Universidad Nacional del Sur

TESIS DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

Técnicas de Aprendizaje Automático y Computación Científica Aplicadas a la Predicción de Parámetros ADME-Tox

Axel J. Soto

BAHIA BLANCA ARGENTINA

2010

PREFACIO

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico de Doctor en Ciencias de la Computación, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Departamento de Ciencias de la Computación durante el período comprendido entre el 27 de septiembre de 2005 y el 26 de Abril de 2010, bajo la dirección del Dr. Ignacio Ponzoni, profesor del Departamento de Ciencias e Ingeniería de la Computación e investigador adjunto del CONICET, y del Dr. Gustavo E. Vazquez, asistente del Departamento de Ciencias e Ingeniería de la Computación e investigador asistente del CONICET.

Axel J. Soto
Bahía Blanca, 26 de Abril de 2010

UNIVERSIDAD NACIONAL DEL SUR
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el …./…./…... , mereciendo la calificación de .....(……)

"Un sólo pensamiento de gratitud hacia el cielo,
es la oración más perfecta"
Gotthold Ephraim Lessing

Agradecimientos

Entiendo ahora por qué los agradecimientos van al principio de un libro. La ayuda de las personas que aquí nombro hicieron posible esta empresa, por lo que ellos también son autores de este trabajo.

En primer lugar agradezco a mis padres por haberme apoyado en mis emprendimientos, así como también por su amor incondicional durante toda mi vida. También a mis hermanos por acompañarme cada vez que necesite de ellos.

A Noelia, futura esposa y eterna novia, quien me acompañó con su amor y comprensión durante los últimos cinco años. Gracias por la paciencia y el apoyo cuando necesité de tiempo para este trabajo.

También quiero agradecer a mis directores Ignacio y Gustavo, quienes me otorgaron esta oportunidad de hacer un doctorado, me mostraron el mundo de la investigación y me formaron ponderando siempre el aspecto humano. También agradezco a Eladio Pardillo-Fontdevilla por ayudarme desinteresadamente durante los primeros años de mis investigaciones. Un cuarto director fue Marc Strickert, quien me brindó su conocimiento y amistad de manera generosa.

No quiero olvidar mencionar a mis amigos y compañeros de trabajo, tanto de la Planta Piloto de Ingeniería Química como del Departamento de Ciencias e Ingeniería de la Computación. A todos ellos mis agradecimientos y mis deseos de poder continuar esta buena relación.

Finalmente, considero muy importante mencionar y valorar la educación pública y gratuita, de la cual hice uso durante toda mi vida.

Axel J. Soto
Bahía Blanca, 26 de Abril de 2010

Resumen

Hace 15 años atrás, el desarrollo de nuevos productos farmacéuticos consistía en un proceso de prueba y error basado mayormente en la búsqueda de un farmacóforo o principio activo. Sin embargo, muchos compuestos eran finalmente descartados en la última etapa del proceso de su desarrollo debido al comportamiento ADME-Tox. Por estos motivos, el interés en la industria y el ámbito académico en las disciplinas de quimioinformática y, en particular, en las técnicas de tipo QSAR ha crecido considerablemente en los últimos años.

El objetivo de esta tesis se centró en desarrollar metodologías para la mejora de los modelos existentes de QSAR mediante el uso de técnicas numéricas y de aprendizaje automático. En este sentido, se obtuvo un potente método para lo que se considera uno de los problemas más importantes del modelado QSAR: la selección de subconjuntos de descriptores relevantes. Para esta tarea se desarrollaron distintos enfoques utilizando computación evolutiva.

Asimismo, otro aspecto central considerado fue el desarrollo de una técnica de identificación de dominio de aplicación para un método de predicción, el cual permite determinar los alcances en las predicciones de un modelo. Para esta técnica se consideró la aplicación de medidas de distancia entre compuestos químicos, usando aprendizaje no supervisado. Finalmente, se desarrolló un método generalizado que permite la proyección de los datos en un espacio de menor dimensión, en donde las distancias entre los datos proyectados guardan relación con las distancias de la propiedad a modelar. Este nuevo espacio permite mejorar la visualización, reducir el conjunto de descriptores en forma embebida y mejorar la precisión de los modelos de predicción.

Abstract

15 years ago, development of new drugs was based on a trial-and-error process that was mostly devoted to the searching of pharmacophore fragments or drug potency. Nevertheless, many compounds were discarded in the latest stages of their development due to poor ADME-Tox behavior. Therefore, chemoinformatics interest given by the scientific and industrial community has grown considerably in the last years.

The thesis' objective is focused on the development of statistical and machine learning techniques aimed at improving current QSAR-based methods' limitations. In this way, a robust method for tackling the problem of selecting relevant descriptor subsets was developed. Selection of descriptors is one of the most important challenges in QSAR. This task was carried out by means of evolutionary computing techniques.

Moreover, another crucial issue was the development of a method for identifying the applicability domain of a given prediction method, in order to estimate the scope of the accuracy of a model. Similarity metrics and unsupervised learning were applied for this task. Finally, a general approach for data projection onto a low-dimensional space was proposed, where distances among projected data are in maximum correlation with distances on the target space. This new subspace projection allows a better visualization capacity, an embedded descriptor selection and an improvement of the prediction capacity.

Lista de Publicaciones

Publicaciones en Revistas Científicas y en la serie Lecture Notes in Computer Science:

- **A.J. Soto**, M. Strickert, G.E. Vazquez. "Adaptive matrix metrics for molecular descriptor assessment in QSPR classification". **Journal of Cheminformatics**. Vol 2 (Suppl 1):P47, 2010.
- **A.J. Soto**, R.L. Cecchini, G.E. Vazquez, I. Ponzoni. "Multi-Objective Feature Selection in QSAR using a Machine Learning Approach". **QSAR & Combinatorial Science**. Vol 28, 11-12, 1509-1523, 2009.
- **A.J. Soto**, I. Ponzoni, G.E. Vazquez. "Segregating Confident Predictions of Chemicals' Properties for Virtual Screening of Drugs". In: S. Omatu; M.P. Rocha; J. Bravo; F. Fernández; E. Corchado; A. Bustillo; J.M. Corchado (Eds.). **Lecture Notes in Computer Science** Vol. 5518, 1005-1012. Springer-Verlag Berlin Heidelberg. Third International Workshop on Practical Applications of Computational Biology and Bioinformatics - IWPACBB 2009. Salamanca, España, 10 - 12 Junio 2009.
- **A.J. Soto**, R.L. Cecchini, G.E Vazquez, I. Ponzoni. "A Wrapper-based Feature Selection Method for ADMET Prediction using Evolutionary Computing". In: E. Marchiori, J.H. Moore (Eds.). **Lecture Notes in Computer Science** Vol. 4973, 188-199. Springer-Verlag Berlin Heidelberg. Sixth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics - EvoBIO 2008. ISSN: 0302-9743. Napoli, Italia, 26 - 28 Marzo 2008.

- **A.J. Soto**, R.L. Cecchini, G.E Vazquez, I. Ponzoni. "An Evolutionary Approach for Feature Selection applied to ADMET Prediction". **Inteligencia Artificial - Ibero American Journal of Artificial Intelligence**. Vol. 12 N°37, 55-63, 2008. ISSN: 1137-3601.

Publicaciones en Conferencias (excluidos LNCS):

- **A.J. Soto**, I. Ponzoni, G.E. Vazquez. "On Designing Confident Statistical QSPR Models". **1st International Meeting of Pharmaceutical Sciences** - RICiFa 2010. 24-25th June 2010, Córdoba, Argentina.
- **A.J. Soto**, J. Martinez, R.L. Cecchini, G.E. Vazquez, Ignacio Ponzoni. "DELPHOS: Computational Tool for Selection of relevant descriptor Subsets in ADMET Prediction". **1st International Meeting of Pharmaceutical Sciences** - RICiFa 2010. 24-25th June 2010, Córdoba, Argentina.
- **A.J. Soto**, J. Martinez, R.L. Cecchini, G.E. Vazquez, Ignacio Ponzoni. "A Prototype Software Tool for Selection of Relevant Descriptors in QSAR Models". **1st Argentine Congress on Bioinformatics and Computational Biology** - CABBC 2010. 12-14th May 2010, Quilmes, Argentina.
- **A.J. Soto**, I. Ponzoni, G.E. Vazquez. "On defining applicability domains for prediction models in chemoinformatics". **1st Argentine Congress on Bioinformatics and Computational Biology** - CABBC 2010. 12-14th May 2010, Quilmes, Argentina.
- M. Strickert, **A.J. Soto**, G.E Vazquez. "Adaptive matrix distances aiming at optimum regression subspaces". **European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2010**. 28-30 Abril 2010, Brujas, Bélgica.

- **A.J. Soto**, M. Strickert, G.E. Vazquez. "A mapping method for linking chemical compounds to biological and physicochemical properties in drug discovery". **ISCB Latin America**. 13-16 Marzo 2010, Montevideo, Uruguay.
- **A.J. Soto**, R.L Cecchini, G.E. Vazquez, I. Ponzoni. "A new method for multi-objective selection of molecular descriptors for QSAR/QSPR". **ISCB Latin America**. 13-16 Marzo 2010, Montevideo, Uruguay.
- M. Strickert, **A.J. Soto**, J. Keilwagen, G.E. Vazquez. "Towards matrix-based selection of feature pairs for efficient ADMET prediction". **10º Argentine Symposium on Artificial Intelligence**, ASAI 2009 - 38th JAIIO, Agosto 24-25 2009, Mar del Plata, Argentina, págs 83-94.
- **A.J. Soto**, R.L. Cecchini, D. Palomba, G.E. Vazquez, I. Ponzoni. "An Evolutionary Approach for Multi-Objective Feature Selection in ADMET Prediction". **XXXIV Latin American Conference on Informatics**, CLEI 2008, Santa Fe, Argentina. 8 - 12 Septiembre 2008, págs 112-121. ISBN 978-950-9770-02;
- **A.J. Soto**, D. Palomba, M. Diaz, G.E. Vazquez, I. Ponzoni. "Aplicaciones de Inteligencia Computacional sobre Clusters de Compuestos Químicos". **XVII Congress on Numerical Methods and their Applications**, ENIEF 2008, San Luis, Argentina. 10 - 13 Noviembre 2008, págs 2429-2442. ISSN: 1666-6070.
- **A.J. Soto**, R.L. Cecchini, I. Ponzoni, G.E. Vazquez. "Computational Intelligence Methods for Physicochemical Property Prediction". **9th Latin American Conference on Physical Organic Chemistry**, CLAFQO 2007. Los Cocos, Argentina. 30 Septiembre - 5 Octubre 2007.
- **A.J. Soto**, I. Ponzoni, G.E. Vazquez. "Predicting Physicochemical Properties for Drug Design Using Clustering and Neural

Network Learning". **Brazilian Symposium on Bioinformatics 2007**. Angra dos Reis, Rio de Janeiro, Brasil. 29-31 Agosto 2007, págs 46-57. ISBN: 978-85-7669-123-5.

- R.L. Cecchini, **A.J. Soto**, G.E. Vazquez, I. Ponzoni. "A Genetic Algorithm for Detection of Relevant Descriptors in ADMET Prediction". **Brazilian Symposium on Bioinformatics 2007**. Angra dos Reis, Rio de Janeiro, Brasil. 29-31 Agosto 2007, págs 62-65. ISBN: 978-85-7669-123-5.
- **A.J. Soto**, R.L. Cecchini, G.E. Vazquez, I. Ponzoni. "Feature Selection for ADMET Prediction using Genetic Algorithms". **9º Argentine Symposium on Artificial Intelligence**, ASAI 2007. ISSN: 1850-2776, págs 77-88.
- **A.J Soto**, I. Ponzoni, G.E. Vazquez. "Análisis Numérico de diferentes criterios de similitud en algoritmos de clustering". **XV Congress on Numerical Methods and their Applications**, ENIEF 2006. págs 993-1011. ISSN: 1666-6070.

Índice general

Índice de figuras

ÍNDICE DE FIGURAS

Capítulo 1

Introducción

1.1. Predicción de parámetros ADME-Tox

Históricamente, cuando se quería desarrollar una nueva droga, se realizaba un proceso en serie donde la potencia y la selectividad de la droga eran examinadas en primer lugar (Selick *et al.* (2002)). Muchos de los compuestos inicialmente seleccionados como candidatos fallaban en las últimas etapas del proceso de desarrollo por tener un comportamiento inaceptable de las propiedades ADME-Tox o ADMET (absorción, distribución, metabolismo, excreción y toxicidad) para el cuerpo humano.

En la actualidad, la tasa de fallos en el desarrollo de una droga candidata antes de que ésta llegue al mercado sigue siendo alta (Gola *et al.* (2006)). El principal problema reside en el desconocimiento de las reglas que gobiernan a las propiedades ADMET (Segall *et al.* (2006)). Por esta razón, el interés en técnicas del tipo QSAR/QSPR (*Quantitative Structure Activity/Property Relationships*), por parte del ámbito académico e industrial, ha ido creciendo de manera considerable en las últimas décadas (Banik (2004); Winkler (2002)). Estas técnicas apuntan a encontrar relaciones entre los parámetros químico-estructurales de una molécula y una determinada propiedad química o biológica.

Por otra parte, un área de creciente importancia dentro de las Ciencias de la Computación es el Aprendizaje Automático (*Machine Learning*), el cual consiste en la identificación de patrones dentro de grandes volúmenes de datos (Bishop

(2006)). En el aprendizaje automático se tienen, básicamente, dos variantes: *supervisado* y *no supervisado*. En el primer caso, se busca deducir el comportamiento de una variable a partir del análisis de observaciones descriptas por otras variables. Por otra parte, en el aprendizaje *no supervisado* el objetivo es identificar características relevantes de los datos presentados. Cabe mencionar que en esta modalidad no se utiliza una variable a modelar, ya que no es el objetivo construir un modelo con fines de predicción. En cualquier caso, es importante contar con información significante y además en suficiente cantidad, de manera que el método aplicado sea capaz de detectar las relaciones existentes entre los datos.

La aparición de nuevas metodologías en el campo del aprendizaje automático en combinación con técnicas de estadística multivariada, provocaron un incremento sustancial en la capacidad de modelado y predicción, en comparación con las técnicas tradicionales de QSAR (Fox & Kriegl (2006); Winkler (2004)). La importancia ganada por este tipo de técnicas computacionales más modernas, logró que se reconociera su utilidad y se le adjudique el término de predicciones *in silico*, en alusión a las experimentaciones *in vivo* e *in vitro* que vienen utilizándose desde hace tiempo atrás en la ciencia de la Farmacia.

Además de la predicción de propiedades ADMET, los métodos *in silico* también pueden ser utilizados para diversas aplicaciones dentro del proceso de desarrollo de nuevos fármacos (Oprea (2005)). La motivación principal para el uso de las técnicas *in silico* es que su utilización representa un ahorro importante desde el punto de vista económico. Su principal ventaja reside en que no se requiere que el compuesto (o el conjunto de compuestos) a analizar sea sintetizado para calcular alguna de sus propiedades. De igual manera, esta innovación tecnológica permite ahorrar tiempo de laboratorio, ya que, dependiendo del experimento a realizar, se pueden testear enormes cantidades de compuestos en tiempos relativamente cortos. Este tipo de técnicas *in silico*, aún en evolución, no pretenden reemplazar por completo a los ensayos *in vivo* e *in vitro*, pero sí, disminuir la cantidad de estos ensayos mediante el descarte previo de compuestos candidatos que no superen estándares mínimos *in silico*.

Sin embargo, muchos autores coinciden en que la capacidad predictiva de los modelos basados en QSAR aún necesita ser mejorada (Gola *et al.* (2006); Hou

& Wang (2008); Tetko *et al.* (2006)). Esta mejora es necesaria para que la precisión en las predicciones de los métodos basados en QSAR sea lo suficientemente aceptable para que su aplicación en casos reales y de gran escala sea efectiva.

1.2. Objetivos y alcance

El trabajo realizado en el marco de esta tesis se centra en la predicción de propiedades relacionadas con los parámetros ADMET basándose en métodos de tipo QSAR/QSPR. En particular, el núcleo central del contenido de esta tesis se origina a partir del análisis y el reconocimiento de falencias y limitaciones de los enfoques existentes. En tal sentido, se detectó que las fuentes de imprecisiones de los modelos de predicción provienen, generalmente, de tres posibles situaciones:

- Elección inapropiada de los datos utilizados para construir el modelo.
- Falencias del método de predicción, errores de diseño o de su estrategia de validación.
- Aplicación del modelo de predicción sobre un dominio inapropiado.

La fiabilidad de los modelos de predicción es tan grande como es la calidad y la relevancia de los datos utilizados en la construcción del modelo. Por tal motivo, la identificación de las características químicas-estructurales (descriptores) que resulten relevantes es un problema central en el desarrollo de métodos basados en QSPR (Downs (2004)). La utilización de modelos con más descriptores o descriptores diferentes de los necesarios, posibilita la aparición de correlaciones espurias carentes de significado real. En definitiva, este tópico resulta crucial para la buena generalización del método de predicción aplicado y, al mismo tiempo, para la interpretación teórica de las relaciones obtenidas.

Asimismo, la construcción de un modelo de predicción requiere un diseño cuidadoso en cuanto a su configuración de parámetros, a fin de que la metodología sea aplicada de manera adecuada. En este contexto resulta importante que, independientemente de la estrategia aplicada, la misma sea validada de manera que su capacidad predictiva sea correctamente estimada (Tropsha *et al.* (2003)). Sin

embargo, muchas de las metodologías propuestas en la literatura no realizan los procedimientos necesarios que aseguren la correcta estimación de la capacidad del método (Jónsdóttir *et al.* (2005)).

Por otra parte, la inmensa cantidad de posibles combinaciones de átomos y enlaces dan lugar a un considerable número de posibles entidades químicas diferentes. En este contexto, un método puede ser útil para cierta clases de compuestos pero no para otros. Esto no constituye un problema *per se*, siempre y cuando se tenga en consideración el dominio de aplicación para el cual un método es generado (Papa *et al.* (2009)). Si este tipo de controles no es tenido en cuenta, la confiabilidad de un modelo de predicción no puede ser correctamente cuantificada.

Por lo tanto, y como consecuencia de la identificación de estas tres grandes fuentes de conflictos, se procedió al estudio de los métodos que podrían resultar interesantes para atacar cada una de estas dificultades. En este contexto, se realizó un estudio cuidadoso de diferentes algoritmos de aprendizaje automático, como así también su aplicación juiciosa para datos de origen químico.

Dentro de este estudio, se abordó la aplicación de métodos basados en algoritmos evolutivos cómo método de optimización automático en la tarea de selección del conjunto de descriptores a ser utilizados por un modelo de predicción. Más específicamente, se desarrolló un nuevo modelo basado en la optimización mono-objetivo y otro en la optimización multi-objetivo, obteniéndose en este último mejores resultados en cuanto se refiere a la relevancia de los subconjuntos de descriptores encontrados. También se realizaron experimentaciones con una técnica embebida basada en la proyección a subespacios, la cual representa una metodología novedosa en este campo de aplicación.

En cuanto se refiere al estudio de los métodos de predicción, se procedió al estudio y aplicación de diferentes metodologías para la construcción de modelos de regresión, en donde los métodos basados en redes neuronales constituyeron una herramienta clave para esta tarea. Se estudió también la combinación de métodos *no supervisados* y *supervisados*, de manera de aplicar una *mezcla de expertos* basada en la identificación de agrupamientos naturales en los datos.

En cuanto a la tercer problemática enunciada, se desarrolló un novedoso método de detección de irregularidades que permite identificar la confiabilidad en la

predicción de un compuesto en función del método de predicción y el conjunto de datos de entrenamiento que fue usado para la construcción del modelo de regresión. Para esta metodología se utilizaron mapas auto-organizativos y ciertas técnicas estadísticas basadas en el control y monitoreo de datos multivariados en la ingeniería de procesos.

Finalmente, el método de proyección a subespacios también se lo utilizó como una técnica de preprocesamiento para la predicción, la cual permite obtener resultados competitivos con otros métodos clásicos, como así también poder ser utilizada para visualización a baja dimensionalidad.

El desarrollo del conjunto de metodologías propuestas para mejorar la calidad de las predicciones aporta herramientas que pueden ser usadas para la mejora del proceso de descubrimiento y desarrollo de nuevas drogas. Estas nuevas propuestas permiten disminuir el tiempo dedicado a compuestos que no resultarán aplicables, como así también reducir los costos de este proceso. Por tal motivo, la presente tesis abarca investigaciones y aplicaciones de carácter interdisciplinario, en donde, a su vez, la proposición de nuevos métodos de inteligencia computacional representan una contribución para la ciencia de la computación.

1.3. Contenido y estructura

La presente tesis está organizada en 8 capítulos. En el primero se introduce brevemente la problemática y se delinean los principales objetivos. En el capítulo 2 se introduce, al lector no familiarizado con el ámbito farmacéutico, en la complejidad inherente del descubrimiento de drogas, así como también las dificultades particulares que la industria farmacéutica posee en el proceso y manufactura de fármacos. Luego, el capítulo 3 presenta todos los conceptos básicos de aprendizaje automático requeridos para la comprensión de los algoritmos propuestos. A continuación, el capítulo 4 brinda un estudio bibliográfico de las metodologías de la Quimioinformática existentes orientadas a la predicción de propiedades ADME-Tox, así como también se analiza las principales virtudes y deficiencias de las propuestas existentes en la actualidad. En los capítulos 5, 6 y 7 se detallan los aportes realizados en el marco de esta tesis. En el capítulo 5, se proponen tres metodologías en pos de resolver el problema de la selección de descriptores. El

capítulo 6 aborda cuestiones relacionadas con la aplicación de medidas de distancias en modelos de predicción utilizando métodos híbridos *supervisados* y *no supervisados*. En este capítulo también se presenta un método basado en mapas auto-organizativos que apunta a determinar en forma más precisa el dominio de aplicación real de un modelo inferido. El capítulo 7 contiene 3 metodologías que describen una secuencia de técnicas de proyección a subespacios usando matrices adaptivas. Finalmente, el capítulo 8 resume las principales conclusiones y aportes de las investigaciones realizadas y se enuncian lineamientos y sugerencias para futuras investigaciones.

Capítulo 2

Proceso de descubrimiento y desarrollo de drogas

2.1. Conceptos preliminares

Comenzaremos el presente capítulo, explicando el significado de conceptos usados dentro de las ciencias de química y farmacia. La química tiene por objetivo el estudio de la materia, entendiendo por ésta como todo lo que existe. La materia se encuentra, comúnmente, formando mezclas de distintos materiales, las cuales pueden ser separadas y diferenciadas de acuerdo a su comportamiento. Por tanto, una *sustancia* es una muestra de materia en estado de pureza, es decir que no puede separarse en otras unidades a partir de medios físicos.

Las sustancias pueden clasificarse en dos grupos; el primero denominado *sustancias simples* o *elementos*, los cuales hacen referencia a la unión de iguales elementos de la tabla periódica. El segundo grupo corresponde a las *sustancias compuestas* o *compuestos* los cuales están formados por la unión de sustancias simples.

En el caso de los elementos, se denomina *átomo* a la unidad mínima de la sustancia, mientras que en los compuestos esta unidad recibe el nombre de *moléculas*. Sin embargo, independientemente de las definiciones estrictas usadas en el marco teórico de la química inorgánica, en el transcurso de esta tesis usaremos los términos *compuesto*, *molécula* o *entidad química* en forma intercambiable y

haciendo referencia al concepto de sustancias compuestas, donde su significado preciso podrá desprenderse del contexto en el que se menciona el término.

Del mismo modo, puntualizaremos algunos conceptos relacionados con la ciencia farmacológica. Un fármaco o droga es toda sustancia destinada para el diagnóstico, curación, atenuación, tratamiento o prevención de una enfermedad o para alterar la estructura o función del cuerpo. Por otra parte, un medicamento es la combinación de uno o más fármacos combinados con excipientes, disponible en una forma farmacéutica determinada (*e.g.* comprimidos, cremas, soluciones) y aptas para ser administradas en personas o animales. Un excipiente se añade a un fármaco en pos de alterar las propiedades organolépticas, para posibilitar la preparación y estabilidad, o también para modificar las propiedades fisicoquímicas y por consiguiente mejorar aspectos tales como el transporte o la biodisponibilidad.

El principio activo se define como la sustancia presente en una droga o medicamento responsable de una acción farmacológica. En colación con los conceptos anteriores, un excipiente carece de principio activo mientras que una droga, puede tener uno o más principios activos.

2.2. Mecanismos de acción de las drogas

Del Proyecto Genoma Humano (Collins *et al.* (2003)) se sabe que hay aproximadamente tres mil millones de pares de bases que conforman la molécula de ADN. Sólo ciertos sectores de esta molécula de ADN codifican proteínas. A estos segmentos se los llama *genes* y se estima que en el cuerpo humano existen alrededor de 30.000. A partir de estos genes varios miles de proteínas son sintetizadas.

Para que una droga funcione, ésta debe interactuar con al menos un *blanco*, *diana* o *target* en nuestro cuerpo e intervenir en la modificación de sus funciones. Los blancos farmacológicos más frecuentes los constituyen las proteínas o las glicoproteínas que conforman las enzimas y receptores, con los que las drogas interactúan. Cuando una droga se vincula con un blanco, ésta puede provocar su efecto terapéutico mediante la activación o inhibición de un proceso biológico.

Exceptuando las drogas que actúan al nivel de los nucleótidos, ARN o genes, los principales blancos dentro del cuerpo humano pueden dividirse en tres categorías:

- Enzimas: hay muchos tipos diferentes de enzimas en el cuerpo humano que realizan un gran número de funciones. Las drogas pueden interactuar con las enzimas y alterar su actividad enzimática.

- Receptores intracelulares: estos receptores se encuentran en el citoplasma o núcleo de las células. Para que una droga interactúe con estos receptores, deben atravesar la membrana lipídica de la célula.

- Receptores de superficie celular: como su nombre lo indica, estos receptores están en la superficie de la célula. A través de estos receptores el citoplasma recibe un estímulo externo que afecta los procesos celulares. Las variantes más comunes de este tipo de receptor lo conforman los receptores acoplados a proteínas G y los canales iónicos.

La vinculación de una droga con un receptor o enzima no es un proceso simple de predecir. En primer lugar es críticamente dependiente de las formas y los tamaños de las moléculas, y para provocar la acción terapéutica, la molécula de droga debe unirse en determinados sitios de unión de la enzima o el receptor. Es importante destacar que una proteína, enzima y en algunos casos también las drogas, suele tener una intrincada estructura en tres dimensiones, que requiere de una alta precisión en el modelado molecular, la cual comúnmente se describe con la analogía de una llave que encaja en una cerradura. Asimismo, la vinculación se puede producir por distintas motivos, tales como por la acción de enlaces covalentes, interacciones electrostáticas o por efectos hidrofóbicos. Además, antes de que la droga se una al receptor, ésta puede sufrir deformaciones a nivel celular proveniente de movimientos termales o vibraciones.

2.3. Farmacodinámica y farmacocinética

Cuando se desarrolla un medicamento, se busca que éste sea potente, eficaz y específico, esto es, se pretende que el compuesto tenga un efecto fuerte en la alteración de un determinado proceso biológico y mínimos efectos en cualquier otro proceso biológico que no se necesite alterar. Estas dos condiciones son muy difíciles de alcanzar en forma simultánea, y su estudio dio lugar a dos grandes

ramas dentro de la farmacología: la *farmacodinámica* (FD) y la *farmacocinética* (FC).

La farmacodinámica investiga cómo es la incidencia de una droga en el cuerpo humano. Esto es, una vez que el complejo droga-receptor se forma, la droga regula al receptor provocándole una activación (droga agonista) o una inhibición (droga antagonista). La farmacodinámica analiza la relación cuantitativa del efecto de la droga, en función de la dosis suministrada. El estudio sobre el efecto de la droga no es sólo en cuanto a su acción terapéutica sino también en cuanto a sus efectos adversos.

El área de la farmacocinética es central en esta tesis y la misma estudia la incidencia que provocan los órganos del cuerpo humano en la droga. Básicamente se refiere al estudio del comportamiento cinético del compuesto, esto es desde que el mismo es administrado, transportado hasta el sitio específico de interacción con el receptor y eliminado del cuerpo. Este comportamiento farmacocinético se lo divide en cuatro etapas: absorción, distribución, metabolismo y excreción (ADME) (Figura 2.1). Estas etapas guardan también una estrecha relación con la farmacodinámica ya que, en conjunción con la dosis, determinan la concentración de una droga en sus sitios de acción y, por consiguiente, la intensidad de sus efectos como función del tiempo.

Hay distintos modos de administrar una droga. Puede ser por via: intravenosa, oral, bucal, sublingual, rectal, subcutáneo, intramuscular, transdermal, tópica e inhalatoria. Exceptuando el modo de administración intravenoso, en el que la droga se inyecta directamente en el torrente sanguíneo, todos los demás modos requieren de una sucesión de condiciones biológicas y fisicoquímicas para que la droga llegue a su sitio de acción. Primero se requiere que la droga sea absorbida para que ésta llegue a la sangre y así sea distribuida al blanco terapéutico deseado. Sin embargo, en este recorrido la droga puede tener "complicaciones". Por ejemplo, el metabolismo puede preceder a la distribución, sobre todo en el caso de la administración oral, y de esta forma transformar el principio activo de la droga. Asimismo, el cuerpo humano posee un mecanismo de eliminación que puede eliminar drogas a través de la excreción antes de que llegue a su sitio de acción.

Antes de introducirnos en cada una de la etapas ADME, daremos algunas nociones sobre los mecanismos de transporte existentes. Los mecanismos de transporte permiten que las moléculas de la droga atraviesen las membranas celulares, desde el exterior al interior de una célula o viceversa, y así alcanzar los sitios de blanco farmacológico. Definiremos a continuación los cuatro tipos de transporte principales.

Difusión pasiva. Es el mecanismo de transporte más habitual. La difusión es un movimiento aleatorio de moléculas en un fluido. Por tanto, si el fluido está separado por una membrana semipermeable, como es el caso de la membrana celular, se produce un mayor movimiento de moléculas a través de la membrana desde el sitio de mayor concentración hacia el sitio de menor concentración. En este mecanismo, no hay un gasto de energía por parte de la célula.

Difusión facilitada. En este caso el transporte se produce gracias a la presencia de portadores, que generalmente son proteínas, los cuales se unen a la droga y juntos atraviesan la membrana celular. También se produce en el sentido de gradiente de carga eléctrica, esto es desde el sitio de mayor concentración hacia el de menor concentración. A diferencia del caso anterior, la tasa de transferencia depende de la existencia de portadores, por lo que este transporte tiene un punto de saturación, en donde mayores cantidades de esta droga, no provoca una mayor tasa de transferencia.

Transporte activo. El transporte activo requiere de energía para transportarse en contra del gradiente de concentración. En este caso la tasa de transferencia depende de la existencia de portadores y al mismo tiempo de energía disponible a través de ciertos procesos biológicos.

Endocitosis-exocitosis. La endocitosis es un proceso en el cual la célula envuelve una molécula del exterior hacia su interior. Para esto, la membrana de la célula forma una concavidad, que luego se desprende de la pared celular y se incorpora al citoplasma. En el caso de la exocitosis, el proceso es análogo, sólo que éste se produce desde el interior hacia el exterior.

En este contexto es importante destacar cómo se produce el fenómeno de la difusión a través de la membrana celular. Esta membrana posee una doble capa lipídica, lo que hace que moléculas más liposolubles (también mencionadas como lipofílicas, hidrofóbicas o no polares) sean capaces de penetrar esta membrana por difusión. En contrapartida, las moléculas más hidrosolubles (hidrofílicas, lipofílicas o polares) no son capaces de penetrar fácilmente la capa lipídica por difusión y, generalmente, requieren de otros mecanismos de transporte.

2.3.1. Absorción

La absorción de una droga dentro del cuerpo humano, depende fuertemente del modo de su administración. En el caso más común, el de la administración oral, para que ésta llegue al torrente sanguíneo debe solubilizarse en los fluidos del tracto gastrointestinal, el cual está recubierto de células epiteliales. Por tanto, la droga debe atravesar las membranas de estas células mediante algún mecanismo de transporte. Su éxito en la absorción dependerá de las propiedades fisicoquímicas de la droga y del medio en el que se encuentra. Por ejemplo, en el estómago el pH es bajo lo que produce que las drogas que son ácidos débiles sean preferentemente absorbidas; en cambio el intestino tiene un pH alto, lo que favorece la absorción de las drogas que son bases débiles.

Las drogas absorbidas en el tracto gastrointestinal pasan a la vena porta, la cual desemboca en el hígado. El hígado metaboliza la droga, provocándole una modificación en su estructura química (ver Sección 2.3.3), lo que conlleva a una disminución o aumento de la disponibilidad de la droga para interactuar con los receptores.

En el caso de la administración intravenosa, la dosis entera queda biodisponible para ser distribuida a los sitios de acción. De este modo, se evita la incertidumbre sobre la biodisponibilidad de la droga debido a los procesos de absorción y transporte. Este modo de administración es especialmente útil para casos de emergencia, en donde en aproximadamente un minuto se cumple una circulación completa de la sangre. Sin embargo, por este mismo motivo, es también el más peligroso, ya que puede ocurrir una reacción adversa casi inmediatamente y una vez inyectada la droga no hay forma de evitar su circulación.

Los restantes modos de administración poseen una absorción que depende del órgano sobre el cual se aplique la droga (piel, pulmón, músculo, mucosa, etc). Una ventaja que presentan la mayoría de estos modos de administración es que permiten sortear el metabolismo, al menos el de primer paso hepático. Las desventajas dependen del modo específico de aplicación. Entre los puntos negativos más comunes se tiene: irritación, limitación en el volumen suministrado, dolor, ardor, menor control de la dosificación absorbida y dificultad de aplicación.

Si bien depende de los fines terapéuticos para los que una droga se diseña y de las propiedades estructurales del medicamento, comúnmente se busca que el modo de administración sea via oral. Entre otras ventajas se tiene que resulta cómodo, sencillo, económico, seguro, menos traumático y además permite la autoadministración.

2.3.2. Distribución

Una vez que la droga entra en el aparato circulatorio, la misma se distribuye a los distintos tejidos y en donde la tasa de distribución dependerá de distintos factores. En primer lugar, dependerá de cuán irrigado está un determinado tejido. Por otra parte, una droga podría unirse a proteínas de la sangre (albúmina y lipoproteínas), perdiendo su forma libre y provocando una reducción en la cantidad de droga biodisponible. El tercer factor influyente en la distribución es la difusión pasiva. Las drogas lipofílicas pueden cruzar la membrana celular de los tejidos para interactuar con sus receptores. En cambio, los compuestos hidrofílicos permanecen en la sangre y su difusión es tanto más difícil cuanto mayor sea el grado de su hidrofilia.

Un tema que requiere especial atención en la distribución de un fármaco es la distribución en la placenta y el cerebro. El caso más crítico es cuando se quiere evitar que la droga entre en estos tejidos, dado que los efectos adversos pueden acarrear peligro o incertidumbre. Ampliaremos más sobre la distribución de drogas en el cerebro en la Sección 2.7.2.

2.3.3. Metabolismo

La mayor parte de las drogas son metabolizadas en el cuerpo humano por enzimas. Esto significa que su estructura química se altera y, generalmente, su actividad farmacológica se reduce. Las reacciones más comunes que se generan en las moléculas incluyen la oxidación, la hidrólisis y la agregación o disminución de subgrupos. Estos cambios generalmente provocan que la molécula se vuelva más polar y por consiguiente sea más fácilmente excretada por el cuerpo.

El metabolismo más fuerte es provocado en el hígado, y se le da el nombre de metabolismo de primer paso. Sin embargo, el metabolismo puede ser producido por los riñones, pulmones, intestinos u otros órganos, o bien por interacciones con otras drogas o alimentos que estén siendo suministradas al mismo tiempo.

2.3.4. Excreción

El cuerpo humano posee un mecanismo de eliminación de sustancias. Los órganos principales que se encargan de esta tarea son los riñones y el intestino. Nuevamente, la polaridad de la molécula juega un papel importante, ya que moléculas más solubles en agua son más fácilmente excretadas.

Sin embargo el cuerpo también puede eliminar sustancias por el sudor, saliva, leche o epitelios descamados. Un término generalmente utilizado para describir el tiempo de permanencia de un medicamento es el de *vida media*, el cual corresponde al tiempo necesario para eliminar el 50 % de la dosis suministrada.

2.3.5. Toxicología

Los estudios en toxicología apuntan a evaluar los efectos (adversos) funcionales y morfológicos que una droga provoca. Estas evaluaciones son esenciales a fin de que se pueda determinar la seguridad en el suministro de un fármaco. La toxicología está en estrecha relación con la farmacodinámica y la farmacocinética, al punto que inicialmente se hacía referencia a las propiedades ADMET o ADME-Tox, enfatizando su consideración como un caso particular de estudio dentro de la farmacocinética. No obstante, su estudio es tan amplio y complejo que en los últimos años se ha consolidado como un área de estudio diferenciado.

En estos estudios se analiza el modo, sitio y grado de acción, relación con la dosis, diferencias de sexo, latencia, progresión y reversibilidad de los efectos adversos. La toxicidad también se evalúa en cuanto al efecto provocado por dosis únicas (toxicidad aguda) y por dosis repetidas (toxicidad crónica). Otras temáticas relacionadas con la toxicología comprenden la carcinogenética, genotoxicidad y la toxicología reproductiva.

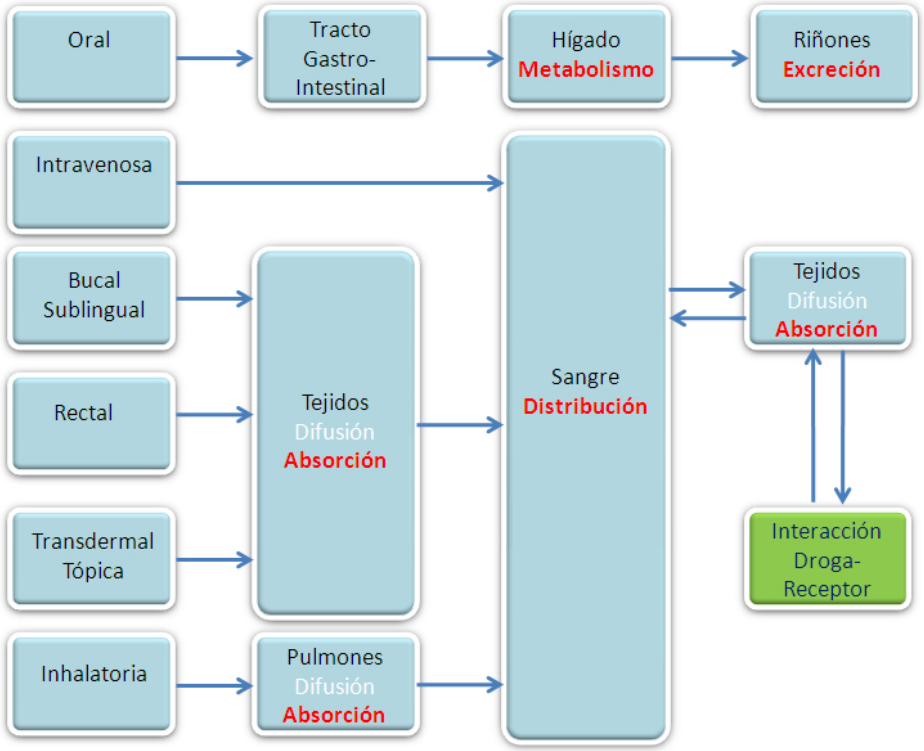


Figura 2.1: Representación esquemática de la farmacocinética de una droga

2.4. Proceso de fabricación de drogas

En esta sección detallaremos las principales etapas del proceso de fabricación de fármacos para así comprender qué actividades de este proceso son de interés mejorar.

