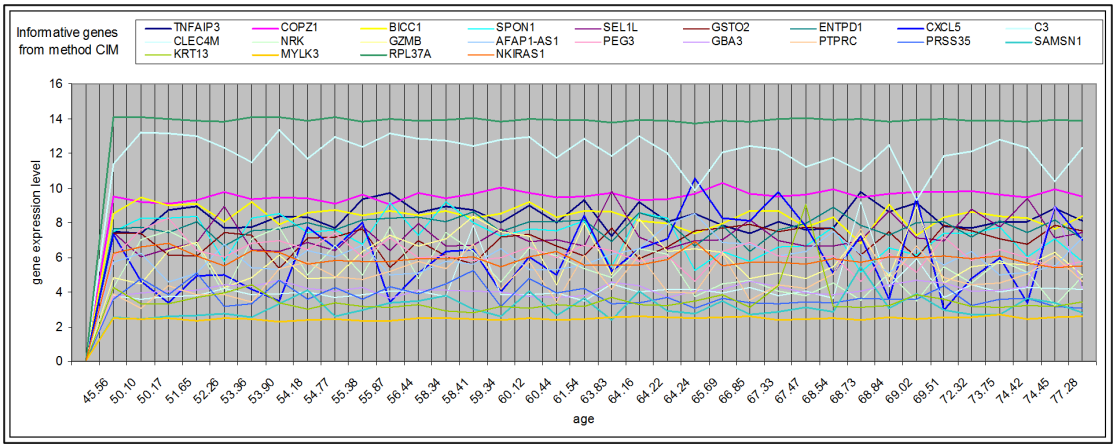

FIGURA 3.10: Niveles de expresión de los genes seleccionados mediante el módulo alternativo MIC ordenados según el factor edad.

módulo MF, se obtuvieron 162 y 128 genes en los módulos alternativos MCJE y MIC, respectivamente.

En la Tabla 3.7 se muestra una comparativa de este método (abreviado más adelante en la Tabla 3.8 como MSC+Red) con los resultados de aplicar el *framework* con la etapa de MF, viéndose los resultados de tasa de acierto alcanzados por el kNN en ambos casos. Para cada uno de estos, la tasa de acierto se ha calculado utilizando validación cruzada de $10 - fold$. Como se puede ver en la Tabla 3.7, la versión del *framework* con cálculo

Método	**Nº de genes**	***K***	**kNN (%)**
ECM sin módulo MF	162	3	84,72
ECM con módulo MF	26	6	90,28
MIC sin módulo MF	128	3	86,11
MIC con módulo MF	22	4	90,28

TABLA 3.7: Comparación de los métodos de selección alternativos del *framework* propuesto con y sin el módulo de selección de genes frontera en el *dataset* de PDAC. Los resultados se han evaluado mediante el clasificador kNN.

de puntos frontera alcanza mejores resultados en kNN. Por otra parte, reduce más el conjunto de genes, lo cual es deseable para facilitar su estudio, probablemente eliminando información redundante que puede afectar a la clasificación. De hecho, alcanza una mayor tasa de acierto utilizando menos características, lo que nos permite ver que los genes en la frontera pueden ser marcadores. De todas formas, nótese que este resultado se obtiene usando un conjunto de genes que ha sido seleccionado para estudiar un factor adicional,

y no se filtra en este caso considerando los genes que mejor diferencian tejido normal respecto de tejido tumoral, sino considerando aquellos que, además de mantener una relación con el cáncer, presentan correlación con el factor de estudio, en este caso la edad. Así, los genes relacionados con la edad, si bien se han seleccionado también de acuerdo al criterio normal-tumoral, no son los más adecuados para discriminar exclusivamente la condición de presencia o ausencia de tumor, pese a lo cual obtienen los mejores resultados.

3.2.2.3. Comparación con otros métodos

Aunque no era el objetivo principal, se consideró apropiado comparar la selección con otros métodos de cara a la clasificación, para comprobar al menos que los genes seleccionados eran importantes, además de para el factor de estudio, también como predictores del cáncer en sí, pues es la clase de genes que se pretende obtener. Además, era preciso comparar la influencia del cálculo de genes frontera en el proceso. Se han empleado, en lugar del *framework* de selección, diferentes técnicas habituales en la bibliografía, algunas de las cuales ya estaban presentes en la comparación de métodos del capítulo anterior. Sin embargo, se juzgó interesante añadir en este caso más técnicas específicas y relevantes empleadas en la selección de genes para datos de expresión. Los parámetros de cada método se han configurado con los valores por defecto y, en el caso de kofnGA, ya que es un algoritmo genético, tras la pertinente optimización de parámetros se ha fijado además una población inicial = 100 y el número de generaciones = 9000.

La Tabla 3.8 muestra una comparativa de la tasa de acierto alcanzada por estos métodos y el tiempo de ejecución (en horas o minutos, según proceda). La tasa de acierto ha sido calculada mediante una validación cruzada de $10 - fold$ estratificada. En este punto, con el doble objetivo de ver si el uso de un reducido número de genes redunda en una mayor aplicabilidad y competitividad del método, se redujo la cantidad de marcadores para clasificación a 2, en el caso del conjunto obtenido mediante la alternativa ECM, y a 3, en el obtenido a través de MIC, considerándose estos marcadores los mejores en cada uno de los dos métodos del *framework*. Esta reducción final se ha realizado utilizando el valor de la función *score* y seleccionando los mejores marcadores que mostrasen además la evolución más clara posible de acuerdo a la edad en los gráficos de coordenadas paralelas en la Figuras 3.9 y 3.10. Se puede ver, por tanto, que los métodos alternativos

Método	Número de genes	K	Tasa de acierto en kNN(%)	Tiempo de ejecución
propOverlap	1123	5	86,11	$0{,}11'$
Boruta	10	1	91,67	$26{,}63'$
SDA	5	7	88,89	$0{,}10'$
Spikeslab	37	3	90,28	$4{,}86'$
kofnGA	5	1	72,22	$14{,}91\,h$
Etapa MSC+Red.	25	1	88,89	$[5{,}04, 10{,}04]\,h$
ECM	2	5	91,67	$[3{,}27, 6{,}27]\,h$
CIM	3	22	91,67	$11{,}52'$

TABLA 3.8: Comparación de los métodos de selección de genes aplicados sobre el *dataset* de PDAC. Los módulos MCJE (basado en ECM) y MIC del *framework* híbrido propuesto son comparados con 6 métodos reconocidos en la bibliografía; su tasa de acierto resultante corresponde al empleo de un clasificador kNN. Se refleja en cada caso el número de genes seleccionado y el tiempo de ejecución requerido.

del *framework* junto con Boruta han alcanzado la mejor tasa de acierto. Cabe destacar que ECM alcanza esta cifra empleando el menor número de genes. Estos genes han sido COL6A3 y ISLR, mientras que los obtenidos a partir de MIC han sido SPON1, CXCL5, C3. Teniendo en cuenta que esta clasificación no estaba prevista en los objetivos iniciales durante la implementación del sistema, y que los genes no se han seleccionado exclusivamente para la discriminación de las clases de tejido control y tumoral (a diferencia de los otros métodos de la comparativa), este método ha demostrado ser competitivo, alcanzando una elevada tasa de acierto. Esto revela que los marcadores seleccionados asociados a la edad sí constituyen a su vez buenos discriminadores de tejido tumoral.

3.2.3. Discusión y conclusiones

En esta sección se ha presentado un *framework* híbrido basado en la selección de genes frontera que ha dado lugar a dos subconjuntos de genes alterados en PDAC. Dichos genes se encuentran estadísticamente relacionados con los grupos establecidos de acuerdo al factor de estudio y nos han permitido evaluar el impacto de la edad en las muestras de PDAC. Se ha analizado además la consistencia biológica de los resultados obtenidos. Las bases de datos y publicaciones relacionadas indican que los genes seleccionados por el *framework* presentan una asociación previamente observada con el cáncer de páncreas, si bien en muchos casos solo se constata la existencia de dicha conexión, en otros se demuestra de forma directa y detallada.

Otro aspecto destacable es que, a pesar de que la bibliografía apoya los genes seleccionados, también se han encontrado algunas alteraciones que no habían sido reportadas con anterioridad en lo que al PDAC respecta. Aunque estos genes sean demasiados como para explicar aquí sus funciones y discutir sus posibles implicaciones en el cáncer, estas fueron estudiadas de cara a validar los conjuntos obtenidos. Por ejemplo, resulta especialmente interesante mencionar casos como el del gen SPON1 (*Spondine 1*). No sólo ha sido elegido por ambos métodos alternativos sino que, además, en uno de ellos varios grupos de sondas representando diferentes transcritos de este mismo gen han sido seleccionadas. Consideramos por ello este gen como uno de los más relevantes de acuerdo a la selección. SPON1 codifica para la síntesis de una proteína de matriz extracelular que, entre otras funciones, contribuye al crecimiento de axones en la médula espinal. En general, los genes que codifican proteínas de matriz suelen estar alterados en PDAC, en relación a la producción de estroma. De hecho, resulta interesante que este gen constituya un buen marcador para PDAC en la medida en que puede denotar la presencia del estroma desmoplásico. Respecto al objetivo de dotar un sistema de selección de genes con la capacidad de extracción de conocimiento biológico, hay que resaltar que los resultados hablan por sí mismos y nos permiten observar que la inclusión de un nuevo factor y el uso de *clustering* jerárquico y el módulo visual ofrecen mayor cantidad de información susceptible de interpretación biológica cuando se compara con el *framework* basado en CBR.

De hecho, en la línea del objetivo principal cabe decir que, al tiempo que se analiza la implicación de la edad en transcriptomas de PDAC, se ha comprobado que, pese a la ligera disminución de la expresión de varios genes según se incrementa la edad observada en las Figuras 3.9 y 3.9, este factor por sí solo no implica un cambio transcriptómico (y hasta donde sabemos, tampoco fenotípico) notable una vez desarrollada la enfermedad. Así, no parece que la edad afecte el transcriptoma hasta el punto de implicar cambios relevantes en el pronóstico de un paciente, si bien podría influir de otras formas.

Nótese que esto no significa que la edad no influya en el desarrollo del cáncer, –hablamos de una relación bien conocida e innegable con los procesos de senescencia celular y acumulación de daño oxidativo–, sino que no marca una diferencia transcriptómica significativa a lo largo del tiempo entre los pacientes ya diagnosticados de PDAC. En este sentido, se puede ver que el *framework* híbrido nos permite determinar fenómenos

como este, pudiéndose realizar el proceso de selección de genes con un fin más orientado a la extracción de conocimiento.

En cuanto a la capacidad de filtrado y el establecimiento de grupos de biomarcadores de tamaño manejable, las salidas de las dos variantes de análisis presentes en el *framework*, una evolutiva y otra de intersección de genes, se componen de 26 y 22 características respectivamente. No obstante, cuando cruzamos ambos resultados observamos que son 10 los genes seleccionados por ambos métodos y que los dos obtuvieron una elevada tasa de acierto incluso si no se había realizado una selección exclusivamente de acuerdo con la discriminación entre sanos y enfermos (pues se contemplaba también el factor edad).

En la Tabla 3.8 podemos ver una comparativa del *framework* híbrido con otros métodos, mostrando su tasa de acierto mediante un clasificador kNN, así como los pertinentes tiempos de ejecución. El tiempo de ejecución del *framework* es mayor que el de varios de los otros métodos estudiados, fundamentalmente debido al módulo MCJE, como era esperable. En este caso el tiempo asignado se fijó entre 2 y 5 horas aproximadamente. Si bien esto es causante de la subida del tiempo medio de ejecución de todo el *framework*, los resultados obtenidos y la comparación con otros métodos justifican su uso.

Cabe destacar que el método Boruta alcanza la misma tasa de acierto que el *framework* híbrido propuesto. No obstante, cuando se observa el número de características empleado, es menor el que utiliza este último. En este sentido, el *framework* parece poseer una capacidad de filtrado satisfactoria, al ser capaz de alcanzar una alta tasa de acierto con un *set* de genes más reducido. La aportación del módulo de genes frontera supone un punto clave en el proceso de selección del *framework*. Este proporciona además la suficiente flexibilidad en su estructura como para permitir diferentes objetivos y/o estudios de selección de genes. Además, su estructura modular permitiría añadir nuevos módulos sin un coste excesivo en caso de necesidad. El sistema implementado cuenta con dos métodos de selección alternativos, ECM y MIC, y ambos han alcanzado buenos resultados en clasificación, seleccionando varios genes en común.

Si bien el módulo que integra ECM ha alcanzado mejores resultados que MIC, (lo cual es comprensible teniendo en cuenta que se alimenta de un principio evolutivo para mejorar las soluciones de otros métodos), hay que destacar que, en el proceso de selección que incluye, la técnica utilizada es escogida por el usuario. Esto es una

desventaja respecto al automatismo completo de MIC, ya que una elección u otra podría cambiar drásticamente el resultado. También es mucho más lento, como es esperable de un método evolutivo. Además, el hecho de que el usuario no pueda intervenir en MIC de ninguna forma impide que introduzca error en las soluciones.

Respecto a los resultados de ambos métodos, si bien es normal que cuando se aplica un filtro muy restrictivo los genes varíen más, también es cierto que esto, junto con la variabilidad génica y todos los problemas mencionados, nos permite comprender la gran repercusión de la metodología empleada en la uniformidad de los subconjuntos de genes seleccionados. Por ello se ha considerado apropiado seleccionar aquellos genes comunes entre ambos grupos, lo cual enlaza con el objetivo de dotar al *framework* con la capacidad de realizar una selección más estable. Así, además de utilizar distintos métodos de *clustering* para obtener más posibles fronteras con las que poder observar mejor su relevancia general en el proceso de selección, también consideramos la intersección de genes seleccionados simultáneamente por dos criterios. Se asume así que estos genes son mas relevantes al ser seleccionados por ambas vías. De hecho, las observaciones realizadas durante el análisis del caso de estudio acerca de la uniformidad en la selección motivan en gran medida el diseño del siguiente *framework*, pues si bien hay diferentes grupos de genes que pueden tener un elevado éxito en clasificación, se adquiere gracias a ello mayor conciencia de la baja uniformidad de resultados y de como resulta conveniente utilizar varios criterios y enfoques distintos a la hora de buscar grupos de marcadores.

3.3. Lidiando con la inestabilidad en el en el proceso de selección: un *framework ensemble* orientado a la búsqueda de subconjuntos de marcadores estables

Tras la aplicación de los *frameworks* anteriores se observó que algunas de las propiedades relevantes para una metodología de selección de genes habían sido alcanzadas y eran abordables desde distintas perspectivas. No obstante, si bien tras la aplicación de un primer framework basado en CBR surge la idea de un enfoque centrado en una selección flexible que permitiese realizar diferentes estudios (abordando así distintas necesidades investigadoras), una vez llevados a cabo los experimentos mediante ambas

metodologías, se observa la necesidad a caballo entre los enfoques de las dos técnicas previas. Esta hace referencia en realidad a uno de los problemas más importantes en la búsqueda de biomarcadores. Se trata, como se ha mencionado, de la inestabilidad en el proceso de selección. El problema subyacente es que a menudo la significancia de los genes seleccionados puede depender en gran medida del método de clasificación empleado, así como de otros aspectos metodológicos que afectan a la recogida de datos. Todo ello puede inducir a ciertos errores, especialmente en aquellos casos en que no se realiza una validación apropiada antes de proponer un grupo de genes como marcadores de una condición biológica. Esto quiere decir que, a menudo, los genes seleccionados en un laboratorio pueden no ser realmente relevantes para la predicción de la enfermedad estudiada. Esta falta de capacidad de generalización de muchos métodos también viene potenciada por la baja cantidad de muestra, problema tratado con anterioridad. A pesar de que, una vez seleccionados los genes, una evaluación con otros conjuntos de datos nos aporta una visión de la capacidad de un grupo de biomarcadores, dicha capacidad será aún limitada y no habrá hecho frente al problema del sesgo metodológico. Más que por los resultados de clasificación en sí mismos, nos interesa evitar esta clase de imprecisiones en la medida en que genera errores de interpretación biológica (y así no engrosar la lista de asunciones en los experimentos que sí que son inevitables). Aunque realmente no se adquiera independencia de un método de selección, podemos decir que se adquiere dependencia de un mayor número de ellos, lo que persigue eliminar características que son exclusivamente dependientes de uno solo y cuya selección está ligada exclusivamente a dicha metodología. Dadas las observaciones realizadas en los casos de estudio anteriores, en este los objetivos prioritarios del *framework* que presenta esta sección son:

- Establecer metodologías de selección con capacidad de generalización y que sean robustas frente a la inestabilidad, priorizando la significancia biológica de los resultados en tamaños muestrales reducidos.
- Minimizar el número de biomarcadores utilizados en diagnóstico maximizando la capacidad de clasificación, estableciendo grupos de genes reducidos y aptos para su estudio molecular en un laboratorio.
- Diseñar metodologías capaces de extraer conocimiento biológico en el contexto del cáncer, persiguiendo obtener información sobre la implicación de genes que

puedan actuar como dianas terapéuticas y permitiendo el estudio de factores en relación con la expresión génica.

Así pues, si bien se persigue una propuesta aplicada, centrada en la clasificación frente a la flexibilidad del *framework* híbrido de la sección anterior, se pretende en esta sección llevar a cabo un proceso de selección y clasificación que priorice, sobre la obtención de elevadas tasas de acierto en situaciones altamente específicas (como es el caso del *framework* CBR en la Sección 3.1), el establecimiento de un conjunto de biomarcadores que dé más importancia al punto de vista de la realidad biológica del transcriptoma. Para comprender este objetivo, pensemos en el resultado final de cualquier proceso de selección, donde contamos con un *set* de genes extraídos de acuerdo a un método de clasificación concreto en un conjunto de datos. Lo deseable sería que el *set* de marcadores obtuviese buenos resultados en una nueva muestra. No obstante, incluso si dispusiésemos de una muestra para realizar la evaluación generada en las mismas condiciones (algo sumamente improbable), desde el punto de vista teórico deberíamos priorizar un conjunto de genes que sea capaz de garantizar una clasificación precisa y eficaz en diferentes clasificadores y *datasets*. Esto es importante porque podemos tener mayor seguridad de que el gen seleccionado, independientemente de su capacidad de clasificación, es relevante para la patología, ya que se ha hecho frente al sesgo metodológico durante su selección. Incluso podría ocurrir que dicho conjunto mostrase menor tasa de acierto en una nueva muestra mediante un clasificador concreto, pero es más probable que su capacidad global lo hiciese biológicamente más relevante. Aunque esto no significa que un gen informativo y alterado en el cáncer no pueda ser seleccionado por ambos enfoques, no podemos estar seguros de si realmente es relevante cuando no es seleccionado por un método que utiliza diferentes técnicas de selección de forma simultánea. Una de las prácticas que se puede adoptar ante este problema es no seleccionar los genes respecto a un solo clasificador o utilizar exclusivamente medidas que no dependan de un modelo de clasificación (aunque ello implica otra clase de limitaciones). El problema de la inestabilidad en selección se afronta aquí desde la perspectiva de reducir el *dataset* de forma que se maximice la tasa de acierto de un conjunto de clasificadores de forma simultánea. El grupo de genes que satisface este requisito se ha denominado "conjunto estable". La presente aproximación a este problema pretende, por tanto, aplicar técnicas híbridas durante el proceso de selección para buscar dicho conjunto.

De esta forma se intenta paliar el problema descrito a varios niveles. En este caso, una metodología *ensemble* resulta apropiada, pues las técnicas de selección englobadas en el *framework* se basan en diferentes principios. Además, se utilizan al mismo tiempo distintos clasificadores, basados igualmente en varios criterios, para encontrar el *set* que maximice la tasa de acierto en todos ellos. Se pretende aumentar así la validez de los genes seleccionados. El *framework* consta de 4 etapas consecutivas en la que cada una lleva a cabo un filtrado diferente para reducir el *dataset* al tiempo que se determinan marcadores estables.

A diferencia de varios trabajos recientes del área, este *framework* se ha implementado para realizar una reducción secuencial del *dataset* priorizando la estabilidad de un grupo de genes mediante filtrados sucesivos y/o paralelos llevados a cabo por distintos métodos, en lugar de centrarse en un *set* a medida del caso de estudio.

Para evaluar la metodología aquí propuesta, se establece un caso de estudio en el que se utilizan muestras de tejido de adenocarcinoma ductal pancreático, con el objetivo de discriminar las muestras tumorales. De esta forma también es posible observar los cambios de selección respecto al *framework* anterior y ver si existe alguna clase de solapamiento de genes relevantes en relación a la edad con los genes obtenidos por el presente método de *ensemble*.

En cuanto a la posibilidad de ampliar este método, se ha implementado de tal forma que no resultaría complejo añadir más clasificadores y métodos de selección, lo cual, de hecho, podría resultar interesante en un futuro para extender la validez de los subconjuntos estables bajo el mismo principio o integrar algoritmos de los otros sistemas que componen la propuesta. En la siguiente subsección se expone la organización y funcionamiento del *framework* a través de las distintas etapas que lo conforman.

3.3.1. Estructura y metodología del *framework ensemble*

El presente *framework* consta de 4 etapas enlazadas de forma secuencial, cada una de las cuales realiza un filtrado del conjunto de genes que constituirá la entrada de la siguiente. La Figura 3.11 muestra un diagrama reflejando esta estructura.

La primera etapa está destinada a realizar el preprocesado (que incluye la normalización y limpieza de los conjuntos de datos), un filtrado estadístico preliminar y un proceso de eliminación de ruido. Tras ello se aplican paralelamente varios métodos de selección compuestos para conformar un *set* de genes resultante de la unión de todos ellos. El proceso continúa mediante la aplicación de métodos de *wrapper* basados en varios clasificadores. Estos *wrappers* crean un conjunto de genes englobando los mejores marcadores según cada clasificador. A continuación se lleva a cabo la búsqueda de un subconjunto dentro del anterior que sea estable, es decir, que alcance buenos resultados en todos los clasificadores. Se realiza posteriormente una evaluación del conjunto estable utilizando un segundo *dataset* para comprobar su validez y significancia, así como la capacidad de generalización del método. Se pretende así conocer en qué medida se está haciendo frente a las dificultades dadas por las diferentes condiciones que marcan la recogida de datos y construcción de los *datasets*.

3.3.1.1. Preprocesado y eliminación de ruido

En esta etapa se aplican técnicas destinadas a normalizar el *dataset* y eliminar ruido, estableciendo un primer filtrado.

En primer lugar, se debe llevar a cabo una normalización del conjunto de datos. Aunque el algoritmo empleado y tratamiento preliminar varía en función de qué clase de datos transcriptómicos estamos utilizando, en el caso de estudio presentado más adelante se utilizan *microarrays* de expresión por lo que se aplica el algoritmo RMA (*Robust Multichip Average*).

Una vez normalizado el *dataset*, se emplean dos filtros sucesivos. En primer lugar, y dado que se contempla la existencia de dos grupos de muestras, se realiza un test de Mann-Whitney. Al igual que en el preprocesamiento expuesto en el capítulo anterior, se utiliza un p-valor = 0,05 como valor de corte para reducir el *dataset* de acuerdo a la hipótesis nula de que los valores de expresión provienen de una misma población [148].

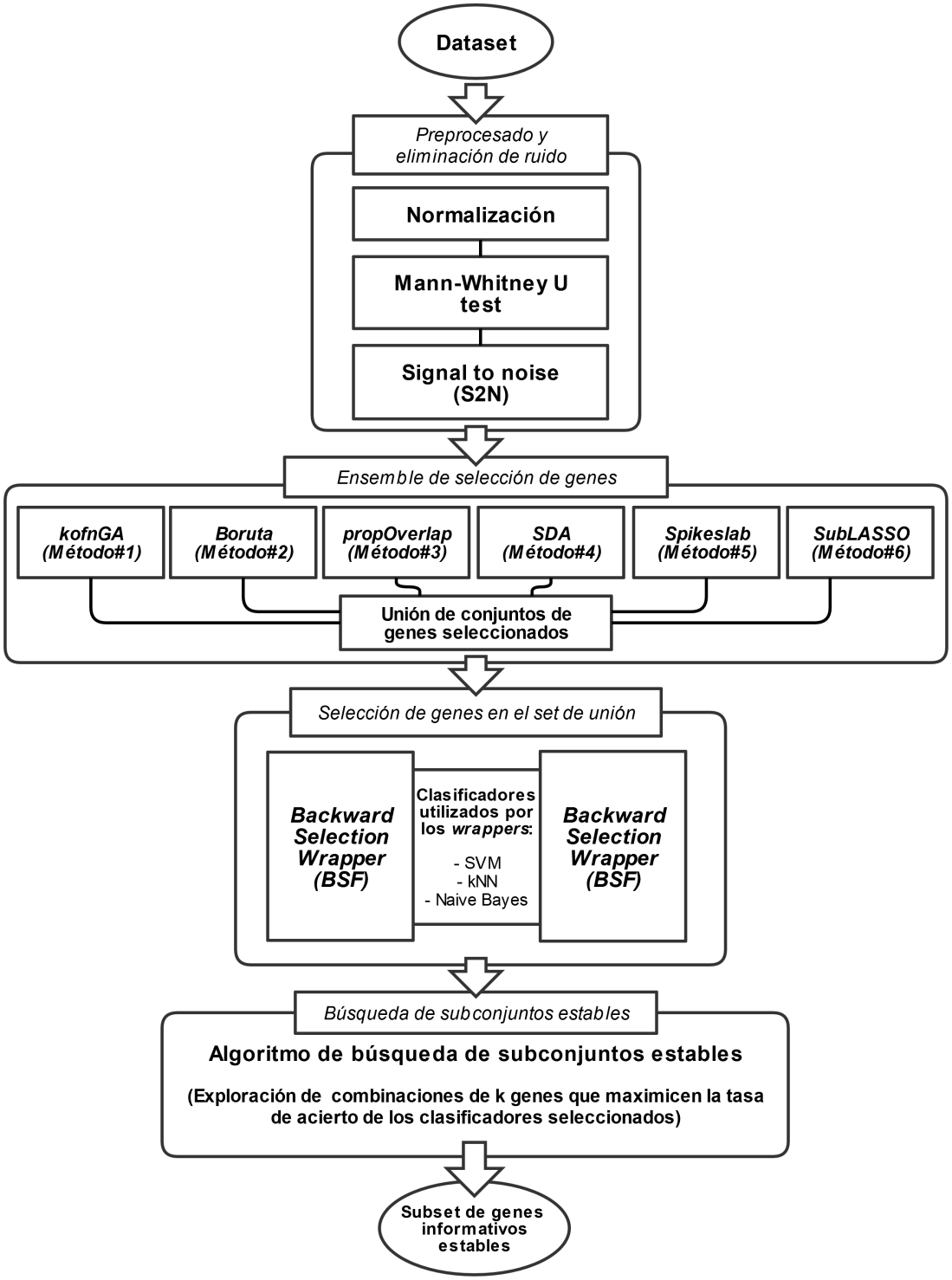

FIGURA 3.11: Diagrama representando las distintas etapas en el proceso de selección de genes llevado a cabo por el *framework ensemble*

Tras aplicar el test, los genes considerados estadísticamente relevantes se someten a un segundo filtro basado en la técnica *Signal-to-Noise ratio* o S2N adaptada a la selección

de genes [116], que pretende refinar la salida de Mann-Whitney, eliminando ruido que pueda no ser relevante para la discriminación de clases.

S2N [16, 94] es un método que define una métrica de la separación relativa de las clases del *dataset*. Mediante esta técnica podemos calcular la correlación de cada gen con cada una de esas clases, una de las cuales ha de ser la negativa y otra la positiva. De esta forma el método asigna valores positivos a los genes correlacionados con la clase positiva y viceversa y considera la distancia entre las medias de dos clases como una medida de separación.

$$S2N(f_i, c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \tag{3.20}$$

donde c es el vector de clase y f_i es el vector de la i-ésima característica.

Para filtrar los genes más significativos de acuerdo a esta técnica, se debe establecer un *threshold* de tal forma que las características seleccionadas tengan los valores más positivos o más negativos. En este caso se utiliza el punto medio del rango de valores de cada una de las clases de tal modo que, en la clase positiva, todas aquellas características superiores al punto medio son seleccionadas, mientras que en la negativa se seleccionan las inferiores al punto medio de la clase. Después de aplicar este filtrado resulta un nuevo *dataset* (libre en gran medida del ruido inicial) que constituye una entrada más manejable de genes relevantes para ser procesados por los métodos integrados en el *ensemble*.

3.3.1.2. *Ensemble* de selección de genes

En esta etapa se aplican individualmente distintos métodos de selección, que constituyen el núcleo del proceso de selección de otras propuestas, sobre el conjunto de datos. Se han escogido 6 técnicas distintas para este propósito: KOfnGA [149, 150], Boruta [91, 92], SDA [4, 5], propOverlap [105, 106], Spikeslab [73] y SubLasso [51, 102]. KofnGA se basa en un algoritmo genético para buscar un subconjunto de tamaño k de entre los enteros $1 : n$, persiguiendo minimizar para dicho subconjunto una función proporcionada el usuario. Boruta es un *wrapper* que utiliza *Random Forest* para seleccionar genes informativos. SDA (*Shrinkage Discriminant Analysis*) es un *framework* que realiza un análisis LDA (*Linear Discriminant Analysis*) y DDA (*Direction Dependence*

Analysis) con el objetivo de seleccionar variables, proporcionando un *ranking* de características mediante CAT *scores* (*correlation adjusted t-scores*). PropOverlap analiza el solapamiento entre los niveles de expresión de dos clases y proporciona un *score* de solapamiento proporcional para cada gen procurando evitar así el efecto de los *outliers*. Spikeslab utiliza, como su propio nombre indica, la técnica Bayesiana *spike & slab* para seleccionar variables en modelos de regresión lineal; y utiliza el método de regresión *elastic net* para la selección. SubLasso es una modificación del algoritmo de ajuste de un modelo de regresión Lasso (*Least Absolute Shrinkage and Selection Operator*), que permite además determinar un subconjunto de genes que deben ser seleccionados en el modelo final, permitiendo así también su uso para la comprobación de la validez de marcadores conocidos.