2. PROCESO DE DESCUBRIMIENTO Y DESARROLLO DE DROGAS

Si bien desde hace cientos de años la humanidad ha consumido sustancias para el tratamiento de enfermedades, se considera que el estudio sistemático de sustancias xenobióticas empezó en los últimos años del siglo XIX. En estos años, la producción consistía en la extracción de pequeñas cantidades de drogas a partir de recursos naturales. La producción de drogas sintéticas y naturales a gran escala comenzó años después de la primera guerra mundial, dando origen a lo que es la industria farmacéutica moderna.

En la actualidad la industria farmacéutica es una de las industrias más complejas, en donde un sinnúmero de personas provenientes de muy diversas disciplinas actúan en conjunto. Entre ellos se tienen farmacéuticos, médicos, químicos, bioquímicos e ingenieros abocados al descubrimiento y desarrollo de los fármacos. También resulta natural esperar que la industria esté sustentada por economistas, empleados de relaciones públicas y humanas como así también abogados. Sin embargo, en los últimas dos décadas la ciencia de la computación y personas del área de estadística han ganado un espacio importante en esta industria.

El proceso clásico de fabricación de nuevos fármacos se lo divide en las siguientes etapas:

Descubrimiento de drogas. En esta etapa la atención se concentra inicialmente en encontrar el blanco terapéutico que causa la enfermedad. Una vez encontrado el blanco terapéutico se requiere buscar compuestos que sean candidatos a acoplarse con el receptor y generar una reacción biológica de interés. Los compuestos candidatos son optimizados en pos de aumentar su actividad farmacológica. Tanto para encontrar el target como así también el compuesto candidato, diversas técnicas y enfoques han sido usadas a lo largo de la historia. Claramente, la metodología aplicada y la precisión obtenida en esta etapa marcarán fuertemente los tiempos de todo el proceso.

Pruebas preclínicas. Una vez que un conjunto de compuestos candidatos son identificados, se les aplican pruebas llamadas preclínicas que consisten en el testeo del compuesto en animales (pruebas *in vivo*) o en tubos de ensayos que contienen células similares a la de los receptores de interés (pruebas *in vitro*). Esta etapa se caracteriza por ser altamente iterativa, ya que las

pruebas suelen repetirse con pequeñas variaciones sobre los compuestos para determinar los parámetros de farmacodinámica, farmacocinética y toxicología que resulten óptimos. Estas pruebas son documentadas y controladas por organismos públicos para asegurar el cumplimiento de consideraciones de calidad y éticas.

Pruebas clínicas. Una vez que las pruebas pre-clínicas se superaron con éxito, se puede dar comienzo a las pruebas clínicas, las cuales se aplican sobre humanos. Las regulaciones y normas establecidas para esta etapa son naturalmente más estrictas que en la anterior. A su vez las pruebas clínicas se dividen en 3 fases (fase I, II y III), en donde en cada fase se aumenta la cantidad de personas a las que se les suministra el fármaco. En cada fase se tiene un objetivo de evaluación diferente.

Manufactura. Una vez completadas satisfactoriamente las pruebas de fase III, toda la información recopilada es enviada al organismo de aprobación correspondiente de cada país. En este informe se debe especificar, entre otras cosas, cuáles es el proceso de manufactura a seguir para desarrollar el fármaco a gran escala. La producción a gran escala posee dificultades adicionales que no aparecen cuando se trabaja a escala de laboratorio, y que en definitiva puede afectar negativamente la calidad del fármaco o hacer inviable su industrialización. Rigurosas auditorías y revisiones externas son aplicadas dentro de la fábrica para asegurar que se cumplan los estándares previstos en la calidad del proceso.

Fármaco vigilancia. Una vez que la droga fue aprobada para su comercialización, existe una última etapa de vigilancia del fármaco en donde la empresa debe seguir recopilando datos sobre los efectos del compuesto a largo término. En caso de que se hallara evidencias sobre algún efecto no deseado, es responsabilidad de la empresa retirar el fármaco del mercado, lo que conlleva una gran pérdida económica y de confianza para la empresa.

2.4.1. Puntos salientes de la industria farmacéutica

Es importante destacar ciertos datos estadísticos de la industria farmacéutica que la hacen muy particular. El establecimiento del monto de dinero necesario para el desarrollo de una droga ha sido motivo de innumerables debates y discusiones, dado que su valor involucra repercusiones de índole ética y política. Una de las mediciones más aceptadas al respecto establece que el monto promedio invertido en una droga desde la etapa de su estudio hasta su aprobación es de US$ 802 millones, y si contamos otros gastos no estimados y gastos post aprobación la suma asciende a los US$ 1500 millones (DiMasi *et al.* (2003); PhRMA (2005)).

Si se miran las inversiones de dinero en investigación y desarrollo desde hace unos 40 años atrás, vemos que éstas crecen vertiginosamente año tras año (Figura 2.2). Sin embargo, el incremento en la cantidad de drogas aprobadas por año en el mundo no tiene el mismo ritmo de crecimiento.

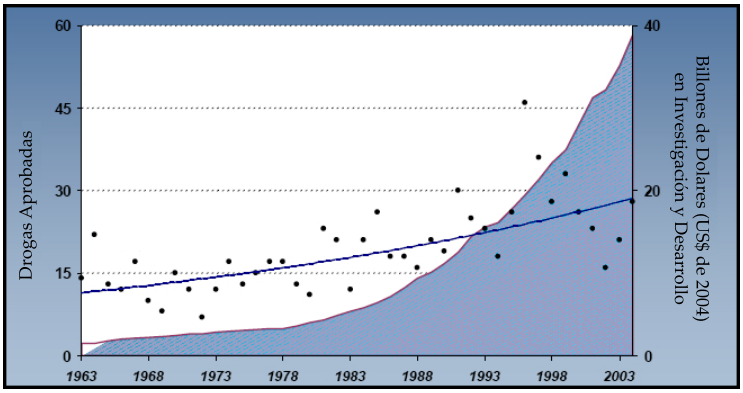


Figura 2.2: Cantidad de aprobaciones de nuevas drogas (puntos negros) y gastos en investigación y desarrollo (curva sombreada)

Está estimado que en promedio el desarrollo de una droga toma entre 10 y 15 años desde el comienzo de las investigaciones hasta la comercialización. Resulta interesante que de cada 5000 compuestos candidatos que se investigan, sólo 250 ingresan a las etapas preclínicas. A su vez de esos 250 compuestos, sólo 5 logran llegar a las pruebas clínicas, y de los cuales sólo 1 resulta finalmente aprobado para su comercialización (PhRMA (2005)). Si se considera además el tiempo

invertido en cada etapa, se advierte el tiempo, dinero y esfuerzo que se pierde en compuestos que no saldrán al mercado. Este proceso además está cargado de una alta incertidumbre ya que para obtener una patente sobre el fármaco se debe demostrar que su aplicación es superior (acción terapéutica, toxicidad) a lo existente en el mercado, así como también se debe asegurar que su manufactura sea viable a gran escala. Se estima que sólo 3 de cada 10 drogas que salen al mercado logran ingresos que superan sus costos de investigación y desarrollo. La Figura 2.3 muestra en forma tabular algunos de los datos expuestos en esta sección.

2.5. Mejoras del proceso de desarrollo de drogas

A partir del análisis anterior, resulta evidente, que cualquier mejora tendiente a disminuir los tiempos de desarrollo o la alta tasa de fallos reduciría los costos de todo este proceso. Este beneficio recaería, no sólo sobre la industria en sí, sino también en todas las personas que hacen uso de medicamentos para tratar sus enfermedades.

A partir de la presión social y empresarial de reducir los costos del proceso de desarrollo de fármacos, hace unos 20 años atrás las empresas comenzaron a analizar cuáles eran las principales causas por las que un compuesto fallaba una vez que había sido lanzado en su proceso de desarrollo. Históricamente las empresas estaban enfocadas mayormente en la potencia de la droga contra su blanco terapéutico, es decir, la capacidad del compuesto de provocar una reacción biológica con la menor dosis posible. También se privilegiaban aquellos compuestos que fueran más selectivos con el blanco, y los que demostraban baja toxicidad en un modelo animal (Gola *et al.* (2006); Selick *et al.* (2002)).

Sin embargo, este enfoque basado prioritariamente en la potencia no es apropiado. Supongamos, por ejemplo, que se quiere desarrollar una droga oral para el cerebro. En este caso, no sólo la potencia, selectividad y toxicidad importa sino que además, la droga debe solubilizarse en el intestino, absorberse en la circulación sistémica a través de la mucosa del intestino, superar el metabolismo del hígado con una cantidad suficiente de droga en sangre para continuar siendo efectiva y por último atravesar la barrera hematoencefálica (ver sección 2.7.2).

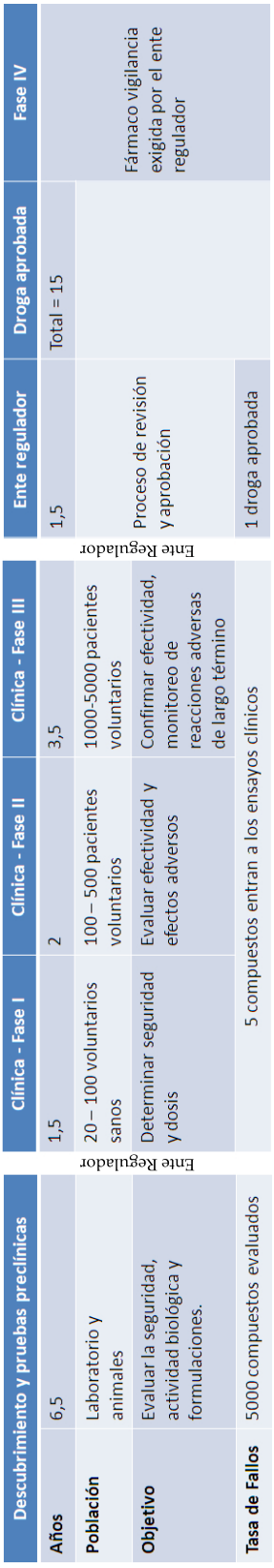


Figura 2.3: Proceso de descubrimiento y desarrollo de nuevos fármacos

Todo esto suponiendo que la droga persistió a los mecanismos de eliminación del cuerpo por excretar o fragmentar el compuesto.

La industria farmacéutica estaba tan cerrada en realizar este proceso en forma serial, y comenzando por la potencia y la selectividad, que no vislumbró que todo el conjunto de pruebas farmacodinámicas (potencia, selectividad), farmacocinéticas (ADME) y toxicológicas importan por igual y por ende, no corresponde su desarrollo en forma serial. Hace no demasiados años atrás, las propiedades ADME eran evaluadas recién al final del proceso de optimización del compuesto, sin embargo era común ver fallas de las propiedades ADME en las pruebas preclínicas e incluso en las pruebas clínicas, etapas para las cuales ya se había invertido mucho tiempo y esfuerzo. El motivo de realizar estas etapas en un estadío posterior al de la potencia y selectividad, era debido al alto costo y el alto tiempo insumido por los experimentos, que mayormente se realizaban *in vivo*.

Durante la década del 90 se conocieron avances importantes en el área de robótica y de síntesis química combinatorial, en donde empezó a ser posible la producción automatizada de grandes cantidades de compuestos. Paralelamente, también aparecieron tecnologías, conocidas con el nombre de *high-throughput screening* (HTS), para experimentar a altas velocidades cuán intensa es la vinculación y actividad de un compuesto con respecto a un receptor. Esto significó un gran avance para la industria farmacéutica ya que se era capaz de sintetizar y testear 100000 compuestos en sólo una semana. Asimismo, algunas empresas comenzaron a aplicar técnicas de ensayos ADME *in vitro* a altas velocidades (*high-throughput ADME*). Sin embargo, a la actualidad, las técnicas de ensayos ADME de alta velocidad son de varios órdenes de magnitud más lentas, que las técnicas HTS convencionales (Tarbit & Bermann (1998)).

El sentimiento general sobre mediados de los años 90, era que aplicando estas técnicas *in vitro* de altas velocidades en paralelo, se reducirían los tiempos y la alta tasa de fallos en las etapas finales del proceso de desarrollo. Sin embargo, esto no fue así. El proceso seguía siendo demasiado empírico y poseía un enfoque de fuerza bruta, ante una cantidad de posibles combinaciones químicas extremadamente grande.

2.6. Pruebas *in silico*

La pregunta natural que surge, es por qué no se puede diseñar una droga pensada para que cumpla con los requisitos de selectividad, potencia y propiedades ADMET, en lugar de necesitar sintetizar y probar empíricamente con una gran cantidad de compuestos hasta encontrar uno apropiado. La respuesta se debe a que no se conoce con exactitud las complejas reglas biológicas y químicas que se conjugan dentro del cuerpo humano y que determinan el comportamiento farmacocinético, farmacodinámico y toxicológico de un compuesto.

Sin embargo lejos de aceptar pasivamente esta imposibilidad de diseño premeditado de las drogas y ante el gran avance tecnológico y científico en la ciencia de la computación, algunos trabajos científicos comenzaron a estudiar la posibilidad de aplicar ciertos filtros o estimaciones de valores utilizando el poder de cálculo de computadoras. A estos métodos por computadora se les dio el nombre de métodos *in silico* o procedimientos de *virtual screening*.

Para el problema de la selectividad y el acoplamiento se han realizado grandes avances en los últimos años, a pesar de que las uniones de tipo no covalentes son prohibitivamente complejas para calcular en forma precisa. Si bien existen simulaciones basadas en conceptos de mecánica clásica y cuántica, generalmente son de una complejidad computacional tan alta que todavía resultan inviables para aplicarlas a gran escala. Este tipo de simulaciones no son abordadas en esta tesis, por lo que el lector puede referirse a Leach (2001) para profundizar en el tema.

Sin embargo, para calcular el comportamiento farmacodinámico y farmacocinético, no resulta posible a la actualidad desarrollar modelos matemáticos completos y precisos basados en las interacciones físicas de los átomos y moléculas. En el caso particular de la farmacocinética, esta necesidad se vuelve más imperiosa dado que las pruebas ADME in vitro son más lentas e inexactas, que las de actividad usando el mismo tipo de prueba. Por lo tanto, se adoptó un modo más indirecto para intentar predecir el comportamiento ADME de un compuesto. Este modo es a partir del cálculo de propiedades fisicoquímicas y biológicas que guarden relación con los procesos farmacocinéticos. Esto es, una vez obtenida una propiedad fisicoquímica de un compuesto, ésta puede brindar información para

comprender cómo va a interactuar en el cuerpo humano. Las propiedades fisicoquímicas más usadas y su relación con las propiedades ADME, se describirán en la Sección 2.7. Existen muchos métodos *in silico* distintos para el cálculo de propiedades, los cuales serán detallados en el Capítulo 4.

2.6.1. Ventajas y desventajas de la predicción *in silico*

Los métodos *in silico* representan una gran oportunidad para la industria farmacéutica. La primer gran ventaja es económica ya que permiten trabajar con moléculas "virtuales", esto es moléculas que no requieren ser sintetizadas para poder ser testeadas. Esto evita además el uso de reactivos y costosa tecnología necesaria para sintetizar y realizar los ensayos.

Los métodos *in silico* no pretenden reemplazar a los métodos *in vitro* o *in vivo*, al menos en el corto término. Sin embargo, permiten establecer un orden de preferencia para los compuestos que son de interés para experimentar en forma *in vivo* o *in vitro* y así evitar sintetizar y experimentar con compuestos que no resultan prometedores para ser utilizados como drogas medicinales. Además, desde el punto de vista de la bioética, resulta alentador cualquier esfuerzo por disminuir la cantidad de experimentaciones sobre animales.

Más aún, las pruebas *in vitro* y las *in vivo* no son infalibles, por lo que agregar una forma más de experimentación de bajo costo resulta altamente importante. En el caso de las pruebas *in vitro*, éstas generalmente son aplicadas utilizando células similares a las que el cuerpo humano usa pero en distintas condiciones, como por ejemplo, distinta concentración de oxígeno o sin contabilizar el efecto de otras macromoléculas del cuerpo humano. Para el caso de las *in vivo* las imprecisiones surgen por las diferencias existentes entre los seres humanos y los animales usados para experimentar.

No obstante, la principal desventaja que presentan los métodos *in silico* es que su precisión y robustez restan por ser mejoradas. Muchos métodos *in silico* parecían ser prometedores pero cuando fueron aplicados en casos de gran escala con compuestos registrados por laboratorios privados, su comportamiento no fue el esperado (Mannhold *et al.* (2008), Tetko & Bruneau (2004)). Estos métodos de predicción son relativamente nuevos y aún tienen mucho camino por delante

para seguir mejorando. El objetivo de esta tesis, se centró en la identificación de algunos de estos problemas y la consecuente proposición de metodologías para solucionarlos. En el Capítulo 4 se analizarán algunos problemas y las soluciones propuestas en la literatura.

2.7. Propiedades fisicoquímicas y biológicas

Cerraremos el presente capítulo describiendo algunas propiedades fisicoquímicas y biológicas que tienen gran importancia dentro del cuerpo humano, y por tanto resultan de interés predecirlas utilizando métodos computacionales.

2.7.1. Hidrofobicidad

La hidrofobicidad (del griego *hydro* y *phobos* que significan agua y miedo respectivamente) es la propiedad física de una molécula de repelerse de una masa de agua. Las moléculas hidrofóbicas tienden a ser no polares, y por ende exhiben una preferencia hacia otras moléculas neutrales y solventes no polares. Por otra parte, la lipofobicidad es la propiedad física opuesta, ya que moléculas de este tipo son polares, se repelen de compuestos lipídicos y tienden a solubilizarse en solventes polares. Los términos lipofilicidad e hidrofilicidad (del griego *filia*, amistad) son usados como sinónimos de hidrofobicidad y lipofobicidad respectivamente, aunque su significado no sea exactamente el mismo. Cuando se usa el sufijo -fobicidad, la palabra connota la capacidad de repulsión, mientras que para el sufijo -filicidad se refiere más bien a la capacidad de solubilizarse en solventes con la misma polaridad.

Esta propiedad es una de las más ampliamente modeladas en forma *in silico*, ya que esta característica guarda estrecha relación con las propiedades ADMET (Caron & Ermondi (2008)). En secciones anteriores, hemos analizado que para que la droga alcance su sitio de acción, generalmente ésta debe primero absorberse y luego distribuirse usando el torrente sanguíneo. El xenobiótico consigue avanzar en el cuerpo, a partir de los mecanismos de transporte vistos en la Sección 2.3. En estos mecanismos, y en la difusión pasiva en particular, la hidrofobicidad juega un papel central ya que ésta determina su capacidad de atravesar o no una membrana

celular. Asimismo, el efecto hidrofóbico es una de las fuerzas más influyentes para la vinculación de moléculas. Esto es, la afinidad a unirse a una proteína que le sirva como transporte o a un receptor que sirva de blanco terapéutico, dependerá, básicamente, de que compartan las mismas características de polaridad.

Las drogas más hidrofóbicas tienen más tendencia a acumularse en compartimientos lipídicos, y por tanto a permanecer en el cuerpo por más tiempo. Esto las convierte en drogas generalmente más tóxicas ya que su excreción es más lenta. Del mismo modo, la propiedad de permanecer en el cuerpo por más tiempo las hace más propensas a metabolizarse más frecuentemente.

La forma más usada para medir el grado de hidrofobicidad de un compuesto es a partir de un coeficiente de partición octanol-agua. Un coeficiente de partición mide la proporción de las concentraciones de un compuesto neutro ubicado en dos solventes no miscibles entre sí (como octanol y agua) en estado estable. Para mantener los rangos de las concentraciones en una escala más conveniente, se utiliza el logaritmo de las proporciones y a esta medición se la conoce con el nombre de logP.

El logaritmo de distribución octanol-agua (logD), permite medir la hidrofobicidad, a un nivel fijado de pH, de una molécula que puede estar cargada (ionizada). Si bien el logD representa una medida más precisa para compuestos ionizables, a partir del trabajo de Hansch y Leo (Leo *et al.* (1971)) el logP ha sido tomado como la unidad de medida para la hidrofobicidad.

El grado de hidrofobicidad buscado para un compuesto que pretende ser utilizado como droga medicinal dependerá del modo de administración y del blanco terapéutico. Generalmente se busca que sea lo suficientemente hidrofílica para que sea soluble en el agua de los líquidos estomacales e intestinales (en el caso de una droga administrada oralmente) y que su excreción sea rápida, pero lo suficientemente hidrofóbica para que pueda vincularse al target deseado o realizar el transporte transcelular. Lipinski *et al.* (2001) propuso la llamada "regla de los 5" ('*rule of five*') en donde establece ciertos límites sobre los cuales debería estar un compuesto para evitar que éste tenga una absorción pobre. Una de las variables usadas en esta regla es el logP.

2.7.2. Barrera hematoencefálica

La barrera hematoencefálica es una estructura celular del sistema nervioso central que sirve de mecanismo de control para evitar que ciertas sustancias ingresen al cerebro o a la médula espinal. Este mecanismo de control permite que metabolitos esenciales, como glucosa y oxígeno, puedan ingresar al sistema nervioso central, al tiempo que otras sustancias, como virus y bacterias, no puedan ingresar. La barrera hematoencefálica recibe en los artículos científicos el nombre de BBB, el cual proviene del inglés *blood-brain barrier* (Ertl (2008)).

La medida más aceptada para cuantificar la capacidad de un compuesto de ingresar al sistema nervioso central es a partir del logaritmo del cociente de la concentración del compuesto en el cerebro sobre la concentración del compuesto en sangre, es decir $logBB = log(C_{brain}/C_{blood})$. El conocimiento sobre el valor de logBB de un compuesto resulta importante no sólo para las drogas que pretenden actuar sobre un receptor localizado dentro del sistema nervioso central, sino que también resulta importante para drogas destinadas a receptores situados fuera del sistema nervioso central no realicen un efecto no deseado.

2.7.3. Absorción intestinal humana

Para la mayor parte de las drogas, la ruta de administración preferida es por via oral. Por tanto para que la droga pueda alcanzar el torrente sanguíneo, el intestino (delgado) juega un papel central. Si la droga es administrada como sólido, ésta debe disolverse primero y luego atravesar una serie de membranas celulares hasta alcanzar el sistema portal (Klamt & Smith (2008)). El mecanismo de absorción intestinal ha sido ampliamente discutido en la literatura, no obstante aun sigue habiendo discrepancias, sobre todo en la velocidad de acceso al torrente sanguíneo (Abraham *et al.* (2002)).

La forma de medir la absorción intestinal es a partir del porcentaje de la dosis que aparece en la vena porta con respecto a la cantidad suministrada. Esta propiedad resulta importante para establecer la disponibilidad de un compuesto, es decir qué porcentaje de la dosis está disponible y cuánto fue excretado (antes de la metabolización). Cuanto más biodisponible sea un compuesto, mejor es, ya que requiere dosis menores y por lo tanto posee menos riesgo de toxicidad.

Capítulo 3

Conceptos de aprendizaje automático

En este capítulo se presentará una revisión de los principales conceptos y técnicas en las áreas de estadística y aprendizaje automático, necesarias para la correcta comprensión de las metodologías propuestas en el marco de esta tesis. Este capítulo busca ser una referencia rápida para el lector de este trabajo, pero no pretende ser una guía completa y acabada de todos los temas propuestos. Para una lectura más avanzada se puede consultar Bishop (2006), Hastie *et al.* (2009) y Gan *et al.* (2007).

3.1. Introducción aprendizaje automático

El aprendizaje es una característica humana natural y esencial. En algún punto, aprender involucra cambiar para mejorar (según algún criterio) cuando alguna situación similar a lo visto o vivenciado suceda. Intuitivamente se advierte que el aprendizaje de memoria no se corresponde con la esencia más pura del aprendizaje.

El aprendizaje automático o *machine learning* en inglés, se corresponde a la tarea de que un modelo computacional pueda aprender. Desde el punto de vista informático, sabemos que las computadoras pueden memorizar fácilmente. Sin embargo, el desafío del aprendizaje automático consiste en que se pueda generalizar un comportamiento frente a una nueva situación.

La idea básica del aprendizaje automático consiste en observar una cantidad finita de datos de entrenamiento, de la cual se debe derivar una relación para un dominio mayor al del entrenamiento. El primer problema que surge es que la cantidad de posibles relaciones es infinita. En la Figura 3.1 se muestra un ejemplo de distintas funciones que vinculan la variable del eje horizontal con la del vertical. La pregunta natural es entonces, ¿cómo debemos elegir una relación cuándo múltiples relaciones son posibles?

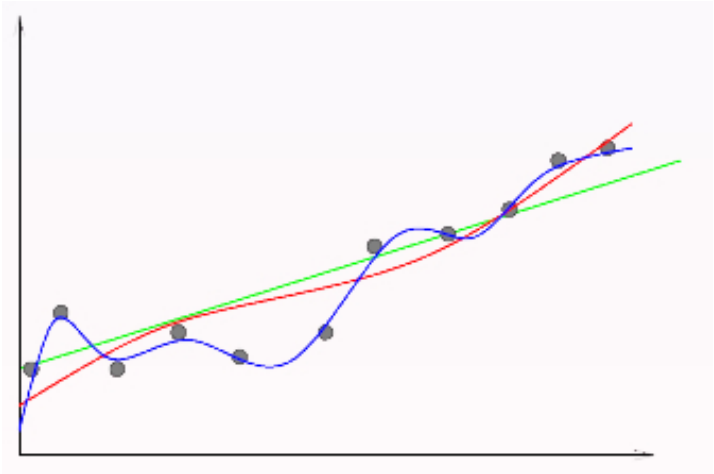


Figura 3.1: Ejemplo de las infinitas relaciones posibles para un modelo.

Existen dos principios que ayudan a contestar esta pregunta para identificar la relación apropiada. El primero de ellos consiste en hacer uso de más datos, ya sea en el mismo conjunto de entrenamiento o en un segundo conjunto, llamado de validación. En este segundo conjunto se puede constatar lo aprendido y refinar o detener el aprendizaje en caso que el conjunto de validación no se condiga con el modelo inferido del entrenamiento. El segundo principio fue tomado de un fraile franciscano del siglo XIV, William de Occam, quien dijo: "*no se debería aumentar, más allá de lo necesario, la cantidad de entidades necesarias para explicar algo*". Esto implica que, cuando existe más de una solución para un mismo problema, se debe elegir la más simple. Aún cuando la simpleza puede ser un concepto subjetivo, se han propuesto distintos enfoques para que un modelo computacional

sea simple, lo cual conlleva significativos avances en cuanto a la generalización del aprendizaje, como se da por ejemplo con la regularización (sección 3.2.5.1).

Estos conceptos se encuadran dentro de lo que se conoce como aprendizaje *supervisado*. Es decir, el aprendizaje supervisado puede ser formalizado como la tarea de inferir una función $\mathcal{Y} = f(\mathcal{X})$ a partir de un conjunto de entrenamiento $T = ((\mathbf{x_1}, y_1), ..., (\mathbf{x_m}, y_m))$, donde cada par $(\mathbf{x_i}, y_i)$ es una observación con su rótulo. Cuando el espacio $\mathcal{Y}$ es continuo e $\mathcal{Y} \in \mathbb{R}$ estamos en el contexto de una regresión. En cambio si $\mathcal{Y}$ es discreto tal que, por ejemplo, $\mathcal{Y} \in \{-1, 1\}$ estamos en un contexto de clasificación de dos clases. En el caso supervisado, $\mathcal{Y}$ se la conoce como variable a modelar, variable dependiente, variable de salida o variable destino (*target*), mientras que $\mathcal{X}$ describe el conjunto de datos u observaciones de entrada. Cada observación del conjunto de entrenamiento debe estar acompañada de su correspondiente valor en la variable a modelar.

Hay otro tipo de aprendizaje que no hemos considerado hasta el momento, que será tenido en cuenta en este trabajo de tesis y se denomina aprendizaje *no supervisado*. En este tipo de aprendizaje no existe o no se tiene en cuenta la variable de salida asociada a cada observación de la entrada. Por lo tanto, el objetivo aquí pasa por descubrir grupos de observaciones similares o patrones dentro de los datos de entrada. Cuando se busca descubrir datos similares dentro del set de datos, se denomina *análisis de agrupamiento* (*clustering*), mientras que si se quiere encontrar el patrón de distribución de los datos, se denomina *estimación de densidad*.

3.2. Modelos supervisados

Comenzaremos entonces a describir los modelos supervisados de relevancia para la comprensión de los métodos aplicados en esta tesis. En cada uno de los métodos se apunta a indicar los principios básicos de funcionamiento, junto con sus ventajas y desventajas.

3.2.1. Modelos de regresión lineal

Los modelos de regresión lineal (RL) también llamados modelos de regresión lineal múltiple (haciendo explícita su aplicación multidimensional) asumen la existencia de una vinculación lineal entre la variable de salida y las variables independientes, tal como se muestra en la Ecuación 3.1, donde $X \in \mathbb{R}^{m \times n}$, $\mathbf{c} \in \mathbb{R}^{n+1}$, $\mathbf{y}$, $\mathbf{e} \in \mathbb{R}^{m}$. En esta ecuación $\mathbf{c}$ es el vector de coeficientes, $\breve{X}$ es el resultado de concatenar horizontalmente a X un vector de '1' y $\mathbf{e}$ es un término de error aleatorio con media cero, un desvío constante σ_e e independiente de una observación a otra.

$$\mathbf{y} = [X\mathbf{1}] \cdot \mathbf{c} + \mathbf{e} = \breve{X} \cdot \mathbf{c} + \mathbf{e} \tag{3.1}$$

El modelo es tradicionalmente ajustado de acuerdo al método de cuadrados mínimos, el cual busca minimizar el cuadrado del error $\mathbf{e}$. Si la inversa de la matriz $\breve{X}^T \cdot \breve{X}$ existe, la solución de este problema de minimización es única y depende explícitamente de $\breve{X}$ e $\mathbf{y}$ (Función 3.2).

$$\mathbf{c} = (\breve{X}^T \cdot \breve{X})^{-1} \breve{X}^T \mathbf{y} \tag{3.2}$$

En contrapartida, RL presenta una limitación más allá de su linealidad. En su forma original, el método no puede ser aplicado cuando el número de variables supera la cantidad de datos. RL asume que las variables son linealmente independientes entre sí, por lo que el rango de la matriz de datos debe ser igual a la cantidad de variables. Si las variables poseen cierta colinealidad, el calculo de los coeficientes se vuelve un problema mal condicionado. Por lo tanto, este método suele ir acompañado de alguna técnica de reducción/selección de variables. Entre las técnicas de selección de variables para RL destacamos las siguientes:

Regresión de componentes principales (PCR - *principal components regression*): Consiste en aplicar RL posteriormente a una transformación de la matriz X según sus componentes principales. La regresión se aplica en el espacio de las componentes principales. De esta forma se puede reducir la cantidad de variables evitando al mismo tiempo la colinealidad.

Cuadrados mínimos parciales (PLS - *partial least squares*): Este método es muy similar a PCR, en el sentido que se hace una transformación de los datos a un espacio de dimensión menor o igual al rango de la matriz X, y sobre ese espacio se hace la regresión. Difiere del anterior, en el sentido que la proyección se realiza al subespacio que mejor rescata la relación lineal con la variable de salida $\mathbf{y}$. El problema de PLS es que las variables del espacio transformado suelen perder su interpretabilidad aún más que en el caso de PCR.

Regresión de a pasos hacia adelante/atrás (*forward/backward stepwise regression*): La idea de este método es agregar o descartar variables de a una y en forma iterativa, de acuerdo a un test de significancia de la variable en cuestión con la variable de salida $\mathbf{y}$. De esta forma el método selecciona la cantidad de variables a usar para la regresión. Sin embargo, este método posee la desventaja de los múltiples test de significancia sobre los mismos datos, que hacen perder su confiabilidad. Además, la relevancia de las variables se mide en forma individual y no globalmente (ver sección 5.1).

3.2.2. Árboles de decisión para regresión

La técnica más simple para ajustar un árbol de decisión (AD) para regresión consiste en particionar recursivamente los datos en grupos más pequeños (llamados nodos) con divisiones binarias que surgen de la división de una variable, por ejemplo: $var_j < c$ y $var_j \geq c$, donde c es un valor escalar. En cada iteración, se evalúan posibles divisiones para cada una de las variables mediante una búsqueda exhaustiva y se elige la mejor división. Existen diversos criterios para definir la mejor división. El criterio de Breiman *et al.* (1984) es definir la mejor partición como aquella que minimiza el error cuadrado total de los G nodos, tal como se muestra en la Ecuación 3.3. En el g-ésimo nodo el valor de predicción $\tilde{y}_i$ puede ser simplemente definido como el valor promedio de los M_g valores de salida de los datos que quedaron en el nodo g.

$$\Sigma_{g=1}^{G}\Sigma_{i=1}^{M_g}(y_i - \tilde{y}_i)^2 \tag{3.3}$$

Para un nuevo dato $\mathbf{x_j}$ se puede calcular su valor de predicción según la Ecuación 3.4, donde $I_g(x_j)$ es el indicador de nodo (1 si la observación $\mathbf{x_j}$ pertenece a al nodo g, 0 en caso contrario) e $\bar{y}_g$ es el valor promedio de salida del nodo g. Existen también otras variantes de este enfoque, en donde en lugar del promedio se puede aplicar una regresión lineal entre todos los datos de un nodo.

$$\tilde{y}_j = \Sigma_{g=1}^{G} \bar{y}_g I_g(\mathbf{x_j}) \tag{3.4}$$

De esta forma, el árbol puede ser recursivamente particionado hasta que todos los nodos terminales (llamados hojas) poseean un solo dato. Sin embargo, un árbol de estas características no es tan deseable ya que es propenso a que sobreajuste los datos. Por lo tanto existen alternativas de "podado" para llevar el árbol a un tamaño optimal. También existen otras variantes de optimización y de división no solamente binaria. El caso para problemas de clasificación es análogo al presentado.

La salida del método es una estructura de árbol matemático donde las ramas están determinadas por las reglas de división y el conjunto de nodos terminales que contienen la respuesta media de su grupo. Resulta una técnica interesante dado que no se requiere información *a priori* del tipo de relación o de cuáles son las variables importantes, por lo que se lo puede considerar como un método de selección de variables embebido (sección 5.1). Además, permite modelar relaciones de tipo no lineal y la lógica del modelo queda explícitamente expresada por el contenido de las reglas determinadas.

3.2.3. *k*-vecinos más cercanos

El método de los k-vecinos más cercanos (kVC) consiste en predecir la variable de salida de un dato a partir del promedio (o promedio ponderado) de los valores de salida de los k datos más cercanos. En este método, las variables no son usadas directamente para establecer una regresión sino para definir la vecindad de la entidad. Más precisamente, para cualquier dato $\mathbf{x_i}$ el valor de predicción se calcula según la Ecuación 3.5, lo que corresponde a la salida promedio de las k observaciones más cercanas, las cuales se encuentran en el subconjunto $G_k(x_i)$.

$$\tilde{y} = \frac{1}{k}\Sigma_{x_i \in G_k(\mathbf{x_i})} y_i \tag{3.5}$$

Aún cuando su teoría es bastante simple, su aplicación requiere de ciertas selecciones de parámetros y métodos que cambian considerablemente el comportamiento del método (Hastie *et al.* (2009)).

Elección de k: esto es el número de datos que pertenecerán en la vecindad. Existe una regla establecida que consiste en adaptar k al número m de observaciones, como por ejemplo $k = \sqrt{m}$. Sin embargo, existen otros esquemas más interesantes que consiste en elegir el k basándose en una validación cruzada con distintos valores de k y analizando con cuál se obtiene el mejor compromiso de sesgo-varianza.

Medida de distancia de las observaciones: algunas posibles opciones, entre otras, son la Euclídea, la de Manhattan y Mahalanobis y otras más complejas, sobre todo para los casos de variables discretas, como por ejemplo las de Dice, Gower y Tanimoto (Johnson & Wichern (1992)).

Método de ponderación: en el caso del promedio todos los datos de la vecindad tienen el mismo peso, pero uno podría variar los pesos de acuerdo a alguna función de la distancia o de acuerdo a un determinado kernel.

kVC es altamente intuitivo y no requiere de conocimiento *a priori* de la función que vincula las variables de entrada con las de salida. Sin embargo, en escenarios donde la dimensionalidad es muy alta, el método no tiene un buen comportamiento, dado que el concepto de distancia se vuelve aún mas incierto.

3.2.4. Redes neuronales

Las redes neuronales son una familia de técnicas de modelado no lineal cuyas numerosas contribuciones en problemas científicos e industriales, las convierten en una de las técnicas de aprendizaje automático más usadas. El trabajo de Rosenblatt (1958) y su modelo *Perceptron* es el origen de esta técnica inspirada en

la configuración y funcionamiento de las neuronas en el cerebro humano. La configuración de estos Perceptrones, nodos o neuronas en múltiples capas, o redes neuronales, permite extender ampliamente la potencialidad del modelo. Una suficiente cantidad de estos nodos permite, en problemas de regresión, el modelado de cualquier función no lineal y, en problemas de clasificación, la separación de cualquier región no linealmente separable.

El funcionamiento básico de una red neuronal de dos capas puede describirse a partir de la Ecuación 3.6, donde f_k es la función a modelar con la red a partir del vector de entrada $\mathbf{x}$ y una serie de matrices $W^{(k)}$ (con $k = 2$ si la red tiene dos capas). Las funciones g_j y G se denominan funciones de activación de las neuronas correspondientes a nodos de la capa oculta y de salida respectivamente. Las matrices $W^{(1)}$ y $W^{(2)}$ corresponden a las matrices de pesos que vinculan los nodos de cada capa y que pondera cada una de las entradas que reciben las funciones de activación de los nodos. En la Figura 3.2 se muestra la estructura en red con la que se representa el modelo. Cada cuadrado corresponde a una de las variables de entrada del problema y los círculos se corresponden con los nodos. Los pesos de la matriz W quedan representados en los enlaces que unen los nodos.

$$f(\mathbf{x}, W) = G\left(\sum_j W_{1,j}^{(2)} g_j\left(\sum_{i=1}^{n} W_{j,i}^{(1)} x_i + w_{j,0}^{(1)}\right) + W_{1,0}^{(2)}\right) \tag{3.6}$$

La potencialidad de las redes de múltiples capas fue reconocida hace un buen tiempo atrás, pero no fue hasta la existencia de un algoritmo de aprendizaje para las redes neuronales que el método atrajo la atención de un gran número de científicos de distintas disciplinas. El primer algoritmo de aprendizaje es el de retro-propagación (*back-propagation*), el cual fue propuesto en forma independiente en Bryson & Ho (1969), Werbos (1974) y Rumelhart *et al.* (1986).

En función del tipo de nodo o función de actividad, arquitectura de la red, tipo de entrenamiento y algoritmo de aprendizaje, se pueden armar múltiples combinaciones de métodos que sería demasiado extenso mencionarlas aquí. Para una revisión completa y profunda en el estudio de las redes neuronales sugerimos el libro de Bishop (1995).

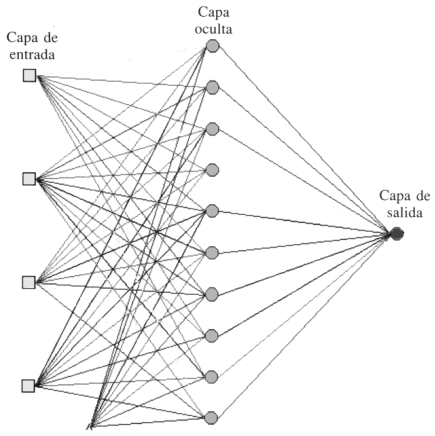


Figura 3.2: Representación de una red neuronal de dos capas de nodos.

3.2.5. Aspectos importantes del modelado

3.2.5.1. Regularización

Dada la gran potencialidad de las redes neuronales para modelar cualquier tipo de relación no lineal, se debe tener cuidado de no caer en el sobreajuste de los datos. El sobreajuste se caracteriza por el modelado muy preciso de los datos del entrenamiento, posiblemente con un modelo excesivamente complejo, que tiende a no ser generalizable ante la presencia de nuevos datos.

En los modelos de regresión lineales, aunque en menor medida, también se tiene este problema. Una de las alternativas para remediarlo se conoce con el nombre de *regresión contraída* (*ridge regression*) el cual consiste en incorporar a la función a minimizar la suma de los cuadrados de los coeficientes de regresión. De esta manera, coeficientes más chicos son favorecidos, con lo que se logran funciones más "suaves". Esta penalización se denomina regularización.