En definitiva, esta etapa engloba el *ensemble* de técnicas de selección. Tras ejecutar todos los métodos, las características seleccionadas se incorporan a un único *set* que denominamos *Unionset* que, como su nombre indica, resulta de realizar la unión de los conjuntos obtenidos. Esta etapa proporciona como salida un conjunto de genes seleccionados mediante diversas estrategias. La razón de construir este *set* es tomarlo como punto de partida para encontrar un subconjunto de genes representativos capaz de maximizar simultáneamente la tasa de acierto de varios clasificadores en un caso de estudio específico. De hecho, esta es la tarea que realiza principalmente la siguiente etapa del *framework*, abordándola mediante una estrategia tipo *wrapper*.

3.3.1.3. Fase de *wrappers*

Como se ha mencionado, en esta fase se buscan varios subconjuntos de genes dentro del *Unionset*, uno para cada clasificador empleado, que debe maximizar la tasa de acierto del mismo. Para ello se ha optado por una estrategia de *wrapper*. De hecho, se han implementado dos estrategias para ser utilizadas de forma paralela. Una de ellas, a la que se ha denominado FSW (*Fordward Selection Wrapper*), lleva a cabo una estrategia basada en la adición de genes, mientras que la otra, que denominamos BSW (*Backward Selection Wrapper*), se basa en la eliminación. Ambas técnicas se aplican tomando todos los clasificadores para maximizar su acierto y obtener los subconjuntos de genes correspondientes a cada uno. En este *framework*, los *wrappers* utilizan los clasificadores kNN, Naive bayes y SVM. El conjunto de genes resultante en cada uno de los métodos de

wrapper, contendrá los mejores genes para cada clasificador de entre los presentes en el *Unionset*. Los métodos BSW y FSW operan de la siguiente forma:

- *Backward Selection Wrapper*: este método precisa del *Unionset* y un clasificador como entradas. Su función consiste en eliminar de forma iterativa un gen del *Unionset* y evaluar la tasa de acierto del conjunto formado por los restantes genes mediante el clasificador empleado. Entonces, si la tasa de acierto del nuevo subconjunto, conformado por un gen menos, es mayor que la alcanzada cuando ese gen estaba incluido en el grupo, el nuevo conjunto reemplaza al *Unionset* como conjunto seleccionado. Si ocurriese lo contrario, es decir, que la tasa de acierto disminuyera, el gen retirado se devuelve al Unionset, puesto que esto indica que se ha eliminado un gen que estaba aportando información útil al proceso de clasificación. En cualquiera de los dos casos, se procedería a retirar un nuevo gen del *Unionset* y evaluar el grupo restante mediante el clasificador. Este mecanismo se repite hasta que todos los genes del *Unionset* han sido descartados o seleccionados. De esta forma, al final queda solamente un pequeño grupo de genes empleado para maximizar la tasa de acierto del clasificador. En nuestro caso, ambos *wrappers* se ejecutan utilizando los tres clasificadores integrados en el *framework* (procurando así el uso de diferentes criterios de selección no solo en la fase *ensemble* de métodos, sino también aquí), por lo que se obtienen varios subconjuntos de genes, uno por cada ejecución del algoritmo con un clasificador de entrada distinto.

- *Forward Selection Wrapper*: este *wrapper* parte de las mismas entradas que el anterior, cambiando su estrategia por una aditiva. A diferencia de BSW, aquí se selecciona en cada iteración un gen del Unionset para formar parte de un nuevo conjunto NS (*New Set*). Este conjunto, formado tras su primera iteración del algoritmo por un único gen, se utiliza sobre el clasificador. Si la tasa de acierto alcanzada por el clasificador mediante NS es mayor que la alcanzada utilizando NS cuando no tenía ese gen, el gen es incorporado al nuevo conjunto. Si esto no ocurre, el gen no es añadido a NS. La ejecución finalizaría cuando no quedaran genes susceptibles de ser integrados en el conjunto para aumentar la tasa de acierto del clasificador. Nuevamente, se ejecuta el algoritmo con los diferentes clasificadores de entrada, obteniéndose varios subconjuntos comprendidos en el Unionset que constituyen candidatos para la selección de biomarcadores, llevada a cabo en la

siguiente etapa. Tras la aplicación de ambas estrategias, los subconjuntos obtenidos son unidos para formar uno solo que englobe todos los genes relevantes y que servirá de entrada al siguiente algoritmo.

3.3.1.4. Búsqueda de subconjuntos estables

Teniendo en cuenta que cada uno de los subconjuntos de genes anteriores han sido seleccionados para maximizar la tasa de acierto de un único clasificador, lo deseable es obtener un grupo que maximice en la medida de lo posible la tasa de acierto de todos los clasificadores empleados aquí de forma simultánea, el denominado "conjunto estable". Con esta finalidad se ha desarrollado un algoritmo para encontrar este subconjunto al que se ha llamado *GeneCombine*, formalizado en Algoritmo 3. Dada una k, realiza una búsqueda de todas las k combinaciones de genes posibles de un conjunto de genes de entrada para encontrar aquella que maximice la tasa de acierto en todos los clasificadores. Es preciso señalar que, para que en la búsqueda de un *set* estable el coste computacional se mantenga dentro de unos límites razonables, el conjunto de genes de entrada debe ser reducido, de tal forma que k debería ser mucho menor que el tamaño del conjunto. Este, además de otros expuestos en el Capítulo 2, es uno de los motivos por los que el *framework* es más restrictivo en sus etapas anteriores.

3.3.1.5. Algoritmo *GeneCombine*

Finalmente, este proceso proporciona k genes estables como resultado. Estos genes deben ser evaluados y validados empleando *datasets* distintos, para comprobar si en efecto son capaces de hacer frente en grado suficiente a los problemas de las diferencias metodológicas, variabilidad natural y errores de medida que afectan a los resultados de clasificación.

3.3.2. Caso de estudio

En consonancia con el capitulo anterior, este caso de estudio se centra en tejidos de pacientes de PDAC. Para analizar la capacidad de generalización de los marcadores seleccionados, se utilizan dos *datasets* distintos de expresión génica obtenida mediante

Algorithm 3 Algoritmo *GeneCombine*

Input: T, a gene subset of the current dataset. k, the number of genes to select from T and L, a list of gene classifier methods.
Output: $\langle S, ma \rangle$, where S is a subset of k genes from T and ma is the mean accuracy from the classifier accuracies given in L.
Required: Accuracy function, computes the mean accuracy for a classifier trained from a gene subset by applying a stratified 10-fold cross-validation.

1: $S := \emptyset, ma := 0$
2: **for each** subset P_k of k genes from T **do** ▷ Select each different gene subset of size k from T
3: $sum := 0$
4: **for each** classifier C_i in L **do** ▷ Evaluate the accuracy of P_k for each classifier in L
5: $sum := sum + Accuracy(C_i(P_k))$ ▷ Add the accuracy of each classifier applied to P_k to sum
6: **end for**
7: $mean := \frac{sum}{length(L)}$ ▷ Compute the mean accuracy of the classifiers evaluated on P_k
8: **if** $mean > ma$ **then** ▷ Update the accuracy and the gene subset
9: $ma := mean$
10: $S := P_k$
11: **if** $ma = 100$ **then**
12: **break** ▷ If the mean accuracy is the maximum accuracy (100 %) then algorithm ends
13: **end if**
14: **end if**
15: **end for**

un mismo modelo de *microarray*, provenientes de diferentes centros. Estos conjuntos de datos, a los que llamaremos a efectos de comodidad PDAC-1 y PDAC-2, están disponibles públicamente en el repositorio del NCBI en [43] y [12], respectivamente. Se ha considerado interesante utilizar este segundo *dataset* para evaluar los resultados, pues es el empleado en el caso de estudio de la sección anterior. El *framework* se aplica sobre el primer *dataset* y los resultados conseguidos son evaluados en PDAC-2 para comprobar si los genes seleccionados son significativos en los dos *datasets* y, en definitiva, para el cáncer de páncreas. El objetivo es, por tanto, comprobar la eficacia de este *framework* en la selección de marcadores que se vean afectados en la menor medida posible por la inestabilidad en los procesos de selección y clasificación, producto del amplio arco de técnicas y diferencias metodológicas a las que se someten los conjuntos de datos. Se

debe emplear así un nuevo conjunto de datos del mismo problema –aquí PDAC-2– para comprobar la capacidad de generalización de los resultados del *framework*. Este proceso se lleva a cabo mediante la selección en el segundo *dataset* de los mismos genes descubiertos en el primero, que nos permite ver también cómo se ve afectada la tasa de acierto en un nuevo *dataset* de la misma patología, lo cual constituye una cuestión recurrente en la bibliografía.

3.3.2.1. Resultados experimentales en PDAC-1

A continuación se detallan los resultados de aplicar el *framework* sobre el conjunto de datos PDAC-1. La Tabla 3.9 muestra además el número de características seleccionadas en cada etapa y el porcentaje de reducción respecto al tamaño del *dataset* en la etapa anterior.

Procesos de filtrado	**Nº de carac. descartadas**	**Reducción (%)**
Preprocesado	76	0.15
Mann-W. test	46182	84.58
S2N	7804	92.73
Fase de *ensemble*	463	75.53
Fase de *wrappers*	137	91.33
Alg. de combinación	10	76.92

Tabla 3.9: Comparación de los porcentajes de reducción de los distintos procesos utilizados en el filtrado, calculados sobre el tamaño del *dataset* tras la aplicación del proceso anterior.

Como se ha mencionado, en la primera etapa se lleva a cabo un filtrado estadístico preliminar que contempla las poblaciones de estudio; control y tumoral. Puesto que en este caso se trata de datos de biochips, una vez realizada la normalización, se retiran las sondas control y se somete a la prueba de Mann-Whitney. Tras aplicar el filtrado a partir del p-valor de corte seleccionado, resultan cerca de 8000 características de las 54675 iniciales.

La técnica S2N asigna un valor positivo o negativo a las características de acuerdo con su grado de pertenencia a la clase positiva o negativa. Se seleccionan, por tanto, los genes más extremos de cada clase, es decir, los valores más positivos y más negativos.

Para establecer un *threshold* (más concretamente dos, uno para cada clase), se selecciona el punto medio de cada una de las dos clases y los valores mayores que ese corte en la clase positiva y menores en la negativa son seleccionados como los más representativos de sus respectivas clases. Esto condujo a la selección de 613 características. Una vez aplicada la eliminación de ruido, se ejecutan los métodos de selección específicos. En este caso se utilizan 6 métodos, requiriéndose en el caso de kofnGA la especificación de parámetros iniciales, fijándose los valores iniciales siguientes: tamaño poblacional = 100, nº de generaciones = 50000, la función de aptitud ha sido la correlación entre genes y los parámetros restantes son los preestablecidos por el algoritmo. Como se ha mencionado con anterioridad, el objetivo de usar estos métodos, algunos de ellos evaluados además en los capítulos anteriores, es precisamente comprobar el poder del *ensemble* frente a los métodos empleados de forma individual y proporcionar al *framework* distintos criterios de relevancia para buscar un conjunto de marcadores estable.

El resultado final de cada método proporciona un conjunto de características que se une con los demás mediante la operación de unión para conformar el *Unionset*.

Los resultados individuales de cata método se muestran en Tabla 3.10., donde se especifica el método aplicado, el número de características seleccionadas y la tasa de acierto media alcanzada en los tres clasificadores (de la siguiente etapa) a través de una validación cruzada estratificada de $10 - fold$. Dichos clasificadores son SVM, Naive Bayes y kNN. Después de llevar a cabo la operación de unión de todos los conjuntos seleccionados se obtuvo un *Unionset* formado por 150 genes. Los resultados alcanzados por este *set* en los métodos de clasificación se muestran también en la Tabla 3.10. En la siguiente etapa, los dos *wrappers* utilizan cada uno de los tres clasificadores y el *Unionset* como entrada para reducir lo máximo posible el número de características empleadas en el diagnóstico sin sacrificar poder de clasificación. Se obtienen, por tanto, al final de esta etapa 6 subconjuntos de genes, tres por cada *wrapper*. La tasa de acierto alcanzada para cada uno de estos casos se refleja en la Tabla 3.11. Se resaltan para cada grupo de genes la tasa de acierto correspondiente al clasificador para el que ese grupo trata de maximizar la tasa de acierto y en torno al cual se ha construido. Nótese que los valores más altos han sido alcanzados por Naive Bayes en ambos métodos de *wrapper*.

Tras ejecutar los *wrappers*, se construye un *set* de genes, formado en este caso por trece genes resultantes de la unión de los genes seleccionados mediante diferentes

Método	**Nº de genes**	**SVM (%)**	**NaiveBayes (%)**	***K***	**kNN (%)**
KofnGA	20	100	100	1	96.87
Boruta	41	96.87	93.75	2	96.87
propOverlap	2	90.62	90.62	3	93.75
SDA	20	100	93.75	2	96.87
Spikeslab	100	93.75	93.75	4	96.87
SubLasso	6	96.87	93.75	1	93.75
Unionset	150	100	100	3	96.87

TABLA 3.10: Comparación de los métodos de selección empleados sobre el conjunto PDAC-1. Se muestra el nombre del método, el número de genes resultantes de su aplicación y la tasa de acierto alcanzada por los tres clasificadores mediante el empleo de dichos genes. También aparecen, en último lugar, los resultados obtenidos del empleo del *Unionset*.

métodos.

En la última etapa se busca un subconjunto contenido en el resultante de la fase anterior que maximice la tasa de acierto en los tres clasificadores. La idea reside en buscar un grupo de genes estable respecto a los clasificadores para encontrar genes no dependientes de una única metodología. El *set* de 13 genes sirve por tanto de entrada para esta fase.

Una de las estrategias para determinar el valor de k en el algoritmo desarrollado para realizar la selección de combinaciones, la empleada en este caso, es definir k como $k = 2$ y después incrementar el valor de k hasta encontrar una combinación que alcance una tasa de acierto media menor que la tasa media del grupo de genes anteriormente seleccionado. Se ha fijado el valor de k en 3 para este caso tras explorar mediante un gráfico de coordenadas los niveles de expresión de los 13 genes resultantes de la etapa anterior. En la Figura 3.12 podemos observar el comportamiento de la expresión de estos genes. En la derecha se distinguen fácilmente las 7 muestras de tejido no tumoral, y se observa un cambio pronunciado en muchos de los genes. Resulta interesante la presencia en el grupo de muchos marcadores cuya expresión se haya en realidad suprimida en cáncer, siendo estos los que presentan diferencias más claras. Desde un punto de vista de los niveles de expresión a través del *dataset*, la gran mayoría de los genes seleccionados presentan variaciones fácilmente distinguibles entre ambas clases y, sobre todo, varios de ellos presentan valores relativamente uniformes a lo largo una misma clase (o que al

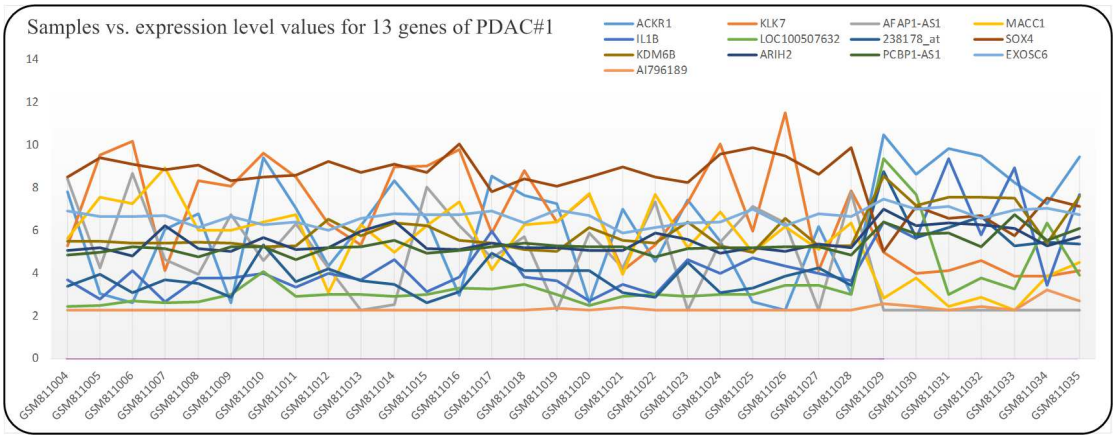

FIGURA 3.12: Gráfico de coordenadas paralelas asociando cada muestra de tejido con el nivel de expresión de cada uno de los 13 genes significativos en PDAC-1.

menos se mantienen dentro de un rango diferenciable de la otra clase). Este hecho llama especialmente la atención porque apoya la idea de que estos marcadores pueden ser estables, que es en definitiva el objetivo del *framework*. En la Figura 3.12 se muestran estos niveles, donde cada línea representa los niveles de expresión de un gen en los diferentes pacientes. Dada esta k, el algoritmo encuentra la mejor combinación de entre los genes del conjunto, cuyos resultados se muestran en la Tabla 3.11. Las tres características seleccionadas son los genes KDM6B y SOX4, y el transcrito LOC100507632. El resultado prueba que estos transcritos son estables ante el uso de distintos clasificadores.

Método	**Nº de genes**	**SVM (%)**	**NaiveBayes (%)**	**k**	**kNN (%)**
BSW	2	100.00	96.87	1	96.87
	5	96.87	100.00	1	100.00
	2	96.87	93.75	1	100.00
FSW	1	100.00	78.12	3	100.00
	3	93.75	100.00	3	100.00
	1	100.00	78.12	3	100.00
GeneCombine	3	100.00	100.00	1	100.00

TABLA 3.11: Comparación del número de genes obtenidos y la tasa de acierto alcanzada por BSW y FSW mediante los tres clasificadores empleados en el estudio.

3.3.2.2. Evaluando en PDAC-2 los genes encontrados en PDAC-1

Este proceso es clave para no perder de vista el objetivo real de este estudio. Se pretende aquí evaluar la estabilidad utilizando los genes descubiertos a lo largo de todo el proceso anterior en un *dataset* distinto del mismo problema, que incluye muestras de tejido tumoral y control. En este sentido el objetivo es analizar la variación en la tasa de acierto de los genes seleccionados en PDAC-1 cuando son utilizados en PDAC-2. De esta forma podemos comprobar la capacidad de generalización del conjunto de genes seleccionado. Estos resultados se reflejan en la Tabla 3.12. De entre los métodos evaluados, los mejores resultados son los alcanzados por esta propuesta de *ensemble*. Pese a una reducción general en la tasa de acierto, dado que existe una diferencia en los valores que toman las características en ambos *datasets*, la cifra conseguida por el *framework* supera el 87 %, alcanzando un valor máximo de 92 en el caso de la máquina de soporte vectorial. Por tanto, el resultado obtenido por el método en PDAC-1 es ciertamente estable cuando se emplean nuevos conjuntos de datos que abordan la misma condición biológica. La Figura 3.13 muestra los niveles de expresión de los tres genes seleccionados en PDAC-1 para los dos conjuntos de datos.

Método	**Número de genes**	SVM (%)	**NaiveBayes** (%)	**k**	**kNN (%)**
KofnGA	20	87.18	87.18	3	82.05
Boruta	41	88.46	84.62	4	91
propOverlap	2	69.23	71.79	5	78.20
SDA	20	84.61	82.05	1	83.33
Spikeslab	100	89.74	84.61	3	88.46
SubLasso	6	83.33	78.21	4	83.33
Unionset	150	89.74	85.90	2	89.74
Stage-V algorithm	3	92.30	87.18	1	91.03

Tabla 3.12: Comparación de los métodos de selección de genes utilizando PDAC-2. Los genes seleccionados por las técnicas de la Tabla 3.11 se han extraído en PDAC-2 y se han evaluado mediante los tres clasificadores empleados en el estudio. Los genes seleccionados como resultado final del sistema propuesto también han sido extraídos y evaluados en PDAC-2.

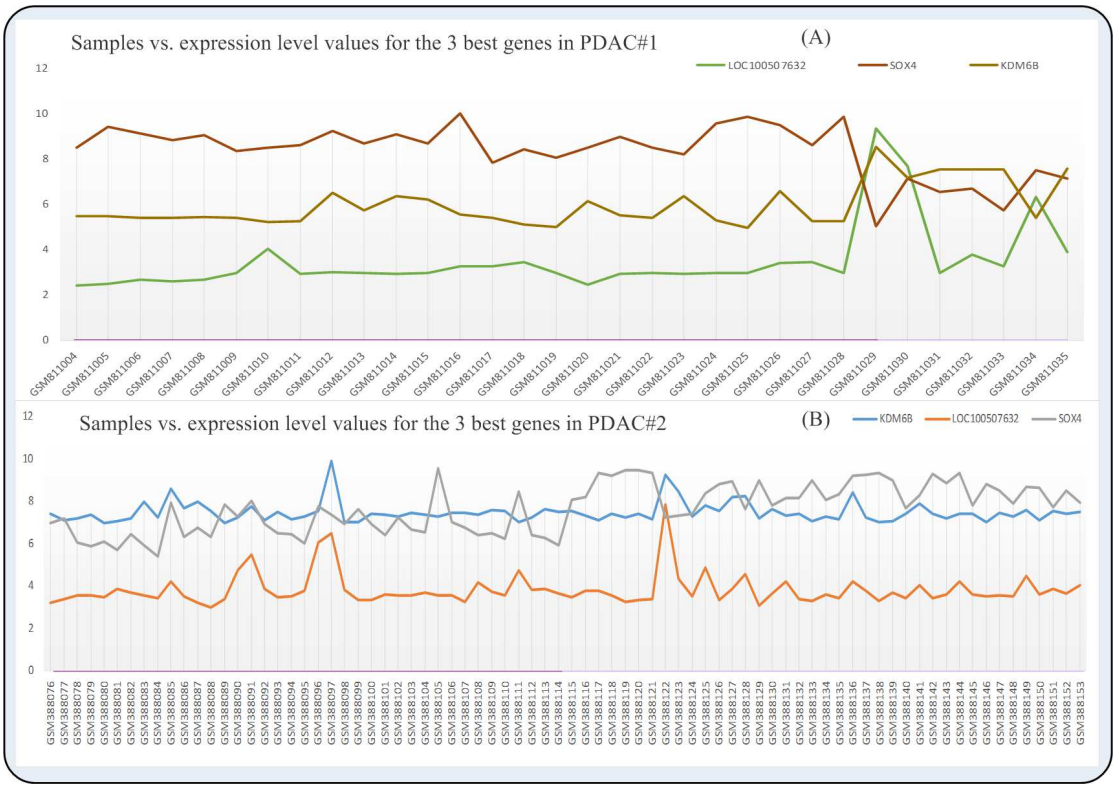

FIGURA 3.13: Gráfico de coordenadas paralelas asociando cada muestra de tejido con el nivel de expresión de cada uno de los tres genes seleccionados en PDAC-1.

3.3.3. Discusión y conclusiones

En este capítulo se ha presentado un *framework* de tipo *ensemble* que considera el problema de la inestabilidad inherente al proceso de selección de genes y la capacidad real de generalización de los biomarcadores seleccionados. Para ello, en primer lugar, lo importante no es sólo utilizar varios métodos, sino contar además entre ellos con diferentes criterios de selección para aumentar la validez del *set* estable.

Si nos centramos en la tasa de acierto cuando se utiliza un nuevo *dataset* distinto al de selección, si bien la alcanzada por el método implementado es alta, podría pensarse que la reducción observada haría más deseable otra aproximación de cara a la capacidad de generalización. Sin embargo, respecto a este objetivo (referente a la inestabilidad), la capacidad de generalización mostrada por el conjunto de marcadores obtenidos se halla más cercana a una realidad médica, y probablemente también a la realidad biológica. Debemos tener en cuenta que, si los marcadores seleccionados por muchas propuestas

que aparentemente presentan una alta capacidad de generalización fuesen evaluados mediante otro método de clasificación, los resultados podrían cambiar drásticamente. De hecho, esto es parte del problema para alcanzar un conjunto consensuado de biomarcadores en diferentes tipos de cáncer. La excesiva especificidad de muchos métodos que compiten por alcanzar la mayor tasa de acierto en un caso concreto, muy a menudo seleccionan marcadores que no serían útiles en clínica, pues dadas las diferencias en los conjuntos de datos, podrían no poseer la misma relevancia en otros casos, dificultando su aplicabilidad. Así pues, la capacidad demostrada por el *set* seleccionado mediante este método, muestra mejores resultados que cualquier otra de las técnicas que componen el *ensemble*, por lo que este *framework*, además de llevar a cabo una selección más robusta, no ve disminuida su capacidad de selección y clasificación frente a otras técnicas comunes en la bibliografía.

Además, se ha observado que el uso de diferentes técnicas y, fundamentalmente el enfoque de la fase de *wrappers*, favorece en gran medida la selección de marcadores complementarios, puesto que en ambas estrategias los genes se seleccionan o descartan de acuerdo a su aportación a la tasa de acierto alcanzada por el grupo en los diferentes clasificadores. Por otro lado, respecto a la reducción de los conjuntos de selección, se ha comprobado que con sólo 3 genes el método alcanza mejores resultados que las técnicas que componen la fase *ensemble* de forma individual, que seleccionan en su mayoría grupos de genes de mayor tamaño. Esto es importante también de cara al objetivo principal del *framework*, pues revela que la extracción de conjuntos de genes estables (desde el punto de vista del proceso de selección) no tiene que traducirse necesariamente en una menor eficacia en clasificación.

En cuanto a posibilitar la extracción de conocimiento, el *framework* no ofrece más información que cualquier método de selección de los consultados en la bibliografía, aunque sí nos permite asumir que dicha información es más fidedigna, en la medida en que las metodologías que combinan diferentes criterios de selección tratan de hacer frente a un sesgo en el filtrado que puede tener efectos muy variables en diferentes propuestas, patologías y conjuntos de datos.

Además de todo lo anterior, podemos señalar que la fase de *wrappers*, si bien no es su objetivo principal, cumple la tarea de encontrar relaciones entre genes y nos permite seleccionar marcadores complementarios, lo que hace de esta una propuesta más

completa y equilibrada en términos de cumplimiento de los distintos objetivos.

Como conclusión, en esta sección se ha propuesto un *framework* que trata de abordar el proceso de selección desde un punto de vista orientado a la clasificación, pero bajo la premisa de que una selección que priorice la estabilidad en dicho proceso repercute positivamente en la relevancia biológica y en el consenso en la elección de marcadores para diferentes conjuntos de datos de un tipo de cáncer dado.

Capítulo 4

Conclusiones finales

Tras el diseño, implementación y aplicación de los sistemas propuestos a tres casos de estudio y considerando las discusiones expuestas en el capítulo anterior, el presente capítulo refleja las conclusiones principales de esta investigación. Así pues, se analizan las tareas y observaciones realizadas en relación con los objetivos previamente establecidos y el orden en el que se señalan en el Capítulo 1. También se exponen las limitaciones de la propuesta y la capacidad mostrada por los sistemas para afrontar los principales obstáculos del área.

En primer lugar, observamos que, en cuanto a la búsqueda de biomarcadores complementarios, el uso del sistema basado en CBR ha demostrado ser una elección válida para seleccionar genes que funcionan bien juntos y afrontar la redundancia en el conjunto de datos. Como se ha explicado con anterioridad, la presencia exclusiva de genes con un elevado poder de clasificación individual en un conjunto seleccionado no implica que el grupo funcione mejor que otros en clasificación. Es fundamental que las características que componen dicho conjunto proporcionen información no redundante para la clasificación. Se ha podido comprobar que el proceso de selección basado en GBRT en la Sección 3.2 ofrece una forma de alcanzar esta meta junto con los *wrappers* BSW y FSW del método *ensemble* (de forma menos eficiente). De acuerdo con estas observaciones, podemos concluir que los métodos embebidos y *wrappers* constituyen la opción más apropiada de entre las técnicas estudiadas para lidiar con la redundancia, lo cual concuerda con las cualidades que se les atribuyen en la bibliografía.

Se ha alcanzado, por tanto, el objetivo de seleccionar características complementarias durante la extracción de genes orientada a la clasificación de tumores. Esto podría

llevarnos a pensar que los métodos embebidos son la opción más viable para la selección, pero aún hay muchos otros factores que deben ser tenidos en cuenta. De hecho, los *wrappers* y los métodos embebidos no favorecen la estabilidad del proceso, que es otro de los objetivos de la propuesta.