Algo similar ocurre con las redes neuronales, en donde distintas estrategias de regularización han sido propuestas en la literatura. Uno de los métodos más conocidos se conoce como *deterioro de los pesos* (*weight decay*), el cual consiste en penalizar las redes con valores altos en los pesos que conectan los nodos. Otra alternativa relacionada consiste en hacer un ajuste "manual" de la cantidad de nodos a usarse en las capas ocultas. Cuantos más nodos y capas tenga una red,

más propensa es la red a generar modelos que sobreajusten los datos.

Otra estrategia muy frecuentemente usada, se conoce con el nombre de *detención temprana* (*early stopping*). El entrenamiento de una red neuronal involucra la reducción iterativa de la función de error obtenida sobre los datos de entrenamiento. Intentar reducir el error a 0 puede llevar a que el modelo se vuelva demasiado complejo y, por ende, poco general. Por lo tanto, la propuesta consiste en detener el algoritmo de aprendizaje de acuerdo a la evaluación de la función de error sobre un conjunto adicional, denominado de validación. La idea es conservar los pesos del estado en el que se obtuvo un mínimo de la función de error sobre el conjunto de validación. La hipótesis que se sigue es que en las iteraciones posteriores no se mejora la capacidad predictiva del modelo, sino que se sobreajustan los datos del entrenamiento.

Por otra parte, la regularización puede llevarse a cabo siguiendo un enfoque Bayesiano, el cual se aplicó por primera vez en MacKay (1992a). Para comprender en forma intuitiva las diferencias con los modelos de aprendizaje no bayesianos, éstos pueden ser vistos como un proceso de dos partes diferenciadas, en donde por una parte se ajustan los parámetros de un modelo y por otra parte se compara su funcionamiento con otros modelos y así se determina cuál es mejor. Por ejemplo, redes neuronales con distinta cantidad de nodos en la capa oculta pueden ajustarse por separado y luego compararse. Sin embargo, a menos que se usen datos adicionales, esto no es simple de evaluar. Los métodos Bayesianos permiten efectuar este ajuste y comparación de las distintas opciones en una forma automática y elegante. En la misma optimización se aplica el principio de Occam sin la necesidad de introducir penalidades *ad hoc* en la optimización.

Explicar todo el funcionamiento de la regularización bayesiana, requeriría una larga exposición de conceptos en el área de probabilidad bayesiana. Recomendamos consultar el trabajo de MacKay (1992b) para una descripción detallada del tema aplicado a redes neuronales. En resumidos términos, el enfoque Bayesiano aplica una normalización de la probabilidad *a posteriori* que no depende del ajuste de los pesos en sí, pero que favorece la selección de las hipótesis de aprendizaje más simples. Esta normalización se realiza mediante el cálculo de la probabilidad de los datos de entrenamiento T dada una determinada hipótesis $\mathfrak{H}_i$, esto es $p(T|\mathfrak{H}_i)$.

3.2.5.2. Comités de modelos

Un comité de modelos puede ser definido como un conjunto de dos o más modelos, que se usan juntos como un meta modelo. La motivación de esta técnica se basa en la idea de usar conocimiento de distintas fuentes antes de tomar decisiones. Existen distintas estrategias para llevar a cabo esta idea. Por un lado existen los enfoques divisivos, que intentan dividir el problema en subproblemas con ninguna o mínima superposición. Para cada subproblema se aplica un método de aprendizaje que pasa a ser el "experto" en ese subproblema. Se necesita, además una función que indique qué método usar ante nuevas instancias. Esta idea se conoce con el nombre de *mezcla de expertos* en la literatura (Jacobs *et al.* (1991)).

Por otra parte, existe otro enfoque que consiste en combinar los resultados de distintos métodos sin la necesidad de reducir el problema. El objetivo de esta idea consiste en eliminar la relación de compromiso sesgo-varianza (Geman *et al.* (1992)). Esto se logra o bien disminuyendo la varianza de predictores poco sesgados, o reduciendo el sesgo de predictores estables.

En el primer caso, se trata de construir muchos predictores y promediar los resultados, técnica que se conoce como *ensemble*. Para que esta técnica tenga sentido, en ciertos casos se debe forzar a que cada modelo tenga un comportamiento diferente. Esto se realiza, por ejemplo, modificando ligeramente el conjunto de entrenamiento que cada conjunto usa o modificando parámetros o funcionamiento del modelo, tal como sucede en las técnicas de *bagging* o *random forest* (Breiman (1996)). En el caso de las redes neuronales convencionales (no regularizadas), no es necesario forzar una modificación, ya que la misma varianza producida por la configuración de los pesos iniciales, produce una gran varianza en los resultados.

En el segundo caso, se trata de incorporar sucesivamente modelos simples, es decir con alto sesgo y poca varianza, haciendo que cada método se concentre en modelar los datos peores modelados por el comité anterior. Esta técnica se conoce con el nombre de *boosting* y su algoritmo *AdaBoost* es su representante más conocido (Freund & Schapire (1997)). Sin embargo, esta última técnica no ha sido atacada en este trabajo de tesis.

Esta idea de la proliferación de modelos fue publicada hace más de cuarenta años atrás por Nilsson (1965), aunque valorada recién en la década del 90, dado que el beneficio se sustenta en el aumento del cómputo asociado. Aunque el concepto usado es simple, los beneficios que proporciona esta técnica de multiplicación de modelos son importantes, en términos de reducción del error y la varianza de las predicciones.

3.2.5.3. Validación de los modelos

Al aplicarse un método de aprendizaje debe ponerse particular atención al mecanismo utilizado para evaluar la capacidad predictiva del método. Hemos dicho que muchos métodos de aprendizaje no lineales permiten ajustar cualquier función posible, por lo que se desprende que la evaluación del error sobre el mismo conjunto de entrenamiento posee una importancia limitada. Por lo tanto, se debe hacer uso de datos adicionales e independientes a los del entrenamiento, si se quiere tener una aproximación no sesgada de la capacidad predictiva de un método. Un conjunto de validación, usado para controlar el grado de ajuste de un modelo, aún cuando no es usado para el modelado, tampoco puede ser usado como evaluador de la calidad predictiva del modelo, ya que no es completamente independiente del proceso de optimización.

En este contexto, podemos diferenciar dos grandes formas de aplicar validación: externa o interna. En el caso de la validación externa, además de los conjuntos de entrenamiento y validación (si es que se usa validación), se hace uso de otro set de datos, denominado de testeo[1], el cual se mantiene ajeno a todo el proceso de entrenamiento. Una vez entrenado el modelo, se constata la calidad predictiva del modelo haciendo uso del conjunto de testeo. Este conjunto debe ser suficientemente grande como para que la evaluación sobre él se considere representativa. Para que esta validación sea verdaderamente independiente, el conjunto de testeo no se debe usar más de una sola vez, dado que de lo contrario, se estaría ajustando el modelo a los datos del conjunto de testeo. Cabe mencionar que existe mucha controversia sobre cómo separar y usar el conjunto de testeo a usarse en un modelo (Hastie *et al.* (2009)).

[1]En algunos trabajos se utiliza una denominación distinta o invertida para los conjuntos de

La validación interna o cruzada, en cambio, si bien también separa un subconjunto de los datos para usarlos como testeo, esta acción se repite varias veces, usando en cada iteración, subconjuntos distintos para entrenar/validar y testear. La forma en que se realiza la partición y la proporción de los datos usados en cada iteración, origina distintas denominaciones del método. Esto es, si sólo se reserva un dato para testear por iteración, se conoce a esta estrategia con el nombre de *leave-one out* (del inglés, *dejar uno afuera*); si se reservan más de un dato para testear, se le da el nombre de *leave-many out* (del inglés, *dejar muchos afuera*). Cada uno de estos subconjuntos separados para testear recibe el nombre de *doblez* o *fold*, por lo que este tipo de técnicas también se conocen con el nombre de validación cruzada de *f-folds*, donde $100/f\,\%$ es el porcentaje de los datos que se reserva en cada fold, siendo f un número entero. Cabe decir que si m es el total de los datos, la validación cruzada de m-folds equivale a la técnica de leave-one out.

La ventaja de la validación interna es que no es tan dependiente de la separación entrenamiento-testeo, y por ende más imparcial y reproducible. No obstante, no se la puede considerar como un método de validación completamente insesgado, ya que la capacidad predictiva final no corresponde a ningún modelo particular sino al de un modelo inexistente que se calcula como promedio de los distintos modelos, produciéndose así un "suavizado" del error final (Gramatica (2007)). Generalmente se estila utilizar validación cruzada para la selección de parámetros de un modelo o cuando la cantidad de datos no es lo suficientemente grande como para separar un conjunto independiente de testeo. Desde nuestro punto de vista, el uso de ambas estrategias de validación resultaría lo óptimo.

3.3. Modelos no supervisados

En esta sección describiremos los métodos de aprendizaje no supervisados que fueron aplicados dentro de las experimentaciones desarrolladas en los capítulos 6 y 7. En términos generales son técnicas que permiten identificar distintas formas de agrupación de los datos. Estas técnicas no requieren suposiciones sobre el

validación y testeo.

número o la estructura de los grupos, sino que el agrupamiento es realizado en base a distintos criterios impuestos por el método. Presentamos por una parte las técnicas de análisis de agrupamiento, la cual consiste en una familia de métodos que se usan desde hace más de 30 años dentro de la estadística multivariada, y por otra parte los mapas auto-organizativos, la cual representa una técnica más moderna usada dentro del aprendizaje supervisado.

3.3.1. Análisis de agrupamientos

El *Análisis de Agrupamientos*, de *Conglomerados* o de *Clusters* tiene como objetivo formar grupos de elementos de manera tal que los que pertenecen a un mismo grupo sean parecidos entre sí y distintos a los pertenecientes a otros grupos. Normalmente se usa para agrupar observaciones, pero también pueden utilizarse para variables.

Existen esencialmente dos tipos de métodos: los jerárquicos y los de partición. Los primeros intentan ordenar los elementos en una gradación según su nivel de similitud, mientras que los segundos meramente asignan cada elemento a un grupo de manera tal de obtener grupos internamente homogéneos. En particular, los métodos jerárquicos aglomerativos (los usados en esta tesis) utilizan información de una matriz de distancias para agrupar los datos en forma sucesiva empezando por los que están más cercanos. Una vez agrupados los dos datos, el grupo se comporta como un nuevo dato, sobre el cual se recalculan las distancias con los restantes datos y luego se repite el paso inicial, hasta que quede un solo grupo. Después de este proceso, se puede armar un dendograma (Figura 3.3) que ilustra a qué distancia se une cada uno de los grupos. Un concepto clave aquí involucra la medida de distancia usada y el criterio de ligamiento, el cual indica cómo medir la distancia entre grupos de datos. Una vez obtenido el dendograma se establece un criterio de corte, como por ejemplo la distancia máxima de agrupamiento o el número máximo de grupos, y así determinar la cantidad de grupos a considerar.

Por otra parte, en los métodos de partición se define primeramente la cantidad de grupos y luego se determinan cómo quedan conformados los grupos. El método más conocido de este tipo se lo conoce con el nombre de las k-medias. En este método se comienza con la definición de k puntos aleatorios iniciales. Luego,

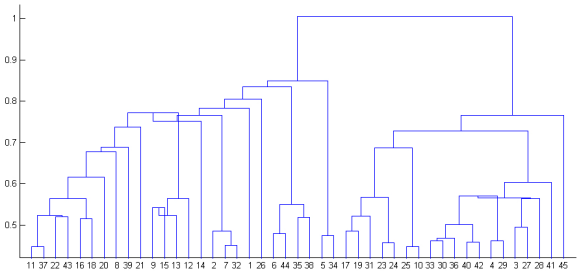


Figura 3.3: Ejemplo de un dendograma.

se calculan las distancias de todos los datos a cada uno de los k puntos, para determinar cuál es el punto más cercano y así formar k grupos. A continuación, y siguiendo algún criterio de optimalidad, se calcula si una nueva reasignación de los datos mejora dicho criterio, y así se continua hasta que ninguna reasignación mejore el criterio impuesto. Uno de los criterios comúnmente usados es la suma de cuadrados dentro de los grupos. Para un estudio detallado en el tema de análisis de agrupamiento recomendamos la lectura de Anderberg (1973).

3.3.2. Mapas auto-organizativos

Dentro de las técnicas más comunes de aprendizaje automático no supervisado se encuentran los mapas auto-organizativos (*self-organizing maps* - SOMs) y la cuantización de vectores (*vector quantization* - VQ) (Kohonen (1997)). Ambas presentan un comportamiento similar, pero nos centraremos en los SOMs, dado que corresponde a uno de los métodos utilizados en este trabajo.

El SOM es un algoritmo generalmente considerado dentro de las redes neuronales. Los nodos o celdas de salida suelen estar dispuestos en formas geométricas, donde cada nodo actúa de modo competitivo ante la presentación de un dato. A su vez, cada nodo posee un vector de pesos, de igual dimensión que los datos de entrada. La Figura 3.4 muestra un esquema de la arquitectura de un SOM con los nodos dispuestos en forma de grilla rectangular, con una configuración de 9×7, y donde los casos presentados poseen n variables.

Cada caso es presentado al SOM, en forma aleatoria y consecutiva, en donde el nodo ganador (NG) resulta de comparar el caso presentado con los vectores

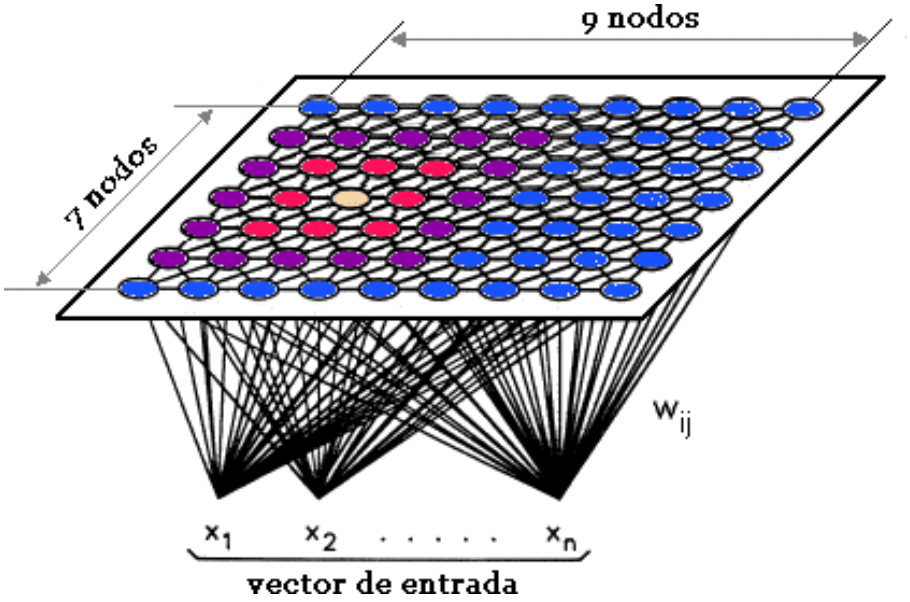


Figura 3.4: Esquema de un SOM con una arquitectura de grilla rectangular.

de pesos de cada nodo de la grilla, definiéndose al ganador como el más cercano siguiendo alguna medida de distancia. Luego, el peso del NG es modificado de manera tal que se disminuya la distancia al caso presentado. La Ecuación 3.7 refleja esta actualización del vector de peso $\mathbf{w}$ para el NG ante la presentación de un dato $\mathbf{x_i}$, en donde t representa el tiempo y φ es la tasa de aprendizaje, la cual decae con el tiempo según la Ecuación 3.8 y φ_0 es la tasa de aprendizaje inicial.

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \varphi(t)(\mathbf{x_i} - \mathbf{w}(t)) \tag{3.7}$$

$$\varphi(t) = \varphi_0 \cdot e^{-\frac{t}{\lambda}} \tag{3.8}$$

Una característica distintiva de los SOMs es la capacidad de preservación de la topología. Cada nodo guarda una relación de vecindad con los nodos próximos de la grilla, de manera que al actualizarse el peso del NG, también se actualizan los pesos de los nodos vecinos. Esto hace que la relación de distancias entre los grupos de datos en el espacio original, se mantenga en el espacio proyectado de

la grilla. La forma de calcular la vecindad varía de un algoritmo a otro, pero generalmente el alcance de la vecindad se achica con el avance del algoritmo. La Ecuación 3.9 muestra una posible formulación para establecer la vecindad σ en el tiempo, donde σ_0 es el alcance inicial de la vecindad. La distancia entre dos nodos se calcula, generalmente, mediante la distancia euclídea, donde las coordenadas de un nodo corresponden a su índice dentro de la matriz.

$$\sigma(t) = \sigma_0 \cdot e^{-\frac{t}{\lambda}} \tag{3.9}$$

Cabe decir que la tasa de aprendizaje de los pesos de los nodos vecinos a un NG, varía en función de la distancia que se encuentre del mismo. Esto significa que dentro del área de vecindad, el cambio en el peso de un nodo es mayor cuanto más cerca se esté del NG. Con la incorporación del concepto de vecindad, la actualización de los pesos para cualquier nodo dentro de $\sigma(t)$, puede ser representado según la Ecuación 3.10, en donde θ representa el grado de influencia de la distancia del nodo al NG (Ecuación 3.11).

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \theta(t)\varphi(t)(\mathbf{x_i} - \mathbf{w}(t)) \tag{3.10}$$

$$\theta(t) = e^{-\frac{dist^2}{2\sigma^2(t)}} \tag{3.11}$$

El entrenamiento puede realizarse también en dos fases diferenciadas, denominada la primera de ellas como fase de ordenamiento y la segunda como fase de ajuste. La distinción principal entre una fase y la otra reside en el cambio de las funciones σ y φ. Lo que se busca es que en la fase de ajuste las actualizaciones de los pesos se vuelvan más sutiles, tanto en el NG como también en los pesos de los nodos vecinos.

3.4. Técnicas de búsquedas de proyecciones

Indicaremos en esta sección una técnica de búsqueda de proyecciones, denominada SARDUX. Este tipo de técnicas apuntan a aplicar una proyección de

los datos a un espacio de menor dimensionalidad, en donde se revelen importantes características en los datos. Hemos agrupado estas técnicas en una sección aparte, ya que las mismas pueden hacer uso de una estrategia supervisada o no supervisada. Por ejemplo, el análisis de componentes principales o los mapas auto-organizativos, son técnicas de proyección no supervisada, mientras que PLS corresponde a una técnica de proyección supervisada. En el caso de PCR (sección 3.2.1) la proyección en sí es no supervisada, pero luego se aplica un método supervisado para armar un modelo de predicción.

Por lo tanto, el término de búsqueda de proyecciones (PP - *projection pursuit*) lo usaremos para los procedimientos que buscan proyecciones interesantes desde espacios de alta dimensionalidad a espacios de baja dimensionalidad. La definición de "proyecciones interesantes" depende de la optimización de un determinado criterio. Si bien en las proyecciones se puede perder información, los espacios de alta dimensionalidad se caracterizan por estar mayormente vacíos y, por ende, díficiles para encontrar relaciones dentro de ellos. Richard Bellman fue uno de los pioneros en el estudio sistemático de problemas en el análisis de datos donde el cociente entre el número de observaciones y el número de variables es chico. El principal problema deriva del hecho que el volumen de un espacio matemático se vuelve exponencialmente más grande con el aumento lineal del número de variables. Esto implica que las observaciones quedan muy alejadas unas de otras y por lo tanto todos los puntos parecen *outliers*. En virtud de las dificultades que esta situación genera, en uno de sus trabajos (Bellman (1961)) se refirió a estas dificultades como "*la maldición de la dimensionalidad*" ("*the curse of dimensionality*"); frase que se sigue utilizando con frecuencia en la literatura actual.

3.4.1. Búsquedas de proyecciones usando matrices adaptivas - SARDUX

La presente técnica guarda una relación cercana al análisis de discriminantes lineal (LDA - *linear discriminant analysis*). LDA es la técnica más antigua y popular para la tarea de encontrar direcciones que mejor separan las clases dentro de un conjunto de datos multiclase (Fisher (1938)). La diferencia esencial

de SARDUX con LDA, reside en que en el primero las proyecciones se calculan mediante la optimización de matrices adaptivas. Esta característica le da más robustez al método que, a diferencia de LDA, no puede calcularse exactamente cuando el rango de la matriz de datos es menor a la cantidad de variables. Recomendamos al lector de esta tesis referirse a McLachlan (2004) para una lectura más profunda en LDA.

Las técnicas de escalado adaptivo supervisadas para ponderación de variables han hecho cierto progreso en la última década (Hammer *et al.* (2004), Sun (2007), Yang & Jin (2006)). En el contexto de la clasificación, las métricas adaptivas basadas en transformaciones matriciales representan técnicas recientes como las presentadas en Schneider *et al.* (2008) y Weinberger & Saul (2008).

La metodología SARDUX (detección de relevancia de atributos en forma supervisada usando comparaciones cruzadas, del inglés *supervised attribute relevance detection using cross-comparisons*) aplica distancias adaptivas para optimizar la separación entre datos agrupados en clases.

Sea un set de datos con c clases diferentes, se quiere obtener una dirección óptima λ en el conjunto de datos X, donde cada dato $\mathbf{x} \in \mathbb{R}^n$ y donde los datos agrupados por clase $(X^1, ..., X^c)$ contienen m_k muestras, con $k = 1...c$. Sea también $\mathbf{x_i^k}$ la i-ésima observación de los datos que pertenecen a la clase X^k. Esta dirección óptima se obtiene mediante la minimización de la Ecuación 3.12 (Strickert *et al.* (2008b)):

$$S_{d^v} = \frac{W(X|d^v)}{B(X|d^v)} \tag{3.12}$$

donde:

$$W(X|d^v) = \sum_{k=1}^{c} \sum_{i=1}^{m_k-1} \sum_{j=i+1}^{m_k} d^v(\mathbf{x_i^k}, \mathbf{x_j^k}|\lambda) \tag{3.13}$$

$$B(X|d^v) = \sum_{k_1=1}^{c-1} \sum_{k_2=k_1+1}^{c} \sum_{i=1}^{m_{k_1}} \sum_{j=1}^{m_{k_2}} d^v(\mathbf{x_i^k}, \mathbf{x_j^k}|\lambda) \tag{3.14}$$

Este función apunta a optimizar la distribución de los datos en donde se minimiza la función S_{d^v}. El numerador $W(X|d^v)$ describe la suma de las distancias entre las observaciones que pertenecen a una misma clase, mientras que $B(X|d^v)$ describe la suma de las distancias entre las observaciones que pertenecen a clases diferentes. Cabe notar que las distancias son evaluadas de acuerdo al vector de parámetros λ, en donde x_i e y_i corresponden aquí a la i-ésima variable del conjunto X.

$$d^v(x, y|\lambda) = (\sum_{i=1}^{n} \lambda_i \cdot (x_i - y_i))^2 = (\mathbf{x} - \mathbf{y})^\intercal \cdot \Lambda \cdot (\mathbf{x} - \mathbf{y}) \tag{3.15}$$

El producto vectorial del vector de parámetros $\lambda = (\lambda_1, ..., \lambda_n)$ define una matriz adaptiva de rango 1, tal que $\Lambda = \lambda \cdot \lambda^\intercal$. La Ecuación 3.15 expresada en forma matricial guarda una analogía con la distancia de Mahalanobis (De Maesschalck *et al.* (2002)). En términos geométricos la dirección λ es la dirección de proyección ortogonal de los datos en donde se minimiza la Ecuación 3.12.

Para la minimización de la función objetivo de la ecuación 3.12, considerando que la misma es una función derivable (Ecuación 3.16), lo más apropiado resulta usar un método de descenso directo basado en gradiente.

$$\frac{\partial d^v(x, y|\lambda)}{\partial \lambda_j} = 2 \cdot (\mathbf{x_j} - \mathbf{y_j}) \cdot \sum_{i=1}^{n} \lambda_i \cdot (\mathbf{x_i} - \mathbf{y_i}) \tag{3.16}$$

3.5. Técnicas evolutivas

Se conoce con el nombre de técnicas evolutivas a una familia de métodos computacionales de búsqueda u optimización estocásticos inspirados en la teoría de la evolución de Darwin. Aunque los orígenes en computación evolutiva pueden ser remontados hacia fines de los años 50 con el trabajos de Bremermann (1958), existe un consenso general en que los trabajos de algoritmos genéticos de Holland (1975), estrategias evolutivas de Rechenberg (1973) y la programación evolutiva de Fogel *et al.* (1966) sentaron las bases de esta disciplina tal como la conocemos hoy en día.

A mediados de los 80, con la aparición de computadoras con mayor poder de cómputo, cambia por completo el panorama de la Computación Evolutiva, dado que comienza a utilizarse estas técnicas para resolver con éxito ciertos tipos de problemas que antes no podían ser tratados eficientemente. En esta época surge el libro de Goldberg (1989), en donde se plantean y resuelven problemas de la vida real. Distintos autores fueron proponiendo diferentes nombres para sus algoritmos inspirados en la evolución natural de las especies, hasta que se consensuó llamar como algoritmos evolutivos a la generalización de los algoritmos genéticos clásicos de Holland.

Los principales beneficios de los técnicas evolutivas consisten en que son flexibles y adaptables a distintos problemas, en combinación con un desempeño robusto y características de búsqueda global. Básicamente, los AEs funcionan mediante la evolución de individuos pertenecientes a una población, los cuales son representados por un cromosoma. Cada cromosoma es una representación de una posible solución del problema que se quiere resolver con el AE. En otras palabras, un individuo representa un punto dentro del espacio de búsqueda del problema a resolver. Para realizar la búsqueda de la solución óptima, los individuos deben ser transformados y/o combinados en sucesivas iteraciones. Las operaciones clásicas de transformación y combinación, son las de mutación y cruzamiento respectivamente. Finalmente, estos individuos son sometidos a un procedimiento de selección basado en su aptitud o *fitness*[1], el cual está dado por una función que depende del problema. El propósito de este último paso es que, de acuerdo a los principios de la evolución, sólo los individuos más aptos deben sobrevivir.

Por lo tanto, un algoritmo evolutivo puede describirse mediante los siguientes pasos:

$t \leftarrow 0$

$P_t \leftarrow inicializar(\mu)$

$F(t) \leftarrow evaluar(P_t, \mu)$

Mientras $(\sim (P_t, \theta_l) \neq$ verdadero) **hacer**

$P'_t \leftarrow recombinar(P_t, \theta_r)$

$P''_t \leftarrow mutar(P'_t, \theta_m)$

[1] Usaremos el término fitness directamente en inglés debido a que su uso se ha generalizado incluso en la comunidad de computación evolutiva hispanoparlante.

$F(t) \leftarrow evaluar(P''_t, \lambda)$
$P_{t+1} \leftarrow seleccionar(P''_t, F(t), \mu, \theta_s)$
$t \leftarrow t + 1$
Fin-Mientras

donde P_t, P'_t, P''_t son poblaciones de individuos y $\mu, \lambda, \theta_l, \theta_r, \theta_m, \theta_s$ son parámetros o funciones de mutación, cruzamiento, evaluación o selección. Para un estudio minucioso, sobre las distintas variantes de estos métodos, recomendamos consultar Goldberg (1989).

3.5.1. Algoritmos evolutivos multi-objetivos

Es importante destacar, que un problema de optimización multi-objetivo plantea la optimización (minimización o maximización) de un conjunto de funciones, posiblemente en conflicto entre sí. Por lo tanto, la existencia de múltiples funciones objetivo plantea una diferencia fundamental con un problema mono-objetivo, ya que, en general no existirá una única solución al problema, sino un conjunto de soluciones que plantearán diferentes compromisos entre los valores de las funciones a optimizar. Esto hace que los problemas multi-objetivo requieran técnicas de optimización especializadas.

Independientemente de las técnicas implementadas, un concepto clave que suele utilizarse para encontrar el conjunto de soluciones en este tipo de problemas es el denominado óptimo de Pareto (Osborne & Rubinstein (1994)). Considerando el caso de un problema de minimización de funciones, un punto $\mathbf{x}^*$ es denominado óptimo de Pareto si para todo $\mathbf{x} \in \Omega$, donde Ω es la región factible del problema, se cumple que $f_i(\mathbf{x}) \geq f_i(\mathbf{x}^*)$, $\forall i$ con $1 \leq i \leq k$, siendo k la cantidad de objetivos a optimizar. Esto significa que no existe un vector factible que sea "mejor" que el óptimo de Pareto en alguna función objetivo sin que empeore los valores de alguna de las restantes funciones objetivo.

Asociada a la definición anterior, se introduce una relación de orden parcial denominada dominancia de Pareto entre vectores solución del problema de optimización multi-objetivo. Un vector $\mathbf{w} = (w_1, w_2, ..., w_k)$ domina a otro $\mathbf{v} = (v_1, v_2, ..., v_k)$ si $w_i \leq v_i$, $\forall i$, con $1 \leq i \leq k$ y $\exists i$ tal que $w_i < v_i$. Esta relación

se nota como $\mathbf{w} \prec \mathbf{v}$. De esta definición de dominancia se desprenden muchos conceptos útiles, como por ejemplo el término *frente* que nuclea todas las soluciones en donde ninguna solución es dominada por ninguna otra solución del mismo frente. Luego, el frente de Pareto, es el conjunto de soluciones óptimas dentro del espacio Ω, y el frente de no dominados contiene las soluciones encontradas que no son dominadas por ninguna otra solución.

La posibilidad de exploración paralela de múltiples soluciones convierten a los algoritmos evolutivos en una técnica muy conveniente para abordar problemas multi-objetivos. Según Coello Coello *et al.* (2007) pueden considerarse, en general, dos tipos principales de algoritmos evolutivos multi-objetivo: los algoritmos que no incorporan el concepto de óptimo de Pareto en el mecanismo de selección y los que jerarquizan a la población de acuerdo a si un individuo es dominado o no de acuerdo al orden parcial definido por la dominancia de Pareto. Dentro de las técnicas no basadas en Pareto destacaremos las funciones de agregación, mientras que entre las basadas en Pareto describiremos a dos de los algoritmos más populares: el NSGA-II y el SPEA2.

3.5.1.1. Técnicas basadas en agregación

Estas técnicas consisten en integrar todos los objetivos en uno solo, mediante suma, multiplicación o cualquier otra combinación de operaciones aritméticas. Por ejemplo, la Ecuación 3.17 es un ejemplo de función de agregación en donde todos los objetivos son ponderados de acuerdo a las constantes c_i. Una vez aplicada la agregación, a los fines prácticos, el problema se transforma en un problema mono-objetivo.

$$min \sum_{i=1}^{k} c_i f_i(\mathbf{x}) \tag{3.17}$$

Entre las principales ventajas de esta técnica se puede mencionar que resulta muy simple implementarla, y su ejecución es muy eficiente. Como desventaja podemos encontrar que, en general, las combinaciones lineales de pesos no suelen funcionar bien en los casos en que el frente de Pareto es cóncavo, más allá de los pesos utilizados. Otro problema de estas técnicas reside en la dificultad de

balancear correctamente la importancia dada a cada objetivo, acción que se vuelve aún más difícil cuantos más objetivos el problema tenga.

3.5.1.2. *Non Dominated Sorting Genetic Algorithm - II* (NSGA-II)

El algoritmo de NSGA-II (Deb *et al.* (2002)) comienza creando una población aleatorias de padres P_0 de tamaño s. Esta población es ordenada basándose en el concepto de no dominancia. A cada individuo se le asigna una puntuación igual a su nivel de no dominancia, esto es, 1 si pertenece al primer frente, 2 si pertenece al segundo frente y así sucesivamente. Luego de puntuar las soluciones, una población Q_0 de s hijos es creada usando selección por torneo, recombinación y mutación. La i-ésima generación sigue tres pasos bien diferenciados. Primero, se genera una población combinada $R_i = P_i \bigcup Q_i$ de tamaño $2s$. Segundo, R_i es ordenada de acuerdo a su no dominancia. Nótese que como todos los padres e hijos están presentes en R_i el elitismo está asegurado. Claramente, los individuos que pertenecen al primer frente, esto es $D_1(R_i)$, son las mejores soluciones de la población combinada R_i. Finalmente, si $|D_1(R_i)| \leq s$, entonces $D_1(R_i) \subseteq P_{i+1}$, esto es, todos los individuos del primer frente de R_i estarán en la población siguiente. Los individuos que faltan para completar el tamaño s en la nueva población, son llenados con los individuos de los frentes subsiguientes en el orden marcado por la puntuación. Si $D_j(R_i)$ fuera el último frente del cual se sacaron individuos, pero no todos sus individuos pudieron ser acomodados, entonces se debe aplicar un criterio de selección de individuos dentro del mismo frente. NSGA-II usa un criterio de selección que favorece las soluciones ubicadas en regiones menos pobladas del frente en el espacio de los objetivos.

3.5.1.3. *Strength Pareto Evolutionary Algorithm 2* (SPEA2)

El algoritmo de SPEA2 (Zitzler *et al.* (2002)) comienza con una población P_0 de tamaño s y una población vacía $\bar{P}_0$ la cual tiene una capacidad máxima $\bar{s}$. La i-ésima generación repite cuatro pasos básicos. En primera instancia, $\bar{P}_i = \bar{P}_i \bigcup D_1(P_i)$, esto es, el conjunto de no dominados de P_i es calculado y copiado a $\bar{P}_i$. En segundo término, todas las soluciones dominadas de $\bar{P}_i$ son descartadas. Si $|\bar{P}_i| > \bar{s}$, es decir, si la población externa es sobrepasada en su

capacidad máxima, $\bar{P}_i$ es podado mediante un algoritmo de análisis de agrupamiento aplicado en el espacio de los objetivos, el cual se realiza a partir de la elección de "representantes". El "representante" es elegido como el individuo con la menor distancia a otros dentro de un mismo frente. En el tercer paso, se calcula el valor de *fitness* de cada individuo dentro de $P_i \bigcup \bar{P}_i$. Luego, sobre la unión de estos conjuntos se seleccionan individuos mediante selección por torneo. Finalmente, se crea una población P_{i+1} de s hijos aplicando recombinación y mutación comúnmente.

Capítulo 4

Estado del arte en QSPR-QSAR

4.1. Quimioinformática

En el Capítulo 2 se ha explicado la necesidad de analizar los efectos de moléculas candidatas sobre un determinado receptor, tanto en cuanto a su vinculación como así también a los mecanismos biológicos que esta vinculación genera. Asimismo se ha reseñado la aparición de modernas tecnologías que permitieron la rápida síntesis química y el testeo a gran velocidad de grandes colecciones de compuestos. Estas tecnologías modernas dieron inicio a un crecimiento importante de bases de datos con información heterogénea proveniente de diversos experimentos aplicados a compuestos químicos.

El crecimiento en el volumen de datos se vio beneficiado por el desarrollo vertiginoso de tecnologías de la información y las comunicaciones, en donde estos grandes repositorios de datos pueden ser accedidos, actualizados y administrados con facilidades que hace 20 años atrás no hubieran sido posibles. Además esta disponibilidad de grandes cantidades de datos, propició el uso de técnicas de aprendizaje automático, minería de datos y herramientas estadísticas para descubrir nuevos patrones y estructuras que sirvan para ganar en conocimiento sobre las relaciones entre la estructura química de un compuesto y sus propiedades físicas y biológicas. Todo este desarrollo tecnológico y metodológico propició la realización de ensayos *in silico* para identificar y priorizar compuestos candidatos a drogas medicinales, para que posteriormente sean validados experimentalmente.

Existen muchas definiciones sobre el concepto de quimionformática (*chemo-informatics* o *cheminformatics*). Sin embargo, la definición de quimioinformática más general y aceptada es la siguiente:

> La quimioinformática es la disciplina científica que estudia el diseño, creación, organización, administración, recuperación, análisis, visualización y uso de información de origen químico (Paris (1999)).

La quimioinformática engloba múltiples problemáticas relacionadas con la química, sin embargo los desarrollos más importantes se han aplicado para el descubrimiento y desarrollo de drogas. Incluso dentro de las aplicaciones farmacéuticas, existe una miríada de áreas sobre el cual se desarrollan diferentes investigaciones de quimioinformática. Diferentes aplicaciones del uso de métodos informáticos en la industria farmacéutica son descriptas en los trabajos de Leach & Gillet (2007) y Varnek & Tropsha (2008).

Por cuestiones prácticas, el contenido de este capítulo se centró sólo en la revisión bibliográfica de trabajos científicos y comerciales que abarquen el uso de enfoques de aprendizaje automático para el desarrollo de modelos cuantitativos de relación estructura-propiedad, comúnmente abreviado como QSPR. Esta revisión pretende incluir, no de forma exhaustiva, las principales temáticas relacionadas con las investigaciones realizadas en el marco de esta tesis.

4.2. Modelos cuantitativos de relación estructura-propiedad o estructura-actividad

Los modelos cuantitativos de relación estructura-propiedad (*quantitative structure-property relationship* - QSPR) establecen relaciones entre las características estructurales de un compuesto químico y una dada propiedad fisicoquímica o biológica. Comúnmente, QSPR y QSAR (*quantitative structure-activity relationship*) se usan como sinónimos, ya que si bien los métodos basados en QSAR relacionan estructura química con actividad farmacológica, la esencia de los métodos es prácticamente la misma. En el transcurso de esta tesis haremos explícita la diferencia de los términos cuando sea necesario, de lo contrario tanto uno como el otro serán usados en forma intercambiable.

En la actualidad los métodos basados en QSPR no solamente relacionan una propiedad a partir de características meramente estructurales, sino que esta relación puede obtenerse a partir de cualquier descriptor, es decir cualquier información numérica sobre un aspecto de una molécula (ver Sección 4.3). En términos matemáticos, un modelo QSPR se plantea como una función $\tilde{y}_i = f(\mathsf{x}_i)$, donde $\mathsf{x}_i = (x_1, x_2, ..., x_d)$ es uno de los m compuestos químicos representado como un vector de d descriptores, $\tilde{y}_i$ es el valor calculado por f, e y_i es una propiedad experimental de x_i. El objetivo es encontrar una función f en donde se minimice $|\tilde{y}_i - y_i|$ para todo i.

En líneas generales, la idea del método consiste en partir de una base de datos con compuestos químicos, en donde se tengan calculados un conjunto de descriptores y para cada uno de los compuestos se tenga información experimental de la propiedad fisicoquímica o biológica que se desea modelar. A partir de este conjunto de datos de referencia o de entrenamiento, se construye la función f. Una vez obtenida esta función se la aplica a compuestos no contemplados en el entrenamiento y sobre los cuales puede desconocerse el valor experimental. De esta manera, se puede predecir *in silico* el valor de una propiedad, a partir del análisis de información de otras experimentaciones. Vale destacar que, en la práctica, el proceso de modelado es más complejo y deben aplicarse ciertos criterios de selección y división de los datos en forma previa a la predicción.

Básicamente, hay cuatro elementos que conforman el desarrollo de los modelos QSPR: los descriptores, el método con el que se construye la función f, la propiedad que se desea modelar y la validación que se realiza sobre el modelo. En las secciones siguientes intentaremos presentar una revisión de las principales técnicas existentes en la literatura, de manera de darle al lector de este manuscrito una visión del estado del arte de los modelos de QSPR.

4.3. Descriptores moleculares

Existe una gran complejidad en la naturaleza de las moléculas y las interacciones entre los átomos. En pos de proveer una base para el análisis y la comunicación de conceptos químicos, se desarrollaron distintos *descriptores moleculares* los cuales brindan información sobre alguna característica estructural de la molécula.

La cantidad de aspectos o características que pueden ser considerados sobre una misma molécula es muy grande, motivo por el cual existen una gran cantidad de descriptores, cada uno de ellos ideados con un determinado fin. Dos grandes aplicaciones en donde los descriptores han sido aplicados, además de QSAR, es en la búsqueda por subestructuras dentro de una base de datos y para ser usados como atributos en cálculos de similaridad entre moléculas (Brown & Martin (1996) y Scsibrany *et al.* (2003)).