La aproximación basada en GBRT también ha sido capaz de abordar con éxito un caso de estudio de clasificación complejo, alcanzando elevadas tasas de acierto durante la validación en un nuevo *dataset*. La redundancia y los bajos tamaños muestrales son los principales problemas que afectan a la clasificación de subtipos tumorales. Sin embargo, podemos concluir que los métodos de la propuesta que nos permiten descubrir relaciones entre genes y, más concretamente, la aproximación basada en GBRT, son apropiados para distinguir casos de mayor complejidad como lo es el de subtipos de cáncer de pulmón de células no pequeñas, ya que recibe más importancia la selección de un grupo de marcadores que la de genes individuales. En este tipo de situaciones es especialmente importante no perder de vista este concepto, pues un gen alterado puede ser síntoma de muchas patologías, pero la combinación de determinadas características puede constituir una firma indicadora de situaciones más específicas. Por otro lado, el impacto del CBR es probablemente más notable en este tipo de casos, ya que el efecto producido por una baja muestra es mas problemático cuando se comparan subtipos tumorales con ciertas similitudes moleculares. Así, vemos que el éxito de la clasificación de subtipos se halla muy fuertemente relacionado con el objetivo de alcanzar un conjunto de genes que se complementan.

Continuando con la capacidad para clasificar clases de tumores con ciertas similitudes, hay que destacar que la selección de marcadores como SFTA2 permite realizar observaciones de interés para la clasificación. Los genes que codifican para la formación de proteínas asociadas a surfactante son marcadores del linaje alveolar de tipo 2 (AT2), que comparte muchas características moleculares con las células de adenocarcinoma [81] (y es el origen de algunos tejidos pulmonares). De hecho, recientemente se ha concluido que las células AT2 son las más frecuentes entre las iniciadoras de adenocarcinoma pulmonar inducido por K-RasG12 [155]. Esto no significa, sin embargo, que las células AT2 sean siempre las iniciadoras de adenocarcinoma de pulmón. Además, la sugerencia aquí realizada de emplear SFTA2 como marcador ha sido apoyada por estudios posteriores. Sería interesante, por tanto, estudiar la expresión de estas proteínas asociadas a surfactante en el adenocarcinoma, puesto que nos pueden estar aportando información acerca

de un nivel de clasificación más profundo. En consecuencia, seleccionar genes que son característicos de un linaje celular o tejido y que se hallan además alterados en cáncer resulta de gran utilidad para la clasificación y caracterización.

Durante la aplicación de los sistemas propuestos en esta investigación se ha observado que muchos genes presentan un poder de discriminación similar. Así, el problema subyacente de maximizar la precisión y minimizar el tamaño de los subconjuntos seleccionados es nuevamente la redundancia en los datos. En este trabajo se ha mostrado que no es necesario contar con un elevado número de genes para discriminar clases tumorales. La propuesta ha afrontado con éxito el problema de la redundancia y cada uno de los *frameworks* alternativos ha obtenido elevadas tasas de acierto utilizando un grupo de marcadores reducido. En las Secciones 3.1 y 3.3 (ambas orientadas a la clasificación), ha sido comprobado que los métodos que consideran las relaciones entre características constituyen la principal forma de combatir la redundancia. No obstante, no se trata de la única vía de hacerlo, pues la Sección 3.2 revela que la selección de genes frontera puede aportar una ayuda interesante en las aproximaciones que hacen uso del *clustering*, al definir un *cluster* mediante una cantidad menor de puntos que lo delimitan. En algunos casos, sin embargo, puede ser recomendable realizar una reducción final del conjunto seleccionado. En tales situaciones no se trata exactamente de filtros que se aplican al final del proceso, sino más bien de retirar genes de forma iterativa y estudiar el impacto en el potencial del grupo persiguiendo un equilibrio entre aplicabilidad y capacidad de clasificación (lo cual es, en definitiva, una especie de *wrapper* final).

En resumen, se ha observado que la reducción de la redundancia, uno de los problemas principales de la selección de genes, puede paliarse en gran medida mediante la apropiada elección de los algoritmos, sus parámetros y su orden de aplicación. Además, la presente investigación ha alcanzado la meta de reducir conjuntos de genes relevantes a cantidades manejables sin sacrificar capacidad de clasificación. Por otro lado, pese a la orientación inicial del *framework* en la Sección 3.2 hacia otros objetivos, su inesperado potencial para clasificar hace interesante estudiar con más detenimiento la implicación de seleccionar genes frontera.

Como se ha explicado, la propuesta incluida en el Capitulo 3 trata también de desarrollar estrategias capaces de afrontar la inestabilidad en el proceso de selección. Durante el diseño experimental se ha podido comprobar que, si bien la redundancia

es uno de los problemas más importantes y más frecuentemente mencionados en la bibliografía, lidiar con la inestabilidad biológica y metodológica supone un objetivo más complejo de alcanzar. Se ha observado que esto es debido a que muchos de los procesos que generan inestabilidad no tienen lugar durante el tratamiento bioinformático, sino en etapas previas. Las diferencias tecnológicas y metodológicas durante la construcción de los *datasets* constituyen así un problema de gran complejidad que, desde el punto de vista de los datos, tiene consecuencias semejantes a la inestabilidad génica: se trata de procesos que están generando variabilidad en el conjunto de datos antes de que este llegue a nuestras manos, momento en el cual todos estos cambios ya están reflejados en los datos y son irreversibles.

Como se ha visto, el problema anterior se halla agravado por el reducido tamaño muestral, que incrementa la dificultad de desarrollar un *framework* que sea capaz de generalizar. En este punto, en las primeras etapas del estudio se pudo observar que otro obstáculo reside en que, durante el análisis, se está añadiendo inestabilidad a los resultados por usar técnicas específicas y por no validar apropiadamente aquellos. Por tanto, no solo es importante pensar en métodos para paliar los errores que ya han afectado a nuestros datos, sino también diseñar nuestra metodología de forma que evitemos en la medida de lo posible la inestabilidad generada en los resultados durante el proceso de selección, pues esta es probablemente la única etapa que somos realmente capaces de controlar. Estos factores acaban afectando de forma significativa a la aplicabilidad de cualquier propuesta de selección. Sin embargo, utilizar únicamente filtros y no otro tipo de técnicas para no depender de un modelo de clasificación sería trasladarse al otro extremo metodológico y tampoco ofrecería una solución, ya que aquellos carecen de determinadas habilidades importantes para afrontar otros problemas (como la redundancia).

En cuanto al objetivo de extraer conocimiento, al discutir acerca del papel de los marcadores seleccionados en clasificación (como en el caso de SFTA2) y acerca de la proposición de nuevas clasificaciones y subgrupos, ya estamos poniendo de manifiesto que la propuesta es capaz de generar conocimiento; y este puede motivar posteriores investigaciones. A este respecto, se ha visto que la selección de biomarcadores complementarios es en realidad uno de los procesos que más información útil ofrece: obtener un grupo de marcadores que se complementan no es sólo necesario para clasificar subtipos tumorales, sino también para establecer nuevas clases y caracterizarlas. Como se

ha explicado, algunos marcadores no son seleccionados por su potencial individual, sino por el poder combinado que ofrecen en el grupo. En este sentido, cuando se exploran los resultados en la Sección 3.1, se puede ver que la reducida capacidad individual de AKR1B10 como marcador se debe al hecho de que los cambios en su nivel de expresión solo se corresponden claramente con el subtipo tumoral en algunas muestras, pero no en todas (Figura 3.5). Aquí se podrían estar dando, por tanto, varias posibilidades. Una de ellas implicaría que este gen no es un buen marcador y que no se encuentra asociado con una clase de tumor, sino con cualquier otra condición o factor enmascarado por la reducida muestra. No obstante, esta opción no parece probable, dado que se cuenta en este caso con mayor muestra en el *dataset* de validación y AKR1B10 sigue contribuyendo a aumentar la tasa de acierto. También existe la posibilidad de que haya una razón biológica relacionada con el tumor para las diferencias en los niveles de expresión, lo que podría estar sugiriendo a su vez que existen subgrupos en el conjunto de datos o que quizás la actual no es la clasificación más adecuada de acuerdo a factores moleculares.

En la línea de la extracción de conocimiento también se ha demostrado que los genes frontera pueden ser marcadores, apoyando la idea de que no es necesariamente mejor escoger sólo genes cercanos al centroide para la detección de aquellos. Además, en la Sección 3.2, estos genes alcanzan una elevada tasa de acierto en clasificación y, cuando se escogen muchos puntos de un mismo *cluster*, el cálculo de fronteras puede constituir una buena opción para paliar el incremento de redundancia en la selección, teniendo en cuenta que definen la estructura espacial del *cluster* y constituyen al mismo tiempo la elección menos redundante de entre los puntos pertenecientes a este. Además, hasta donde se ha podido comprobar, la aproximación desarrollada en este trabajo incluye la primera aplicación que se hace del cálculo de puntos frontera a la selección de genes.

Podemos concluir, por tanto, que cada sistema alternativo de la propuesta permite extraer conocimiento biológico y estadístico, si bien el *clustering* jerárquico en el *framework* híbrido de la Sección 3.2 es claramente capaz de alcanzar una mayor profundidad en el estudio del conjunto de datos. Además se ha observado que el conocimiento más fiable y útil que podemos extraer mediante selección de genes es acerca de las relaciones entre características y grupos de patrones de expresión, puesto que aquellas pueden sugerir clasificaciones alternativas para determinados tumores.

En cuanto a la aplicabilidad en tareas de clasificación, hay que decir que los distintos sistemas ofrecen resultados satisfactorios y que, aunque no dispongamos de la posibilidad de realizar inmunohistoquímicas, el *framework ensemble* y basado en CBR son más apropiados en este aspecto, al menos desde el punto de vista metodológico. Esto se debe a que la aproximación empleada en la Sección 3.2 podría ofrecer mayor variabilidad en sus resultados debido al uso de determinados procedimientos específicos del *framework*, principalmente la selección de un nivel en el *clustering* jerárquico (si bien la selección de genes frontera ha arrojado resultados positivos). Sin embargo, la capacidad de generalización de este *framework* aún no se ha validado en distintos *datasets*. El sistema basado en GBRT parece arrojar mejores resultados de clasificación, pero el *ensemble* es sin duda más robusto, por lo que en realidad el primero podría obtener marcadores menos aplicables a clínica pese ser más preciso en los casos estudiados. Debemos señalar, no obstante, que la integración del sistema CBR y su etapa de características sería capaz de mejorar el *framework ensemble* expuesto en la Sección 3.3.

Si consideramos todas estas observaciones, es inevitable preguntarse cuál de los métodos de selección es más apropiado o soluciona un mayor número de problemas. Cuando se aplican filtros estadísticos, surgen determinadas preguntas que debemos hacernos, teniendo en cuenta que en los últimos años se ha incrementado de forma notable la discusión en distintos estudios acerca de la significancia de los genes seleccionados por estos métodos. Sin embargo, un estudio en mayor profundidad nos permite concluir que la mayoría de problemas señalados a este respecto concierne con mayor frecuencia al uso de un p-valor que al del algoritmo en sí mismo. Las grandes preguntas –¿existe un p-valor? y ¿hay un *threshold* óptimo?– permanecen sin responder. Sin embargo, es una realidad que los filtros, si bien presentan carencias, ofrecen muchas otras ventajas en selección de genes. A pesar de que pudiese parecer trivial, es importante ser conscientes de que, aunque no conocemos el punto en que deberíamos cortar un conjunto de datos en muchos casos, el hecho de que se trate de un *ranking* es de vital importancia, pues nos permite comparar la relevancia de las características. Dado el uso que se ha hecho en este estudio de los filtros, hemos podido observar que el problema del *threshold* podría ser menos notable cuando se aplica el filtro en la fase inicial, en la que contamos con decenas de miles de características en el conjunto de datos. Al hacer esto, se asume, por supuesto, que los genes seleccionados al final de todo el proceso son suficientemente significativos desde el punto de vista estadístico como para que, en el caso de que hubiera

habido pequeñas variaciones del valor de corte, su selección no se hubiera visto afectada. En el caso de la Sección 3.2, se revisaron los genes finalmente seleccionados para comprobar que el p-valor que alcanzaron en los *tests* iniciales no se hallaba cerca del punto de corte. Se observó que, pese a presentar distintas significancias, todas las características distaban mucho de acercarse al *threshold* establecido. En resumen, la utilidad de los filtros es innegable pero, conociendo sus limitaciones, es preciso ser consciente de las distintas clases de errores que podemos cometer al establecer un *threshold* durante su uso. Asimismo, sería interesante estudiar el impacto de emplearlos en distintos puntos del proceso de selección.

En este estudio se ha procedido aplicando tres sistemas por separado para evaluar de forma apropiada cada metodología, comprendiendo su capacidad individual para alcanzar los objetivos propuestos. A este respecto surge inevitablemente la pregunta de si los *frameworks* aquí expuestos son combinables y si es apropiado combinarlos. La respuesta corta sería que muchos de los algoritmos son integrables en determinados puntos de otro de los *frameworks*, aunque no todos ellos. A pesar de que la Sección 3.1 alcance mejores resultados de clasificación en el *dataset* de evaluación, de cara a la aplicabilidad y a la extracción de conocimiento sería más apropiado considerar más métodos en un entorno de tipo *ensemble*. La aproximación basada en GBRT podría, entonces, mantener su eficacia si los datos siempre se obtuviesen de la misma forma y se reprodujesen las mismas condiciones (lo cual no es una situación realista). Por ese motivo podría ser demasiado dependiente del criterio del algoritmo y quizás existan otros marcadores con mayor potencial. Así pues, hay que considerar como una prioridad la obtención de características relevantes para el contexto biológico (que sean más estables y seleccionables por distintos métodos), porque de otro modo la aplicabilidad sería mucho menor.

En suma, la cuestión que se pone aquí de manifiesto indirectamente es que los objetivos de clasificación y diagnóstico han sido a menudo confrontados con los de extracción de conocimiento, pero se trata en realidad de una división falsa. Si bien es cierto que dar prioridad a una u otra vía puede dar lugar a *frameworks* de análisis diferentes, en clínica existen muchos factores que no controlamos y por los que la eficacia de un método específico podría verse muy afectada. Así, como se ha visto, en realidad un *ensemble* que alcance menor tasa de acierto podría también ver su capacidad de clasificación menos

resentida al aplicarse a clínica, ya que su selección es más estable y por ello, más relevante biológicamente. En consecuencia es más probable que otros métodos consideren importantes los marcadores que utiliza este *ensemble*. Tratar de buscar, por tanto, una solución específica de clasificación maximizando la tasa de acierto a cualquier coste dista mucho de ser una solución eficaz; otros sistemas con una capacidad aparentemente menor podrían ser mucho más aplicables, al haber realizado una selección estadística y biológicamente significativa. De otra forma se pierde de vista el objetivo último de la investigación en el área, pudiendo seleccionar características que permitan alcanzar tasas de acierto elevadas a medida de la situación y posiblemente irrelevantes.

Por tanto, en relación con lo anterior, la selección basada en GBRT podría integrarse entre los métodos del *ensemble* de la Sección 3.3 y el CBR ha demostrado ser una metodología aplicable a este mismo *framework*. Por otro lado, el *clustering* jerárquico implica el problema de decidir un *threshold* y escoger una división realmente significativa, pero a su vez permite realizar otro tipo de averiguaciones. Pese a ello, hay que resaltar que la selección de genes frontera ha ofrecido mejores resultados de los esperados en clasificación. En vista de las observaciones realizadas en cada método, cabe mencionar que, por el momento, un *framework* híbrido que integre un *ensemble* permite abarcar un mayor número de problemas. Así, algunos de los módulos de los distintos *frameworks* propuestos podrían ser combinados o integrados en el *framework* de la Seccion 3.3.

Se ha expuesto, por tanto, que los *wrappers* y los métodos embebidos ofrecen ciertas ventajas para el proceso de selección distintas a las de los filtros, y que el *ensemble* es la mejor respuesta para afrontar la inestabilidad del proceso.

La conclusión es que un *framework* híbrido, que contenga tanto una fase de filtrado como un *ensemble* de distintas técnicas, es por el momento la opción más apropiada para la selección. Aun así, todavía no todos los problemas pueden ser solucionados por un solo *framework*.

En el caso específico de la Sección 3.2, la estocástica de los niveles de expresión y baja muestra ha ofrecido ciertas dificultades, pues establecer grupos de edad en tan pocos datos implica muchos retos y no hay que perder de vista que se trata de tejidos de distintos pacientes a lo largo del tiempo, lo cual también podría afectar a la validez del análisis de resultados. Sin embargo, este es un problema específico de los datos y del uso de la edad, no del *framework* cuando contempla un factor de estudio adicional. No

obstante, sería conveniente añadir una alternativa que no precise de discretización en los datos para este tipo de variables.

Finalmente, cabe destacar que no todos los problemas son de naturaleza científica, pues algunos de los más difícilmente paliables de cara a la aplicabilidad de las propuestas de selección residen en la comunicación interdisciplinar.

En términos generales, se han ofrecido en este trabajo soluciones que permiten el análisis de conjuntos de datos de expresión génica, y se han alcanzado resultados positivos en clasificación, lo que posibilita extraer conocimiento que alimente futuros estudios. Así, si bien es cierto que hoy día existe una constante e intensa revisión y crítica de los métodos de selección –lo que pone de manifiesto la increíble dificultad de escoger cada algoritmo, sus parámetros y la forma apropiada de combinarlos–, hasta ahora muchos descubrimientos de relevancia médica se han realizado por medio de estas técnicas. Los objetivos principales se han visto satisfechos, pero durante el proceso se ha podido comprobar cuáles son los problemas de más difícil solución a nivel general en el área. Por eso es necesario sumarse a la autocrítica y, mientras se continúan buscando y encontrando soluciones para las diversas carencias en la selección de genes y caracterización de tumores, hay que poner al alcance de otros investigadores las posibilidades que ofrecen estas técnicas, pues no podemos ya concebir el futuro de la biomedicina sin ellas.

Como se ha expuesto, una de las dificultades de la selección de genes en cáncer es clasificar una diversidad inconmensurable, pues no existen dos cánceres iguales y durante mucho tiempo nos hemos basado en esquemas demasiado generalistas. Clasificar es una forma instintiva de aprender que ha sido frecuentemente utilizada en la historia de la medicina, tratando de imponer un orden en una diversidad caótica a nuestros ojos [138]; si bien la clasificación puede ser de gran utilidad, no debemos perder de vista que es un artificio, igual que lo es compartimentar la investigación.

Conclusions *(English version)*

After designing, implementing and applying the proposed systems to three real case studies and having also considered the discussions in previous sections, this chapter contains the main conclusions from this research. The accomplished tasks and the observations provided are analyzed in relation to the established objectives. Limitations and other issues affecting the performance of the proposal have been also considered.

First, regarding the search of complementary biomarkers, the use of GBRT has proved to be an appropriate choice for selecting genes that work well together, facing the redundancy in data. As previously explained, the presence in a marker subset of only genes with high individual discrimination power does not imply that the group will perform better in classification. Features in the selected group must provide non redundant information for classification tasks. GBRT based selection process in Section 3.2 provides a way to achieve this goal, as also do both backward and forward selection wrappers in Section 3.3 (in a less effcient way). We hence can conclude that wrappers and embedded methods deal better with redundancy (which is coherent with the properties of such methods reported in the literature).

Thus, the objective of selecting tumor classification oriented complementary features has been accomplished. This could lead to think that embedded methods are the most suitable option for selection, but there are many other factors to consider. In fact, it has been shown that these methods do not favor the stability of the selection process, which is one of the other goals of this proposal.

When dealing with complex classification cases, the GBRT based approach has also performed successfully, offering high accuracy results when validated in a different dataset. Redundancy and low sample sizes are the main problems affecting tumor subtypes classification. However, we can conclude that the analysed in the proposal allowing the

discovery of relations between genes and more concretely the GBRT approach, are suitable to distinguish high complexity cases like the one with NSCLC subtypes. This is due to the fact that the selection of a biomarker group is receiving more attention than individual genes. In this scenario is especially important to keep track of this "group" concept, as an altered gene can be a symptom of many disorders, but the combination of more altered features may constitute a signature for specific situations. In addition, the impact of CBR is probably more notable in this case, since the effect produced by a low sample size is even more severe when comparing tumor types with certain molecular similarities. Thus, the classification success is closely related to finding complementary genes.

Further analyzing the subtype discrimination capability, it should be noticed that the selection of some markers such as SFTA2 leads us to make some interesting considerations for classification. Genes encoding surfactant proteins are markers for Alveolar type 2 lineage (AT2), which shares many molecular characteristics with adenocarcinoma cells (and originates many lung tissues) [81]. In fact, Xu *et al.* [155] more recently concluded that AT2 cells are the predominant cancer-initiating cells of K-RasG12D induced lung adenocarcinoma. Also the suggestion of SFTA2 as a potential biomarker has been also supported in later works. Nevertheless, AT2 cells are not always the initiating cells of LUAD [46], and it would be interesting to further study the expression patterns of SFTA2 in LUAD, as they could provide us with information of a deeper classification level. Thus, it is very useful for discrimination and characterization purposes to select altered genes related to a certain cell lineage or tissue.

While selecting genes, it has been observed that many of them present a similar discrimination power. Hence, the problem underlying the process of minimizing the size of a selected subset while maximizing the accuracy turns out to be, once more, the redundancy in the data. In this work, it has been shown that it is not necessary to have big groups of features available on order to distinguish between tumor classes. The proposal has successfully dealt with the redundancy problem and every alternative framework has achieved high accuracy results by using a reduced marker set. In Sections 3.1 and 3.3 (both classification oriented), we have observed that the use of methods considering relation between features is the main way to fight redundancy. However, this is not the only way of facing this problem; in this sense, Section 3.2 reveals that the use of boundary genes can provide an advantage in those frameworks based on

clustering, by defining each cluster through a smaller set of genes shaping it. However, in final stages of the selection process a final reduction of the results provided by these methods may be necessary. By this, we actually do not mean apply filters to the final set but iteratively remove relevant genes from it to understand their contribution, and find the balance between applicability and accuracy (which, in fact, could be considered as some kind of wrapper).

In summary, it has been seen that redundancy reduction, one of the main problems of gene selection, can be faced by properly choosing the algorithms, their order of application and their parameters. Moreover, the present proposal has reached the goal of reducing the selected subsets to manageable sizes without sacrificing discrimination capability. Furthermore, the unexpected potential of the framework in Section 3.2 in classification tasks (as it was initially oriented to focus on other objectives) motivates further research about the implications of selecting genes in boundaries.

The proposal exposed in Chapter 3 also aims to develop strategies capable of facing instability in selection process. While carrying out the experimental design, it was noticed that dealing with redundancy may be one of the main issues in the area, nevertheless solving both biological and methodological instability probably conforms a more challenging task to accomplish. This is due to the fact that many of processes generating instability do not take place during the bioinformatic data treatment, but at previous stages. The technological and methodological differences during the generation of the data constitute a very complex problem that, in terms of data treatment, has similar consequences as biological instability: there are processes generating variability before the dataset is built and given to the analyst. At this point, all of these changes are already reflected in the data and they are irreversible.

The small sample size aggravates this problem, as it increases the difficulty of developing a framework capable of generalizing. At this point, another issue was observed in the fact that, during the analysis, we add more instability to results while using specific techniques, or by not properly validating the results. Consequently, it is not only important to think about methods to compensate those errors that have already conditioned our data, but also to design our methodology in a way that avoids the instability during the selection process, as this is probably the only part we are able to control. Eventually, this can greatly affect the applicability of any proposal. However, as it has been shown,

using only filters to not to depend on a classification model is not a solution, as such models lack some necessary abilities to face other problems (like redundancy).

Regarding the knowledge extraction, it should be noticed that, when discussing about the role of selected markers in classification (like in the case of SFTA2), we are already showing that this proposal is able to generate new knowledge that can motivate further research. Concerning this, we can see that the selection of complementary biomarkers is one of the processes offering more valuable information. It has been observed that they are not just useful for subtype classification, but for establishing new tumor subtypes and characterize them. As explained, some biomarkers are selected not because of their individual classification potential, but for their combined power. For example, when exploring the results in Section 3.1, it can be noticed that the limited individual power of AKR1B10 is due to the fact that its expression level only corresponds clearly to the tumor subtype in some samples. There are different possibilities for this phenomenon. One of them is that such a gene is not a good marker because it is not related to cancer but to other condition or factor masked by the small sample sizes. However, this does not seem to be the case because in this instance a biggest sample size is available in the validation dataset and AKR1B10 still contributes to increase accuracy. Another possibility is that there is some biological reason underlying the differences in expression inside a tumor class, which can be supporting the existence of subgroups or just showing that this classification is not the most adequate according to molecular factors.

It has been also shown that boundary genes can be biomarkers, showing that it is not necessarily better for biomarker detection to choose genes near the centroid. Indeed, as seen in Section 3.2, such genes reach very high accuracy in classification. Furthermore, they can be a good option in order not to increase redundancy during selection when choosing several points from a same cluster, since they determine the spacial structure of the cluster while being the less redundant option among the points in such cluster. Moreover, to the best of our knowledge, the approach presented in the present work is the first one in applying the boundary points computation to gene selection.

We can hence conclude that biological and statistical knowledge can be extracted from every alternative system, but the hierarchical clustering hybrid framework in Section 3.2 clearly provides a deeper insight in the data. Moreover, it has been observed that the most reliable knowledge that can be extracted from gene selection regards relation

between features and distribution of values, which may suggest a different classification for some tumors.

Regarding the applicability to classification tasks, the three alternative systems offer satisfactory results and, even if we are not able to perform immunohistochemical analysis, CBR and ensemble approaches have proven to fit better to such goals, at least from a methodological point of view. This is due to the fact that the approximation analyzed in Section 3.2 could provide higher variability results because of certain specific processes, namely the selection of a clustering level. Nevertheless, boundary gene selection has reached promising results, but generalization capabilities of this framework have not been validated yet. The GBRT based system seems to get better results when classifying but the ensemble is more robust. Thus, regardless of the higher precision shown by the first in the case study, it could be more difficult to apply it to clinic. However, we must highlight that integrating the CBR system and its feature selection module could greatly improve the ensemble framework performance in Section 3.3.

When considering all these observations it is unavoidable to wonder which of the selection methods used in this research is more appropriated or which one solves more problems. During the application of statistical filters, some questions to consider arise, as in recent years there has been an increasing discussion across different studies regarding the significance of selected genes by these methods. Nevertheless, a deeper study allows us to understand that the main addressed issue is not normally the filter algorithm itself, but the use of a p-value. The main questions of ranking methods –is there an optimal threshold? or is there a p-value?–, remain unanswered. However, filter methods have many advantages for gene selection. It may sound trivial, but it is important to consider that even if we do not know where to "cut" a dataset, having a ranked list is of great relevance as it makes it possible to compare features. Giving the use made of filters in this research, it has been possible to observe that this threshold problem may be less notable when applying the filter in initial stages. By doing this, we are of course assuming that final selected genes are statistically significant enough not to be affected by small variations in the threshold. In the Section 3.2 case, selected genes were revised in order to check if the p-value they reached in initial filters was not close to the threshold value. It was observed that, regardless of the difference between the significance of selected genes, all of them were far from the cut-off point. In summary, utility of filters is undeniable but, considering their limitations, we must be aware of the

errors we could make when reducing the dataset. In addition, it would be interesting to study the impact of using such methods in different stages of the selection process.

To properly evaluate each methodology, the present work applies three individual systems to study their capability in order to reach the established objectives separately. In this regard, one can wonder if the methods introduced are combinable and if it is recommendable to do so. The short answer would be that many of the algorithms of each framework could be added to others, but not all. In spite of Section 3.1 reaching the best classification results when evaluated, it would be more appropriate to use more methods in an ensemble environment for applicability and knowledge extraction purposes. The method in Section 3.1 could then maintain its efficiency as long as the data were always obtained in the same way and the experimental conditions were exactly the same (which is definitely not a realistic situation). Because of this, such a method could be too dependent on the algorithm criterion, and other markers with higher potential may exist. Consequently, obtaining relevant features (stable and selectable by different methods) according to the biological context must be considered a priority. Otherwise, applicability would be greatly limited.

The issue exposed here is the fact that diagnosis and classification objectives have been often confronted with knowledge extraction in the past. However, this is, in fact, a false division. It is true that focusing on one of this objectives can result in different frameworks but, in clinic there are many factors that we do not control and thus the capacity of a specific method could be greatly affected. As exposed, an ensemble system reaching smaller accuracy values could be also less affected when applied to diagnosis, since the selection it performs is more stable and probably more relevant from a biological point of view. Because of this, it is more likely that other methods will consider the markers used by this ensemble system more relevant.

Searching for a specific classification solution while maximizing accuracy at any cost is far from being an effective and applicable solution. Also, other systems with an apparent lower capacity could be much more useful as they perform a significant selection from both statistical and biological points of view. Otherwise, we could be loosing sight of the main research objective, while selecting features (possibly irrelevant) that allow to reach high accuracy values by fitting an specific situation, while other systems with

an apparent lower capacity could be much more useful, as they perform a significant selection from both statistical and biological points of view..

In relation to the previous, GBRT based selection could be integrated in the group of methods forming the ensemble in Section 3.3 and CBR has shown to be an appropriated methodology that could be also applied to the latter. On the other hand, hierarchical clustering, though involving the issue of establishing a threshold to choose a significant cut. However, it allows to make other findings. In spite of this, boundary genes selection has provided better results than the expected in classification tasks. Given the observations made in each system, we can conclude that a hybrid framework involving an ensemble allows to address a wider group of different issues. Thus, some of the methods in Sections 3.1 and 3.2 could be integrated in the framework in Section 3.3.

It has been exposed that both wrapper and embedded methods are providing some advantages to the selection process which differ from those given by filters. In general terms, ensemble methods seem to be the best option to face the instability in the selection process.

We can conclude that hybrid frameworks containing an enssemble of different techniques may be the most appropriate choice for selection at the moment. Even then, not all issues can be solved by just one framework.