A lo largo de este desarrollo en la teoría sobre descriptores moleculares se han identificado ciertas características que son deseables que los descriptores posean. Una de ellos es que a diferentes estructuras, correspondan distintos valores de descriptores, incluyendo la capacidad de diferenciar isómeros[1]. Otra característica importante es que no hayan sido definidos a partir de experimentaciones o modelos cuya exactitud sea cuestionada. Esto genera modelos ambiguos y que atentan contra uno de los principios que se enunciarán en la sección 4.5. Esta variabilidad no deseable de los resultados de predicción en función del método de cálculo de los descriptores es descripta en los artículos de Papa *et al.* (2005) y Gramatica (2007). Siguiendo con las características deseables de los descriptores, es importante que sean interpretables en lo que su significado representa (Cronin & Schultz (2003)). La velocidad de cálculo también se incluye como un factor deseable, aunque hoy en día no representa éste un criterio tan importante como los anteriores, pues para cualquier entidad química este cálculo se realiza una sola vez y puede ser almacenado.

Hay muchas formas de establecer taxonomías entre los distintos descriptores. Una de ellas, es separar los descriptores en locales y globales, en donde los primeros hacen referencia a una parte de la molécula (*e.g.* los átomos exteriores de un compuesto) mientras que los últimos toman en consideración toda la molécula (*e.g.* cantidad de átomos de hidrógeno en la molécula). Otra forma de discriminar las distintas familias es en base al método usado para calcular el valor: observación directa a partir de su estructura, cálculo a partir de propiedades físicas, cálculo a partir de otros descriptores u obtención en forma experimental. Una forma muy

[1]En química, los isómeros son moléculas con la misma fórmula química y el mismo tipo de enlaces entre átomos, pero en el que los átomos están dispuestos de diferente forma, por ejemplo con una diferencia en los ángulos que forman los distintos enlaces.

común de agrupar a las distintas familias es a partir de su dimensionalidad, es decir qué tipo de representación geométrica de la estructura se hace para obtener información del compuesto. Este será el enfoque que seguiremos en nuestra taxonomía:

0-dimensionales (0D): En este grupo se encuentran los descriptores de tipo constitucional, como por ejemplo el peso molecular o también la presencia o número de átomos o enlaces de determinado tipo. Estos descriptores son simples de entender, rápidos de calcular y muy útiles en muchas aplicaciones (ver Sección 2.7 Regla de Lipinski). La desventaja que presentan estos descriptores es que no son capaces de diferenciar isómeros o quiralidad.

Unidimensionales (1D): Estos descriptores tienen mucha similitud con la clase anterior. La idea aquí es que a partir de una codificación de la molécula como un string de caracteres (código SMILES, Weininger (1988)), puede distinguirse la presencia de fragmentos o grupos funcionales con sus distintos tipos de interacciones, a diferencia del caso anterior en donde sólo se distingue la presencia de átomos.

Bidimensionales (2D): En este caso se construyen tablas de conexión en donde se mantiene información 2D, sobre cómo es la conformación y conectividad de los distintos fragmentos o átomos de la molécula. A partir de esta información, se construyen matrices o grafos a los que se le aplican diferentes cálculos para obtener información sobre los distintos índices de conectividad, forma y/o flexibilidad. Entre los índices más comunes destacamos, el índice de conectividad de Kier y Hall (Kier & Hall (1976)), el índice Wiener (Wiener (1947)), el índice Zagreb (Gutman *et al.* (1975)) y el índice Randic (Randic (1975)). Las ventajas de estos descriptores es que son rápidos de calcular, ya que no necesitan de la estructura tridimensional, y además proporcionan un mejor conocimiento de la estructura, diferenciando isómeros por ejemplo. La desventaja es que hay muchos métodos distintos y en buena parte de ellos su interpretación no es sencilla.

Tridimensionales (3D): Estos descriptores necesitan de la optimización de las posiciones de los átomos para obtener la información de su conformación

en 3 dimensiones. Esta optimización busca conformaciones geométricas en donde la energía de interacción sea un mínimo local y por lo tanto, represente una conformación estable y posible. A partir de esta optimización pueden derivarse distintas clases de descriptores entre los que se distinguen los de superficie, los volumétricos y los geométricos (Fontaine (2005), Todeschini & Gramatica (2006)). Los descriptores 3D permiten obtener información sobre los ángulos y las distancias geométricas entre los átomos. Requieren de tiempo para optimizar la molécula y su comprensibilidad depende del tipo específico del descriptor.

Cuatridimensionales (4D): Estos descriptores no sólo representan los aspectos 3D de la molécula sino que también pueden representar la disposición de los campos de fuerza responsables de las distintas interacciones. Los distintos tipos de interacciones consideradas son las interacciones estéricas, electroestáticas e hidrofóbicas. El artículo pionero en el el uso de modelos QSAR-4D fue el de Hopfinger *et al.* (1997).

En la actualidad existen distintos programas de computación que permiten el cálculo de descriptores asociados a un conjunto de compuestos químicos, donde el más popular de todos ellos es DRAGON (2003). Hace 10 años atrás en Todeschini & Consonni (2000) había unos 3000 descriptores y en la actualidad se calcula que ese número supera los 8000. En el sitio web de Dragon (Talete (2007)) se puede acceder a la lista de los descriptores de Dragon, la cual contiene prácticamente todos los descriptores utilizados en el transcurso de esta tesis. En caso que se utilicen descriptores que no figuran en esta lista, se indicará explícitamente la fuente.

Para una lectura más profunda en el tema de descriptores moleculares el lector puede recurrir a los siguientes artículos: Livingstone (2000) y Downs (2004).

4.4. Revisión de propiedades y modelos basados en QSAR/QSPR

En esta sección intentaremos abordar los diferentes enfoques de modelado basado en QSAR que se encuentran publicados en la literatura. Sería prácticamente

inviable citar todos los trabajos en el tema, por lo que únicamente enunciaremos algunos trabajos de cada aspecto del modelado que guarde alguna relación con los contenidos de esta tesis.

4.4.1. Predicción del coeficiente de partición octanol-agua: logP

La hidrofobicidad posee un papel fundamental dado que su determinación influye en buena parte de los diferentes procesos farmacocinéticos. Sin ir más lejos, las otras dos propiedades experimentales que atacaremos en esta tesis también guardan relación con el valor de hidrofobicidad.

La predicción de propiedades basadas en el enfoque QSPR comenzó hace poco más de 40 años atrás con los trabajos de Fujita *et al.* (1964) y Hansch & Fujita (1964) con modelos para el cálculo de la hidrofobicidad. Sin embargo, en aquel momento el espíritu de QSAR era muy distinto, ya que no se contaba con el desarrollo que hoy existe en los métodos estadísticos de predicción, y donde además, sólo se usaban unos pocos descriptores. De aquel primer momento a la actualidad, muchos enfoques diferentes para atacar el mismo problema han surgido, los cuales agruparemos según la taxonomía propuesta en Erös *et al.* (2002). Esta taxonomía, si bien fue definida para la predicción de la hidrofobicidad es igualmente aplicable a la predicción de otras propiedades:

Método de sustituto π: Calcula el valor de una propiedad mediante el reemplazo de un átomo de hidrógeno de un compuesto padre cuyo valor experimental es conocido, por un sustituto π cuyo valor también es conocido. Asume que la propiedad tiene un comportamiento "aditivo" (Fujita *et al.* (1964)), por lo que el nuevo valor se calcula con una suma aritmética de lo que esa sustitución involucra.

Métodos basados en fragmentos: A partir de una gran base de datos de propiedades experimentales se determinan las contribuciones promedios de cada uno de los fragmentos de la molécula. Luego el valor experimental para

un nuevo compuesto se puede calcular haciendo la suma de las contribuciones de los fragmentos y ciertos factores de corrección (Rekker & Manhold (1992)).

Métodos basados en contribuciones atómicas y/o area superficial: Esta categoría es similar a la anterior, aunque usa fragmentos atómicos y datos de área superficial en lugar de fragmentos de grupos químicos (Iwase *et al.* (1985)).

Métodos basados en propiedades moleculares: Estos métodos estiman un valor experimental como función de diferentes propiedades moleculares (descriptores). Uno de los primeros trabajos que siguió este paradigma es el de Bodor *et al.* (1989).

La última clase corresponde al enfoque dado en nuestras investigaciones. Una gran cantidad de trabajos científicos han desarrollado y/o estudiado modelos basados también en este enfoque para el cálculo de logP. Entre algunos trabajos citamos: Duprat *et al.* (1998), Erös *et al.* (2002), Tiño *et al.* (2004), Liao *et al.* (2006), Ghasemi & Saaidpour (2007) y Yang *et al.* (2009).

4.4.1.1. Softwares comerciales y académicos

Existen muchos softwares comerciales y académicos que permiten calcular logP, algunos de los cuales se muestran en la Tabla 4.1. Una comparación entre todos estos desarrollos es muy difícil de realizar en forma justa, ya que cada uno de ellos parte de distintos sets de entrenamiento, con tamaños y diversidad muy distintos entre sí. Por ejemplo, aquél con el set de entrenamiento más grande tiene más chances de salir favorecido, como también tendrá más chances de ganar la comparación aquél que se evalúe con los compuestos más similares a los que fueron considerados en su entrenamiento.

De todos modos, distintos *benchmarks* realizados (Martin (2007) y Mannhold *et al.* (2008)) coinciden en afirmar que el problema no está resuelto y que existen ciertos factores moleculares que afectan la calidad de la predicción. Entre otros factores se enuncian el tamaño molecular (medido en número de átomos no-hidrógenos) y la ionizabilidad de los compuestos. En Mannhold *et al.* (2008) se

concluye que la mayoría de los métodos muestra una precisión baja, en especial al ser testeados con un conjunto de 95809 compuestos registrados por el laboratorio Pfizer.

4.4.2. Predicción de la penetración en la barrera hemato-encefálica: logBB

A pesar de la importancia en la predicción del grado de penetración en la barrera hemato-encefálica (seccion 2.7.2), los primeros trabajos sobre el modelado de esta propiedad *in silico* empezaron a partir del año 2001, de la mano de Platts y Hou, en los trabajos Platts *et al.* (2001) Hou & Xu (2002). De estos dos trabajos, podemos identificar dos tipos de modelos de predicción de logBB: los basados en los descriptores LFER (*linear free-energy relationship*) y aquellos que no usan estos descriptores. Estos descriptores LFER parecen tener una estrecha relación con la propiedad logBB. Sin embargo una de las críticas que reciben estos descriptores es su dificultad de cómputo para todo tipo de compuesto candidato con diferentes estructuras. En Konovalov *et al.* (2007) se propone una comparación de ambos enfoques usando como representantes los trabajos de Abraham *et al.* (2006) y Narayanan & Gunturi (2005).

Otros trabajos relacionados al modelado de esta propiedad pueden encontrarse en: Hou & Xu (2003), Garg & Verma (2006), Guerra *et al.* (2008), Konovalov *et al.* (2008).

4.4.3. Predicción de la absorción intestinal humana: logHIA

Hemos visto en la sección 2.3 que la administración oral es usualmente el modo más deseable de administrar una droga, y por lo tanto el efecto terapéutico queda supeditado al transporte o absorción al torrente sanguíneo. La primera barrera contra la biodisponibilidad es la absorción intestinal humana (HIA). Desarrollar modelos computacionales para esta propiedad es de alto interés dado que los métodos *in vitro* disponibles son costosos o, en su defecto, imprecisos.

Podemos identificar dos tipos de modelos. Los primeros en orden cronológico, generalmente, consideraban un sólo factor, el cual era el grado de hidrofobicidad

medido según la propiedad logP (Camenisch *et al.* (1998), Testa *et al.* (1996)). Si bien se obtenían predicciones interesantes con compuestos homólogos, la generalización se veía afectada al modificarse la diversidad estructural. Luego, los trabajos subsiguientes comenzaron a incorporar otros descriptores en el modelo tales como el peso molecular, tamaño, forma y/o área de superficie de van der Waals polar. Entre ellos podemos destacar los trabajos de: Polley *et al.* (2005), Gunturi & Narayanan (2007), Hou *et al.* (2007) y Jung *et al.* (2007).

4.4.4. Métodos de predicción usados

Los primeros enfoques de modelos basados en QSAR eran aplicados usando técnicas convencionales tales como regresión lineal múltiple - RL - (Topliss (1979), Cronin & Schultz (2003)) y cuadrados mínimos parciales - PLS - (Dunn *et al.* (1984), Eriksson *et al.* (2003)). Sin embargo, estos modelos, aunque simples e interpretables, fallan en el modelado de ciertas relaciones estructura-propiedad, dado que las mismas suelen tener una vinculación no lineal (Tiño *et al.* (2004)).

Este escenario fue el marco propicio para empezar a aplicar modelos de aprendizaje automático en el dominio químico. Una gran cantidad de trabajos se han publicado y sólo haremos hincapié en el uso de algunas técnicas. Una revisión más completa en el tema puede encontrarse en Goldman & Walters (2006).

Entre todas las técnicas presentes destacamos primero a las redes neuronales, las cuales corresponden a los métodos de aprendizaje automático más utilizados en el área de quimioinformática. Desde el trabajo de Aoyama *et al.* (1989) -el cual es quizás el primer artículo que sugiere el uso de redes neuronales para QSPR- hasta la actualidad, las redes neuronales han sido usadas en forma constante. Winkler (2004) y Taskinen & Yliruusi (2003) contienen dos excelentes revisiones de las principales contribuciones de las redes neuronales para la predicción de propiedades.

Creemos importante hacer una mención especial a las redes neuronales con regularización bayesiana (sección 3.2.5.1), que fueron por primera vez usadas en el dominio químico en el trabajo de Burden & Winkler (1999). En este trabajo se demuestran propiedades interesantes de este tipo de red, la cual coincide con nuestra experimentación realizada, en cuanto a que se trata de una de las redes

que mejor se comporta en situaciones de no linealidad evitando el sobreajuste. Las mismas permiten entrenar el modelo sin hacer uso de datos de validación, y tampoco requieren de un ajuste del número óptimo de nodos. Además, como en la optimización se favorece la selección de modelos menos complejos, las funciones suelen tener una variabilidad baja, independientemente de los pesos iniciales usados. Estos son algunos de los trabajos en los que se usa este tipo de regularización en las redes para QSPR: Winkler & Burden (2004), Polley *et al.* (2005), Bruneau & McElroy (2006)

También destacamos a los árboles de decisión (sección 3.2.2) como una de las técnicas de interés, dado que permiten una ponderación automática de las variables, son aptos para grandes sets de datos y permiten una interpretación transparente (Izrailev & Agrafiotis (2001), Tong *et al.* (2003)). Su principal desventaja es que su precisión para modelos de regresión es inferior en comparación con otros modelos, como por ejemplo las redes neuronales. Algo similar ocurre con el método de k-vecinos más cercanos (sección 3.2.3) con respecto a la precisión en la predicción. Sin embargo, se tiene a favor que el espíritu del método captura la hipótesis básica de QSAR, esto es, que compuestos molecularmente similares se comporten en forma similar (Hoffman *et al.* (1999), Zheng & Tropsha (2000)). En este tipo de método resulta central la elección de los descriptores y la medida de distancia aplicada.

En base a nuestra revisión y experimentación, también destacamos fuertemente el uso de comités de modelos de aprendizaje (sección 3.2.5.2). Los comités de modelos han sido usados en numerosas propuestas en quimioinformática, dadas sus ventajas en la reducción de sesgo y varianza de las predicciones, como por ejemplo en: Agrafiotis *et al.* (2002), Svetnik *et al.* (2003), Arodź *et al.* (2006).

Finalmente, mencionamos también que, aunque con menor frecuencia que los modelos de regresión, se encuentran en la literatura trabajos de clasificación. La motivación es la misma, sólo que la variable a modelar suele tener categorías del tipo 'alta', 'baja' e incluso también 'media'. Por ejemplo el trabajo de Clark (1999) caracteriza la absorción intestinal humana en tres categorías: 'buena', 'media' y 'pobre'. El principal problema con este enfoque reside en que las categorías no existen naturalmente, y por lo tanto el criterio humano de lo que una 'buena'

absorción significa, depende del umbral definido, el cual a su vez puede depender del científico y/o del contexto de aplicación del compuesto.

Entre las técnicas más comunes de clasificación encontradas destacamos al análisis de discriminantes lineal, el cual se ha empleado por ejemplo en Dutta *et al.* (2007) y Molina *et al.* (2004). Con el crecimiento de las técnicas de aprendizaje automático surgieron trabajos que aplican clasificación usando, generalmente, comités de modelos. Entre otras metodologías podemos encontrar que se aplican: árboles de decisión (Simmons *et al.* (2008)), k-vecinos más cercanos (Mattioni & Jurs (2002)) o redes neuronales (Manallack *et al.* (2003)).

4.4.5. Metodologías para selección de descriptores

Como se enunció en la sección 4.3 la cantidad de descriptores disponible es muy grande. Una de las mayores dificultades en cualquier aplicación de quimioinformática es encontrar el conjunto de descriptores apropiado para cada modelo, dado que no existe un consenso sobre cuáles son los descriptores más influyentes para cada propiedad. Este tema será abordado ampliamente en el capítulo 5 y en donde se explicará claramente la problemática y las dificultades de esta tarea. Enunciaremos aquí solamente algunas de las propuestas en la literatura que apuntan a resolver este problema.

Los primeros trabajos en el área sólo apuntaban a eliminar variables redundantes o correlacionadas (Whitley *et al.* (2000)), y si bien esta eliminación es necesaria e importante, según se explica en Liu & Motoda (2008) no resulta suficiente. Un amplio número de trabajos han investigado diferentes enfoques para seleccionar descriptores en QSAR mediante distintas alternativas: (Burden & Winkler (2009); Dutta *et al.* (2007); Fröhlich *et al.* (2004); Godden & Bajorath (2003); Konovalov *et al.* (2008); Liu (2004)).

Existen enfoques basados en optimizaciones del tipo *greedy* o *stepwise* las cuales son deterministas y generalmente poseen una buena performance en tiempo de ejecución (Fröhlich *et al.* (2004), Figueiredo (2003)). Sin embargo, estas técnicas poseen la desventaja que pueden obtener subconjuntos que representan sólo óptimos locales del problema. Asimismo, su determinismo hace que se produzca una tendencia hacia ciertas elecciones en situaciones en donde puede que no

haya sólo una única mejor opción. En el trabajo de Horvath *et al.* (2007) existe un interesante debate acerca de las diferencias en aplicar enfoques estocásticos y deterministas.

Asimismo, algunas propuestas en la literatura utilizan algoritmos evolutivos (Lin *et al.* (2006); Nicolotti & Carotti (2006); Shen *et al.* (2008)) dado que éstos permiten realizar una búsqueda estocástica y paralela de los posibles subconjuntos de descriptores. En este sentido, estas técnicas son menos propensas a converger en óptimos locales. Algunos trabajos, como es el caso de So & Karplus (1996) aplican algoritmos evolutivos, pero utilizan métodos de evaluación tan intensivos computacionalmente, que su utilización sólo puede restringirse a la selección de un pequeño subconjunto de descriptores.

4.4.6. Metodologías para chequeo de dominio de aplicación

Es importante enfatizar que no importa cuan robusta y significativamente un modelo QSPR haya sido validado, no se puede pretender que un mismo modelo de predicción sea confiable para cualquier compuesto químico posible (Tetko *et al.* (2006), Tropsha *et al.* (2003)). Para tener una idea de la inmensidad del espacio químico combinatorial, el número de compuestos químicos posibles que se pueden dar usando 10 substitutos en los cuatro lugares libres de un anillo bencénico desustituido asimétricamente es aproximadamente 10000.

Por lo tanto, antes que un modelo QSAR sea puesto a disposición para la selección virtual de compuestos candidatos a drogas, se debe definir para qué compuestos el modelo va a ser preciso y para cuáles no, esto es, se debe determinar su dominio de aplicación, y solamente los compuestos que estén dentro de este dominio podrían ser evaluados en forma confiable.

Veremos en los principios de la OECD descriptos en la sección 4.5, que la definición del dominio de aplicación debería acompañar la especificación de un método. Desde nuestro punto de vista, y en coincidencia con estos principios, consideramos fundamental la definición del dominio de aplicación de un método. Si esto no fuera así, la calidad de predicción de un compuesto no usado en el entrenamiento o validación de un modelo será una incógnita. Sin embargo, la

definición del dominio de aplicación no está presente en la mayoría de los métodos presentados en la literatura científica y mucho menos en el ámbito comercial.

Las razones de esta carencia se da por dos motivos fundamentales. El primero de ellos es que, aún cuando claramente importante, resulta un concepto nuevo del cual se empezó a considerar someramente en los artículos de quimioinformática recién en el año 2003 (Eriksson *et al.* (2003), Nikolova-Jeliazkova & Jaworska (2003), Jaworska *et al.* (2003)) y con más profundidad a partir del año 2005 (Netzeva *et al.* (2005), Nikolova-Jeliazkova & Jaworska (2005), Jaworska *et al.* (2005)). El otro motivo reside en la dificultad de la determinación sobre qué es un dominio de aplicación, y al mismo tiempo cómo deber ser medido.

Según Netzeva *et al.* (2005) se define el dominio de aplicación como "*el alcance y las limitaciones de un modelo*". Comúnmente, se asocia el concepto de dominio de aplicación con una medida de similaridad hacia el conjunto de entrenamiento. Sin embargo, el concepto de similaridad tiene aristas subjetivas y en donde, al momento de aplicarse, surgen diferencias de acuerdo a la métrica usada. Haremos a continuación una revisión de los principales enfoques que han sido propuestos para el cálculo de dominio de aplicación siguiendo la taxonomía propuesta en Schroeter *et al.* (2007).

Métodos basados en rango: revisan si los descriptores de los compuestos del testeo exceden el rango de los descriptores usados para el entrenamiento (Sheridan *et al.* (2004), Tropsha (2006)). Esta técnica parte de que cada conjunto de entrenamiento define un hiper-cubo en el espacio de los descriptores y se analiza si el nuevo compuesto a predecir está dentro o fuera de este hipercubo. Existen otros esquemas similares más apropiados que generan el volumen convexo mínimo que encierra todos los datos del entrenamiento (Dimitrov *et al.* (2003)). Es importante notar de estos enfoques, que con el aumento lineal de la cantidad de descriptores, el volumen considerado aumenta exponencialmente. Además, los espacios vacíos que quedan dentro del volumen no son detectados.

Métodos basados en distancias: también conocidos como distancias de extrapolación, se han propuesto este tipo de enfoques en diferentes artículos: Bruneau & McElroy (2006), Sheridan *et al.* (2004), Eriksson *et al.* (2003),

Tropsha *et al.* (2003), Tetko *et al.* (2006). En este caso, se utilizan distintas medidas de distancia (Euclídea, Mahalanobis/leverage) para medir la distancia desde uno de los compuestos del testeo: al compuesto del entrenamiento más cercano, a un radio de distancia (contando cuántos compuestos quedan dentro) o a todo el conjunto de entrenamiento.

Métodos basados en distribuciones de densidad de probabilidad: en este caso el objetivo pasa por estimar la probabilidad de densidad de los datos de entrenamiento (Netzeva *et al.* (2005)). Luego, las predicciones de compuestos que pertenecen a áreas de alta densidad, son propensas a ser más precisas. La ventaja principal de esta técnica es que permite identificar áreas vacías dentro del volumen de los datos. Sin embargo, la estimación de la probabilidad de distribución de densidad no es una tarea sencilla. Esta tarea se vuelve más compleja con el aumento del número de descriptores. Por tales motivos, esta clase de métodos no ha sido aún abordada extensamente para QSAR.

Métodos basados en comités: se basa en la idea de usar un conjunto de métodos diferentes para comparar las predicciones obtenidas por cada uno de ellos. Así, los modelos tenderán a acordar en sus predicciones en áreas bien representadas y se diferenciarán cuando predigan sobre áreas que no fueron contempladas en el entrenamiento. Estas técnicas han sido usadas en distintos trabajos (Manallack *et al.* (2003), Göller *et al.* (2006), Bruneau & McElroy (2006)), aun cuando no siempre se hiciera mención de ellas como técnica de definición de dominio de aplicación.

Librerías de datos: Kühne *et al.* (2006) proponen usar una estrategia de selección de modelos, en donde se elige el que sea más propenso a tener el resultado más confiable. Para esto utiliza una librería de datos en donde se tiene información de nuevas mediciones experimentales de compuestos. De esta forma, al tener que predecir un nuevo compuesto, se busca en la librería cuáles son los compuestos más similares y la elección del modelo se elige de acuerdo a cuál es el que tiene las predicción más cercana a lo que la librería indica.

Métodos Bayesianos: las técnicas que utilizan modelos bayesianos de predicción (Schroeter *et al.* (2007), Fechner *et al.* (2010)) consideran que los parámetros provienen de distribuciones probabilísticas y por ende las salidas son también distribuciones probabilísticas. De esta forma, el grado de incertidumbre de la salida queda automáticamente reflejada en su función de distribución. La dificultad de estas técnicas reside en la definición inicial de las probabilidades.

4.4.7. Metodologías de proyección de datos a subespacios

En la sección 3.4.1 hemos visto un esquema para obtener proyecciones que optimizan un determinado objetivo. Este tipo de técnicas posee ciertas ventajas interesantes. En primer lugar, obtienen una reducción del espacio de variables, realizando una selección embebida, y al mismo tiempo codifican el espacio original en un nuevo espacio donde todos los datos quedan comprimidos. El objetivo de estas técnicas suele ser el de facilitar la búsqueda de relaciones y mapeos, ya sea entre los datos de entrada entre sí, o bien entre los datos de entrada con los de salida.

Entre los distintos tipos de métodos de predicción podemos separar a los que trabajan en forma *no supervisada* (sólo usan la información de los descriptores) de los que lo hacen en forma *supervisada* (tienen en cuenta la variable a modelar al momento de realizar la proyección). Sin embargo, estas últimas técnicas no son muy populares en quimioinformática. Entre aquellas que pertenecen al primer grupo, podemos nombrar a la regresión por componentes principales y al SOM, de las cuales ya hemos hablado en secciones anteriores.

El segundo grupo, al cual corresponde el método de SARDUX, contiene al grupo de técnicas que suelen denominarse con el término de '*Projection Pursuit*'. Este tipo de técnicas no ha sido aún bien explorado en el ámbito de la predicción QSAR. Solamente hemos encontrado un artículo en el tema de la tesis (Liang *et al.* (2007)) y dos en el área de relaciones cuantitativas de estructura-retención (Du *et al.* (2002a), Du *et al.* (2002b)).

4.5. Consideración de las técnicas QSAR-QSPR en la regulación internacional

La Comunidad Europea lanzó en Junio del 2007 una nueva regulación, conocida con el nombre de REACH (*Registration, Evaluation, Authorisation and restriction of Chemical Substances*) (REACH (2007)), la cual son una serie de medidas para regular el uso seguro de nuevas entidades químicas. En la misma, en el Anexo VI del Artículo 10, se indica explícitamente que el registrante de un nuevo compuesto químico "*debería incluir información de fuentes alternativas (por ejemplo, Relaciones Cuantitativas Estructura-Propiedad (QSAR) [...]) las cuales podrían asistir en la identificación de la presencia o ausencia de propiedades riesgosas de la sustancia y que en ciertos casos podrían reemplazar los resultados de pruebas sobre animales*".

Obviamente, dentro del marco legal propuesto en las REACH, resulta esencial que los modelos basados en QSAR provean estimaciones confiables, lo que implica que los modelos deben ser correctamente validados. Por tal motivo, se consideró importante desarrollar un conjunto de principios de reconocimiento internacional para la validación de modelos QSAR y para promover la aceptación académica y legal de estos modelos estadísticos.

En este contexto, diferentes principios, conocidos como los "Principios Setubal", fueron propuestos en 2002 durante un conferencia internacional (Jaworska *et al.* (2003)). Los mismos sirvieron de punto de partida para los principios propuestos por la OECD (*Organisation for Economic Co-Operation and Development*) en 2004 sobre validación de modelos QSAR (OECD (2004)). Estos principios, aunque generales, son simples y claros:

Principio 1: Claridad en la definición de la variable experimental que está siendo modelada. Esto implica claridad de los protocolos usados y de las condiciones en las que se aplicó dicho protocolo.

Principio 2: Transparencia del algoritmo usado para el modelo. Esto es importante para permitir una evaluación independiente del modelo, acción que en ciertos softwares comerciales no es posible de realizar.

Principio 3: Necesidad de la definición de un dominio de aplicación, de manera de conocer las limitaciones del modelo. Sobre este principio se alienta a la promoción de nuevos trabajos científicos que ayuden a definir los dominios de aplicación de los métodos.

Principio 4: Apropiadas medidas de evaluación para la bondad de ajuste, robustez y capacidad de predicción de un modelo. En este sentido, se diferencia la importancia de la validación interna (bondad de ajuste y robustez) de la validación externa (capacidad de predicción). También se alienta en este principio a la promoción de trabajos sobre validación externa.

Principio 5: Importancia desde el punto de vista científico de la interpretación física del modelo. Sin embargo, también se enfatiza que la ausencia de esta interpretación no significa que el modelo no pueda ser útil.

4.6. Conclusiones de la revisión

Como conclusiones de nuestra revisión, destacamos la abundancia de métodos de predicción de propiedades siguiendo el enfoque QSAR/QSPR, tanto en métodos publicados por la comunidad científica, como así también software comerciales. Aun cuando parece que la predicción de propiedades ha sido ampliamente estudiada, advertimos la presencia de muchas deficiencias desde el punto de vista estadístico, las cuales generalmente son acompañadas de una pobre validación (Tropsha *et al.* (2003), Gramatica (2007)). Otros artículos de QSAR advierten cómo estos errores han perjudicado la precisión y la credibilidad sobre los métodos basados en QSPR (Johnson (2008), Hou & Wang (2008), Doweyko (2008)). Los principios de la OECD justamente apuntan a cambiar estas malas prácticas y desarrollar los modelos del futuro.

Nuestra postura en esta tesis se centró, por lo tanto, en la aplicación de técnicas computacionales para mejorar la predicción de los métodos. Aún cuando hemos programado y/o evaluado una gran cantidad de métodos de predicción, sólo nos centraremos en explicar aquellos desarrollos que contengan un aspecto novedoso, ya sea desde el punto de vista de ciencias de la computación o de la aplicación en quimioinformática.

La primer contribución de esta tesis se enfocó en el problema de selección de descriptores, el cual se detalla extensamente en el capítulo 5. Consideramos que esta tarea constituye un paso fundamental, ya que define la calidad predictiva del modelo, así como también la interpretabilidad del mismo. En este contexto, advertimos que no existía un estudio detallado sobre las diferentes opciones de diseño de los métodos que comúnmente se aplicaban para seleccionar los descriptores. Por otra parte, algunas propuestas resultan imprecisas o inviables con grandes bases de datos. Por estos motivos, apuntamos a realizar un fuerte estudio de diferentes enfoques para atacar este problema en forma efectiva.

En segundo lugar, advertimos el problema de la capacidad de generalización de un modelo a nuevos datos, y la definición del dominio de aplicabilidad, los cuales pueden pensarse como dos caras del mismo problema. En nuestra revisión se diferenció en clases las distintas propuestas que apuntan a resolver el problema de la identificación del dominio del método, indicando en cada una sus ventajas y desventajas. De nuestro estudio en el tema, surgió una alternativa híbrida que aplica distintos aspectos de cada una de las diferentes estrategias, buscando así confluir distintas ventajas en un mismo método.

Finalmente, los métodos basados en proyecciones a subespacios no han sido vastamente aplicados en quimioinformática, por lo que creemos que es importante su experimentación, dados los beneficios enunciados. Por este motivo, hemos aplicado y desarrollado dos métodos novedosos para sacar provecho de este tipo de proyecciones para los modelos QSAR.

Tabla 4.1: Softwares comerciales y académicos para cálculo de logP.

Nombre	Empresa / Universidad	URL / e-mail
AB/LogP v. 2.0	Pharma Algorithms, Lituania/Canada	http://www.ap-algorithms.com
ABSOLV, LSER	Pharma Algorithms, Lituania/Canada	http://www.ap-algorithms.com
ACD/logP v. 11	Advanced Chemistry Development, EEUU	http://www.acdlabs.com
ALOGP (DragonX 1.4)	Talete SRL, Milano, Italia	http://www.talete.mi.it
ALOGP98	Accelrys Software Inc., EEUU	http://www.accelrys.com
ALOGPS v. 2.1	Virtual Computational Chemistry Laboratory, Alemania	http://www.vcclab.org
CLIP	University of Geneva, Suiza	pierre-alain.carrupt@pharm.unige.ch
CLOGP v. 4.3 (v. 5.0)	BioByte Inc., EEUU	http://www.biobyte.com
COSMOFrag v. 2.3	COSMOlogic GmbH & Co. KG, Alemania	http://www.cosmologic.de
CSlogP v. 2.2.0.0	ChemSilico LLC, EEUU	http://www.chemsilico.com
GBLOGP	Max Totrov, EEUU	max@molsoft.com
HINT	EduSoft, LC, EEUU	http://www.edusoft-lc.com
KowWIN v. 1.67	Syracuse Inc., EEUU	http://www.syrres.com
LSER UFZ	Helmholtz Centre for Environm. Research UFZ, Alemania	ralph.kuehne@ufz.de
MiLogP v. 2.2	Molinspiration Chemoinformatics, República Eslovaca	http://www.molinspiration.com
MLOGP (DragonX 1.4)	Talete SRL, Milano, Italia	http://www.talete.mi.it
MLOGP(S+), ADMET 2.3	Simulations Plus, Inc., EEUU	http://www.simulations-plus.com
MolLogP	MolSoft LLC, EEUU	http://www.molsoft.com
NC+NHET	Virtual Computational Chemistry Laboratory, Alemania	http://www.vcclab.org
OsirisP	Actelion, Suiza	http://www.actelion.com
QikProp v. 3.0	Schrödinger, LLC, EEUU	http://www.schrodinger.com
QLOGP	University of Miami, EEUU	PBuchwald@med.miami.edu
QuantlogP	Quantum Pharmaceuticals, Rusia	http://q-pharm.com
S+logP, ADMET 2.3	Simulations Plus, Inc., EEUU	http://www.simulations-plus.com
SLIPPER-2002	Institute of Physiol. Active Compounds, Rusia	http://camd.ipac.ac.ru
SPARC	University of Georgia, EEUU	http://ibmlc2.chem.uga.edu
TLOGP	Upstream Solutions, Suiza	http://www.upstream.ch
VEGA	University of Milan, Italia	http://www.ddl.unimi.it
VLOGP	TOPKAT, Accelrys Software Inc., EEUU	http://www.accelrys.com
XLOGP2	Inst. of Physical Chemistry, Peking University, China	ftp://ftp2.ipc.pku.edu.cn
XLOGP3	Institute of Organic Chemistry, Shanghai, China	http://sioc-ccbg.ac.cn

Capítulo 5

Selección y reducción del conjunto de descriptores

En las áreas de las ciencias de la computación y de la estadística multivariada hay un sinnúmero de técnicas que se encuadran dentro de los métodos de selección o reducción de variables. Comenzaremos el presente capítulo brindando una taxonomía de las diferentes técnicas, explicando brevemente los fundamentos de cada metodología. Luego enfatizaremos la importancia de este tipo de técnicas para su aplicación en métodos basados en QSAR. Finalmente, describiremos los principales aportes realizados en el marco de esta tesis, los cuales consisten en la proposición de dos metodologías de selección de descriptores para el posterior uso en un modelo QSPR. Ambas metodologías utilizan algoritmos evolutivos, sin embargo la primera utiliza un enfoque mono-objetivo, mientras que la segunda utiliza un enfoque multi-objetivo. En el capítulo 7 se describirá otro método que también permite la selección de descriptores pero, por poseer un enfoque teórico completamente diferente, se presentará en un capítulo distinto al presente.

5.1. Selección de variables

El proceso de selección de un subconjunto de variables dentro de un espacio multivariado es una tarea en creciente estudio dado que en las aplicaciones de hoy en día resulta común el acceso a librerías de datos, que son especificadas por gran cantidad de variables. Tal es el caso, por ejemplo, de los datos provenientes

de *micro-arrays* en donde resulta importante contar con técnicas que seleccionen sólo un subconjunto de genes que estén relacionados con un determinado proceso biológico (Deb & Raji Reddy (2003); Guyon *et al.* (2002); Zhu *et al.* (2007)).

El estudio de este tipo de técnicas dentro de las ciencias de la computación recibe el nombre de selección de características (*feature selection*). Algunos autores hacen una diferenciación del concepto de "variables" y "características", asumiendo que estos últimos corresponden a variables transformadas. Sin embargo, en el transcurso de esta tesis, los términos "variables" y "características" serán usados en forma intercambiable.

En cuanto a la taxonomía, los métodos de selección de características pueden clasificarse en función de si se aplican en forma *no supervisada* o *supervisada*. Los primeros se centran en la necesidad de aplicar métodos de proyección o agrupamiento a espacios de menor dimensionalidad y sin tener en cuenta variable de predicción alguna. Los segundos buscan seleccionar un subconjunto de variables que resulten influyentes para predecir una determinada variable de predicción. En el transcurso de este capítulo nos centraremos en el segundo tipo de técnicas, mientras que recomendamos a cualquier lector interesado en las técnicas *no supervisadas* a referirse a los siguientes trabajos: Dy (2008); Liu & Yu (2005).

En forma paralela con la clasificación anterior, un método de selección de variables puede clasificarse en *filtros*, *wrappers* y *métodos embebidos* (Guyon & Elisseeff (2003)). La principal diferencia entre los filtros y los *wrappers* es en cómo se mide la relevancia del subconjunto de variables seleccionadas. Los primeros aplican procedimientos matemáticos deterministas, como ganancia de información (*information gain*) o pruebas χ^2 (*chi-cuadrado*), en pos de evaluar la utilidad de un determinado subconjunto (Duch (2006); Liu (2004)). Los métodos *wrappers* se caracterizan por usar un método de aprendizaje automático, el cual es entrenado usando la información de un subconjunto de variables, y donde la capacidad predictiva de este subconjunto se evalúa, siguiendo algún método de validación. Finalmente, los *métodos embebidos* corresponden a cualquier método de predicción que aplica selección de variables como parte de su proceso de entrenamiento, como sucede por ejemplo en el caso de los árboles de decisión con procedimientos de podado.

Por otra parte, los métodos supervisados pueden aplicarse de manera individual o grupal, dependiendo de si la evaluación de la relevancia de las variables se realiza en forma individual para cada variable o de si se realiza en el contexto de la pertenencia a un grupo. La primera variante resulta ampliamente superada por la segunda, dado que una variable puede parecer no relevante por sí sola, pero puede resultar relevante cuando se la analiza en presencia de otras variables (Guyon & Elisseeff (2003)). De igual modo, dos variables pueden resultar individualmente relevantes, pero analizadas juntas podrían ser mutuamente redundantes. Por lo tanto, el análisis grupal de variables resulta necesario, aún cuando la complejidad de estas técnicas es ampliamente superior al de las técnicas individuales.

Profundizaremos un poco más en los métodos *wrapper* dado que corresponden a la estrategia de selección de variables elegida para el presente capítulo, mientras que el método presentado en el capítulo 7 corresponde a un *método embebido*. Un método *wrapper* puede dividirse internamente en dos partes bien diferenciadas. Llamaremos a una de estas partes *Buscador de Variables* y a la otra *Evaluador de Variables*. La primera es la responsable de realizar la búsqueda combinatorial entre los posibles subconjuntos de variables, mientras que la segunda se encarga de evaluar la relevancia de los subconjuntos y de este modo guiar al buscador de variables hacia una selección de descriptores relevantes. Usaremos los términos *funciones de búsqueda* o *métodos de búsqueda* para referirnos a los métodos con que se implementa el *Buscador de Variables*. De la misma manera, usaremos *funciones de evaluación* o *métodos de evaluación* para referirnos a los métodos usados para el *Evaluador de Variables*. La Figura 5.1 esquematiza las dos partes de un *wrapper*, junto con los métodos aplicados en la sección 5.3.