In the specific case of Section 3.2, the stochastic in expression levels and small sample size have been especially challenging when establishing age groups. We should take into account that we are comparing different patients across time, which can affect the validity of result analysis. However, this is an issue concerning the data and age values, but not related to the framework considering an additional factor. In any case, it would be interesting to integrate an alternative which does not require data discretization.

Finally, it should be noticed that not every problem is scientific in nature, as some of the most difficult to solve when speaking about applicability lie on interdisciplinary communication.

In general terms, this research has offered solutions for analyzing gene expression data, reaching positive results in classification tasks, which allow to extract knowledge to feed future works. Nowadays there exist an increasing review and critic of the selection methods, exposing the high complexity behind the algorithm election and the way to

combine them. However, many relevant medical findings have been made by means of these algorithms. The main objectives of this research have been reached, but it has also been possible to determine which problems the most difficult to solve within the area. Thus, it is necessary to join the self-critic and provide others researchers with the power offered by these methods, while the search for new solutions continues, since the future of biomedicine cannot be conceived without these techniques anymore.

As explained, one of the main obstacles in cancer gene selection lies on classifying an immeasurable diversity, as there are not two equal cancers and we have been based in very generalist schemes for too long. To classify is an instinctive way of learning, commonly used in medicine history, for trying to impose an order to a chaotic diversity [138]. This can be of great utility, but we cannot forget that it is something artificial, as it is to compartmentalize research.

Bibliografía

[1] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398.

[2] Agnar, A. and Plaza, E. (1994). Case-Based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59.

[3] Aguinis, H. (1993). Action research and scientific method: Presumed discrepancies and actual similarities. *The Journal of Applied Behavioral Science*, 29(4):416–431.

[4] Ahdesmäki, M. and Strimmer, K. (2012). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *Annals of Applied Statistics*, 4(1):503–519.

[5] Ahdesmaki, M., Zuber, V., Gibb, S., and Strimmer, K. (2015). *sda: Shrinkage Discriminant Analysis and CAT Score Variable Selection.* R package version 1.3.7, http://CRAN.R-project.org/package=sda.

[6] Anaissi, A., Goyal, M., Catchpoole, D. R., Braytee, A., and Kennedy, P. J. (2015). Case-based retrieval framework for gene expression data. *Cancer Informatics*, 14:21–31.

[7] Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2016). Supervised, unsupervised and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5):737–742.

[8] Argon, A., Nart, D., and Veral, A. (2015). The value of cytokeratin 5/6, p63 and thyroid transcription factor-1 in adenocarcinoma, squamous cell carcinoma and non-small-cell lung cancer of the lung. *Turkish Journal of Pathology*, 31(2):81–88.

[9] Arshadi, N. and Jurisica, I. (2004). Maintaining case-based reasoning systems: A machine learning approach. In *European Conference on Case-Based Reasoning*, pages 17–31.

[10] Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

[11] Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., and Napolitano, A. (2012). A review of the stability of feature selection techniques for bioinformatics data. In *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012*, pages 356–363.

[12] Badea, L., Herlea, V., Dima, S. O., Dumitrascu, T., and Popescu, I. (2008). Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepato-Gastroenterology*, 55(88):2016–2027.

[13] Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A. M., Gingras, M. C., Miller, D. K., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52.

[14] Bandara, I. A., Baltatzis, M., Sanyal, S., and Siriwardena, A. K. (2018). Evaluation of tumor M2-pyruvate kinase (Tumor M2-PK) as a biomarker for pancreatic cancer. *World Journal of Surgical Oncology*, 16(1):56.

[15] Beca, F. and Polyak, K. (2016). Intratumor heterogeneity in breast cancer. In *Novel biomarkers in the continuum of breast cancer*, pages 169–189. Springer.

[16] Berrar, D. P., Dubitzky, W., and Granzow, M. (2003). *A practical approach to microarray data analysis*. Kluwer Academic Publishers, New York.

[17] Bhaw-Luximon, A. and Jhurry, D. (2015). New avenues for improving pancreatic ductal adenocarcinoma (PDAC) treatment: Selective stroma depletion combined with nano drug delivery. *Cancer Letters*, 369(2):266–273.

[18] Bind, S., Tiwari, A. K., and Sahani, A. K. (2015). A survey of machine learning based approches for parkinson disease prediction. *International Journal of Computer Science and Information Technologies*, 6(2):1648–1655.

[19] Birnbaum, D. J., Bertucci, F., Finetti, P., Birnbaum, D., and Mamessier, E. (2018). Molecular classification as prognostic factor and guide for treatment decision of pancreatic cancer. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1869(2):248–255.

[20] Bø, T. and Jonassen, I. (2002). New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 4(4):research0017.1–research0017.11.

[21] Botling, J., Edlund, K., Lohr, M., Hellwig, B., Holmberg, L., Lambe, M., Berglund, A., Ekman, S., Bergqvist, M., Pontén, F., König, A., Fernandes, O., Karlsson, M., Helenius, G., Karlsson, C., Rahnenführer, J., Hengstler, J. G., and Micke, P. (2013). Biomarker discovery in non-small cell lung cancer: Integrating gene expression profiling, meta-analysis and tissue microarray validation. *Clinical Cancer Research*, 19(1):194–204.

[22] Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79.

[23] Caiado, F., Silva-Santos, B., and Norell, H. (2016). Intra-tumour heterogeneity – going beyond genetics. *FEBS Journal*, 283(12):2245–2258.

[24] Callea, M., Albiges, L., Gupta, M., Cheng, S.-C., Genega, E. M., Fay, A. P., Song, J., Carvo, I., Bhatt, R. S., Atkins, M. B., Hodi, F. S., Choueiri, T. K., McDermott, D. F., Freeman, G. J., and Signoretti, S. (2015). Differential expression of PD-L1 between primary and metastatic sites in clear-cell renal cell carcinoma. *Cancer Immunology Research*, 3(10):1158–1164.

[25] Castellanos-Garzón, J. A. and Díaz, F. (2013). An evolutionary computational model applied to cluster analysis of DNA microarray data. *Expert Systems with Applications*, 40(7):2575–2591.

[26] Castellanos-Garzón, J. A., García, C. A., Novais, P., and Díaz, F. (2013). A visual analytics framework for cluster analysis of DNA microarray data. *Expert Systems with Applications*, 40:758–774.

[27] Castellanos-Garzón, J. A., Ramos González, J., López-Sánchez, D., de Paz, J. F., and Corchado, J. M. (2018). An ensemble framework coping with instability in the gene selection process. *Interdisciplinary Sciences: Computational Life Sciences*, 10(1):12–23.

[28] Chan, A., Diamandis, E. P., and Blasutig, I. M. (2013). Strategies for discovering novel pancreatic cancer biomarkers. *Journal of Proteomics*, 81:126–134.

[29] Chan, B. A. and Hughes, B. G. M. (2014). Targeted therapy for non-small cell lung cancer: Current standards and the promise of the future. *Translational Lung Cancer Research*, 4(1):36–54.

[30] Chan, W. H., Mohamad, M. S., Deris, S., Zaki, N., Kasim, S., Omatu, S., Corchado, J. M., and Al Ashwal, H. (2016). Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme. *Computers in Biology and Medicine*, 77:102–115.

[31] Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S. H., and Zhang, M. Q. (2002). Evaluation and comparison of clustering algorithms in anglyzing es cell gene expression data. *Statistica Sinica*, pages 241–262.

[32] Cheng, D., Kannan, R., Vempala, S., and Wang, G. (2007). A divide-and-merge methodology for clustering. *ACM Transactions on Database Systems*, 31(4):1499–1525.

[33] Chipman, H. and Tibshirani, R. (2006). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, 7:302–317.

[34] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73.

[35] Choi, S. K., Pandiyan, K., Eun, J. W., Yang, X., Hong, S. H., Nam, S. W., Jones, P. A., Liang, G., and You, J. S. (2017). Epigenetic landscape change analysis during human EMT sheds light on a key EMT mediator TRIM29. *Oncotarget*, 8(58):98322–98335.

[36] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699–705.

[37] Crnogorac-Jurcevic, T., Chelala, C., Barry, S., Harada, T., Bhakta, V., Lattimore, S., Jurcevic, S., Bronner, M., Lemoine, N. R., and Brentnall, T. A. (2013). Molecular analysis of precursor lesions in familial pancreatic cancer. *PLoS ONE*, 8(1):e54830.

[38] Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466.

[39] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20):1920–1930.

[40] Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006a). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3):1–13.

[41] Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006b). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3):1–13.

[42] Ding, C. and Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 29(1):185–205.

[43] Donahue, T. R., Tran, L. M., Hill, R., Li, Y., Kovochich, A., Calvopina, J. H., Patel, S. G., Wu, N., Hindoyan, A., Farrell, J. J., et al. (2012). Integrative survival-based molecular profiling of human pancreatic cancer. *Clinical Cancer Research*, 18(5):1352–1363.

[44] Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*, 93(23):13429–13429.

[45] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., Loubes, J. M., Massart, P., Madigan, D., Ridgeway, G., Rosset, S., Zhu, J. I., Stine, R. A., Turlach, B. A., Weisberg, S., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

[46] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. B. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868.

[47] Fang, Z., Du, R., and Cui, X. (2012). Uniform approximation is more appropriate for wilcoxon rank-sum test in gene set analysis. *PLoS ONE*, 7(2):e31505.

[48] Fdez-Riverola, F., Díaz, F., Borrajo, M. L., Yáñez, J. C., and Corchado, J. M. (2005). Improving gene selection in microarray data analysis using fuzzy patterns inside a CBR system. In *International Conference on Case-Based Reasoning*, pages 191–205.

[49] Fonseca, C. M. and Fleming, P. J. (1995). An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation*, 3:1–16.

[50] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407.

[51] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[52] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

[53] Fujimoto, J. and Wistuba, I. I. (2014). Current concepts on the molecular pathology of non-small cell lung carcinoma. *Seminars in Diagnostic Pathology*, 31(4):306–313.

[54] Fukumoto, S. I., Yamauchi, N., Moriguchi, H., Hippo, Y., Watanabe, A., Shibahara, J., Taniguchi, H., Ishikawa, S., Ito, H., Yamamoto, S., Iwanari, H., Hironaka, M., Ishikawa, Y., Niki, T., Sohara, Y., Kodama, T., Nishimura, M., Fukayama, M., Dosaka-Akita, H., and Aburatani, H. (2005). Overexpression of the aldo-keto reductase family protein AKR1B10 is highly correlated with smokers' non-small cell lung carcinomas. *Clinical Cancer Research*, 11(5):1776–1785.

[55] George, G. V. S. and Raj, V. C. (2011). Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. *International Journal of Computer Science & Engineering Survey*, 2(3):16–27.

[56] Gnana, A. A., Balamurugan, S., and Leavline, J. E. (2016). Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications*, 136(1):9–17.

[57] Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Publishing Company.

[58] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

[59] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422.

[60] He, Z. and Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, 34(4):215–225.

[61] Hechenbichler, K., Schliep, K., and Wilson, A. (2004). Weighted k-nearest-neighbor techniques and ordinal classification. *Sonderforschungsbereich 386*, 399:1–17.

[62] Heim, D., Budczies, J., Stenzinger, A., Treue, D., Hufnagl, P., Denkert, C., Dietel, M., and Klauschen, F. (2014). Cancer beyond organ and tissue specificity: Next-generation-sequencing gene mutation data reveal complex genetic similarities across major cancers. *International Journal of Cancer*, 135(10):2362–2369.

[63] Hernandez, Y. G. and Aimee, L. L. (2016). MicroRNA in pancreatic ductal adenocarcinoma and its precursor lesions. *World Journal of Gastrointestinal Oncology*, 8(1):18–19.

[64] Hezel, A. F., Kimmelman, A. C., Stanger, B. Z., Bardeesy, N., and Depinho, R. A. (2006). Genetics and biology of pancreatic ductal adenocarcinoma. *Genes & Development*, 20(10):1218–1249.

[65] Hidalgo, M., Cascinu, S., Kleeff, J., Labianca, R., Löhr, J. M., Neoptolemos, J., Real, F. X., Van Laethem, J. L., and Heinemann, V. (2015). Addressing the challenges of pancreatic cancer: Future directions for improving outcomes. *Pancreatology*, 15(1):8–18.

[66] Hira, Z. M. and Gillies, D. F. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 15:1–13.

[67] Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944.

[68] Huang, H., Liu, Y., Yuan, M., and Marron, J. S. (2015). Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, 24(4):975–993.

[69] Huang, M. L., Hung, Y. H., Lee, W. M., Li, R. K., and Wang, T. H. (2012). Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *Journal of Medical Systems*, 36(2):407–414.

[70] Hurria, A., Muss, H. B., and Cohen, H. J. (2016). Cancer and aging. In *Holland-Frei Cancer Medicine*, pages 750–759. PMPH-USA.

[71] Inza, I., Larrañaga, P., Blanco, R., and Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, 31(2):91–103.

[72] Irizarry, R. A., Bolstad, Benjamin, M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix Genechip probe level data. *Nucleic Acids Research*, 31(4):e15.

[73] Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *Annals of Statistics*, 33(2):730–773.

[74] Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386.

[75] Joergensen, M. T., Heegaard, N. H., and Schaffalitzky De Muckadell, O. B. (2010). Comparison of plasma Tu-M2-PK and CA19-9 in pancreatic cancer. *Pancreas*, 39(2):243–247.

[76] Jurisica, I. and Glasgow, J. (2004). Applications of case-based reasoning in molecular biology. *AI Magazine*, 25:85–96.

[77] Kang, M. W., Lee, E. S., Yoon, S. Y., Jo, J., Lee, J., Kim, H. K., Choi, Y. S., Kim, K., Shim, Y. M., Kim, J., and Kim, H. (2011). AKR1B10 is associated with smoking and smoking-related non-small-cell lung cancer. *Journal of International Medical Research*, 39(1):78–85.

[78] Kaufman, L. and Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, New Jersey.

[79] Kerr, G., Ruskin, H. J., Crane, M., and Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in Biology and Medicine*, 38(3):283–293.

[80] Kerr, M. K. and Churchill, G. A. (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98(16):8961–8965.

[81] Kim, I. J., Quigley, D., To, M. D., Pham, P., Lin, K., Jo, B., Jen, K. Y., Raz, D., Kim, J., Mao, J. H., Jablons, D., and Balmain, A. (2013). Rewiring of human lung cell lineage and mitotic networks in lung adenocarcinomas. *Nature Communications*, 4:1701.

[82] Kimes, P. K., Liu, Y., Neil Hayes, D., and Marron, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics*, 73(3):811–821.

[83] Kohonen-Corish, M. R., Tseung, J., Chan, C., Currey, N., Dent, O. F., Clarke, S., Bokey, L., and Chapuis, P. H. (2014). KRAS mutations and CDKN2A promoter methylation show an interactive adverse effect on survival and predict recurrence of rectal cancer. *International Journal of Cancer*, 134(12):2820–2828.