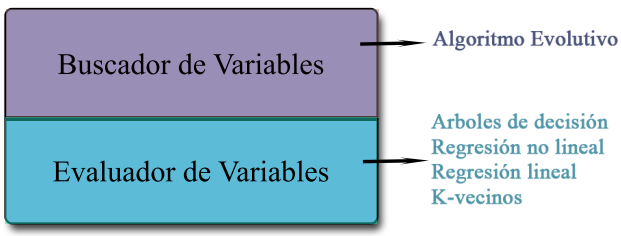


Figura 5.1: Componentes de un *wrapper* y métodos aplicados

5.2. Importancia de la selección de descriptores en QSAR

En el diseño de métodos QSAR, uno de los primeros y más importantes pasos es la selección de descriptores relevantes que relacionen la información química y molecular con la actividad o propiedad de interés. Comúnmente, la tarea de selección de descriptores no puede ser realizada completa y manualmente por expertos en biología o química, dado que estas relaciones estructura-actividad son usualmente complejas, altamente no-lineales y en muchas ocasiones desconocidas.

En QSAR existen dos dificultades coexistentes que hacen el problema particularmente difícil de resolver. Por un lado, existe una inmensidad de descriptores disponibles n que cuantifican información sobre una molécula, y al mismo tiempo existe poco conocimiento sobre cuáles y cuántos descriptores son necesarios para correlacionar con una dada variable de salida. El uso de un método de selección secuencial exhaustivo, nos exigiría probar $\binom{n}{1} + \binom{n}{2} + ... + \binom{n}{n}$ subconjuntos de variables, lo que se vuelve completamente inviable para n no demasiado grandes, ya que su orden de ejecución es de $\mathcal{O}(2^n)$. En Davies & Rusell (1994) se demuestra que el problema de selección de variables es NP-completo, lo que hace necesaria la aplicación de un método estocástico para poder encontrar un subconjunto de variables apropiado en un tiempo razonable. Asimismo, la selección de un subconjunto de variables a partir de un gran número de variables disponibles puede dar lugar a la ocurrencia de correlaciones por chance, tal como se describe en Topliss (1979), Topliss & Costello (1972), Baumann (2005) y Rücker *et al.* (2007).

Por otra parte, dado que las relaciones estructura-actividad o estructura-propiedad son frecuentemente no lineales, los métodos necesarios para analizar estas relaciones suelen ser costosos desde el punto de vista computacional. En este sentido, es deseable que el método usado para el *Evaluador de Variables* sea capaz de evaluar la capacidad predictiva considerando cualquier tipo de relación no lineal, pero al mismo tiempo que este método sea computacionalmente económico, de manera que pueda realizar una evaluación rápida de cada subconjunto de variables presentado. Esta última característica es especialmente importante en los métodos *wrapper* para que éstos sean capaces de realizar una gran exploración del espacio de búsqueda. Claramente, podemos advertir que existe una situación

de compromiso entre la precisión del método de evaluación y el tiempo necesario para generar o entrenar dicho modelo.

5.2.1. Limitaciones de las propuestas existentes para QSAR

Hemos visto en la sección 4.4.5 que se han propuesto distintos enfoques para atacar el problema de la selección de descriptores en QSAR. No obstante la existencia de estos enfoques no existía un estudio detallado sobre las diferentes opciones de diseño de los *wrappers* que comúnmente se aplicaban para seleccionar los descriptores. Especialmente, no había un análisis claro sobre las diferentes opciones de métodos para el *Evaluador de Variables*. En pos de cubrir esta necesidad se realizaron una serie de experimentaciones las cuales se describen en las secciones 5.3 y 5.4.

Asimismo, en la literatura proveniente de las Ciencias de la Computación, surgieron ciertos métodos de selección de características los cuales presentaban un enfoque multi-objetivo (Handl & Knowles (2006); Oliveira *et al.* (2003)). La motivación principal de este tipo de técnicas surge de que las predicciones usando grandes cantidades de descriptores son propensas a sobreajustar los datos y a generar correlaciones por chance. Ante la ausencia de este enfoque en el campo de la selección de descriptores para QSAR, la propuesta presentada en la sección 5.4 apuntó a definir el primer trabajo en el área de la Quimioinformática. Esta propuesta fue realizada con la misma filosofía del enfoque mono-objetivo, es decir, proveyendo un análisis de las diferencias en las distintas opciones de diseño. Cabe destacar que para ambas propuestas se utilizó un enfoque de dos fases. Este enfoque, el cual será explicado en detalle más adelante, representa una manera sensata de dividir el proceso en dos partes. En la primera fase se realiza una búsqueda rápida, aunque relegando cierta precisión, y en la segunda fase se analizan con mayor precisión las soluciones encontradas en la fase anterior. Este enfoque de dos fases fue paralelamente desarrollado por otros grupos de investigación (Gharagheizi (2008); Yang *et al.* (2009)). Una ventaja adicional que posee esta estrategia, es que el uso de distintas metodologías de regresión aplicadas en una fase y en otra, evita que se elija un subconjunto de descriptores que resulte optimal sólo para un método en particular. Este problema es identificado y

explicado en forma más precisa en el trabajo de Dutta *et al.* (2007). Además, el procedimiento de la segunda fase permite efectuar una evaluación rigurosa e independiente.

5.3. 1º Propuesta: selección de descriptores utilizando algoritmos evolutivos mono-objetivo

Para la tarea de selección de descriptores se optó por un método *wrapper*. Los *wrappers* son una opción apropiada dado que son flexibles, muy eficaces en la reducción de la dimensionalidad del problema y al mismo tiempo buscan una buena capacidad predictiva de los subconjuntos seleccionados (Handl & Knowles (2006); Kohavi & John (1997)). Sin embargo, pueden provocar sobreajuste si no son aplicados correctamente (Loughrey & Cunningham (2005)).

En esta primera propuesta, y teniendo en cuenta los métodos propuestos en otras investigaciones (sección 4.4.5), nuestro *Buscador de Variables* fue implementado usando un algoritmo evolutivo. Asimismo, ante la falta de un análisis al respecto, decidimos alternar con distintos métodos para el *Evaluador de Variables*. Describiremos a continuación los distintos componentes de nuestro algoritmo.

5.3.1. Primera Fase: *Evaluador de Variables*

El objetivo del *Evaluador de Variables* es el de cuantificar la relevancia de los subconjuntos de variables presentados. Para esto, los métodos utilizados deben ser capaces de establecer una regresión, tomando como variables independientes un subconjunto de descriptores, provisto por el método de búsqueda. Además, para que el algoritmo de selección de descriptores funcione de manera aceptable, es necesario que el método de regresión aplicado sea capaz de modelar relaciones complejas de la manera más rápida posible. Por tal motivo se decidió experimentar usando Árboles de Decisión (AD), k-Vecinos más Cercanos (kVC), un modelo de regresión no lineal (RNL) y otro lineal (RL).

Los AD fueron utilizados como árboles de regresión. Se utilizó el índice de diversidad de Ginni como criterio de separación (Breiman *et al.* (1984)), siempre y cuando hubiera más de 10 datos para separar. Se tomó el promedio de la variable a

modelar de los datos como el valor de predicción de cada nodo. El segundo método aplicado fue kVC, con $k = 3$ y usando la distancia Euclídea estandarizada como medida de distancia. Tanto AD como kVC habían sido elegidas previamente por otros artículos para modelar datos de naturaleza semejante (Guha & Jurs (2004); Trevino & Falciani (2006),Li *et al.* (2001)).

Para la RNL se utilizó una combinación lineal de funciones bases no lineales (de hasta grado 4) donde los coeficientes $\beta_{i,j}$ son ajustados según el criterio de cuadrados mínimos, donde la optimización se obtiene usando el método de Gauss-Newton (Madsen *et al.* (2004)). La ecuación 5.1 muestra la fórmula de regresión elegida para optimizar, donde p es la cantidad de variables a usar en el modelo y β_0 es el término independiente de la ecuación.

$$RNL: \quad \sum_{i}^{p}(\sum_{s=1}^{4}\beta_{i,s}x_i^s) + \beta_0 \tag{5.1}$$

Finalmente, RL fue aplicado con el fin de analizar si éste, a pesar de su incapacidad para representar relaciones no lineales, podía resultar una alternativa viable.

La cuantificación de la calidad predictiva del j-ésimo subconjunto de descriptores, se definió según la ecuación 5.2:

$$F(\mathcal{P}_{Z_1^j}, Z_2^j) = \frac{1}{m_2}\sum_{(\mathbf{x_i},y_i)\in Z_2^j}(y_i - \mathcal{P}_{Z_1^j}(\mathbf{x_i}))^2 \tag{5.2}$$

La ecuación anterior computa el error cuadrado medio (MSE, del inglés *mean square error*) evaluado en un subconjunto de compuestos no usado durante el entrenamiento, donde:

- Z es una matriz que representa la base de datos con todos los compuestos, donde las filas y las columnas corresponden a compuestos y descriptores respectivamente. La última columna de Z almacena la información de la propiedad experimental a predecir para cada compuesto. Este último vector columna se lo denota como y.

- Z_1 y Z_2 son matrices de compuestos, los cuales son usados como conjuntos de entrenamiento y validación respectivamente. Z_1 es de dimensión $m_1 \times n$ mientras que Z_2 es $m_2 \times n$.
- $\mathcal{P}_{Z_1}$ es un método de regresión entrenado con el set de datos Z_1.
- El superíndice j, como en el caso de Z_1^j, corresponde al conjunto de compuestos de Z_1 usando sólo los descriptores incluidos en el j-ésimo subconjunto. Este subconjunto es provisto por el *Buscador de Variables* del *wrapper*. En nuestra implementación corresponde a la selección codificada en el j-ésimo individuo de nuestro algoritmo evolutivo (sección 5.3.2.1).
- $\mathbf{x_i}$ es un vector, donde cada componente representa los descriptores asociados al i-ésimo compuesto de un set de datos. Del mismo modo, $\mathcal{P}_{Z_1}(\mathbf{x_i})$ es el valor de predicción para el compuesto $\mathbf{x_i}$ usando el método $\mathcal{P}$ entrenado con el set de datos Z_1.
- y_i es el valor experimental para el i-ésimo compuesto de un dado conjunto de datos.

5.3.2. Primera Fase: *Buscador de Variables*

En la presente subsección detallaremos las decisiones de diseño que aplicamos para desarrollar nuestro *Buscador de Variables* utilizando algoritmos evolutivos. Tal como se comentó, el espacio de búsqueda al cual se enfrenta nuestro algoritmo es muy amplio. Por tal motivo, para esta primera propuesta, se decidió que la búsqueda sea restringida a encontrar la mejor selección de un subconjunto de descriptores de cardinalidad p. Dado que no se tiene conocimiento *a priori* sobre el valor recomendado de este parámetro p, se realizaron distintos experimentos variando su valor.

5.3.2.1. Representación

Se utilizaron vectores binarios como representación de los cromosomas de los individuos. Cada vector posee n componentes, es decir un bit por cada descriptor disponible. Un valor no nulo en la i-ésima posición del cromosoma, significa que

el i-ésimo descriptor es elegido dentro de la selección del individuo. En forma contraria, un valor nulo en la i-ésima posición del cromosoma, significa que el i-ésimo descriptor no es elegido dentro de la selección del individuo. El cromosoma de la Figura 5.2 representa una selección de un subconjunto de descriptores en donde el segundo, cuarto, r-ésimo y s-ésimo descriptor fueron seleccionados. La población inicial es constituida en forma aleatoria, teniendo en cuenta que el número de descriptores seleccionados para cada individuo sea siempre igual a p.

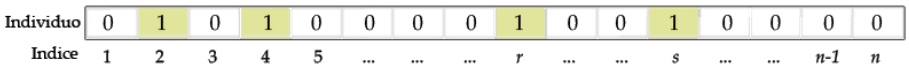


Figura 5.2: Representación binaria de los cromosomas.

5.3.2.2. Función de fitness

La función de fitness tiene por objetivo cuantificar la aptitud de un individuo. En el caso particular de esta aplicación la función de fitness está estrechamente relacionada con el *Evaluador de Variables* del *wrapper*. Por tal motivo, la función de fitness está regida por la Ecuación 5.2, en donde $\mathcal{P}$ dependerá del método de evaluación usado. En esta propuesta, a diferencia de la variante multi-objetivo que presentaremos en la sección 5.4, no existen diferencias entre la función usada para el *Evaluador de Variables* y la función de fitness del *Buscador de Variables*.

5.3.2.3. Selección y operadores genéticos

A fin de determinar la estrategia de selección de individuos del algoritmo evolutivo, se experimentó con diferentes propuestas y se concluyó que la estrategia del torneo resulta la más apropiada. Además este método es altamente recomendado por tener un orden de complejidad $O(n)$ (Goldberg & Deb (1991)).

Con respecto a los operadores genéticos se utilizó un cruzamiento de un punto para la recombinación. Individuos no factibles, es decir que no cumplen con la restricción de la cantidad de descriptores seleccionados, pueden producirse posteriormente a la aplicación del cruzamiento. En este caso, se setean o resetean tantos *loci* como sean necesarios en el cromosoma para cumplir con la restricción.

Dado que este esquema de cruzamiento incorpora una alteración aleatoria de los bits, nos abstuvimos de aplicar adicionalmente otra operación de mutación.

5.3.3. Parametrización y funcionamiento general

Describiremos ahora cuál fue la parametrización y el diseño de experimentos usados para la ejecución de la primera fase. Esta parametrización no es exclusiva de la propuesta presentada, sino que sólo representa un comportamiento adecuado del algoritmo, que nos permitió además obtener conclusiones sobre las distintas alternativas de métodos de evaluación aplicables a la propuesta.

El tamaño de la población fue establecido en 45 individuos, la probabilidad de cruzamiento fue de 0,8, el tamaño del torneo fue igual a 3 y cada generación conservaba 2 miembros de elite por generación. Tal como se recomienda en la literatura (Safe *et al.* (2004)), se utilizó un criterio de parada fenotípico, ya que el algoritmo finalizaba cuando el individuo con fitness más alto de la población no mejoraba durante una cierta cantidad de generaciones o cuando la mejora en el fitness promedio de la población era menor que un valor mínimo establecido.

5.3.4. Segunda Fase: refinamiento y evaluación de los subconjuntos encontrados

Los métodos de predicción utilizados en la primera fase no resultan la mejor alternativa para construir modelos de regresión de precisión. Sin embargo, constituyen una alternativa interesante, dado que son capaces de construir modelos de predicción sin una carga computacional demasiado costosa. Por tal motivo, el procedimiento de selección de descriptores aquí presentado hace uso de una segunda fase en donde los subconjuntos de descriptores encontrados en la primera fase son reevaluados utilizando un método de predicción más preciso. En este caso se utilizaron comités de 5 redes neuronales (CRN), donde el algoritmo de aprendizaje es el de retro-propagación elástica (*resilient back-propagation* Anastasiadis *et al.* (2005)), para ajustar los pesos que conectan los 5 nodos de la capa oculta.

Además, la estrategia de utilizar dos fases favorece la confianza sobre la calidad de los subconjuntos seleccionados. Esto se debe a que cada subconjunto

encontrado no solamente resulta de relevancia según un método en particular, sino que el criterio de selección es capturado por metodologías diferentes. En el trabajo de Dutta *et al.* (2007) se detallan los fundamentos de esta afirmación.

5.3.5. Resultados

Usando el enfoque de selección de variables mono-objetivo se realizaron dos experimentaciones diferentes. En la segunda de ellas se mejoran ciertos detalles de diseño de la primera experimentación.

5.3.5.1. Primera experimentación

Para el primer grupo de experimentaciones se utilizaron los primeros 1200 compuestos (ordenados por CAS) del set de datos descripto en B.1. A su vez, sólo se trabajó con los 47 descriptores de la familia de los constitucionales (Todeschini & Consonni (2000)). El objetivo de esta primera experimentación era la de, utilizando una gran cantidad de compuestos, analizar cuáles son los 10 descriptores más relevantes, comparar el impacto del uso de distintas metodologías en el *wrapper* y determinar si ésta estrategia resulta viable. Es importante mencionar que los descriptores constitucionales usados en esta primera experimentación, son considerados como potencialmente relevantes para la predicción del logP.

El procedimiento de validación se realizó del siguiente modo: 50 % de los compuestos fueron utilizados para entrenamiento (Set 0), 16 % para validación (Set 1) y el 34 % restante para testeo. El porcentaje de testeo fue dividido a su vez en dos grupos (Set 2 y Set 3), dado que ambos grupos presentaban distribuciones de frecuencias para la hidrofobicidad muy diferentes entre sí.

La elección de los parámetros de los métodos usados para el *Evaluador de Variables* fueron elegidos de acuerdo a distintas configuraciones evaluadas sobre 15 corridas, aplicando validación cruzada sobre el Set 0. Durante la ejecución del *wrapper* se usó el Set 0 para obtener los modelos y el Set 1 para evaluar los errores. A partir del objetivo propuesto se estableció como restricción que el algoritmo sólo busque subconjuntos de 10 descriptores, en otras palabras, soluciones con 10 bits seteados. En esta primera experimentación se decidió analizar los siguientes métodos de evaluación: AD, RL y RNL. Además, se analizó el desempeño en la

Tabla 5.1: Errores absolutos *promedios* obtenidos sobre los conjuntos de testeo (Set 2 y Set 3).

	Todos	Aleatorio		AD		RL		RNL	
	——	Mejor	Promedio	Mejor	Promedio	Mejor	Promedio	Mejor	Promedio
Set 2	1,3037	1,2860	1,4318	1,1249	1,3398	1,2851	1,3599	1,2849	1,3213
Set 3	1,0414	1,0787	1,1571	1,0603	1,1227	1,0331	1,1090	1,0255	1,1256

segunda fase cuando en la primera fase se hace una selección aleatoria y cuando se eligen todos los descriptores disponibles.

Es importante destacar que por la naturaleza estocástica del algoritmo evolutivo, el *wrapper* no siempre arriba a la misma selección de descriptores. Por lo tanto, el funcionamiento de cada configuración de los distintos métodos de evaluación se analizó a partir de 15 corridas del *wrapper*. A su vez, la mejor selección de cada corrida es tomada por el CRN en la segunda fase. Dada la naturaleza estocástica de los CRN debida a la configuración inicial de sus pesos, las redes eran reentrenadas 7 veces. En consecuencia, cada alternativa es testeada basada en un promedio sobre 105 réplicas (15×7); mientras que la mejor selección de variables encontrada para cada alternativa está basada en 7 réplicas. Llamaremos al promedio basado en 105 réplicas selección '*promedio*', mientras que el promedio basado en las mejores 7 réplicas lo denominaremos selección '*mejor*'.

En la Tabla 5.1 se detallan los errores absolutos promedios obtenidos según distintas alternativas de selección de los descriptores. La columna 'Todos' muestra el resultado de cuando no se realiza procedimiento de selección de descriptores, es decir, cuando se toman los 47 descriptores constitucionales. En este caso, dado que hay una única posible selección, no se hace una distinción de la corrida 'Promedio' y 'Mejor'. En la columna 'Aleatorio' se tomaron subconjuntos aleatorios con 10 descriptores.

Se puede apreciar que en primer lugar hay una diferencia considerable entre las predicciones realizadas sobre el Set 2 y el Set 3, siendo éste último mejor modelado que el primero.

Analizando las *mejores* selecciones encontradas con cada variante (incluyendo la columna de 'Todos'), vemos que los AD tienen un comportamiento superior al resto para el caso del Set 2, mientras que para el otro conjunto de testeo existe una

pequeña diferencia a favor de la regresión no lineal. Resulta entonces importante analizar en forma estadística si hay o no diferencias significativas.

Realizando los tests ANOVA[1], donde cada factor se refiere a las diferentes estrategias utilizadas, encontramos que para el Set 2 (Tabla 5.2) existen diferencias significativas con una probabilidad p =0,0001, mientras que para el Set 3 (Tabla 5.3) no hay evidencia de que existan diferencias significativas con $p \leq$ 0,05. Al realizar una prueba de comparaciones múltiples del tipo Dunnet (Dunnet (1955)), descubrimos que para el Set 2 los AD superan las otras alternativas con un error global inferior al 1 % (Tabla 5.4) (lo mismo sucede aplicando las pruebas de Bonferroni o Tukey). En el caso del Set 3, como el test ANOVA nos anticipa, no hay evidencia de diferencias significativas globales entre las diferentes alternativas (Tabla 5.5), aunque se ha podido encontrar una diferencia entre RNL y la selección aleatoria, utilizando un test DMS al 5 % de error individual.

Tabla 5.2: Tabla ANOVA para las *mejores* selecciones aplicadas sobre el conjunto de testeo Set 2.

F. de V.	S.C.	G.L.	C.M.	*F*	***p***
Entre	0,15414844	4	0,03853711	35,941	**0,0001**
Dentro	0,03216699	30	0,00107223		
Total	0,18631543	34			

Tabla 5.3: Tabla ANOVA para las *mejores* selecciones aplicadas sobre el conjunto de testeo Set 3.

F. de V.	S.C.	G.L.	C.M.	*F*	***p***
Entre	0,013058371	4	0,003264593	1,3541	**0,2732**
Dentro	0,072325361	30	0,002410845		
Total	0,085383732	34			

En el análisis de la performance *promedio* de las distintas alternativas, debemos recordar que aún con una misma configuración de parámetros el *wrapper* puede dar como resultado distintas selecciones en distintas corridas. Desde el punto de vista estadístico, esto introduce una fuente adicional de varianza provocada

[1]ANOVA (*analysis of variance*) o análisis de la varianza, es una técnica para medir la significancia en la diferencia de distintas medias a partir de la dispersión de las mismas. Para más información sobre las distintas técnicas de ANOVA y de comparaciones múltiples, sugerimos referirse al libro de Winer *et al.* (1991)

Tabla 5.4: Comparación múltiple de *mejores* resultados usando el test Dunnet para el Set 2. *ns* probabilidad mayor a 5 %, * probabilidad menor a 5 %, ** probabilidad menor a 1 %.

Método	AD	RNL	RL	Aleatorio	Todos
Estadístico	—	9,1373	9,1476	9,1988	10,2126
Significancia	—	**	**	**	**

Tabla 5.5: Comparación múltiple de *mejores* resultados usando el test Dunnet para el Set 3. *ns* probabilidad mayor a 5 %, * probabilidad menor a 5 %, ** probabilidad menor a 1 %.

Método	RNL	RL	Todos	AD	Aleatorio
Estadístico	19,865	15,505	10,759	—	10,550
Significancia	ns	ns	ns	—	ns

por cada corrida. Por tal motivo, para el análisis del comportamiento *promedio* utilizamos un test ANOVA anidado, donde el método es el factor principal, cada corrida del *wrapper* es el factor anidado y la variación de las redes se cuantifica con el factor *dentro*. Las Tablas 5.6 y 5.7 muestran los tests ANOVA anidados aplicados sobre los Set 2 y 3 respectivamente. Podemos apreciar que existe una diferencia muy significativa para el Set 2 ($p = 0,0015$), mientras que las diferencias para el Set 3 no son tan probables ($p = 0,0823$). Podemos observar también, que la fuente de variación debida a las distintas corridas del *wrapper* tienen una contribución importante en la variación de los promedios.

Tabla 5.6: Tabla ANOVA para las selecciones *promedio* aplicadas sobre el conjunto de testeo Set 2.

F. de V.	S.C.	G.L.	C.M.	*F*	***p***
F. Principal	0,737076291	3	0,245692097	5,8253	**0,0015**
F. Anidado	2,361903692	56	0,042176851	30,65412	**0,0001**
Dentro	0,495322256	360	0,001375895		
Total	3,594302239	419			

Para determinar si hay beneficios en aplicar el *wrapper*, podemos aplicar una prueba Dunnet tomando como testigo la estrategia *Aleatoria*. En el caso del Set 2 (Tabla 5.8) se evidencia que existen diferencias significativas en aplicar la estrategia de selección de variables en los casos no lineales (RNL y AD). Sin embargo, para el Set 3 (Tabla 5.9) vemos que no hay diferencias a nivel global. Si utilizamos

Tabla 5.7: Tabla ANOVA para las selecciones *promedio* aplicadas sobre el conjunto de testeo Set 3.

F. de V.	S.C.	G.L.	C.M.	F	p
F. Principal	0,16849875	3	0,05616625	2,3485	**0,0823**
F. Anidado	1,33926114	56	0,023915377	18,38	**0,0001**
Dentro	0,468360687	360	0,001301001		
Total	1,976120577	419			

un test con error individual como el de DMS para el Set 3, existen diferencias entre la estrategia lineal y la *Aleatoria* ($p \leq 0,05$), y con menor nivel de certeza entre AD y *Aleatoria* ($p \leq 0,15$) (tabla no mostrada).

Tabla 5.8: Comparación múltiple de resultados *promedio* usando el test Dunnet para el Set 2. *ns* probabilidad mayor a 5 %, * probabilidad menor a 5 %, ** probabilidad menor a 1 %.

Tratamientos	RNL	AD	RL	Aleatorio
Estadísticos	2,922224941	2,352033219	2,044792892	—
Resultados	**	*	ns	—

Tabla 5.9: Comparación múltiple de resultados *promedio* usando el test Dunnet para el Set 3. *ns* probabilidad mayor a 5 %, * probabilidad menor a 5 %, ** probabilidad menor a 1 %.

Tratamientos	RL	AD	RNL	Aleatorio
Estadísticos	1,711	1,0697	0,7647	—
Resultados	ns	ns	ns	—

Como comentarios finales de esta primera experiencia, advertimos que esta propuesta sufre de ciertas deficiencias. En primer lugar, la misma es demasiado rígida en cuanto a su restricción de la cantidad de descriptores considerados p. Esto involucra que en caso de desconocerse el número exacto de descriptores necesarios (como sucede en la mayoría de los casos), el *Buscador de Variables* está confinado a buscar en un espacio que puede no ser el óptimo.

También se aprecia de las experimentaciones que los resultados con el Set 2 y con el Set 3 son bastante diferentes, como así también es diferente el comportamiento de las diferentes alternativas. La restricción de mantener siempre 10 bits seleccionados hace que los individuos muten frecuentemente para satisfacer esta

restricción. Esto puede generar una falta de convergencia en el algoritmo. Si bien, podemos destacar a los AD como aquellos con el funcionamiento más estable, los resultados no son del todo convincentes teniendo en cuenta que el comportamiento *promedio* de las diferentes alternativas del *wrapper*, suele ser peor que el comportamiento utilizando todos los descriptores. No obstante, es importante decir que todos los descriptores disponibles son potencialmente relevantes para la propiedad logP y al mismo tiempo ciertas familias de descriptores posiblemente relevantes no fueron incluidas (por ejemplo, descriptores de grupos funcionales). Por otra parte, el conjunto de datos utilizados no permite la comparación directa con otras estrategias de selección de descriptores.

Los puntos salientes de este enfoque mono-objetivo fueron publicados en Soto *et al.* (2008a).

5.3.5.2. Segunda experimentación

Para esta segunda experimentación, se utilizó el conjunto de compuestos detallado en el Apéndice B.2. La elección de este set de datos se basó en que los resultados admiten una comparación con otro trabajo (Yaffe *et al.* (2002)). Además, los 442 compuestos usados poseen una heterogeneidad de familias químicas que lo convierten en un conjunto de compuestos de interés para ser modelado.

En el trabajo de Yaffe *et al.* (2002) se describen cuáles son los 12 descriptores seleccionados por su método de selección de características, pero no el conjunto de descriptores originales sobre el cuál aplican el método de selección. Por tal motivo, y a fin de recrear la situación original de los autores de este trabajo, procedimos a calcular 61 descriptores más que son comúnmente empleados en otras propuestas, tales como las que se usan en Duprat *et al.* (1998); Lipinski *et al.* (2001); Wang *et al.* (1997). Para esto, utilizamos la versión web de Dragón (Tetko *et al.* (2005)), seleccionando descriptores de las siguientes familias químicas: constitucionales (41), grupos funcionales (16), propiedades (2) y empíricos (3).

A fin de realizar una búsqueda más completa y de explorar el impacto de distintos valores de p se decidió experimentar con 3 valores diferentes: $10, 20$ y 30. Se realizaron 45 corridas por cada valor de p y este mismo procedimiento se

probó para los siguientes métodos de evaluación de variables: AD, kVC y RNL. Por lo tanto, esto representa un total de 405 corridas para el *wrapper*.

Con el propósito de obtener una única selección para cada una de las 45 corridas, se procedió a realizar un esquema basado en rankings. Al comienzo de la segunda fase y para cada valor de p, se establece un ranking de todos los descriptores en función de la cantidad de veces que resultan elegidos en las diferentes corridas del algoritmo evolutivo. A partir de cada uno de los rankings, se tomaron los primeros d descriptores, y con ellos se entrenaron los CRN, donde se varió la arquitectura en función de d. Por lo tanto, se decidió experimentar con los siguientes valores de d: $11, 12, 15, 20, 25, 30, 40, 50, 60$ y 73. Cada comité consistió de 3 redes neuronales de tipo *back propagation*, donde además se aplicó regularización por detención temprana (sección 3.2.5.1). El entrenamiento con cada CRN era repetido 5 veces. El motivo de utilizar distintos valores para d se basa en la necesidad de intentar determinar la cantidad de descriptores apropiados para el modelo con el que se quiere modelar la propiedad experimental. Este enfoque es similar al utilizado en el software GALGO de Trevino & Falciani (2006).

Se dividió el conjunto de compuestos en los mismos subconjuntos de entrenamiento, validación y testeo, tal como se indica en el artículo de Yaffe *et al.* (2002), de manera de poder realizar una comparación directa con este trabajo. De esta forma, vemos que en comparación con la red neuronal propuesta en Yaffe *et al.* la cual utiliza 12 descriptores y obtiene un error absoluto medio (MAE) igual a 0,23, en la Tabla 5.10 vemos que nuestra red neuronal con la asistencia del *wrapper*, mejora la precisión en la predicción, incluso usando un descriptor menos (RNL y $p = 10$).

Tabla 5.10: Errores de predicción en términos de MAE y MSE en las 5 corridas de los CRN

	Método	*d = 11*	*d = 12*	*d = 15*	*d = 20*	*d = 25*	*d = 30*	*d = 40*	*d = 50*	*d = 60*	*d = 73*
MAE	*RNL, p = 10*	0,1972	0,1922	0,1896	0,1928	0,1708	0,172	0,1737	0,1855	0,2021	0,184
	AD, p = 20	0,2331	0,2362	0,2428	0,2114	0,1928	0,1742	0,1656	0,1626	0,1733	0,1847
	kVC, p =30	0,2562	0,2771	0,2986	0,2325	0,2079	0,1866	0,1925	0,1782	0,1827	0,1858
MSE	*RNL, p = 10*	0,074	0,0658	0,0738	0,0773	0,0566	0,0551	0,0593	0,0705	0,0895	0,065
	AD, p = 20	0,1013	0,116	0,1173	0,0903	0,0745	0,0539	0,0522	0,051	0,0521	0,0676
	kVC, p =30	0,1166	0,1266	0,151	0,1066	0,0781	0,073	0,0706	0,0613	0,0664	0,0665

De la Figura 5.3 vemos que, de las funciones de evaluación usadas, los AD tienen su mejor comportamiento en su variante con $p = 20$. Sin embargo, cuando

se los usa con valores chicos de d su comportamiento dista de ser bueno. En el caso de los kVC, valores bajos para p no resultan convenientes, al menos cuando se utilizan menos de 20 descriptores para el valor de d. Algo similar sucede con la RNL, cuando se usa $p = 30$. Como es de preveer, todos los modelos poseen resultados similares cuando más de 25 descriptores son considerados para la segunda fase. Esto se debe a que el conjunto intersección de dos subconjuntos con más de 25 descriptores tiene muchos de ellos en común, dado que son sólo 73 descriptores en total.

Considerando el mejor valor de p para cada alternativa de evaluación de variables (Figure 5.3-d). Destacamos el buen funcionamiento de la alternativa de RNL. La misma tiene un comportamiento bastante similar para todos los valores de d, tomando su valor mínimo para $d = 25$. El comportamiento de los kVC es similar al de RNL, excepto para los valores más bajos de d. En el caso de los AD, si bien se obtienen buenas predicciones para valores grandes de d, la mala performance con valores de d más chicos no hace a los AD tan interesantes como técnica de selección de descriptores, al menos para el presente ejemplo.

A fin de formalizar las conclusiones anteriores, analizamos si existen discrepancias entre las diferentes alternativas mediante la aplicación de un test de ANOVA doble (Tabla 5.11). Los dos factores involucrados en el ANOVA doble son: el método de evaluación y la elección de d. Nuestra comparación se centró en encontrar diferencias significativas sobre los métodos cuando se usan pocos descriptores para el CRN ($d = 11$, $d = 12$ y $d = 15$), ya que es para estos valores donde resulta más interesante analizar el comportamiento de las distintas alternativas. Al no haber una evidencia considerable de un factor de interacción (p =0,38), se está en condiciones de analizar ambos factores por separado. El test ANOVA muestra que no hay evidencias de diferencias en usar 11, 12 o 15 descriptores para un mismo método de evaluación (p =0,81), pero sí hay diferencias significativas entre la elección del método de evaluación usado en el wrapper. Aplicando la prueba de comparación múltiple de Bonferroni, se puede afirmar con un error global de 0,03 que todos los métodos difieren entre sí.

Por otra parte, a fin de evidenciar las ventajas y las diferencias en la aplicación de una técnica de selección de descriptores, analizamos los resultados obtenidos

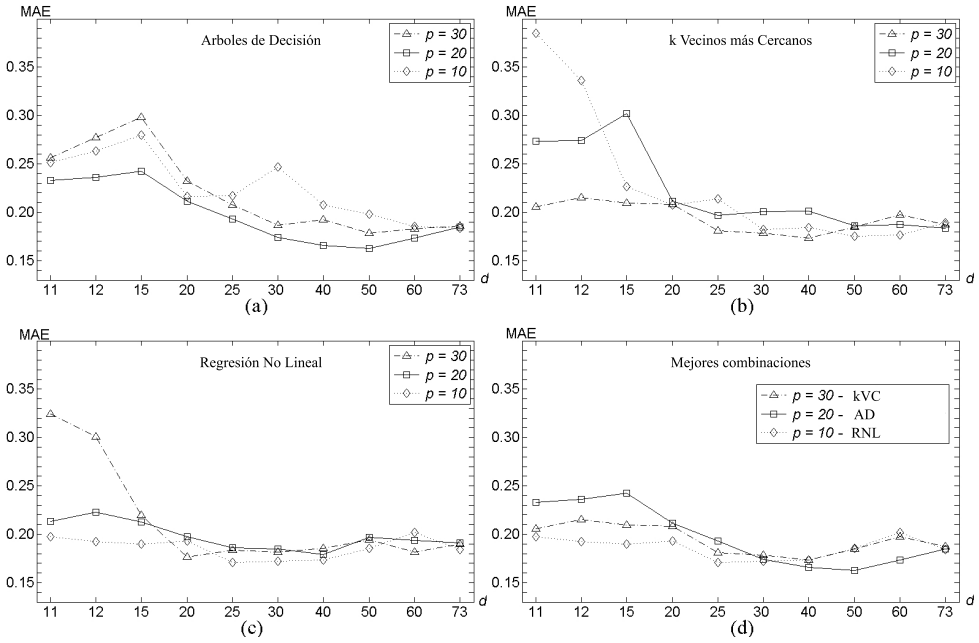


Figura 5.3: Error de predicción en términos de MAE considerando distinta cantidad de descriptores y con distintas alternativas para el *Evaluador de Variables*.

Tabla 5.11: ANOVA doble. Un factor es representado por el MAE de las tres mejores alternativas (subfigura d) de la Figura 5.3, mientras que el otro factor es el impacto en el uso de distintos valores de d ($d = 11$, $d = 12$ y $d = 15$)

F. de V.	S.C.	G.L.	C.M.	F	p
Entre	0,015629	8	0,0019536	14,7950419	2,622E-09
Factor d	0,000056	2	0,0000279	0,21117041	0,8106
Factor eval.	0,015003	2	0,0075014	56,8094222	7,306E-12
Interacción	0,000570	4	0,0001426	1,0797875	0,3809
Dentro	0,004754	36	0,0001320		
TOTAL	0,020383	44			

cuando las selecciones se hacen en forma aleatoria (Figura 5.4). Como era de esperar, el error en la predicción decrece cuantos más descriptores son considerados para el CRN. Esto es así ya que, nuevamente, todos los descriptores disponibles son potencialmente relevantes para la propiedad logP y además la regularización aplicada a los CRN, evita que se empeore la calidad de la predicción. Sin embargo, el error es considerablemente más alto en comparación con las distintas estrategias del *wrapper* cuando se usan menos de 25 descriptores.

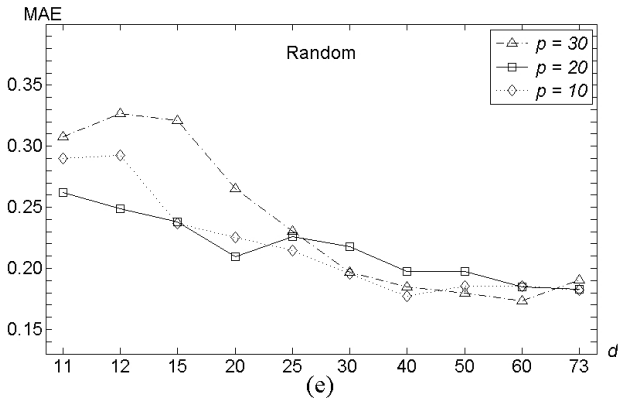


Figura 5.4: Error de predicción en términos de MAE considerando distinta cantidad de descriptores y distintas selecciones aleatorias.

A modo de conclusión sobre esta última experimentación mono-objetivo, destacamos que esta versión mejora ciertas deficiencias de la anterior. En la sección 5.4 veremos algunas limitaciones de esta propuesta, las cuales dieron lugar a

la versión multi-objetivo. Los puntos salientes de esta segunda experimentación mono-objetivo fueron publicados en Soto *et al.* (2008b).

5.4. 2º Propuesta: Selección de descriptores utilizando algoritmos evolutivos multi-objetivo

En base a la experiencia obtenida en el diseño del algoritmo evolutivo mono-objetivo, se procedió a realizar una mejora sustancial de la metodología anterior. Dado que esta propuesta constituye la mejor alternativa presentada en esta tesis para la selección de descriptores, realizaremos una descripción más detallada de la técnica, como también ofreceremos un análisis más exhaustivo de los resultados alcanzados. Un explicación detallada de esta metodología multi-objetivo junto con los resultados alcanzados puede consultarse en Soto *et al.* (2009b).

La primera desventaja de la propuesta anterior, es la necesidad de fijar el número p con el que el algoritmo debe trabajar. Esta característica obligaba a probar con distintos valores de p a fin de poder armar los diferentes rankings. Sin embargo, si este valor no se fija el espacio combinatorial disponible es demasiado grande y el proceso no converge a soluciones interesantes.

Por otra parte, la hipótesis de tomar diferentes cantidades de descriptores a partir de un ranking, además de obligar a realizar una importante cantidad de corridas puede derivar en situaciones indeseables. Esto se debe a que el ranking sólo tiene en cuenta su frecuencia individual, en lugar de ser considerada junto con el grupo de variables con la que fue seleccionada. Para ejemplificar esta situación, supongamos que tenemos dos variables que son relevantes pero altamente correlacionadas (en forma lineal o no lineal). Ambas pueden haber sido elegidas en diferentes oportunidades y, por lo tanto, tener una buena posición en el ranking. Sin embargo, sólo una de ellas debería estar en el modelo final de predicción. En definitiva queremos enfatizar que, para tareas de predicción, un descriptor no debe ser evaluado en forma independiente, sino que se debe considerar todo el subconjunto en su totalidad.

Por tal motivo, ambas fases de la propuesta anterior fueron sustancialmente modificadas. La primera fase consiste ahora en un método *wrapper* multi-objetivo,

el cual apunta a maximizar la capacidad predictiva de los subconjuntos de descriptores seleccionados y al mismo tiempo a minimizar la cantidad de descriptores seleccionados. Este enfoque multi-objetivo brinda dos ventajas simultáneas. Primero, favorece la selección de subconjuntos con menor cardinalidad. Esto es importante ya que puede demostrarse que el incremento lineal en el número de variables seleccionadas conlleva a un incremento exponencial en el número de hipótesis de aprendizaje posibles (Liu & Motoda (2008)). En segundo lugar, la cantidad de descriptores a seleccionar queda determinada de manera automática y sin la necesidad de probar iterativamente con diferentes tamaños de subconjuntos. En la Figura 5.5 podemos apreciar un esquema del nuevo *wrapper* multi-objetivo. Nótese la diferencia con el esquema de la propuesta anterior (Figura 5.1).

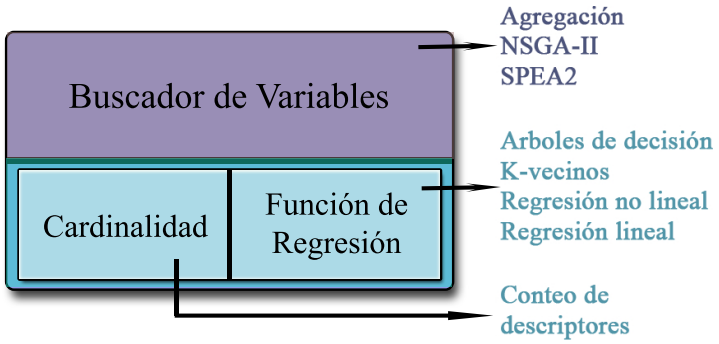


Figura 5.5: Componentes del *wrapper* multi-objetivo.

La segunda fase, al igual que en el caso anterior, es aplicada para establecer una evaluación más precisa sobre cuál es la capacidad predictiva de un subconjunto de descriptores dado. Sin embargo, en este caso un porcentaje de las selecciones obtenidas en la primera fase van a ser reevaluadas por el método de predicción de la segunda fase, pero con la diferencia que los descriptores no se los extrae en forma aislada, sino que se evalúa todo el subconjunto en su totalidad, tal como fueron seleccionados en la primera fase.

5.4.1. Primera Fase: *Evaluador de Variables* Multi-Objetivo

Tal como se comentó en la sección anterior, se introduce en el *Evaluador de Variables* multi-objetivo la minimización del subconjunto de descriptores selec-

cionados como un objetivo adicional. Por lo tanto, el objetivo relacionado a la evaluación de la calidad predictiva es el mismo que el descripto para el caso mono-objetivo. El objetivo nuevo consistirá simplemente en minimizar la función que cuenta la cantidad de variables seleccionadas. Definiremos F_1 como la función a minimizar para el primer objetivo (Ecuación 5.2), y F_2 como la función que cuenta la cantidad de variables consideradas en una selección. En el caso de nuestro enfoque multi-objetivo y utilizando la representación binaria presentada en la sección 5.3.2.1, la función F_2 simplemente contará la cantidad de locaciones con valor uno para cada cromosoma.

En tal sentido, usaremos todos los métodos de regresión antes presentados para el cálculo de F_1, es decir: AD, kVC, RNL y RL. Del mismo modo, realizamos en esta primera fase una separación de los datos de entrenamiento (Z_1) y los usados para validación (Z_2). La evaluación de la capacidad predictiva es siempre reportada considerando el desempeño sobre este último subconjunto. De esta manera se evita que el *wrapper* sobreajuste los datos usados para entrenar. Esta separación se realiza en forma aleatoria, y se repite una vez por cada corrida del algoritmo genético.

5.4.2. Primera Fase: *Buscador de Variables* multi-objetivo

Teniendo en cuenta la naturaleza multi-objetivo del nuevo enfoque, el *Buscador de Variables* debe ahora poder ser guiado por más de un objetivo. Dentro de los métodos de búsqueda multi-objetivo, los enfoques más frecuentemente utilizados son los basados en algoritmos evolutivos (Coello Coello (2006)). Siguiendo los conceptos de la literatura, un algoritmo evolutivo multi-objetivo puede aplicarse siguiendo una estrategia de Agregación o un enfoque basado en dominancia de Pareto (Deb (2004)). Una diferencia importante entre estos métodos reside en cómo se establece la aptitud global de un individuo, el cual representa un subconjunto de descriptores en este caso, en función de los multiples objetivos considerados.

En el caso de los de Agregación, los múltiples objetivos se integran en uno solo (sección 3.5.1.1). Este enfoque está basado en un antiguo método de optimización y que puede derivarse de las condiciones descriptas en Kuhn & Tucker (1951). Este

tipo de técnicas suelen ser aplicadas cuando se pretende atacar ciertos problemas de optimización combinatorial (Coello Coello *et al.* (2007); Jaszkiewicz (2004)).

Para el presente estudio, proponemos la Función 5.3 como función de fitness de agregación. En esta ecuación utilizamos un parámetro α que pondera cada uno de los objetivos, con $0 \leq \alpha \leq 1$, y donde p_m es un parámetro que representa un límite superior para la cardinalidad máxima de un subconjunto. El primer término de la función de fitness devuelve el error de predicción obtenido para una determinada elección de $\mathcal{P}$ y el segundo término refleja la proporción de descriptores elegidos con respecto al total p_m escalado por F_1. Vale destacar que esta estrategia de agregación puede también ser vista como un procedimiento de regularización que efectúa un balance entre la complejidad del modelo (caracterizado por el número de descriptores usados) y la precisión en la predicción. Además podemos ver que α controla la cardinalidad de los subconjuntos elegidos.

$$F_{AG} = \alpha F_1 + (1 - \alpha) F_1 \frac{F_2}{p_m} \tag{5.3}$$

El otro enfoque multi-objetivo aplicado utiliza el concepto de dominación. La dominancia es un orden parcial que puede ser establecido entre vectores definidos en un espacio $\mathbb{R}^k$ (sección 3.5.1), donde k es el número de objetivos a optimizar. En nuestro caso, cada individuo está asociado a un vector en $\mathbb{R}^2$ tal que su primera componente es F_1 y la segunda F_2. A los algoritmos de optimización multi-objetivo que utilizan el concepto de dominación en el mecanismo de selección de individuos para mover la población hacia el frente de Pareto se los denomina algoritmos basados en dominancia de Pareto.

En este contexto, para los métodos de búsqueda basados en Pareto, se decidió experimentar con los dos algoritmos multi-objetivo basados en Pareto que son tomados como referentes en la literatura: NSGA-II (Deb *et al.* (2002)) y SPEA2 (Zitzler *et al.* (2002)). En el capítulo 3.5.1 se encuentra detallado el funcionamiento de estos algoritmos.

En conclusión, para esta propuesta multi-objetivo se tienen tres métodos de búsqueda diferentes, a saber: agregación, NSGA-II y SPEA2. Es importante destacar la diferencia en la relación de la función de fitness con cada una de las

metodologías de búsqueda. En el caso de la estrategia de agregación ambos objetivos (F_1 y F_2) son combinados en una única función de fitness, mientras que en el caso de las basadas en Pareto, existe una función de fitness por cada uno de los objetivos.

5.4.2.1. Diseño del algoritmo evolutivo multi-objetivo

La población inicial es generada al azar, en donde el número de bits no nulos debe ser forzado a estar entre 0 y p_m. Para los operadores genéticos se aplicó el cruzamiento de un punto y el operador de mutación clásico. Cuando como resultado de una de estas operaciones el número de descriptores seleccionados de un individuo es mayor que p_m, se seleccionan al azar bits no nulos para setearlos a cero, hasta que el número de descriptores seleccionados sea igual a p_m. Este tipo de restricciones de dominio es comúnmente aplicado en problemas de optimización, cuando se requiere disminuir el espacio de búsqueda y evitar desperdiciar ciclos de CPU en soluciones que no van a resultar interesantes para el problema (Horvath *et al.* (2007); Michalewicz & Schoenauer (1996)).

El esquema de selección aplicado depende del método de búsqueda usado en el *wrapper* multi-objetivo. Para la estrategia de agregación, por iguales motivos que los establecidos en el caso mono-objetivo, se utilizó el método del torneo. Para las estrategias basadas en Pareto, los operadores de selección corresponden con los establecidos para NSGA-II y SPEA2. Vale destacar que en todos los casos se utiliza el concepto de elitismo.

5.4.3. Segunda Fase: refinamiento y evaluación de los subconjuntos encontrados

Luego de aplicarse una combinación de métodos de búsqueda y evaluación para el *wrapper* multi-objetivo, se conforma un frente de individuos no dominados. Es importante aclarar que la conformación de este frente de soluciones no dominadas es llevada a cabo en forma independiente de si el método de búsqueda es basado en Pareto o no. En tal sentido, todos los subconjuntos no dominados obtenidos en una misma corrida son tratados como el conjunto de soluciones más

interesantes obtenidas por el *wrapper* en esa corrida. Esta selección se indica en la Figura 5.6 como 'Selección gruesa de subconjuntos'.

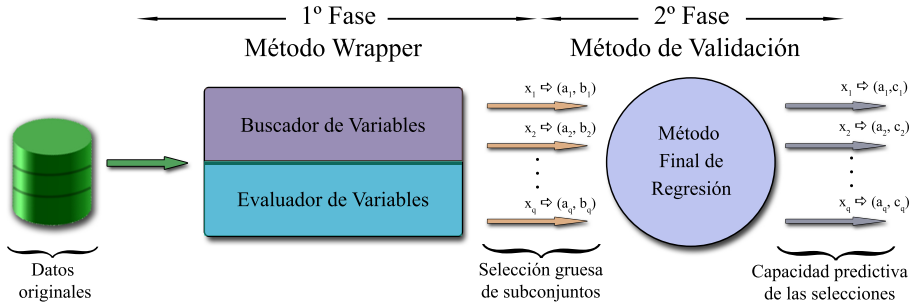


Figura 5.6: Esquema de selección de variables de 2 fases.

Cada subconjunto de descriptores contenido en el frente es evaluado con el método de la segunda fase. En este caso también utilizamos CRN, al igual que en la primera propuesta, con la misma finalidad: que la precisión obtenida usando las redes neuronales sea mayor que la precisión de los métodos de regresión del *wrapper*. Con fines comparativos, para esta segunda fase también utilizamos un método de regresión lineal simple.

Por otra parte, si bien el uso de CRN en esta segunda fase otorga más precisión en la predicción, esta fase representa una evaluación más rigurosa de la capacidad predictiva, ya que el mismo método de regresión es aplicado en forma repetida para un mismo subconjunto de descriptores. Este proceso de evaluación involucra la aplicación de una validación cruzada de f-particiones (*folds*) aplicada r veces, donde en cada iteración cada *fold* es obtenido en forma aleatoria. Este procedimiento también es conocido como validación cruzada de Monte Carlo. A fin de disminuir la carga computacional asociada a esta evaluación de cada subconjunto del frente de no dominados, se aplica una heurística que realiza esta validación de manera iterativa. En la primera iteración, la capacidad predictiva de todos los subconjuntos de la selección gruesa es evaluada de la manera anteriormente descripta. En la iteración siguiente, sólo un porcentaje formado por los subconjuntos que resultaron más relevantes, es nuevamente validado y así sucesivamente. La característica saliente de este proceso es que r es inicializado con un valor bajo y es incrementado en cada iteración.

Es importante destacar que la aplicación de grandes cantidad de réplicas en un procedimiento de validación cruzada por parte de los métodos de regresión del *wrapper* no resultaría viable por razones de costo computacional. Por tal motivo, esta replicación sólo puede ser aplicada para la segunda fase.

5.4.4. Resultados

5.4.4.1. Diseño de los experimentos

Aplicamos la presente metodología de selección de descriptores a los conjuntos de datos, indicados en el apéndice como B.3, B.4 y B.2, los cuales llamaremos DS1, DS2 y DS3 respectivamente. Los parámetros del algoritmo evolutivo se mantuvieron iguales para todos los sets de datos y funciones de búsquedas. El tamaño de la población fue fijado a 145 individuos, la probabilidad de cruzamiento se seteó en 0,75 y la probabilidad de mutación se estableció en $2/n$. Al igual que en el caso anterior, se utilizó un criterio fenotípico para parar el algoritmo, el cual consiste en detenerlo si durante 15 generaciones la variación de la función de fitness promedio de toda la población es menor que un valor de tolerancia ($\xi = 10^{-16}$). Además, el número máximo de generaciones se seteó en 200. En particular para la estrategia de agregación el tamaño del torneo fue de 4 y el número de individuos de elite por generación fue fijado a un valor de 5.

Podemos apreciar que el número de posibles combinaciones de *wrappers* multi-objetivo es de 12, el cual proviene de las tres métodos de búsqueda (estrategia de agregación, NSGA-II, SPEA2) y de las cuatro métodos de regresión (AD, kVC, RNL, RL). Realizamos 10 corridas del *wrapper* para cada una de estas posibles combinaciones. En cada corrida, sólo retuvimos las soluciones que pertenecían al frente de los subconjuntos no dominados (tal como se explicó en la sección 5.4.3). En la segunda fase, tomamos cada subconjunto del frente de no dominados y le aplicamos el método de regresión de la segunda fase (CRN y RL) utilizando una validación cruzada de f-particiones (*folds*), repitiendo esta validación tres veces y obteniendo el promedio de estas tres corridas. Las 20 mejores selecciones (aquellas con valor de MSE más bajo) eran retenidas y, nuevamente, se le aplica una validación cruzada de f-particiones, pero repitiendo esta validación 50 veces,

de manera que se logre una estimación más confiable de la capacidad predictiva de los subconjuntos de descriptores seleccionados.

El objetivo de estos experimentos era comparar nuestros modelos con los modelos reportados en los artículos de donde extrajimos los datos (Konovalov *et al.* (2008); Yaffe *et al.* (2002)). Además, también nos interesaba comparar la capacidad predictiva de nuestros subconjuntos de descriptores con la de los descriptores obtenidos a partir de un método de selección de variables bayesiano el cual usa un método de regresión con una probabilidad *a priori sparse* (en este caso la probabilidad *a priori* de Jeffreys) tal como se describe en Figueiredo (2003). Este enfoque fue elegido por sus buenos resultados y por ser un método usado recientemente para la selección de descriptores en QSAR (Burden & Winkler (2009)).

Para esta segunda fase utilizamos un comité de 5 redes neuronales. La arquitectura de cada red dependió del tamaño del conjunto de datos y de su complejidad. Todas las redes usadas para esta experimentación fueron entrenadas usando el algoritmo de Levenberg-Marquardt con un procedimiento de regularización bayesiana (sección 3.2.5.1).

En relación a la arquitectura de las redes neuronales, utilizamos redes de dos capas donde el número de nodos ocultos era establecido de acuerdo a la optimización para el mejor subconjunto de descriptores reportado en los artículos de donde extrajimos la información de los sets de datos (Konovalov *et al.* (2008); Yaffe *et al.* (2002)). En consecuencia, para un mismo set de datos, el número de nodos ocultos se mantuvo constante mientras se buscaban los subconjuntos de descriptores. Si bien una arquitectura de red fija podría considerarse como un error de diseño, más que proveer un modelo de regresión perfectamente optimizado, la idea es mantener la posibilidad de comparar y al mismo tiempo ahorrar tiempo de diseño mientras se obtiene el objetivo de determinar cuáles serían los subconjuntos de descriptores más relevantes. Es importante destacar que el diseño de la arquitectura de las redes con regularización bayesiana tiene un impacto mínimo en la calidad de predicción, siempre que la cantidad de neuronas ocultas sea suficiente (Burden & Winkler (1999)). Como resultado de todo este proceso, se utilizaron redes con dos nodos en la capa oculta para DS1 y DS2, mientras que fueron 5 los nodos de la capa oculta para DS3. Asimismo, las entradas y

las salidas presentadas a la red neuronal fueron estandarizadas de manera que la regularización bayesiana pueda ser correctamente aplicada (a diferencia del resto de las redes neuronales en donde sólo la entrada es estandarizada).

En el trabajo de Konovalov *et al.* (2008) se reporta la capacidad predictiva de los subconjuntos encontrados en función del error de predicción usando un método de regresión lineal. Realizar una comparación directa de nuestros resultados usando CRN con los de Konovalov *et al.* puede pensarse que no sería del todo justo. Sin embargo, es importante destacar que los CRN pueden usarse en nuestra estrategia, debido a que primero se hace una preselección rápida de los subconjuntos potencialmente relevantes, y sólo un número reducido de subconjuntos son dejados para la evaluación minuciosa con el CRN. Aplicar CRN en otras estrategias de selección de variables como la de Konovalov *et al.* (*i.e.* en una única fase de búsqueda y evaluación) sería inviable computacionalmente. Por este motivo decidimos también incluir RL en la segunda fase a fines comparativos.

Con respecto a la comparación con el método de Figueiredo (2003), se aplicó un procedimiento análogo al aplicado para nuestra propuesta evolutiva. Se empleó un conjunto de validación (el equivalente a Z_2) a fin de determinar el parámetro σ, el cual controla la ralidad del método. Se realizaron 20 corridas de este algoritmo, donde una nueva separación de los conjuntos de entrenamiento-validación se aplicaba antes de cada corrida. No se aplicó transformación alguna a la matriz de datos.

Para el set de datos DS1 se tomó una consideración especial para garantizar que el descriptor 'Iv' esté siempre considerado para cualquier modelo. Esta decisión se basa en el hecho que, según Konovalov *et al.* (2008), existe una diferencia sistemática de aproximadamente 0,5 unidades de logBB entre los valores experimentales obtenidos *in vivo* e *in vitro*. Sobre la base de los subconjuntos reportados en este último trabajo, seteamos $p_m = 20$. En concordancia con el trabajo de Konovalov *et al.*, para las corridas del *wrapper*, el tamaño del conjunto Z_1 se fijó igual al tamaño de Z_2. Para el procedimiento de validación cruzada de Monte Carlo aplicado en la segunda fase del método, el número de particiones se seteó en 2 (50 % para entrenamiento, 50 % para testeo). Las mismas consideraciones y parámetros fueron mantenidos para el set de datos DS2.

Para el tercer set de datos, se calculó que el número estimado de descriptores necesarios sería más alto que en los dos casos anteriores, por lo tanto se fijó $p_m = 50$. En el trabajo de Yaffe *et al.* (2002) se usó una partición fija para el testeo con el 17 % de los datos. Se puede demostrar que esta partición está prácticamente contenida dentro del volumen descripto por los datos de entrenamiento/validación, lo cual no permite una evaluación no sesgada de la capacidad predictiva del modelo. Por lo tanto, intentando mantener una analogía con el trabajo de referencia, se aplicó una validación cruzada de 5 particiones para evaluar los subconjuntos de descriptores en la segunda fase (es decir manteniendo un 20 % de los datos en cada iteración para testeo). Del mismo modo para el *wrapper*, Z_1 y Z_2 comprenden el 80 % y el 20 % de los datos respectivamente.

5.4.4.2. Análisis de los mejores subconjuntos encontrados

En las Tablas 5.12, 5.13 y 5.14 se muestra la información de los mejores subconjuntos seleccionados, donde en cada columna se describe: su método de obtención, la cardinalidad, los errores obtenidos en la segunda fase (MSE y q_2), el número de pesos y el número de pesos efectivos de la red neuronal promedio; discriminado por set de datos. En cada tabla se muestra el mejor subconjunto de descriptores reportado en los artículos sobre los cuáles se extrajo el set de datos (primera fila), el mejor subconjunto encontrado aplicando el método de Figuereido (segunda fila) y los mejores subconjuntos obtenidos con nuestro método multiobjetivo. Para el caso de nuestro método, se muestran dos selecciones aplicando la estrategia de agregación (tercera y cuarta fila) y dos aplicando el enfoque basado en Pareto (quinta y sexta fila). En la Tabla 5.15 se enumeran todos los descriptores seleccionados para las Tablas 5.12, 5.13 y 5.14.

Para el caso del primer conjunto de compuestos DS1 (Tabla 5.12) se ve que el subconjunto III tiene una capacidad predictiva superior a la reportada en el trabajo de Konovalov (subconjunto I) usando la misma cantidad de descriptores. El subconjunto obtenido usando el método de Figuereido (subconjunto II) es ligeramente superior al subconjunto I, cuando se usa una RL para la validación, pero el número de descriptores usados es mayor que en cualquier otro subconjunto considerado. Usando más descriptores que en el subconjunto I, encontramos que

Tabla 5.12: Comparación de los mejores subconjuntos de descriptores obtenidos para DS1. Las columnas de MSE y q^2 se refieren a los resultados obtenidos sobre la partición de testeo en la segunda fase. N_w/N_{eff} es el número de pesos de la red neuronal y el número de pesos efectivos en el modelo.

Subconjunto	Buscador de Vars.	Evaluador de Vars.	Cardinalidad	Método de Validación	MSE	q^2	N_w/N_{eff}
I	MCVS	RL	6	CRN	0.1265	0.6450	17/13
				RL	0.1225	0.6752	—
II	Priori de Jeffreys	RL	20	CRN	0.1302	0.6528	45/34
				RL	0.1210	0.6757	—
III	MO-Agreg, α=0,3	RL	6	CRN	0.1205	0.6816	17/15
				RL	0.1281	0.6525	—
IV	MO-Agreg, α=0,7	RL	15	CRN	0.1103	0.7198	35/31
				RL	0.1113	0.7030	—
V	NSGA-II	RL	8	CRN	0.1140	0.6993	21/18
				RL	0.1178	0.6727	—
VI	NSGA-II	RL	11	CRN	0.1052	0.7352	27/23
				RL	0.1124	0.6821	—

los subconjuntos IV, V y VI poseen una capacidad predictiva superior, independientemente de si son validados con un CRN o con una RL.

En la Tabla 5.13 vemos que, aunque no se encontraron subconjuntos con una capacidad de predicción estrictamente superior que la propuesta por Konovalov (subconjunto VII), se encontraron subconjuntos igualmente interesantes. Los subconjuntos IX, X y XII tienen una capacidad predictiva levemente inferior que la del VII, sin embargo esta última utiliza muchos menos descriptores. El subconjunto XI tiene una capacidad de predicción comparable al VII usando un descriptor menos. Cuando se aplica RL para evaluar la capacidad predictiva, todas las predicciones son ligeramente peores en comparación con el MSE usando CRN. El subconjunto obtenido usando el método de Figuereido (subconjunto VII) fue superado por todos los subconjuntos presentados para este set de datos.

Es importante decir que a diferencia del trabajo de Konovalov *et al.* (2008), no preseleccionamos ningún descriptor, excepto por 'Iv' en el primer set de datos, el cual debe necesariamente ser considerado por alterar las condiciones sobre las cuales se midió la variable experimental.

Los resultados para el tercer set de datos son motivantes ya que la complejidad del modelo para predicción de la hidrofobicidad en términos de los descriptores seleccionados es mayor que la presente en los dos sets de datos anteriores. Sin

Tabla 5.13: Comparación de los mejores subconjuntos de descriptores obtenidos para DS2. Las columnas de MSE y q^2 se refieren a los resultados obtenidos sobre la partición de testeo en la segunda fase. N_w/N_{eff} es el número de pesos de la red neuronal y el número de pesos efectivos en el modelo.

Subconjunto	Buscador de Vars.	Evaluador de Vars.	Cardinalidad	Método de Validación	MSE	q^2	N_w/N_{eff}
VII	MCVS	RL	8	CRN	0,1191	0,6813	21/17
				RL	0,0900	0,7532	—
VIII	Priori de Jeffreys	RL	4	CRN	0.1715	0.5733	13/7
				RL	0.1380	0.6404	—
IX	MO-Agreg, α=0,1	RL	3	CRN	0.0984	0.7421	11/9
				RL	0.1282	0.6500	—
X	MO-Agreg, α=0,3	AD	3	CRN	0.1055	0.7092	11/9
				RL	0.1512	0.5700	—
XI	NSGA-II	RL	7	CRN	0.0915	0.6459	19/16
				RL	0.1112	0.6623	—
XII	NSGA-II	kVC	2	CRN	0.1013	0.6174	9/7
				RL	0.1374	0.6186	—

embargo, las comparaciones con el trabajo original son difíciles de establecer dado que diferentes métodos de predicción y de validación fueron considerados en cada caso. Además, incorporamos descriptores no considerados en el trabajo original, por lo tanto los descriptores de DS3 no pueden ser directamente comparados con nuestros subconjuntos de descriptores. Sin embargo, a fin de poder cuantificar las diferencias entre ambas propuestas, hemos tomado los descriptores elegidos en el trabajo de Yaffe *et al.* (2002) (subconjunto XIII) y aplicamos la misma validación y los mismos métodos de predicción que aplicamos a nuestros subconjuntos.

Se puede observar que cualquiera de los subconjuntos de descriptores propuestos por el algoritmo multi-objetivo (subconjuntos XV, XVI, XVII y XVIII) mejora la capacidad predictiva del subconjunto XIII (Tabla 5.14). Si bien es cierto que nuestros subconjuntos aumentan la cantidad de descriptores usados, la diferencia en términos de MSE es importante. La capacidad predictiva del subconjunto XIV (usando el método de Figuereido) con el CRN es superior al de Yaffe, pero considerablemente inferior a los subconjuntos obtenidos usando el enfoque evolutivo.

Para este último set de datos, se puede apreciar también que el MSE usando RL en la segunda fase es ampliamente superior al MSE usando CRN. En consecuencia, puede inferirse que la relación entre los descriptores involucrados y la

Tabla 5.14: Comparación de los mejores subconjuntos de descriptores obtenidos para DS3. Las columnas de MSE y q^2 se refieren a los resultados obtenidos sobre la partición de testeo en la segunda fase. N_w/N_{eff} es el número de pesos de la red neuronal y el número de pesos efectivos en el modelo.

Subconjunto	Buscador de Vars.	Evaluador de Vars.	Cardinalidad	Método de Validación	MSE	q^2	N_w/N_{eff}
XIII	AG	RL	12	CRN	0.247	0.884	71/61
				RL	0.2900	—	—
XIV	Priori de Jeffreys	RL	16	CRN	0.2052	0.9097	91/80
				RL	0.2724	0.8804	—
XV	MO-Agreg, α=0,9	RL	24	CRN	0.1540	0.9297	131/95
				RL	0.2860	0.8795	—
XVI	MO-Agreg, α=0,1	RL	13	CRN	0.164	0.9317	76/64
				RL	0.2617	0.8698	—
XVII	SPEA2	RL	15	CRN	0.1778	0.9135	86/74
				RL	0.2990	0.8649	—
XVIII	NSGA-II	RL	20	CRN	0.1696	0.9240	111/94
				RL	0.3426	0.8496	—

variable experimental logP es altamente no lineal. Esto mismo no sucede de la misma manera para los sets de datos anteriores, en donde la diferencia de error de predicción entre los distintos métodos de la segunda fase no son tan grandes.

5.4.4.3. Análisis y comparación de los diferentes *wrappers* multi-objetivos aplicados

Presentaremos ahora una comparación del desempeño entre las distintas alternativas para el *wrapper* multi-objetivo. La Tabla 5.16 muestra cuál es la estrategia de evaluación y de búsqueda que mejor funcionó para detectar subconjuntos con buena capacidad predictiva para cada set de datos. A fin de eliminar el efecto negativo de las selecciones con muy mala capacidad predictiva, se decidió analizar solamente el percentil 50 que conteniene las mejores selecciones. Por cuestiones prácticas, no expondremos todas las tablas de los resultados estadísticos (ANOVA, comparaciones múltiples) sino que simplemente enunciaremos los resultados obtenidos, haciendo mención del método estadístico utilizado.

En primer lugar, aplicamos pruebas de comparaciones múltiples para poder comparar el desempeño de las distintas combinaciones de métodos. Utilizamos para esta tarea, la prueba de Tukey-Kramer con un error global del 5 %. Al

Tabla 5.15: Descriptores moleculares seleccionados para cada subconjunto de las Tablas 5.12, 5.13 y 5.14. Los nombres completos de los descriptores pueden encontrarse en el sitio web de E-DRAGON (2007). Los corchetes denotan descriptores preseleccionados antes de aplicar el método correspondiente.

Subconjunto	Descriptores
I	[Iv] , [TPSA(NO)], [Ic], SRW09, BELv4, HATS7v / HATS8e
II	[IV], TPSA(NO), MLOGP2, SRW09, nROH, EEig05d, C-034, nRCOOH, O-057, nArNR2, nArCOOH, H-051, Psychotic-80, nRCOOR, HATS8u, nN(CO)2, Infective-50, HATS7u, nArCOOR, Deppresant-50
III	[Iv], TPSA(NO), ALOGP, Mor21v, EEig12r, Ic
IV	[Iv], TPSA(NO), SRW07, O-057, MATS2p, nOHp, R3u, RDF020p, T(N..I), nArOH, Psychotic-80, RDF050m, HATS8u, Cl-087, G3u
V	[Iv], TPSA(NO), ALOGP, Mor16v, ISIZ, nN(CO)2, BELm4, Ic
VI	[Iv], TPSA(NO), ALOGP, R7u+, ARR, H4p, Mor20v, TE2, BELe3
VII	[ALOGP], LAI, Neoplastic-80, RDF045m, R5v+, DDI, N-074, IDE
VIII	Hy, GVWAI-80, Infective-80, nP(=O)O2R
IX	MLOGP, TPSA(NO), RDF130m
X	ALOGP, C-011, nPyridines
XI	ALOGP, Mor13e, RDF045p, Vs, nArCONHR, R5u+, PW5
XII	ALOGP, R1v+
XIII	MW, D_P, D_H, D_S, E2, EX, ELC, IP, PO, VMC1, VMC2, VMC4
XIV	D_H, IP, PO, Mv, nH, nN, nCL, nR06, nCp, nCaR, nOHp, nOHs, nROR, nRSR, nHDon, PSA
XV	D_S, E2, IP, Sv, Se, nAT, nBM, nDB, nAB, nH, nC, nO, nF, nCL, nI, nR03, nR06, nR11, nCp, nCs, nOHp, nOHs, nOHt, ARR
XVI	D_H, IP, PO, nBT, nBM, nAB, nO, nF, nX, nR06, nCaR, nHAcc, Ui
XVII	MW, E2, AMW, Mv, Mp, nBT, SCBO, nAB, nC, nBR, nCp, nCaR, nOHs, nHDon, ARR
XVIII	MW, E2, PO, VMC2, AMW, Mv, nBM, SCBO, nDB, nH, nC, nBR, nR03, nR06, nCp, nCaR, nOHp, nOHs, nHDon, ARR

analizar los sets de datos DS1 y DS2, encontramos que el método de evaluación de RL fue significativamente superior que los restantes y que, en la mayoría de los casos, las estrategias basadas en Pareto tenían un mejor desempeño que las basadas en agregación. Asimismo, no se encontraron diferencias entre NSGA-II y SPEA2 para DS1.

Al analizar las estrategias de búsqueda para DS3, encontramos que la estrategia de agregación es la mejor independientemente del método de regresión aplicado. Por otra parte, cuando examinamos los métodos de evaluación, detectamos que RL es el mejor cuando se aplica NSGA-II; mientras que RL y kVC son significativamente mejores que el resto (sin diferencias significativas entre ellos) cuando los restantes métodos de búsqueda son aplicados.

Tabla 5.16: MSE promedio del percentil 50 de todas los subconjuntos seleccionados por el *wrapper* (*i.e.* Selección gruesa en Figura 5.6).

Set de Datos	Buscador de Vars.	AD	kVC	RNL	RL
DS1	Agregación	0,1504	0,1486	0,1462	0,1261
	NSGA-II	0,1437	0,1382	0,1454	0,1277
	SPEA2	0,1385	0,1361	0,1368	0,1269
DS2	Agregación	0,1211	0,1285	0,1212	0,1161
	NSGA-II	0,1049	0,1052	0,105	0,101
	SPEA2	0,1018	0,1104	0,1064	0,0982
DS3	Agregación	0,1881	0,1855	0,1877	0,1787
	NSGA-II	0,2645	0,2592	0,3080	0,2222
	SPEA2	0,2120	0,2067	0,2073	0,1963

Por otra parte, analizamos también cuál de las partes del *wrapper* tiene mayor impacto en el desempeño del algoritmo, es decir que, deseamos determinar si el *Evaluador de Variables* o el *Buscador de Variables* tiene una mayor influencia en la capacidad predictiva de los subconjuntos de descriptores obtenidos. Para esto, aplicamos pruebas de ANOVA doble para analizar la contribución a la varianza del error en la predicción que cada factor tiene. Claramente, un factor corresponde a los métodos de evaluación, mientras que el otro factor lo constituye las diferentes estrategias de búsqueda. Para los sets de datos DS1 y DS2, detectamos que la selección del método de regresión tiene el impacto más fuerte (con una contribución a la varianza de 95,8 % y 56,1 % para DS1 y DS2 respectivamente). En el caso de DS3, la selección de la estrategia de búsqueda es la parte que más influye en la capacidad predictiva de los subconjuntos (con una contribución a la varianza del 55,3 %).

En vista del hecho que los métodos basados en Pareto se comportan mejor sólo para los primeros dos sets de datos, entendemos que la causa de esta situación se debe al hecho que DS1 y DS2 requieren menos descriptores para obtener un buen modelo de predicción que en el caso de DS3. Esto sucede porque las estrategias basadas en Pareto buscan minimizar cada objetivo por separado, independientemente del valor alcanzado para los otros objetivos. En otras palabras, un subconjunto que tiene menos descriptores que cualquier otro individuo de la población estará en el frente de no dominados, incluso cuando su capacidad predictiva sea baja. Esta particularidad hace a estos métodos más propensos a encontrar subconjuntos de baja cardinalidad. Para ejemplificar este hecho, la car-

dinalidad promedio de todos los subconjuntos obtenidos usando NSGA-II para DS3 es 12,71, mientras que usando el enfoque de agregación y α=0,1 (el valor que más pondera la obtención de subconjuntos con baja cardinalidad) es 19,98. En este sentido, cuando la cantidad de descriptores "óptima" necesaria es alta, los métodos basados en Pareto tienen que lidiar con un espacio de posibles selecciones óptimas mucho más amplio en relación con el conjunto de posibles selecciones óptimas que la estrategia de agregación considera.

Por otra parte, teniendo en cuenta los buenos resultados obtenidos con RL en el *wrapper* multi-objetivo, este resultado no es sorprendente para los sets de datos DS1 y DS2, ya que la relación entre los descriptores y sus valores experimentales es altamente lineal. Esto se deduce a partir del buen desempeño obtenido con RL en la segunda fase, la cual no dista demasiado de la capacidad predictiva alcanzada por los CRN. Sin embargo para DS3, el buen desempeño de la RL en el *wrapper* es interesante, ya que demuestra que métodos lineales pueden ser útiles para detectar relevancia en forma rápida para la primera fase, incluso cuando la relación con la variable a predecir sea no lineal.

Finalmente, queremos ilustrar cómo la elección del parámetro α de la estrategia de agregación influye en la cardinalidad y la capacidad predictiva de los subconjuntos encontrados. En la Figura 5.7 se ilustra el comportamiento promedio de los subconjuntos de descriptores seleccionados por el *wrapper* y evaluados en la segunda fase con el CRN (primera iteración de su proceso de validación). Se puede apreciar que existe un situación contrapuesta entre estos dos objetivos, pero considerando que la capacidad de predicción es más importante que la cardinalidad en un modelo (siempre que esté validado correctamente), la elección del valor de α debería estar basada en este primer objetivo. Esta figura también permite analizar el rango de valores de α en donde los subconjuntos de descriptores más interesantes pueden ser encontrados. Asimismo, también se puede apreciar que los subconjuntos de descriptores obtenidos con α=1 (*wrapper* mono-objetivo) tiene siempre menor capacidad predictiva que cuando se usa cualquier otro valor con $0 < \alpha < 1$. Esto lleva a afirmar que modelos conformados por subconjuntos con menor cardinalidad son más propensos a ser modelos más generalizables y, en consecuencia, con mejor capacidad predictiva.

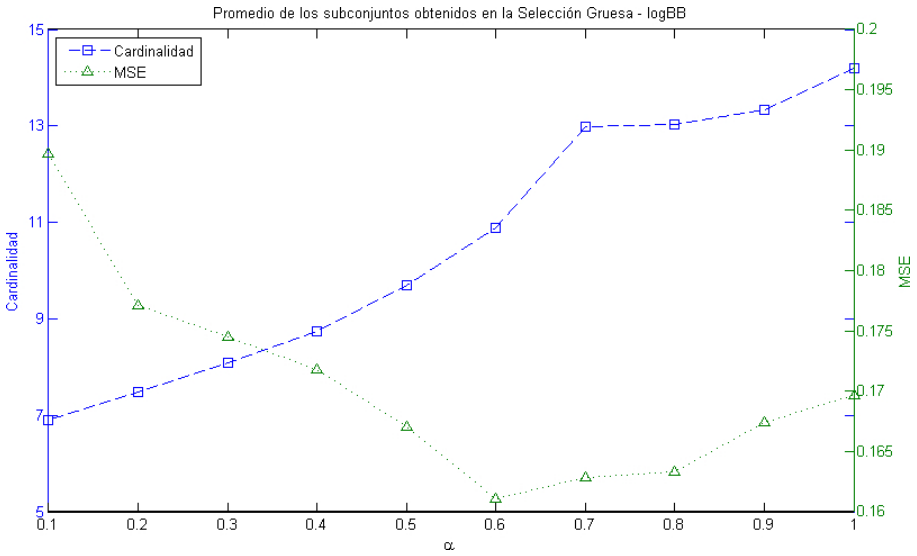


Figura 5.7: Análisis de la elección de α para DS1.

5.4.4.4. Análisis químico de los subconjuntos de descriptores encontrados

Si bien la presente tesis se centra en el área de las ciencias de la computación, deseamos analizar también la relevancia de los resultados obtenidos desde el punto de vista químico/farmaceútico. Por tal motivo, en esta sección presentaremos un breve análisis sobre la validez de los subconjuntos de descriptores encontrados.

Es importante notar que de acuerdo a la naturaleza estocástica de nuestra propuesta, los subconjuntos de descriptores obtenidos no son necesariamente los mismos en las diferentes repeticiones del método. En Horvath *et al.* (2007) se enfatiza que este rasgo distintivo de los métodos de selección de variables basados en búsquedas estocásticas no es necesariamente una limitación, sino que representa la posibilidad de ofrecer más de un subconjunto relevante para la construcción de un potencial modelo de predicción.

No obstante, como se puede observar en la Tabla 5.15 varios descriptores comunes son seleccionados en distintos subconjuntos, y buena parte de ellos, o bien se han elegido en otros trabajos de selección de características, o bien resultan de relevancia por el conocimiento teórico que se tiene de ellos (se puede consultar la lista completa de todos los descriptores seleccionados en Soto *et al.* (2009a)). En

el caso del primer set de datos, el descriptor 'TPSA(NO)' (área superficie polar topológica usando contribuciones polares de oxígeno y nitrógeno) es frecuentemente seleccionado en la mayoría de los subconjuntos. Esta selección repetida es de esperar, teniendo en cuenta la importancia del área superficial polar en la predicción del grado de penetración en la barrera hematoencefálica (Ertl (2008)). Generalmente, los compuestos lipofílicos son capaces de atravesar la barrera hematoencefálica, por lo que también es esperable que descriptores relacionados con la hidrofobicidad como 'ALOGP' (coeficiente de partición octanol-Agua de Ghose-Crippen) o descriptores relacionados con los grupos de ácidos carboxílicos como 'nCOOH' (número de ácidos carboxílicos alifáticos) o 'Ic' (número de ácidos carboxílicos alifáticos y aromáticos) sean frecuentemente elegidos.

De igual modo, no sorprende encontrar que nuevamente descriptores relacionados con la solubilidad en agua tal como 'ALOGP' o MLOGP (coeficiente de partición octanol-Agua de Moriguchi) estén frecuentemente seleccionados en los subconjuntos cuando se quiere predecir la absorción intestinal humana en el segundo set de datos. Asimismo en el caso de DS3, podemos detectar que ciertos descriptores conocidos como relevantes para la propiedad logP se eligen en forma repetida, entre otros: 'MW' (peso molecular), descriptores relacionados con la existencia de carbonos (*e.g.* 'nC' número de átomos de carbono; 'nCar' suma de todos los átomos que pertenecen a una estructura aromática o heteroaromática; 'nCs' número de átomos de carbono secundarios), descriptores relacionados con el momento dipolar ('D_H' dipolo total, hibridización; 'D_S' dipolo total, hibridización + punto de carga).