[84] Koorstra, J.-B. M., Hustinx, S. R., Offerhaus, G. J. A., and Maitra, A. (2008). Pancreatic carcinogenesis. *Pancreatology*, 8(2):110–125.

[85] Korc, M. (2007). Pancreatic cancer-associated stroma production. *American Journal of Surgery*, 194(4):S84–S86.

[86] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

[87] Kumari, B. and Swarnkar, T. (2011). Filter versus wrapper feature subset selection in large dimensionality microarray: A review. *International Journal of Computer Science and Information Technologies*, 2:1048–1053.

[88] Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., Schnabel, P., Warth, A., Poustka, A., Sültmann, H., et al. (2009a). Global gene expression

analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung cancer*, 63(1):32–38.

[89] Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., Schnabel, P., Warth, A., Poustka, A., Sültmann, H., and Hoffmann, H. (2009b). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer*, 63(1):32–38.

[90] Kunovsky, L., Tesarikova, P., Kala, Z., Kroupa, R., Kysela, P., Dolina, J., and Trna, J. (2018). The use of biomarkers in early diagnostics of pancreatic cancer. *Canadian Journal of Gastroenterology and Hepatology*, 2018:10.

[91] Kursa, M. and Rudnicki, W. (2016). *Wrapper Algorithm for All Relevant Feature Selection*. Package Boruta, Version 5.1.0, https://m2.icm.edu.pl/boruta/.

[92] Kursa, M. B. and Rudnicki, W. R. (2015). Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13.

[93] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., De Schaetzen, V., Duque, R., Bersini, H., and Nowé, A. (2012a). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1106–1119.

[94] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., De Schaetzen, V., Duque, R., Bersini, H., and Nowé, A. (2012b). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1106–1119.

[95] Li, Q., Hou, L., Ding, G., Li, Y., Wang, J., Qian, B., Sun, J., and Wang, Q. (2015). KDM6B induces epithelial-mesenchymal transition and enhances clear cell renal cell carcinoma metastasis through the activation of SLUG. *International Journal of Clinical and Experimental Pathology*, 8(6):6334–6344.

[96] Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332.

[97] Liu, C., Huang, X., Hou, S., Hu, B., and Li, H. (2015). Silencing of tripartite motif (TRIM) 29 inhibits proliferation and invasion and increases chemosensitivity to

cisplatin in human lung squamous cancer NCI-H520 cells. *Thoracic Cancer*, 6(1):31–37.

[98] Liu, H., Motoda, H., Setiono, R., and Zhao, Z. (2010). Feature selection: An ever evolving frontier in data mining. In *Feature Selection in Data Mining*, pages 4–13.

[99] Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293.

[100] Lu, Y. and Han, J. (2003). Cancer classification using gene expression data. *Information Systems*, 28(4):243–268.

[101] Luo, G., Xiao, Z., Long, J., Liu, Z., Liu, L., Liu, C., Xu, J., Ni, Q., and Yu, X. (2013). CA125 is superior to CA19-9 in predicting the resectability of pancreatic cancer. *Journal of Gastrointestinal Surgery*, 17(12):2092–2098.

[102] Luo, Y., Meng, Q., Ge, R., Mai, G., Liu, J., Zhou, F., and Zhou, M. F. (2014). Package 'sublasso'.

[103] Ma, S. and Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*, 9(5):392–403.

[104] MacNaughton-Smith, P., Williams, W. T., Dale, M. B., and Mockett, L. G. (1965). Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature*, 202:1034–1035.

[105] Mahmoud, O., Harrison, A., Perperoglou, A., Gul, A., Khan, Z., Metodiev, M. V., and Lausen, B. (2014). A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. *BMC Bioinformatics*, 15(1):274.

[106] Mahmoud, O., Harrison, A.and Perperoglou, A., Gul, A., Khan, Z., and Lausen, B. (2015). *propOverlap: Feature (gene) selection based on the Proportional Overlapping Scores*. R package version 1.0, http://CRAN.R-project.org/package=propOverlap.

[107] Maitra, R., Melnykov, V., and Lahiri, S. N. (2012). Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, 107(497):378–392.

[108] Marzec, J., Dayem Ullah, A. Z., Pirrò, S., Gadaleta, E., Crnogorac-Jurcevic, T., Lemoine, N. R., Kocher, H. M., and Chelala, C. (2018). The pancreatic expression database: 2018 update. *Nucleic Acids Research*, 46(D1):D1107–D1110.

[109] McGranahan, N. and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, 27(1):15–26.

[110] McNiff, J. (2013). *Action research: Principles and practice*. Routledge.

[111] Mittal, R. A., Hammel, M., Schwarz, J., Heschl, K. M., Bretschneider, N., Flemmer, A. W., Herber-Jonat, S., Königshoff, M., Eickelberg, O., and Holzinger, A. (2012). SFTA2-a novel secretory peptide highly expressed in the lung-is modulated by lipopolysaccharide but not hyperoxia. *PLoS ONE*, 7(6):e40011.

[112] Monirul Kabir, M., Monirul Islam, M., and Murase, K. (2010). A new wrapper feature selection approach using neural network. *Neurocomputing*, 73(16-18):3273–3283.

[113] Moorthy, K. and Mohamad, M. S. (2012). Random forest for gene selection and microarray data classification. In *Communications in Computer and Information Science*, pages 174–183. Springer-Verlag Berlin Heidelberg.

[114] Nascimento, A., Bousbaa, H., Ferreira, D., and Sarmento, B. (2015). Non-small cell lung carcinoma: An overview on targeted therapy. *Current Drug Targets*, 16(13):1448–1463.

[115] Natarajan, A. and Ravi, T. (2014). A survey on gene feature selection using microarray data for cancer classification. *International Journal of Computer Science & Com munication*, 5(1):126–129.

[116] Nguyen, T., Khosravi, A., Creighton, D., and Nahavandi, S. (2015). Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification. *PLoS ONE*, 3(10):1–23.

[117] Novotný, I., Dítě, P., Dastych, M., Žáková, A., Trna, J., Novotná, H., and Nechutová, H. (2008). Tumor marker M2-pyruvate-kinase in differential diagnosis of chronic pancreatitis and pancreatic cancer. *Hepato-Gastroenterology*, 55(85):1475–1477.

[118] Pandey, B. and Mishra, R. B. (2009). Knowledge and intelligent computing system in medicine. *Computers in Biology and Medicine*, 39(3):215–230.

[119] Pappa, G. L., Freitas, A. A., and Kaestner, C. A. (2002). A multiobjective genetic algorithm for attribute selection. In *The Fourth International Conference on Recent Advances in Soft Computing*, pages 116–121.

[120] Park, H. L., Yoo, I. R., Boo, S. H., Park, S. Y., Park, J. K., Sung, S. W., and Moon, S. W. (2019). Does FDG PET/CT have a role in determining adjuvant chemotherapy in surgical margin-negative stage IA non-small cell lung cancer patients? *Journal of Cancer Research and Clinical Oncology*, 145(4):1–6.

[121] Pascal, L. E., True, L. D., Campbell, D. S., Deutsch, E. W., Risk, M., Coleman, I. M., Eichner, L. J., Nelson, P. S., and Liu, A. Y. (2008). Correlation of mRNA and protein levels: Cell type-specific gene expression of cluster designation antigens in the prostate. *BMC Genomics*, 9(1):246.

[122] Paulovich, A. G., Whiteaker, J. R., Hoofnagle, A. N., and Wang, P. (2008). The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline. *Proteomics - Clinical Applications*, 2(10-11):1386–1402.

[123] Penfold, C. A. and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870.

[124] Plathow, C., Aschoff, P., Lichy, M. P., Eschmann, S., Hehr, T., Brink, I., Claussen, C. D., Pfannenberg, C., and Schlemmer, H. P. (2008). Positron emission tomography/computed tomography and whole-body magnetic resonance imaging in staging of advanced nonsmall cell lung cancer - Initial results. *Investigative Radiology*, 43(5):290–297.

[125] Ramos González, J., Castellanos-Garzón, J. A., de Paz, J. F., and Corchado, J. M. (2018). A data mining framework based on boundary-points for gene selection from DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study. *Engineering Applications of Artificial Intelligence*, 70:92–108.

[126] Ramos-González, J., López-Sánchez, D., Castellanos-Garzón, J. A., de Paz, J. F., and Corchado, J. M. (2017). A CBR framework with gradient boosting based feature

selection for lung cancer subtype classification. *Computers in Biology and Medicine*, 86:98–106.

[127] Rekhtman, N., Ang, D. C., Sima, C. S., Travis, W. D., and Moreira, A. L. (2011). Immunohistochemical algorithm for differentiation of lung adenocarcinoma and squamous cell carcinoma based on large series of whole-tissue sections with validation in small specimens. *Modern Pathology*, 24(10):1348–1359.

[128] Reza Abbasifard, M., Ghahremani, B., and Naderi, H. (2014). A survey on nearest neighbor search methods. *International Journal of Computer Applications*, 95(25):39–52.

[129] Rivera, M. P., Mehta, A. C., and Wahidi, M. M. (2013). Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer. *Chest Journal*, 143(5 suppl):e142S–e165S.

[130] Runkle, E. A. and Mu, D. (2013). Tight junction proteins: From barrier to tumorigenesis. *Cancer Letters*, 337(1):41–48.

[131] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

[132] Salem, H., Attiya, G., and El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. *Applied Soft Computing Journal*, 50:124–134.

[133] Sanchez-Palencia, A., Gomez-Morales, M., Gomez-Capilla, J. A., Pedraza, V., Boyero, L., Rosell, R., and Fárez-Vidal, M. E. (2011). Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer*, 129(2):355–364.

[134] Shapiro, J. A. (2009). Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178(1):6–28.

[135] Shraddha, S., Anuradha, N., and Swapnil, S. (2014). Feature selection techniques and microarray data: A survey. *International Journal of Emerging Technology and Advanced Engineering*, 4(1):179–183.

[136] Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2015. *CA: A cancer journal for clinicians*, 65(1):5–29.

[137] Singh, R. K. and Sivabalakrishnan, M. (2015). Feature selection of gene expression data for cancer classification: A review. *Procedia Computer Science*, 50:52–57.

[138] Song, Q., Merajver, S. D., and Li, J. Z. (2015). Cancer classification in the genomic era: five contemporary problems. *Human Genomics*, 9(1):27.

[139] Suzuki, R. and Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542.

[140] Swords, D. S., Firpo, M. A., Scaife, C. L., and Mulvihill, S. J. (2016). Biomarkers in pancreatic adenocarcinoma: Current perspectives. *OncoTargets and Therapy*, 9:7459–7467.

[141] Takamochi, K., Ohmiya, H., Itoh, M., Mogushi, K., Saito, T., Hara, K., Mitani, K., Kogo, Y., Yamanaka, Y., Kawai, J., Hayashizaki, Y., Oh, S., Suzuki, K., and Kawaji, H. (2016). Novel biomarkers that assist in accurate discrimination of squamous cell carcinoma from adenocarcinoma of the lung. *BMC Cancer*, 16(1):760.

[142] Toronen, P. (2004). Selection of informative clusters from hierarchical cluster tree with gene classes. *BMC Bioinformatics*, 5(1):32.

[143] Torres, C. and Grippo, P. J. (2018). Pancreatic cancer subtypes: a roadmap for precision medicine. *Annals of medicine*, 50(4):277–287.

[144] Tyagi, V. and Mishra, A. (2013). A survey on different feature selection methods for microarray data analysis. *International Journal of Computer Applications*, 67(16):36–40.

[145] Waddell, P. J. and Kishino, H. (2000). Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Informatics*, 11:129–140.

[146] Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., and Mewes, H. W. (2005). Gene selection from microarray data for cancer classification. A machine learning approach. *Computational Biology and Chemistry*, 29(1):37–46.

[147] Webb, C. P. and Pass, H. I. (2004). Translation research: From accurate diagnosis to appropriate treatment. *Journal of Translational Medicine*, 2(1):35.

[148] Weiss, P. (2005). Applications of generating functions in nonparametric tests. *The Mathematica Journal*, 9(4):803–823.

[149] Wolters, M. A. (2015a). A genetic algorithm for selection of fixed-size subsets with application to design problems. *Journal of Statistical Software*, 68(1):1–18.

[150] Wolters, M. A. (2015b). *A Genetic Algorithm for Fixed-Size Subset Selection*. R-Package kofnGA, Version 1.2.

[151] Wong, D. and Yip, S. (2018). Machine learning classifies cancer. *Nature*, 555(7697):446–447.

[152] Wu, X. and Zhang, K. (1991). A better tree-structured vector quantizer. In *Data Compression Conference*, pages 392–401.

[153] Xiao, J., Lu, X., Chen, X., Zou, Y., Liu, A., Li, W., He, B., He, S., and Chen, Q. (2017). Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma. *Oncotarget*, 8(42):71759–71771.

[154] Xu, R., Hu, J., Zhang, T., Jiang, C., and Wang, H.-Y. (2016). TRIM29 overexpression is associated with poor prognosis and promotes tumor progression by activating Wnt/β-catenin pathway in cervical cancer. *Oncotarget*, 7(19):28579–28591.

[155] Xu, X., Rock, J. R., Lu, Y., Futtner, C., Schwab, B., Guinney, J., Hogan, B. L. M., and Onaitis, M. W. (2012). Evidence for type II cells as cells of origin of K-Ras-induced distal lung adenocarcinoma. *Proceedings of the National Academy of Sciences*, 109(13):4910–4915.

[156] Yang, K., Cai, Z., Li, J., and Lin, G. (2006). A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7(1):228.

[157] Yao, B. and Li, S. (2010). ANMM4CBR: A case-based reasoning method for gene expression data classification. *Algorithms for Molecular Biology*, 5(14):1–11.

[158] Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. (2003). Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5):R34.

[159] Yu, L. and Liu, H. (2004). Redundancy based feature selection for microarray data. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–742. ACM.

[160] Zappa, C. and Mousa, S. A. (2016). Non-small cell lung cancer: Current treatment and future advances. *Translational Lung Cancer Research*, 5(3):288–300.

[161] Zhan, C., Yan, L., Wang, L., Sun, Y., Wang, X., Lin, Z., Zhang, Y., Shi, Y., Jiang, W., and Wang, Q. (2015). Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma. *Journal of Thoracic Disease*, 7(8):1398–1405.

[162] Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1):e78644.

[163] Zhou, J., Foster, D., Stine, R., and Ungar, L. (2005). Streaming feature selection using alpha-investing. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, pages 384–393.

[164] Zhou, X. and Tuck, D. P. (2007). MSVM-RFE: Extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, 23(9):1106–1114.

[165] Zhou, Z. Y., Yang, G. Y., Zhou, J., and Yu, M. H. (2012). Significance of TRIM29 and β-catenin expression in non-small-cell lung cancer. *Journal of the Chinese Medical Association*, 75(6):269–274.