5.4.4.5. Evaluación de la probabilidad de correlaciones por chance

Un punto igualmente importante a la tarea de selección de descriptores es si, en la aplicación de este tipo de técnicas, existe la probabilidad de que se tengan correlaciones por chance. Básicamente, las correlaciones por chance son propensas a ocurrir cuando se elige un subconjunto de descriptores a partir de un conjunto grande. Este problema fue inicialmente estudiado en QSPR para modelos de regresión lineal (Livingstone & Salt (2005); Topliss (1979); Topliss &

Costello (1972)); sin embargo el problema está igualmente presente en cualquier método de predicción.

Analizaremos pues, cuál es la probabilidad de obtener correlaciones por chance aplicando la metodología de selección de descriptores multi-objetivo. En los trabajos de Baumann (2005); Rücker *et al.* (2007) se enfatiza que la probabilidad de la existencia de correlaciones por chance es mayor cuanto menor sea la proporción de cantidad de compuestos sobre el número de variables elegidas. A partir de esta conclusión, podemos afirmar que es poco probable la existencia de correlaciones por chance en nuestra metodología, ya que, primero, en todas nuestras experimentaciones usamos data sets con más de 100 compuestos y, segundo, la técnica multi-objetivo prioriza la selección de subconjuntos de descriptores con cardinalidad mínima.

De todos modos, aplicamos una estrategia denominada *y-randomization*, la cual consiste en una mezcla aleatoria de la variable a modelar (Rücker *et al.* (2007)). El objetivo de esta técnica es determinar si luego de aplicar la aleatorización, el método de selección de descriptores es igualmente capaz de encontrar una selección con alta calidad predictiva. En caso afirmativo, dado que la relación encontrada es claramente espuria, se puede afirmar que existen altas probabilidades de correlaciones por chance.

Aplicamos *y-randomization* a los subconjuntos DS1 y DS2 ya que estos subconjuntos son los que tienen la relación cantidad de datos sobre la cantidad de variables más desfavorable. Ejecutamos 10 corridas utilizando diferentes aleatorizaciones en cada corrida, y para cada corrida ejecutamos 10 veces el algoritmo de selección de descriptores. Los resultados nos muestran que el MSE y el q^2 para DS1 fue 0,3652 y 0,00685 respectivamente, mientras que para DS2 se obtuvo $0,3713$ y $0,018261$ (Tabla 5.17). Estos resultados confirman nuestra suposición de que el método no es propenso a generar correlaciones por chance.

5.4.4.6. Análisis de la complejidad del tiempo en el peor caso

El primer aspecto a tener en cuenta para el análisis del tiempo de complejidad en el peor caso, es que el orden del peor caso estará delimitado por la complejidad

Tabla 5.17: 10 corridas del método aplicando *y-randomization*

	DS1-Agg+DT		**HIA-Agr+MLR**	
Corrida	MSEP	q2	MSEP	q2
1	0,36405741	0,00299829	0,36056883	0,05335769
2	0,36678449	0,00408746	0,37630056	0,0091366
3	0,3692595	0,00361936	0,36745353	0,00973019
4	0,36811486	0,00554326	0,37539955	0,00957474
5	0,36491194	0,00258601	0,37276905	0,01107089
6	0,36606204	0,00449392	0,37306781	0,00960255
7	0,36401141	0,02682486	0,38359021	0,01453971
8	0,36520048	0,00348746	0,36201705	0,03755967
9	0,36497273	0,00365222	0,37172438	0,01004058
10	0,3666346	0,01176187	0,37025715	0,0180009
Prom	0,36600095	0,00690547	0,37131481	0,01826135

de la primera fase, *i.e.* del wrapper multi-objetivo. Este hecho se deduce fácilmente dado que corresponde a la fase con mayor costo computacional. En particular, los métodos basados en Pareto son más intensivos computacionalmente que los basados en agregación, ya que los primeros requieren algunas tareas adicionales: poseen otra población paralela, tienen que ordenar los individuos según el criterio de no dominancia y deben calcular ciertas medidas de distancia entre los individuos.

Según Deb (2004) el orden del tiempo de complejidad del peor caso para NSGA-II es $O(ks^2)$, donde k es el número de objetivos a optimizar (2 en nuestra metodología) y s es el tamaño de la población (145 en nuestra metodología). El orden del tiempo de complejidad para SPEA2 es $O((s+\bar{s})^3)$, donde $\bar{s}$ es el tamaño de la población externa (también 145). Deb (2004) afirma que este límite es bastante pesimista y que el tiempo de complejidad del caso promedio $O((s+\bar{s})^2 log(s+\bar{s}))$ es mucho más realista. Estos órdenes están definidos para una única generación asumiendo que no hay costo de complejidad en el cálculo de la función de fitness.

El orden del tiempo de ejecución para las funciones de fitness se reducen a calcular el orden del tiempo de ejecución de los métodos de regresión del *Evaluador de Variables*. De observaciones empíricas, la RNL es la que tiene el costo computacional más intensivo, el cual es $O(c^3)$ donde c es el número de coeficientes a ser ajustados ($4p+1$ en nuestro caso, donde p es el número de descriptores en el subconjunto evaluado).

En conclusión, y a fin de proveer una forma general del orden del tiempo de ejecución de nuestra metodología, definiremos $O(G)$ como el orden de ejecución de una generación de una estrategia de búsqueda y $O(F)$ como el orden de ejecución de la función de fitness del wrapper. Dado que la función de fitness se computa al comienzo de una generación, el tiempo de complejidad de una sola generación es $O(max(s \cdot O(F), O(G)))$. El tiempo general de una corrida del wrapper multi-objetivo es, por consiguiente, el orden de complejidad de una sola generación multiplicado por la cantidad de generaciones. La cantidad de generaciones resulta difícil de establecer de antemano, ya que dependerá fuertemente del criterio de parada y/o convergencia.

5.4.4.7. Análisis de nuestra metodología utilizando validación externa

En esta propuesta multi-objetivo, cuando nos referimos a la capacidad predictiva de los subconjuntos de descriptores encontrados, nos hemos referido siempre a los errores de predicción utilizando el CRN de la segunda fase. Tal como hemos mencionado, el propósito de esta segunda fase es proveer un método de regresión más preciso que los empleados en la primera fase, pero no por eso pretende ser una metodología de validación que esté estimando la capacidad predictiva real de un modelo QSAR obtenido con los descriptores seleccionados. Podemos observar que esta segunda fase, reporta su validación usando datos que fueron anteriormente usados para la selección de los descriptores de la primera fase. Este hecho podría considerarse como un error de diseño que introduce una estimación sobreoptimista de la verdadera capacidad predictiva del subconjunto. Sin embargo, demostraremos que, incluso cuando esta segunda fase no es un procedimiento de validación puramente estricto, es lo suficientemente confiable para evaluar la capacidad de predicción de los subconjuntos.

Para demostrar nuestra afirmación, aplicamos un procedimiento de validación externa a fin de cuantificar cuán diferente es el error de predicción con respecto a nuestro procedimiento de validación interna. Para esto, separamos un conjunto Z_3 (con 20 % de los datos) antes de aplicar la primera fase. Luego aplicamos nuestro procedimiento, incluyendo la primera y segunda fase, al 80 % restante, de donde extrajimos los 20 subconjuntos de descriptores más relevantes. La capacidad de

predicción de estos subconjuntos fue evaluada de dos formas: con nuestra estrategia de validación interna (tal como se explica en la sección 5.4.3) y también evaluándola sobre el conjunto Z_3 que fue inicialmente separado.

El objetivo es analizar para estos 20 subconjuntos si los errores promedios obtenidos en Z_3 son significativamente peores que los errores obtenidos utilizando nuestra estrategia de validación interna. El procedimiento usual para comparar estadísticamente las medias de ambas estrategias de validación es aplicando la prueba de comparación de t-student para muestras correlacionadas. De esta manera, podemos determinar intervalos de confianza para los promedios de la diferencia entre los errores promedios de ambas estrategias de validación. Se pueden obtener tres situaciones diferentes:

Situación 1: el error de validación externa es significativamente más bajo que el error de validación interna,

Situación 2: no hay evidencia estadística que el error de validación externa sea superior al error de validación interna,

Situación 3: otra situación diferente a las presentadas en 1 y 2.

La situación 1, puede ser claramente identificada (manteniendo una probabilidad de error menor o igual a 0,05 ó 0,01). Sin embargo, no es probabilísticamente posible cuantificar la segunda situación (error de Tipo-II). El procedimiento convencional para identificar la situación 2 se produce cuando no se puede rechazar la hipótesis de diferencia aún cuando la probabilidad de error es mayor o igual a 0,25, o bien cuando el error de validación interna es mayor que el error promedio en Z_3 y la situación 1 no sucede. En cualquier otra situación distinta a las dos anteriores se asume que error de validación externa es superior al error de validación interna (situación 3).

Este procedimiento fue aplicado a los tres sets de datos sobre los que se calcularon los resultados de las secciones anteriores. Para poder tener en cuenta el efecto de diferentes separaciones de datos para Z_3, se replicó esta experimentación 10 veces, y en donde, sin pérdida de la generalidad, se utilizó RL y la estrategia de agregación para esta etapa.

Como resultado de esta experimentación, se obtuvo que en sólo 3 de las 30 corridas se da la situación 3 (una vez para DS1 y dos veces para DS3); mientras que en 11 de las 30 corridas se da la situación 1 y en los restantes 16 casos se produce la situación 2. Estos resultados nos llevan a pensar que el procedimiento de validación aplicado en la segunda fase constituye un buen estimador de la capacidad predictiva de los subconjuntos. La razón principal de estos resultados surgen del hecho que, cuando nuestro procedimiento de validación interna es aplicado, sólo una parte de los datos restantes es usada para entrenamiento (Z_1) en cada iteración del procedimiento de validación cruzada. En cambio, cuando el procedimiento de validación externa es aplicada, todos los datos restantes (80 %) son usados para entrenar el modelo. No hay nada incorrecto en usar todos los datos restantes para entrenar las redes Bayesianas, ya que ha sido demostrado que estos modelos no tienden a sobreajustar los datos usados para el entrenamiento (Burden & Winkler (1999)) y, por lo tanto, no necesitan de un conjunto de datos de validación, tal como se ve en los resultados obtenidos sobre Z_3.

Si bien la validación externa es el estándar máximo cuando se trata de evaluar un modelo de predicción basado en QSAR, en este trabajo no fue utilizado para la presentación de los resultados, ya que es dependiente de la separación de los datos en entrenamiento/testeo. Para ilustrar este punto, se aplicó una prueba ANOVA de efectos aleatorios, donde cada una de las 10 corridas con distintas separaciones fue considerado como un factor aleatorio que tiene la capacidad predictiva de los 20 subconjuntos más relevantes como sus observaciones. Nuestro objetivo es determinar cuán importante es el efecto de la separación de los datos, de acuerdo a la contribución de su varianza. Estos experimentos (Tablas 5.18, 5.19, 5.20) revelan que con una probabilidad de error menor a $0{,}5\times10^{-5}$ hay una fuente de varianza debido a la separación de los datos para los tres conjuntos, en donde esta contribución representa el 34,5 %, 17,85 % y 60,08 % de la varianza total para DS1, DS2 y DS3 respectivamente. Este efecto sobre la varianza debido a las diferentes separaciones aleatorias de los datos de entrenamiento/testeo hace necesario que, de usarse este tipo de validación, se realicen un gran número de corridas a fin de que los resultados no queden afectados por la gran varianza producida. Además, un alto número de corridas para cada combinación del *wrapper*

Tabla 5.18: Prueba ANOVA con efectos aleatorios sobre DS1 donde el factor aleatorio son 10 corridas con distinto conjuntos de testeo sobre el cual se evalúa el error de predicción promedio.

F. de V.	S.C.	G.L.	C.M.	F	p
Entre	0,044525436	9	0,004947271	11,5354	$\leq$0,0001
Dentro	0,081486649	190	0,000428877		
Total	0,126012086	199			

Tabla 5.19: Prueba ANOVA con efectos aleatorios sobre DS2 donde el factor aleatorio son 10 corridas con distinto conjuntos de testeo sobre el cual se evalúa el error de predicción promedio.

F. de V.	S.C.	G.L.	C.M.	F	p
Entre	0,151520572	9	0,016835619	5,3464	$\leq$0,0001
Dentro	0,598303118	190	0,003148964		
Total	0,74982369	199			

multi-objetivo (tal como se realizó en el marco de esta tesis) se volvería computacionalmente intratable. Por otra parte, nuestro esquema de validación interna es computacionalmente menos costoso y, al mismo tiempo, comparable con los resultados obtenidos usando validación externa.

5.5. Conclusiones

En el presente capítulo hemos expuesto diferentes alternativas para abordar el difícil problema de la selección de descriptores para modelos basados en QSAR. Consideramos también que esta clase de estrategias constituye una acción inicial en el proceso de desarrollo de un modelo de predicción de propiedades biológicas o fisicoquímicas en donde no sea clara su vinculación con las características mo-

Tabla 5.20: Prueba ANOVA con efectos aleatorios sobre DS3 donde el factor aleatorio son 10 corridas con distinto conjuntos de testeo sobre el cual se evalúa el error de predicción promedio.

F. de V.	S.C.	G.L.	C.M.	F	p
ENTRE	0,458824117	9	0,050980457	31,1028	$\leq$0,0001
DENTRO	0,311428569	190	0,001639098		
TOTAL	0,770252687	199			

leculares. Asimismo, es de interés resaltar que contar con un método confiable de selección de descriptores es fundamental para mejorar la predicción de un modelo, como así también ganar en comprensión de las relaciones subyacentes entre descriptores y la propiedad experimental a modelar.

Las tres estrategias expuestas en este capítulo siguen un modelo basado en un *wrapper*. Las dos primeras lo hacen siguiendo un enfoque mono-objetivo, mientras que la última de ellas lo hace siguiendo un enfoque multi-objetivo. Los mismos han sido producto de una secuencia de desarrollos que se corresponde con el orden en el que fueron presentados. El tercer enfoque claramente mejora los dos anteriores, tanto en su funcionamiento, el cual permite un uso más automatizado, como así también en los resultados alcanzados. Por tal motivo, consideramos a la estrategia multi-objetivo como la mejor alternativa, y es la que proponemos para futuras experimentaciones. Por tal motivo las conclusiones de este capítulo estarán mayormente centradas en los resultados y consecuencias de aplicar este último enfoque.

Los principales aportes de este capítulo se centran en dos aspectos principales. En primer lugar, la propuesta de una metodología de selección de descriptores de dos fases, que apunta a combinar una búsqueda amplia dentro del espacio de posibles descriptores con métodos de evaluación de precisión como los CRN. La evaluación rigurosa se realiza sólo a los subconjuntos preseleccionados obtenidos con el *wrapper* en la primera fase. La segunda fase, además de mejorar la rigurosidad en la evaluación, permite una independencia con los métodos de evaluación de la primera fase, y por ende la selección de los subconjuntos de variables no queda sesgada de acuerdo a los métodos de evaluación usados. Además, si bien otros métodos de dos fases ya habían sido propuestos en la literatura, no se presentaba un análisis comparativo de diferentes alternativas como sí se muestra en nuestras investigaciones: Soto *et al.* (2008a), Soto *et al.* (2008b) y Soto *et al.* (2009b).

La segunda contribución de relevancia para la quimioinformática representa la proposición de una estrategia multi-objetivo para la selección de descriptores. De acuerdo a nuestro conocimiento, antes del trabajo publicado en Soto *et al.* (2009b) no se encuentra registro del uso de técnicas multi-objetivo para la selección de descriptores en QSAR o QSPR. Asimismo, otros trabajos en el área de selección de variables en forma multi-objetivo (Oliveira *et al.* (2003), Emmanouilidis *et al.*

(2000)), no presentan una comparación ni un análisis sobre la diferencia de aplicar distintas estrategias de búsqueda y evaluación. Este enfoque multi-objetivo favorece la obtención de subconjuntos de cardinalidad mínima. En consecuencia, además de favorecer la interpretación de los resultados, se disminuye en forma exponencial el número de posibles hipótesis de aprendizaje sobre los descriptores (Liu & Motoda (2008)), colaborando de esta manera a reducir la probabilidad de correlaciones por chance.

De la comparación entre los diferentes métodos de búsqueda, creemos que tanto las propuestas basadas en Pareto como las basadas en agregación, constituyen modos recomendables de aplicar selección de variables. Las diferencias en el desempeño entre un enfoque u otro dependerá de la cardinalidad del subconjunto de descriptores "ideal". En el caso del enfoque basado en agregación se deben establecer dos parámetros, α y p_m, lo que puede resultar engorroso y poco automatizado. Sin embargo, los teoremas de *No Free Lunch* indican que no hay un único mejor método de selección de variables para cualquier set de datos (Wolpert & Macready (1997)). Por lo tanto, la incorporación de parámetros y restricciones, es una práctica común y efectiva para realizar un método a medida de las características del problema. Por otra parte, los métodos basados en Pareto no disponen de parámetros que ponderen los diferentes objetivos. En contrapartida, estos métodos no son la mejor opción cuando el número "ideal" de descriptores es alto (sección 5.4.4.3).

Es importante resaltar que la aplicación de nuestro *wrapper* multi-objetivo no constituye una problema de optimización clásico de tipo multi-objetivo. Los mejores subconjuntos de descriptores encontrados por el *wrapper* (o aquellos incluidos en el frente de Pareto) no son necesariamente selecciones optimales, ya que se utilizan métodos de evaluación de limitada precisión para la regresión. Sin embargo, como se demuestra con los resultados del CRN, proveen subconjuntos, los cuales tienden a ser relevantes y no redundantes. Podemos decir que son relevantes, ya que en general poseen una capacidad de predicción alta, según se verifica con la validación de la segunda fase, y además los subconjuntos encontrados son relevantes en términos de la relevancia química, tal como se muestra en la sección 5.4.4.4. Del mismo modo, y teniendo en cuenta que uno de los objetivos del *wrapper* es minimizar la cardinalidad, podemos esperar que tengan una baja

redundancia, dado que si un subconjunto tuviera uno o más descriptores redundantes, ese subconjunto tendría baja probabilidad de sobrevivir durante todas las generaciones del algoritmo evolutivo.

Por otra parte, sería injusto no mencionar que nuestra estrategia evolutiva tiene como principal desventaja su alta demanda en tiempo computacional, en comparación con otras propuestas de selección de variables (Guyon & Elisseeff (2003), Fröhlich *et al.* (2004), Figueiredo (2003)). Sin embargo, quisiéramos enfatizar que el tiempo de cómputo no es un aspecto crucial, teniendo en cuenta que nuestra metodología puede ser ejecutada en un tiempo polinomial razonable, tal como se explica en la sección 5.4.4.6. Asimismo, se debe tener en cuenta que dado un conjunto de entidades químicas a la cuales se les desea aplicar el algoritmo propuesto, la selección de descriptores no es algo que deba ser aplicado en tiempo real ni tampoco que deba ser ejecutado en numerosas oportunidades.

Dado que nuestro método de selección de características de dos fases consiste de un conjunto de métodos estadísticos funcionando en serie, se coincide con un aspecto importante de la teoría del aprendizaje automático, el cual establece que mejores resultados son obtenidos cuando se combinan diferentes estrategias de aprendizaje en lugar de cuando se usa un único modelo (Arodź *et al.* (2006), Gama & Brazdil (2000)). Incluso ciertos trabajos de selección de características, como Dutta *et al.* (2007) y Baumann (2005), destacan la utilización y combinación de diferentes técnicas de aprendizaje automático en un mismo método de selección de variables.

En el desarrollo de los métodos presentados en este capítulo, un subconjunto de descriptores es considerado como relevante dependiendo de su capacidad predictiva, la cual se determina aplicando una medida de error (MAE o MSE) dentro de un esquema de validación cruzada y utilizando el subconjunto en cuestión. Si bien otros autores también defienden esta forma de medir error de predicción en sets de datos de pequeño a mediano tamaño (Konovalov *et al.* (2008), Tropsha *et al.* (2003), Hawkins *et al.* (2003)), hay quienes consideran que medir la relevancia de un subconjunto de descriptores únicamente por métodos estadísticos no es algo deseable (Johnson (2008)). En nuestra opinión, un método de selección de descriptores no pretende dar un solución definitiva al problema de inferir cuál es el mejor subconjunto de descriptores que controla el valor de una propiedad o

actividad biológica. Sin embargo, en ausencia de procedimientos teóricos, creemos que los métodos de aprendizaje automático para la selección de variables brindan una ayuda valiosa cuando no se conocen las reglas que gobiernan una dada actividad o propiedad. De este modo, estos métodos representan una herramienta de utilidad para los científicos del área, quienes a su vez pueden contribuir con su conocimiento en la toma de decisiones.

Otro de los cambios realizados en la propuesta multi-objetivo, es la no consideración de un ranking de descriptores ordenados por relevancia. De acuerdo a nuestra experiencia, un descriptor no es relevante de por sí, sino que su importancia debe ser evaluada en conjunción con todo un subconjunto de descriptores del cual es parte.

En nuestras experimentaciones, más de un único subconjunto de descriptores es reportado como resultado de la aplicación del método de selección de descriptores. No brindar un único subconjunto de descriptores no constituye una desventaja, sino que permite a un potencial usuario del método contar con más de un subconjunto, todos ellos de relevancia comparable. El usuario del método, a su vez, podría tomar una decisión en base a otros criterios, como por ejemplo: la interpretación teórica, la evaluación de los subconjuntos en nuevos datos de compuestos no utilizados por el método de selección de características, etc.

Otra conclusión importante que se desprende de las experimentaciones realizadas, es que para los sets de datos DS1 y DS2, se obtuvieron resultados que son mejores o comparables a los obtenidos cuando se aplica la 'Selección de Variables de Monte Carlo' aplicada al mismo conjunto de datos en Konovalov *et al.* (2008). En el caso del set de datos DS3, se tiene como dificultad adicional que la relación entre los descriptores considerados y el valor de logP es no lineal. Las experimentaciones realizadas, si bien son altamente satisfactorias, no permiten hacer una comparación directa con el trabajo original por diversos motivos (sección 5.4.4.2). No obstante esta incapacidad, se destaca el desempeño del método para funcionar satisfactoriamente en forma independiente de la linealidad o no linealidad del modelo a construir. Asimismo, destacamos la comparación realizada con un método determinístico actual y competente (Figueiredo (2003)), el cual es sobrepasado por nuestra propuesta ante los sets de datos con los que hemos trabajado.

Por estos motivos, consideramos que nuestra estrategia evolutiva de dos fases es una metodología recomendable para la selección de descriptores en modelos de QSAR/QSPR.

5. SELECCIÓN Y REDUCCIÓN DEL CONJUNTO DE DESCRIPTORES

Capítulo 6

Influencia del aprendizaje no supervisado en la predicción e identificación de dominio de aplicación

Los modelos de predicción basados en el enfoque QSPR buscan establecer una relación entre parámetros moleculares y una dada propiedad o actividad. Hemos visto en los capítulos 3 y 4 que los métodos de aprendizaje automático brindan una herramienta muy valiosa en esta tarea. Sin embargo, aún asumiendo la tarea de la identificación de los descriptores relevantes como algo resuelto, el desafío de un método de predicción reside en la capacidad de obtener modelos lo suficientemente generales como para que el modelo funcione sobre compuestos moderadamente distintos a con los que fue entrenado. Ciertos modelos de predicción son muy propensos a sobreajustar o a sobreentrenar el conjunto de datos con el que se lo entrena (Tetko *et al.* (1995), Hawkins (2004)) y, por lo tanto, lo que se obtiene como resultado es un modelo que solamente "memorizó"los datos del entrenamiento.

Al mismo tiempo, esta generalización no puede alcanzarse de una manera absoluta, es decir no puede lograrse una alta precisión para cualquier compuesto químico que pueda ser sintetizado. Por lo tanto, sería muy importante conocer cuando un nuevo compuesto que necesita ser predecido, tendrá o no una preci-

sión semejante a la reportada por el método en su validación. En otras palabras, resulta interesante poder identificar el dominio de aplicación de un método de predicción, es decir el dominio químico en donde un compuesto poseerá una predicción confiable.

Estas dos cuestiones relacionadas a la calidad de los modelos de QSPR, han sido motivo de unas serie de discusiones, en donde se ponía en duda la utilidad de los modelos QSPR que provinieran de un enfoque puramente estadístico (Doweyko (2008), Beresford *et al.* (2004), Hou & Wang (2008), Johnson (2008), Gola *et al.* (2006), Tetko *et al.* (2006)). Por lo tanto, el presente capítulo tiene por objetivo reportar las investigaciones realizadas en estos temas, las cuales básicamente se centraron en estudiar la aplicación de distancias y agrupamientos por similitud dentro del espacio químico. Más específicamente, se estudiaron distintas estrategias de agrupamientos de datos, que sirvieran para favorecer la precisión de los modelos de predicción. Si bien estos estudios no concluyeron en el desarrollo de un método integral como los del capítulo anterior, el conocimiento generado a partir de ellos sirvió de base para un método de identificación del dominio de aplicación, el cual está basado en mapas auto-organizativos (SOM), y sobre el que sí se hizo un desarrollo integral.

6.1. 1º Propuesta: modelos de predicción ajustados a cada grupo

El conjunto de entrenamiento de un método será uno de los aspectos que definirá la capacidad de generalización y el alcance del método. Uno de los problemas que puede surgir en la utilización de datos no homogéneos es que ciertos tipos de datos estén mejor representados que otros. Esto implica que se puede producir un mayor ajuste en los datos mejor representados por sobre los que no están tan bien representados. Por ejemplo, si la mayoría de los compuestos son del grupo de los alcoholes y sólo hay unos pocos alcanos, es de esperar que las predicciones de compuestos nuevos que sean alcoholes sean más precisas que aquellas que se hagan con alcanos. Jónsdóttir *et al.* (2005) mencionan este problema y también

coinciden en resaltar que esta redundancia afecta negativamente la estimación de la predicción de un método.

Otra hipótesis interesante a tener en cuenta es que las reglas que gobiernan las relaciones estructura-propiedad, pueden ser diferentes en función de la familia química a la cual pertenezca el compuesto o a los grupos funcionales que el mismo contenga. Si bien la información disponible en los descriptores permite decodificar, ya sea la familia química o el grupo funcional al que un compuesto pertenece, poder realizar modelos diferenciados en función del tipo de compuesto, reduciría la complejidad de los modelos, dado que la cantidad de descriptores necesarios sería menor y, por ende, las relaciones serían más sencillas.

6.1.1. Idea del método

La primer propuesta que presentaremos en este capítulo consiste en la utilización de medidas de distancia (o similitud) para detectar agrupamientos en los datos, y así realizar modelos de predicción ajustados para cada grupo. Esta técnica se la conoce en la literatura como *mezcla de expertos* (sección 3.2.5.2). La hipótesis que sigue el método es que compuestos con características similares puedan ser puestos juntos y que sean entrenados de acuerdo a las particularidades del grupo. Este trabajo se inspiró en otros métodos que también aplican medidas de similaridad en el campo de la quimioinformática (Sheridan *et al.* (2004), Kühne *et al.* (2006) y Martin *et al.* (2002)).

En la Figura 6.1 se muestra un bosquejo de la metodología propuesta. Inicialmente los compuestos son separados aleatoriamente en un conjunto de entrenamiento y otro de testeo. Sobre el primero de ellos se buscan, siguiendo algún criterio de agrupamiento, cuáles son los grupos T_i que emergen de los datos. A cada uno de estos grupos se les aplica un modelo de predicción supervisado P, quedando tantos modelos como números de grupos hayan sido encontrados. A cada modelo entrenado con el grupo T_i lo llamaremos P_{Ti}. Por otra parte, a cada uno de los compuestos que fueron inicialmente separados para el testeo se les verifica a cuál de los grupos conformados pertenece. Finalmente, a fin de obtener el valor a predecir, se le aplica el método de predicción que corresponde a su grupo.

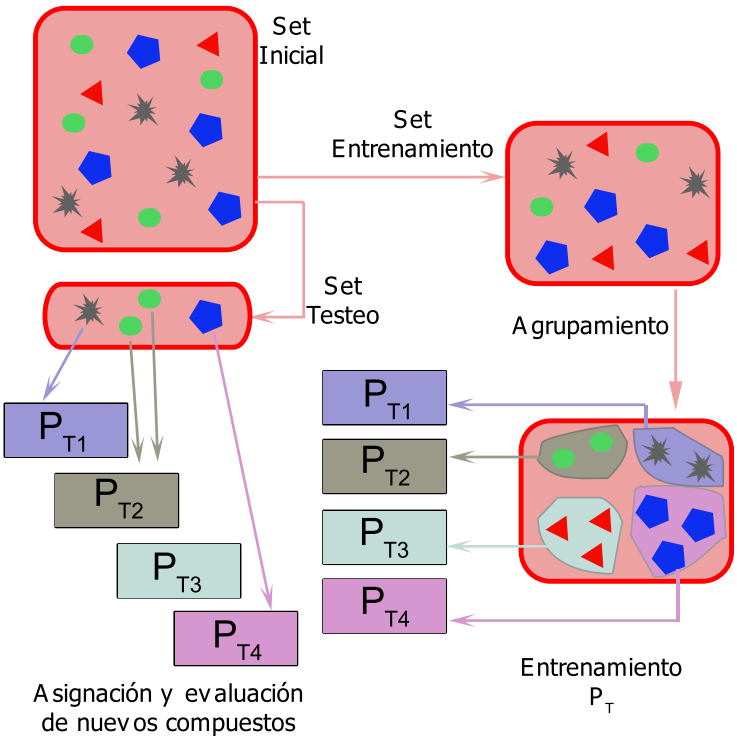


Figura 6.1: Esquema del procedimiento de entrenamiento por grupos.

6.1.2. Diseño de los experimentos

Para evaluar estas metodologías decidimos trabajar con una gran cantidad de compuestos, ya que necesitábamos que en cada grupo quedara un número apropiado de moléculas. Por tal motivo decidimos usar el set de datos descripto en el Apéndice B.1 del cual tomamos los primeros 4778 compuestos. Sobre ellos decidimos calcular 168 descriptores provenientes de todas las familias de los descriptores constitucionales y grupos funcionales (Todeschini & Consonni (2000)).

Cabe mencionar que para este trabajo no se utilizó ninguna técnica de selección de descriptores. Por tal motivo, previo a la tarea de agrupamiento y predicción, se aplicó una transformación por PCA (utilizando matriz de correlación), a fin de trabajar con un set de descriptores más reducido que no contuviera información redundante.

Para el análisis de agrupamiento, utilizamos el método jerárquico aglomerativo (sección 3.3.1) aplicando distancia coseno con un ligamento completo. La misma distancia y ligamento fue usada para asignar un nuevo compuesto a uno de los grupos.

Para la predicción se utilizó un comité de redes neuronales (CRN) de 10 redes. Al momento de esta experimentación, no habíamos incursionado aún en las redes neuronales bayesianas, por lo que aplicamos un algoritmo de entrenamiento de *back propagation* con una optimización de los pesos que se basa en un método de descenso por gradiente con momento y tasa de aprendizaje adaptiva. Además, se eligió un 20 % de los compuestos de manera aleatoria estratificada, los cuales fueron usados como testeo.

6.1.3. Resultados

Luego de aplicar PCA sobre los datos nos quedamos con 27 componentes, algunas de las cuales eran descartadas después de dividirse los datos en los grupos encontrados. Para el análisis de agrupamiento, se utilizó un umbral de 1,8, quedando así 7 grupos tal como se muestra en la Figura 6.2.

La Figura 6.3 muestra el gráfico de las predicciones obtenidas, en donde el eje de las abscisas corresponde a los valores experimentales y el de las ordenadas a los valores de predicción obtenidos con los diferentes modelos. En la parte superior

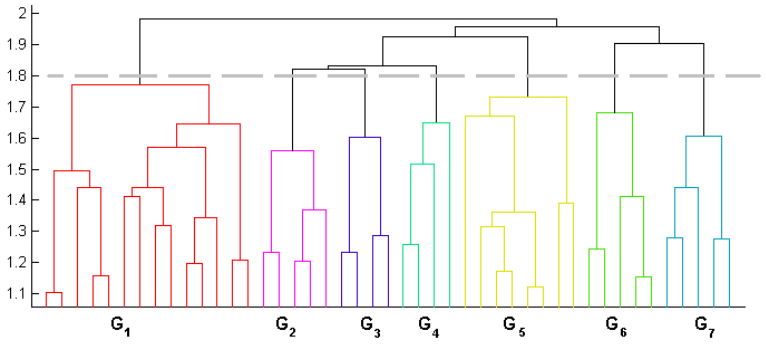


Figura 6.2: Dendograma de los grupos conformados.

de este gráfico, se muestran los valores de predicción cuando no se aplica ninguna estrategia de agrupamiento y, en la parte inferior, al aplicarse la estrategia de predicción por grupos (abajo). Los gráficos de la izquierda muestran los resultados de predicción para los compuestos del entrenamiento, mientras que los del lado derecho muestran las predicciones para los compuestos del testeo.

Podemos apreciar que la métrica de predicción utilizada $RMSE$ ($RMSE = \sqrt{MSE}$) mejora notablemente cuando se usan los modelos ajustados para cada grupo. Evidenciamos además que la complejidad del modelo único en términos de cantidad de pesos, así como en tiempo de entrenamiento, es ampliamente mayor que en los casos de los modelos de cada grupo. En contrapartida, advertimos que los resultados presentaban una gran variabilidad en función del algoritmo de agrupamiento elegido. Además, si la separación de los datos del testeo no se hace de manera estratificada, no se encuentran diferencias significativas en la comparación de ambas estrategias.

6.2. 2º Propuesta: corrección de la predicción usando información de grupos

En esta segunda propuesta apuntamos a desarrollar un metodología que también usara el concepto de agrupamiento entre los datos, pero con un rol distinto al del caso anterior. El objetivo aquí es analizar el comportamiento de los grupos

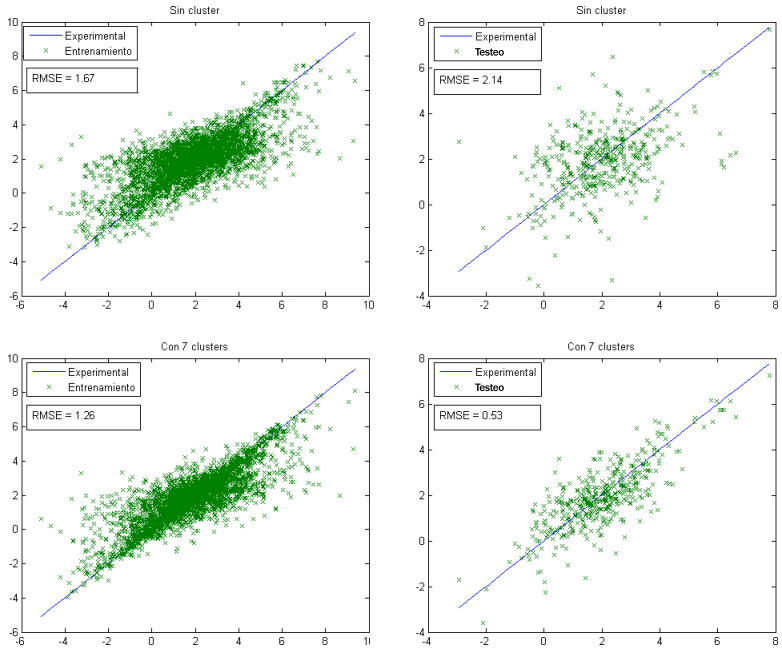


Figura 6.3: Comparación de las predicciones utilizando el modelo único (arriba) y utilizando un modelo diferente por cada grupo encontrado (abajo), para los compuestos del entrenamiento (izquierda) y para los del testeo (derecha).

de compuestos conformados, para identificar información que pudiera ser usada como factor de corrección en futuras predicciones.

En este caso, hemos decidido experimentar con dos nuevas formas de armar grupos entre los datos de entrenamiento. La primera se trata de un método de aprendizaje no supervisado, mientras que la segunda corresponde al criterio de un humano experto que agrupa los compuestos de acuerdo a un criterio químico, como puede ser la pertenencia a una familia química o su acción terapéutica.

6.2.1. Idea del método

La idea de la metodología se basa en el hecho de que cualquier método de predicción posee un sesgo y una varianza que hacen que las predicciones no sean precisas. La idea es identificar los grupos de datos, para luego intentar detectar los sesgos existentes en cada uno de los distintos grupos, para poder finalmente usarlos como factor de corrección en nuevos datos. La varianza de los grupos no será tenida en cuenta en estas experimentaciones, pero sí en el método de la sección 6.3.

Esta propuesta guarda cierta relación con el enfoque de Kühne *et al.* (2006) y con las redes asociativas de Tetko (2002), en donde la información de los compuestos del entrenamiento y su similiaridad son usadas para mejorar las predicciones. El rasgo distintivo existente en esta propuesta es que las distancias se miden al grupo más cercano, en lugar de medir similaridad en términos de compuestos individuales.

6.2.2. Desarrollo de la metodología

Se utilizaron dos formas para identificar grupos de datos. La primera de ellas, consistió en usar SOMs. La segunda forma aplicada para agrupar los datos, consistió en un análisis humano que buscó características químicas en común, de acuerdo a la presencia o ausencia de ciertos grupos funcionales. De esta forma, se armaron 7 grupos diferentes. En el primer grupo, se concentraron sustancias que sólo estaban constituidas por átomos de carbono e hidrógeno. Para el segundo y tercer grupo, se consideraron sustancias que contuvieran grupos carboxilo y grupos carbonilo respectivamente. Al cuarto grupo, se asignaron las moléculas que

constaban de un grupo oxhidrilo unido a una cadena abierta o cíclica; o directamente unida a un átomo de carbono aromático (SP2). En el grupo 5, aglomeramos a todos los compuestos con azufre junto a los éteres, debido a que ambos mantienen una similaridad estructural. El sexto grupo, se conformó por moléculas que contienen un átomo de nitrógeno unido a hasta tres grupos alquilo, o bien a un anillo aromático. Las sustancias que tienen al menos un átomo halógeno unido a un grupo alquilo o a un átomo de carbono aromático fueron asignadas al grupo 7. Finalmente, el último grupo se confeccionó con aquellas moléculas que contienen un anillo formado por más de un tipo de átomo (nitrógeno y oxígeno, además de carbono).

El objetivo de la metodología consiste en analizar primero si la distribución en grupos guarda alguna relación con el valor de la variable que se quiere modelar o no. Esto se hace analizando los promedios y las desviaciones estándar de los valores experimentales de los compuestos del entrenamiento asignados a cada grupo. La característica ideal sería que todos los grupos tengan una desviación estándar baja y valores de promedio bien diferenciados. Esto significaría que cada grupo capturó ciertas características que marcan un rasgo distintivo con respecto a la variable a modelar.

En segundo lugar, otro objetivo es evaluar el desempeño del modelo de predicción usado. Para esto, un único modelo de predicción con todos los datos de entrenamiento es entrenado. Luego, se evalúa el desempeño del mismo, obteniendo la media y el desvío estándar de las predicciones, discriminadas por cada uno de los grupos. De esta forma comparamos si estas métricas se mantienen similares con respecto a las obtenidas con los valores experimentales, y así determinar si existe un sesgo o no. Finalmente, se evalúa la posibilidad de corrección de la predicción de un compuesto $\mathbf{x_i}$ a partir del sesgo que su grupo genera. La Función 6.1 muestra como se aplica esta corrección, donde $P_T(\mathbf{x_i})$ es el valor calculado por el método de predicción P entrenado sobre el conjunto T, $\bar{U}_T(\mathbf{x_i})$ es la media experimental de los compuestos del entrenamiento que pertenecen al mismo grupo que $\mathbf{x_i}$ y $\hat{P}_T(\mathbf{x_i})$ corresponde al valor de predicción corregido:

$$\hat{P}_T(\mathbf{x_i}) = P_T(\mathbf{x_i}) + \frac{\bar{U}_T(\mathbf{x_i}) - P_T(\mathbf{x_i})}{2} \tag{6.1}$$

6.2.3. Diseño de los experimentos

El SOM utilizado en esta experiencia fue entrenado durante 500 generaciones, en donde el entrenamiento fue dividido en dos fases. En la primera fase, la tasa de aprendizaje comienza con 0.9 (φ_0 en Ecuación 3.8) hasta el comienzo de la segunda fase, en donde la tasa se fija en 0.02. Algo similar sucede con el alcance de la vecindad σ_0 (Ecuación 3.9), donde en la primera fase es igual a la máxima distancia entre dos nodos, y a partir de la segunda fase σ se mantiene en 1. Se experimentó con celdas de diferentes tamaños, pero la mayor parte de los resultados reportados se obtuvieron con una grilla cuadrada de 7×7 (en los resultados obtenidos usando otro tipo de grilla, se aclara explícitamente la grilla usada).

Los datos de los compuestos usados corresponden a los descriptos en el Apéndice B.2, en donde un 80 % de los datos se usó para entrenar y el 20 % restante para testeo. Se utilizó un conjunto de 36 descriptores usando el método descripto en la sección 5.3.5.2, en donde se buscó un conjunto que fuera relevante pero no minimal. Con los datos del entrenamiento se construyó un SOM, como así también un CRN de 3 redes, en donde el algoritmo de aprendizaje usado en este caso es el de Levenberg-Marquardt con regularización bayesiana (3.2.5.1).

6.2.4. Resultados

En primer lugar analizaremos cómo quedan agrupados los compuestos en el SOM. La Figura 6.4 muestra el color, y su valor asociado, del promedio del valor experimental de todos los compuestos que son asignados a cada nodo. Podemos apreciar que existe una cierta distribución homogénea de los valores experimentales, esto es, en el borde inferior derecho vemos que se han agrupado compuestos hidrofílicos, mientras que en el borde superior izquierdo lo han hecho los hidrofóbicos. Además la transición de colores entre un extremo y otro sucede de manera gradual. Esto resulta muy interesante teniendo en cuenta que el valor de logP no fue considerado al momento del entrenamiento del SOM y aún así, hay cierta configuración espacial relacionada con la variable experimental.

De la misma manera, la Figura 6.5 muestra la misma información pero del desvío estándar en lugar del promedio. Según se aprecia, con excepción de algunos

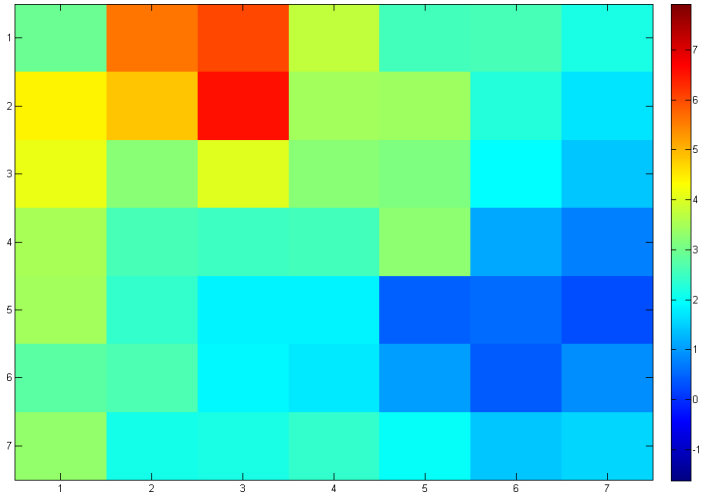


Figura 6.4: Promedio del logP de los compuestos del entrenamiento discriminado por celda.

pocos nodos, la dispersión en cada grupo es baja, lo que implica que, en general, dentro de cada grupo se tienen valores experimentales similares. No obstante, el alto desvío estándar de un nodo puede darnos información de utilidad, de la cual tomaremos ventaja en el último método de este capítulo.

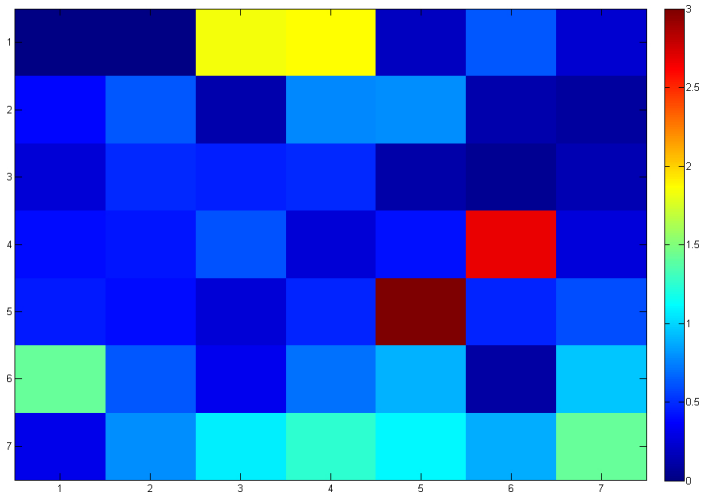


Figura 6.5: Desvío estándar del logP de los compuestos del entrenamiento discriminado por celda.

El mismo análisis puede hacerse para mapas de distintas cantidades de nodos. En la parte superior de la Figura 6.6 se muestran los valores promedios de los compuestos entrenados con mapas de dimensión 10 y 15 respectivamente. Como se ve, la continuidad de valores entre nodos cercanos se mantiene. Es importante notar, que los mapas han tomado otra distribución en cuanto a la ubicación de los valores más altos y más bajos. Esto no invalida el método, y sucede debido a la aleatoriedad existente tanto en los pesos iniciales como en el orden de elección de compuestos durante el entrenamiento del SOM. Asimismo, vemos que a medida que se aumenta el tamaño de la grilla aparecen más nodos que no quedan asignados a ningún compuesto (pintados en blanco). Finalmente, complementamos el análisis anterior con los desvíos de estos mapas, indicados en la parte inferior de la Figura 6.6. Nuevamente se aprecia que, en general, la dispersión dentro de cada celda es baja.

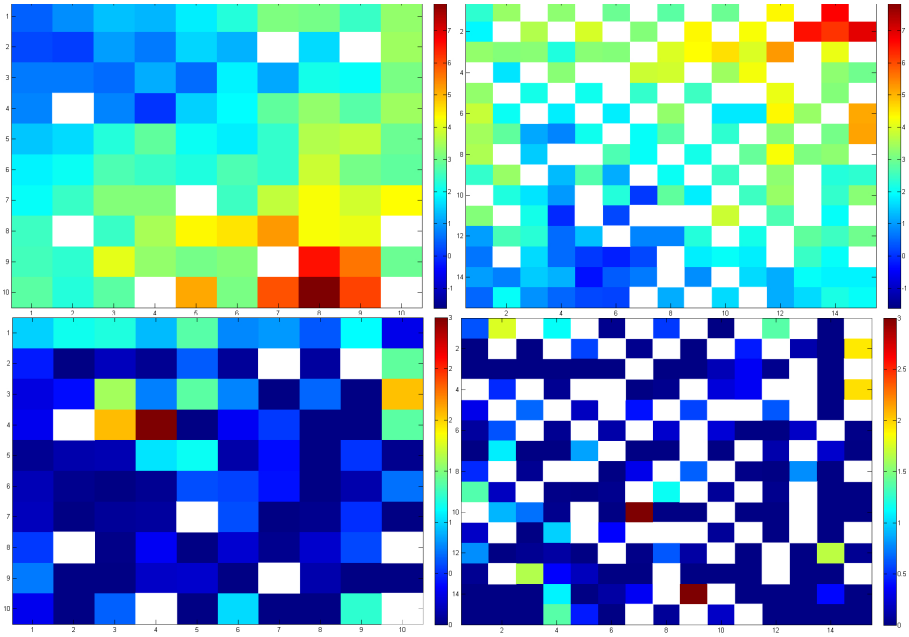


Figura 6.6: Promedio (arriba) y desvío estándar (abajo) del logP de los compuestos del entrenamiento discriminados por celda, utilizando mapas de 10 ×10 (izquierda) y 15×15 (derecha).

En cuanto a la separación por clases químicas, el mismo análisis no arrojó un

resultado tan alentador como el realizado con el SOM. En la Tabla 6.1 se muestran los promedios y desviaciones estándar obtenidos para cada grupo con esta separación manual. Como se ve, esta agrupación por clases químicas no guarda relación con el grado de hidrofobia de los compuestos, ya que los promedios obtenidos son bastantes similares (salvo en las clases 1, 7 y 9) y el desvío estándar siempre es superior a 0,86. Por tal motivo, y luego de una serie de intentos fallidos, abandonamos nuestra experimentación con esta división manual.

	Cantidad Comp.	Promedio logP	Desvío Est. logP
Grupo 1	117	3,2326	1,0690
Grupo 2	12	1,5658	1,4089
Grupo 3	32	1,4119	1,15
Grupo 4	57	1,8100	0,9575
Grupo 5	26	2,0392	1,2593
Grupo 6	37	1,4986	0,9921
Grupo 7	119	3,2050	1,6054
Grupo 8	24	1,3738	1,0199
Grupo 9	16	0,6556	0,8628

Tabla 6.1: Promedio y desviaciones de las clases químicas.

Una vez que el CRN fue entrenado, asociamos el valor de predicción de cada compuesto del entrenamiento a su nodo ganador y se calculó el promedio por celda (Figura 6.7). Visualmente, podemos contrastar que se obtiene una disposición semejante a la de la Figura 6.4. Esto nos indica, en forma aproximada, que al menos los compuestos del entrenamiento son modelados por la red correctamente. Sin embargo, es importante analizar cómo predice el CRN para los compuestos reservados para el testeo.

En la Tabla 6.2 se muestra el resultado de las predicciones realizadas con el CRN sobre el conjunto de testeo. Aquellos errores mayores a 0,5 unidades de logP se muestran en negrita. A juzgar por los errores promedios cuadrados (MSE) y los errores promedios absolutos (MAE), podemos alegar que la red tiene un comportamiento aceptable: MSE= 0,1867 y MAE= 0,3018. No obstante proponemos tomar aquellos compuestos con los que tuvimos un error en la predicción mayor a 0,5 unidades de logP e intentar mejorar estas predicciones con la ayuda del SOM. Esta proposición no la consideramos como una metodología en sí, ya que involucra conocer los valores de testeo, sino que fue utilizada como un método exploratorio para la identificación de situaciones de error.

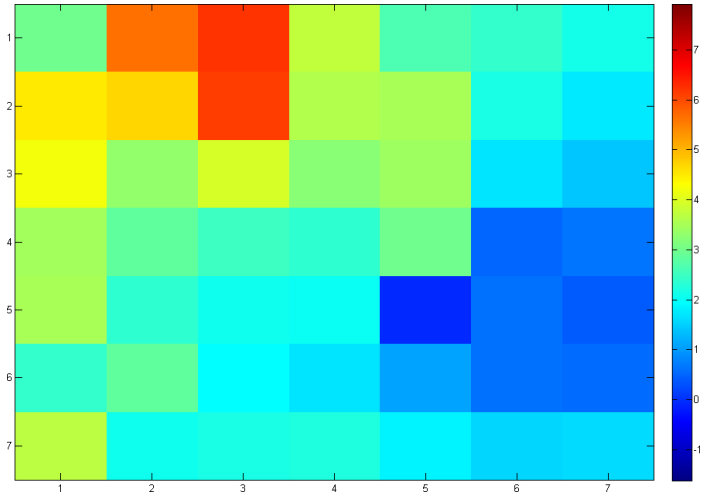


Figura 6.7: Promedio de las predicciones de logP de los compuestos del entrenamiento discriminado por celda.

Para esto, analizamos primero cuáles son los nodos que pertenecen a cada uno de los compuestos con mala predicción. Luego, hacemos tender la predicción obtenida por el CRN al valor promedio experimental de esa celda según la Ecuación 6.1. La Tabla 6.3 nos muestra esta información resumida con el valor de predicción corregido ($\hat{P}_T(\mathbf{x_i})$). Podemos ver que en este proceso de corregir las predicciones en sentido del promedio del SOM, se disminuyó el error en la mayoría de los casos (Error 2). Los errores generales logrados ahora son: MSE= 0,1624 y MAE= 0,2791.

6.3. Método híbrido de identificación de dominio de aplicación

En la sección 4.4.6 se describió la importancia de la definición de un dominio de aplicación para un método basado en QSAR y la problemática asociada a su correcta estimación. En la presente sección presentamos una metodología para la tarea de determinar la confiabilidad asociada a la predicción de un nuevo compuesto. La presente metodología, presenta ciertas características distintivas

Ind.[a]	logP$_{exp}$	CRN	Error	Ind.	logP$_{exp}$	CRN	Error
62	4,57	4,2767	-0,2933	288	1,21	1,1201	-0,0899
34	4,38	4,3406	-0,0394	426	2,52	1,6341	**-0,8859**
309	4,11	3,7555	-0,3545	425	1,81	1,9266	0,1166
135	2,34	3,4080	**1,0680**	171	2,34	2,4683	0,1283
67	1,77	1,7066	-0,0634	418	0,4	0,6393	0,2393
344	3,13	3,1905	0,0605	433	1,44	1,6735	0,2335
132	2,72	3,1232	0,4032	346	3,13	3,7784	**0,6484**
249	4	3,8185	-0,1815	438	0,27	0,4762	0,2062
427	2,69	2,5138	-0,1762	357	2,31	2,1065	-0,2035
341	3,44	3,4488	0,0088	350	2,06	1,9062	-0,1538
329	2,11	1,7346	-0,3754	355	4,22	4,0842	-0,1358
198	5,07	4,9692	-0,1008	22	2,37	2,4189	0,0489
279	4,14	4,5585	0,4185	105	2,14	1,9661	-0,1739
286	3,15	3,0897	-0,0603	43	2,1	2,0841	-0,0159
63	3,5	3,5943	0,0943	89	5,15	5,3265	0,1765
254	2,86	2,7136	-0,1464	120	2,77	3,4289	**0,6589**
59	0,65	0,8186	0,1686	429	3,2	1,3818	**-1,8182**
247	4,01	3,8748	-0,1352	207	2,3	3,0049	**0,7049**
129	3,21	3,0340	-0,1760	16	2,82	2,6862	-0,1338
21	1,99	2,0062	0,0162	38	2,75	2,6593	-0,0907
10	4	4,3054	0,3054	199	2,92	3,9554	**1,0354**
259	1,72	2,3476	**0,6276**	6	0,75	0,7928	0,0428
250	3,45	3,5250	0,0750	326	3,42	3,0776	-0,3424
301	-0,54	-0,3606	0,1794	7	1,24	0,7864	-0,4536
261	4,09	3,2498	**-0,8402**	178	4,61	4,6218	0,0118
197	3,58	3,4629	-0,1171	102	1,32	1,6476	0,3276
196	2,42	3,0746	**0,6546**	24	2	2,4709	0,4709
208	1,88	2,3387	0,4587	30	4,11	3,3222	**-0,7878**
192	-1,38	-1,1113	0,2687	103	2,58	2,4402	-0,1398
253	0,81	0,8381	0,0281	114	1,9	1,7202	-0,1798
399	2,19	2,3337	0,1437	227	2,99	2,6152	-0,3748
362	1,95	1,7794	-0,1706	177	3,33	3,0110	-0,3190
252	1,23	1,2533	0,0233	81	5,02	5,2585	0,2385
391	2,94	3,0198	0,0798	15	1,48	2,7586	**1,2786**
189	3,23	2,8301	-0,3999	432	1,97	2,3688	0,3988
61	3,15	2,3426	**-0,8074**	387	1,56	1,4534	-0,1066
188	1,94	1,8228	-0,1172	361	1,85	1,9972	0,1472
328	-0,66	-0,4611	0,1989	331	0,18	0,5640	0,3840
312	0,59	0,6851	0,0951	320	1,78	1,8777	0,0977
165	1,16	1,2516	0,0916	382	1,26	1,5178	0,2578
176	2,43	2,2360	-0,1940	407	-0,22	0,3407	**0,5607**
121	1,71	1,8857	0,1757	293	-0,3	-0,0447	0,2553
294	1,04	1,2158	0,1758	411	0,46	0,7713	0,3113

Tabla 6.2: Predicciones del CRN para el conjunto de testeo.[a] Índice del compuesto según Yaffe *et al.* (2002)

sobre otros enfoques de la literatura y representa una técnica de interés a juzgar por los resultados obtenidos.

Con respecto a la taxonomía descripta en la sección 4.4.6 advertimos que nuestro enfoque no se ajusta exactamente a las clases descriptas, aunque tiene ciertos aspectos de los métodos basados en medidas de distancias. A diferencia de los enfoques mostrados anteriormente, nuestra metodología no sólo verifica que haya compuestos similares en el entrenamiento, sino que además identifica

Ind.	logP Exp	CRN	Error	Nodo Asoc.	$\hat{P}_T(\mathbf{x_i})$	Error 2
135	2,34	3,4080	1,0680	46	2,8113	0,4713
259	1,72	2,3476	0,6276	42	1,4564	-0,2636
261	4,09	3,2498	-0,8402	15	3,7387	-0,3513
196	2,42	3,0746	0,6546	15	3,6511	1,2311
61	3,15	2,3426	-0,8074	29	2,9283	-0,2217
426	2,52	1,6341	-0,8859	33	0,7738	-1,7462
346	3,13	3,7784	0,6484	8	4,1494	1,0194
120	2,77	3,4289	0,6589	26	3,1998	0,4298
429	3,2	1,3818	-1,8182	5	2,0094	-1,1906
207	2,3	3,0049	0,7049	25	2,6929	0,3929
199	2,92	3,9554	1,0354	11	3,7698	0,8498
30	4,11	3,3222	-0,7878	26	3,1465	-0,9635
15	1,48	2,7586	1,2786	21	2,0825	0,6025
407	-0,22	0,3407	0,5607	41	0,4687	0,6887

Tabla 6.3: Predicciones mejoradas con el SOM.

problemas del modelo mismo y de los datos usados para modelar. Esto se debe a que aún dentro del espacio de los compuestos de entrenamiento existen distintos niveles de complejidad del modelo (Caruana *et al.* (2000)) y por lo tanto no siempre en todo entorno cercano a los compuestos del entrenamiento el método es igualmente preciso.

6.3.1. Idea del método

Tal como hemos mencionado, nuestra metodología busca identificar los compuestos del conjunto de testeo cuyas predicciones resultaran *confiables* y aquellas que resultaran *no confiables*. Usaremos el término *confiable* para la predicción de un compuesto cuando se obtiene un error 'bajo' de predicción al utilizarse un dado P_T, donde P_T corresponde al método de predicción P entrenado usando un conjunto de entrenamiento T. En forma análoga, definiremos el término *no confiable* para un compuesto cuando se obtiene un error 'alto' en su predicción al usarse un P_T. Aún cuando esta separación pareciera ser de tipo binaria, dado la complejidad de los modelos de predicción definiremos el término de confiabilidad *no categorizada* cuando no podamos determinar su grado de confiabilidad de manera segura.

En base a los términos anteriormente definidos, un conjunto de testeo H lo dividiremos en tres subconjuntos:

S_1**:** Compuestos que se espera que sean *confiables*.

S_2**:** Compuestos que se espera que sean *no confiables*.

S_3**:** Compuestos con confiabilidad *no categorizada*.

La idea de nuestra metodología es, por lo tanto, establecer una técnica que permita asignar todo compuesto de un conjunto de testeo a uno de los 3 subconjuntos anteriores. El método fue diseñado para que tenga un comportamiento conservador, en el sentido que no queremos que cometa errores al asignar un compuesto a la clase S_1 o S_2 y ante la posibilidad de error o alta incertidumbre en la determinación, el compuesto debería ser asignado a la clase S_3. Tal como se planteó el método, las siguientes propiedades se cumplen en nuestros tres subconjuntos: $S_1 \cap S_2 = \emptyset$ y $S_3 = H - (S_1 \cup S_2)$.

Analizando más profundamente este tema, resaltamos que a un compuesto químico no se lo asigna como *confiable* o *no confiable* por sí mismo, sino que el grado de confiabilidad es determinado por la precisión en la predicción de su variable a modelar cuando un método P_T es usado. Sin embargo, en un escenario real la variable experimental de un compuesto nuevo no es conocida de antemano. Por lo tanto, la dificultad claramente está en la necesidad de predecir la confiabilidad de un compuesto cuando no se tiene información de su variable a modelar. Nótese también que existe un similitud muy cercana entre los conceptos de predicción de la confiabilidad y la definición del dominio de aplicación.

La técnica que describiremos en esta tesis intenta determinar para un compuesto $\mathbf{x_i}$ perteneciente al espacio de posibles compuestos $X \subset \mathbb{R}^n$, donde n es la cantidad de descriptores moleculares usados, tal que $\mathbf{x_i} \notin T$, cual va a ser su confiabilidad cuando se lo intenta predecir con un método P_T, es decir cuando se calcula $P_T(\mathbf{x_i})$. Esta técnica se apoya en las siguientes tres conjeturas para determinar que un compuesto $\mathbf{x_i}$ es considerado como *confiable*:

Conjetura 1 (C_1). $\mathbf{x_i}$ debería ser lo suficientemente similar a algún subconjunto de compuestos del set de entrenamiento (los cuales llamaremos como vecinos

de $\mathbf{x_i}$). Además (*Conjetura 1b*), el número de vecinos de $\mathbf{x_i}$ debe ser lo suficientemente grande.

Conjetura 2 (C_2). Los compuestos vecinos de $\mathbf{x_i}$ deberían tener una baja dispersión de sus valores experimentales.

Conjetura 3 (C_3). Cuando al conjunto de compuestos vecinos de $\mathbf{x_i}$ se les calcula su variable de predicción usando P_T, la media y la dispersión de los valores calculados deberían ser similares a la de sus valores experimentales. Además (*Conjetura 3b*), los errores obtenidos para el subconjunto de compuestos vecinos deberían tener un error de predicción lo suficientemente bajo.

Claramente se observa que el modo con el que está descripta cada conjetura es general e imprecisa, ya que los términos tales como 'suficientemente similar' o 'suficientemente bajo' son orientativos y dependerá finalmente de las distribuciones específicas de los datos y de qué criterio se tome para evaluar las conjeturas. La similaridad mencionada, la cual se usa para calcular los vecinos de $\mathbf{x_i}$, debe medirse en algún espacio relacionado con el espacio de descriptores de T. Otra observación importante es que los compuestos vecinos de $\mathbf{x_i}$ pertenecen siempre al conjunto de entrenamiento.

Cada conjetura surge de las diferentes debilidades que pueden estar presentes en cualquier modelo. La *conjetura 1* se relaciona con la falta de datos similares, ya sea por cercanía o cantidad, en el conjunto de entrenamiento con respecto al dato $\mathbf{x_i}$. Es claro que la predicción de un compuesto nuevo no resulta confiable si no se tiene un buen número de compuestos similares en el conjunto de entrenamiento. Esto parte de un principio fundamental del aprendizaje automático, en donde lo que se infiere debe guardar cierta relación con lo que se conoce. Es importante aclarar que esto no justifica la separación intencional de compuestos similares a los usados en el entrenamiento para ser usados en un set de testeo, de manera de asegurarse así tener buenas métricas de predicción en los resultados (Leonard & Roy (2005)).

En la segunda conjetura, la existencia de un alto valor de dispersión de la variable a modelar en un entorno de $\mathbf{x_i}$ indicaría que el conjunto de descriptores

usados para entrenar el modelo de predicción no son apropiados, al menos para la región determinada por los vecinos de $\mathbf{x_i}$. Esta conjetura parte de la suposición que para que un set de descriptores sea relevante, en un entorno acotado no debería haber un alta variabilidad de la variable a modelar. Por otra parte, este concepto de que la influencia o la validez de un subconjunto de descriptores puede estar acotada a un subdominio del espacio de descriptores resulta sensato de imaginar considerando que, por ejemplo, los descriptores relevantes para predecir una propiedad experimental pueden ser distintos en función del tipo de compuesto que se modela. Este concepto se corresponde con lo analizado en la sección 6.1.

Finalmente, la tercer conjetura parte de la suposición que un modelo puede tener determinados subdominios en donde no se capturó correctamente la lógica del modelo. Cuando los criterios de esta conjetura no se satisfacen, esto sería un indicador que el método P_T posee un sesgo (inexactitud) o una varianza diferente a la que debería tener, al menos en el entorno determinado por los vecinos de $\mathbf{x_i}$. Este concepto resulta novedoso, ya que la capacidad de predicción se evalúa generalmente utilizando una métrica que engloba todos los compuestos utilizados y no se analiza si el método posee diferentes niveles de precisión de acuerdo al tipo de compuesto o de acuerdo al subespacio que se esté modelando (Caruana *et al.* (2000)). La debilidad de un método en un determinado entorno puede deberse al método en sí mismo o, nuevamente, a que los descriptores no son lo suficientemente descriptivos en ese entorno del espacio de datos. En este sentido, la alta dispersión en la predicción de los vecinos de $\mathbf{x_i}$ puede estar asociada a un sobreajuste de esos datos por parte del método. Asimismo, el sesgo puede ser ocasionado por la incapacidad del método de ajustar en esa región de los datos, ya sea esto por incapacidad del método o por usar descriptores no suficientemente relevantes.

De las tres conjeturas, identificamos a la C_1 como la conjetura más fuerte en el sentido que, de no cumplirse, introduce el mayor grado de probabilidad de error.

6.3.2. Descripción de la metodología

Describiremos nuestra técnica como una función Υ que toma un compuesto $\mathbf{x_i}$ para ser modelado con un método P_T, el cual fue entrenado con un conjunto

de entrenamiento T, realiza una evaluación de las conjeturas antes presentadas y genera una salida tal como se muestra en la Ecuación 6.2. En esta sección ahondaremos sobre los detalles de implementación desarrollados para nuestro método de detección de dominio de aplicabilidad de un método.

$$\Upsilon(\mathbf{x_i}|P_T, T) = \begin{cases} 1 & : \mathbf{x_i} \in S_1 \\ 2 & : \mathbf{x_i} \in S_2 \\ 3 & : \mathbf{x_i} \in S_3 \end{cases} \tag{6.2}$$

La técnica aquí desarrollada está basada en el uso combinado de métodos de aprendizaje no supervisado y ciertos conceptos del campo de la estadística multivariada. Esta metodología utiliza la información de cuáles son los compuestos usados para el entrenamiento (T) y el modelo de predicción que pretende ser usado como estrategia de modelado (P_T). Vale destacar que tanto T como P_T provienen de un modelo ya existente y por lo tanto no son parte del método de confiabilidad aquí presentado.

El método de aprendizaje no supervisado que usaremos son los SOMs. La elección de esta técnica se centró en dos hechos concretos: su capacidad de detección automática de agrupamientos en los datos y su preservación de las relaciones topológicas entre los grupos encontrados (sección 3.3.2).

Asumiremos que el SOM, al cual denominaremos con la letra U, tiene ℓ nodos. Sin pérdida de generalidad, podemos establecer un orden entre los ℓ nodos. Cuando U es entrenado con los compuestos de T, nos referiremos a este modelo como U_T. El nodo ganador de un compuesto es aquel que posee el peso más similar a la entrada presentada. En este sentido, el nodo ganador de un compuesto $\mathbf{x_i}$ que surge al aplicarse U_T, lo indicaremos como $U_T(\mathbf{x_i}) = k$, donde k, donde $1 \leq k \leq \ell$. Finalmente, llamaremos $T_k \subseteq T$ al subconjunto de compuestos del conjunto de entrenamiento T que tienen al k-ésimo nodo como nodo ganador.

6.3.2.1. Evaluación de las conjeturas

El funcionamiento de nuestra metodología dependerá del modo en que se evalúen las conjeturas, ya que las mismas fueron planteadas de una forma abstracta. Describiremos ahora nuestra propuesta de evaluación de conjeturas, las

cuales no pretenden ser la manera óptima de evaluarlas sino solamente una forma sensata.

La conjetura C_1 es cuantificada haciendo uso del estadístico T^2 de Hotelling (De Maesschalck *et al.* (2002)), el cual está basado en la distancia de Mahalanobis (Ecuación 6.3). Básicamente, este estadístico T^2 permite medir la distancia de un dato a un grupo de datos, y para esto utiliza la inversa de la matriz de covarianza del grupo. Por lo tanto, para el compuesto de testeo $\mathbf{x_i}$, el grupo al cual se le mide su proximidad es T_k (los vecinos de $\mathbf{x_i}$), donde $U_T(\mathbf{x_i}) = k$. En caso que la matriz de covarianza de T_k esté mal condicionada, su inversa ya no será confiable, por lo tanto se deben agregar más datos para poder hacer la comparación viable. Estos datos son obtenidos de los compuestos que tienen como nodo ganador a los nodos vecinos de k. Esta elección es posible, ya que los SOMs preservan la relación topológica entre los pesos de nodos vecinos de un mismo mapa. De esta forma, la probabilidad β_i de ser similar al grupo T_k puede ser calculada usando la relación existente entre el estadístico T^2 y la distribución $\mathcal{F}$, la cual se muestra en la Ecuación 6.4. Finalmente la conjetura C_{1b} se analiza de acuerdo a la cantidad de vecinos de $\mathbf{x_i}$ esto es $|T_k|$

$$\mathrm{T}^2 = (\mathbf{x_i} - \bar{T_k}) \cdot cov(T_k)^{-1} \cdot (\mathbf{x_i} - \bar{T_k})' \tag{6.3}$$

$$\frac{|T_k|^2 - n|T_k|}{n|T_k|^2 - n} \cdot \mathrm{T}^2 \sim \mathcal{F}_{(\alpha, n, |T_k|-n)} \tag{6.4}$$

El caso de la evaluación de C_2 es bastante simple. Para esta evaluación se utilizó la desviación estándar de la variable a modelar de los compuestos que pertenecen a T_k como medida de la dispersión en la vecindad del compuesto a evaluar. Llamaremos γ_k a este desvío estándar. La Ecuación 6.5 nos indica la forma de calcular γ_k en donde se considera la función $y(T)$, la cual devuelve la variable a modelar del conjunto T.

$$\gamma_k = std(y(T_k)) \tag{6.5}$$

La conjetura C_3 es evaluada mediante dos criterios. Uno de ellos es calcular la diferencia (θ_k) entre la media de la variable a modelar de los compuestos que pertenecen a T_k con la media de la predicción de la variable a modelar de T_k usando P_T (Ecuación 6.6). El otro criterio consiste en calcular la diferencia (ϵ_k) entre la desviación estándar de la variable a modelar de los compuestos que pertenecen a T_k y la desviación estándar de los valores calculados con el método de predicción P_T al predecir los compuestos de T_k (Ecuación 6.7). Finalmente, la conjetura C_{3b} es evaluada de acuerdo al error absoluto medio (ζ_k) obtenido al predecir cada uno de los compuestos de T_k usando P_T (Ecuación 6.8).

$$\theta_k = y(\bar{T}_k) - P_T(\bar{T}_k) \tag{6.6}$$

$$\epsilon_k = std(y(T_k)) - std(P_T(T_k)) \tag{6.7}$$

$$\zeta_k = \frac{y(T_k) - P_T(T_k)}{|T_k|} \tag{6.8}$$

6.3.2.2. Estableciendo umbrales para la evaluación de las conjeturas

Las ecuaciones dispuestas en el apartado anterior necesitan valores de referencia para establecer si se está satisfaciendo o no cada conjetura. Nuevamente, los umbrales establecidos no pretenden ser óptimos ni definitivos, aunque resaltamos que como todo criterio de corte, siempre se tiene un grado de arbitrariedad en la elección. Por lo tanto, para un compuesto $\mathbf{x_i}$ que tiene a los compuestos T_k como sus vecinos, hemos establecido las siguientes condiciones de evaluación para las conjeturas:

- C_1 no se satisface si $\beta_i < 0{,}05$.
- C_{1b} no se satisface si $|T_k| < \frac{|T|}{\ell}$, esto es si la cardinalidad de T_k es menor que el tamaño promedio de todos los grupos.

- C_2 no se satisface si $\frac{\gamma}{2} < \gamma_k$, donde γ es el desvío estándar de la variable a modelar de todos los compuestos del entrenamiento T.
- C_3 no se satisface si $\frac{\delta}{2} < \theta_k$, donde δ es el error de predicción promedio de todos los compuestos de T cuando se le aplica P_T, o si $\frac{\gamma}{3} < \epsilon_k$.
- C_{3b} no se satisface si $\zeta_k < \frac{3}{2}\delta$.

Basándose en estas condiciones, nuestro método Υ (Ecuación 6.2) determina la confiabilidad de la predicción de un compuesto. Si C_1 no se satisface, $\Upsilon = 2$ y por lo tanto la predicción se considera *no confiable*. Por otra parte, si C_1 así como C_{1b}, C_2, C_3 y C_{3b} se cumplen, entonces $\Upsilon = 1$. Finalmente, cuando ninguno de los dos casos anteriores se cumple, al no poder considerarse una predicción segura aún cuando se encuentra dentro de un entorno en el que compuestos similares fueron modelados, se determina $\Upsilon = 3$. La Figura 6.8 muestra un diagrama de flujo de la salida del método en función de las distintas evaluaciones de las condiciones.

6.3.3. Diseño de los experimentos

Cuatro conjuntos de datos fueron usados aquí a fin de evaluar el funcionamiento de nuestro método. Para cada uno de ellos utilizamos conjuntos de descriptores con distinto grado de relevancia con respecto a la variable de modelado. La idea de esta decisión consiste en que tanto el modelo P_T como los datos T no necesariamente deben ser los óptimos.

Los primeros tres sets de datos, a los cuales llamaremos DS1, DS2 y DS3, se encuentran descriptos en el Apéndice B.3, B.4 y B.2 respectivamente. El cuarto set de compuestos usados, DS4, fue tomado del conjunto de compuestos químicos enunciado en el Apéndice B.1. De ahí se seleccionaron 1939 compuestos de diferentes familias químicas de las cuales teníamos información sobre el valor experimental de logP.

Luego de aplicar la estrategia de selección de descriptores multi-objetivo del capítulo anterior, decidimos utilizar 7, 7, 13 y 17 descriptores para DS1, DS2, DS3 y DS4 respectivamente. Es importante destacar que para DS4 se partió de un pool de descriptores en donde ciertos descriptores relevantes para el cálculo

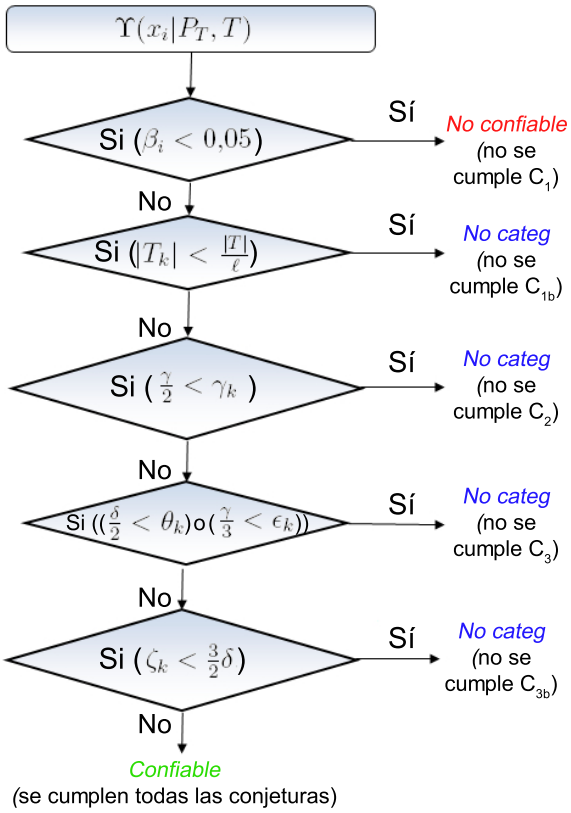


Figura 6.8: Diagrama de flujo del testeo de la confiabilidad de un compuesto $\mathbf{x_i}$.

de logP, relacionados con la carga eléctrica de los compuestos, fueron descartados de antemano. Con esto se pretende evaluar nuestra metodología en un caso donde el modelo de predicción no sea el óptimo y por ende se tenga una mayor incertidumbre sobre los resultados de la predicción. Asimismo, este cuarto set de datos posee una cantidad de compuestos altamente superior a los restantes, por lo que los resultados obtenidos serán mucho más significativos desde el punto de vista estadístico.

Antes de la aplicación de los métodos de selección de descriptores y de modelado, cada uno de los sets de datos fueron divididos en dos conjuntos en forma aleatoria: el 75 % de los compuestos se usó para el entrenamiento (T) y el 25 % restante se usó como testeo (H). La presente metodología fue aplicada 10 veces a cada uno de los cuatro sets de datos mencionados. En cada réplica se aplicaba una nueva separación de los datos en subconjuntos de entrenamiento y testeo.

Como método de predicción P_T se usaron redes neuronales con regularización Bayesiana (3.2.5.1). Destacamos nuevamente que la elección del método es indistinta de nuestra metodología de identificación de dominio de aplicabilidad.

Para el caso del SOM, se utilizó una grilla 2D rectangular para su arquitectura, y donde el número de nodos del mapa fue establecido de acuerdo a la cantidad de compuestos a ser entrenados. Más específicamente, utilizamos una topología de 5×5 para DS1 y DS2, una de 7×7 para DS3 y una de 15×15 para DS4. Cada uno de los SOMs fue entrenado durante 1500 épocas divididas en dos fases, comenzando la segunda luego de las primeras 500 épocas. La tasa de aprendizaje se decrementa de 0,9 a 0,52 al pasar de una fase a la otra. Todos los nodos vecinos son afectados al comienzo del algoritmo y esta cantidad se reduce en forma lineal, hasta llegar a afectar solamente a los vecinos de la grilla que están a 1 unidad de distancia Euclídea al comienzo de la fase de ajuste.

6.3.4. Resultados

Las Tablas 6.4 y 6.5 muestran los errores promedios absolutos (MAE) para los conjuntos de entrenamiento T, testeo H y los subconjuntos S_1 y S_2. Los valores de probabilidad de la última fila indican la probabilidad de cometer un error al determinar que el error en la predicción del subconjunto S_1 o S_2 es distinto al

Tabla 6.4: MAE obtenido para los conjuntos T, H, S_1, S_2 para los sets de datos DS1 y DS2.

	DS1				**DS2**			
Corrida	T	H	S_1	S_2	T	H	S_1	S_2
1	0,23923	0,24876	0,22696	0,32790	0,16418	0,21272	0,19426	0,28761
2	0,24114	0,22884	0,17468	0,38423	0,1584	0,22913	0,17731	0,22724
3	0,23719	0,28427	0,23519	0,31092	0,17395	0,27029	0,14325	0,45516
4	0,23773	0,23202	0,19686	0,42413	0,16708	0,17632	0,15067	0,14894
5	0,2242	0,30957	0,27815	0,46923	0,17264	0,18374	0,16779	0,30403
6	0,23586	0,24661	0,22259	0,36485	0,15899	0,23270	0,21263	0,25933
7	0,22141	0,27986	0,24652	0,34205	0,18424	0,15309	0,17651	0,13548
8	0,2195	0,29556	0,24965	0,49678	0,16314	0,20013	0,13950	0,33147
9	0,23577	0,25263	0,22858	0,31034	0,17006	0,18748	0,16285	0,19435
10	0,24271	0,25491	0,28548	0,40802	0,16871	0,20947	0,16365	0,34812
p	–	–	0,00387	0,00014	–	–	0,01641	0,02655

Tabla 6.5: MAE obtenido para los conjuntos T, H, S_1, S_2 para los sets de datos DS1 y DS2.

	DS3				**DS4**			
Corrida	T	H	S_1	S_2	T	H	S_1	S_2
1	0,12594	0,30057	0,19886	0,41265	0,51727	0,74804	0,50978	0,89252
2	0,16572	0,28278	0,26864	0,54839	0,52721	0,77970	0,54755	1,10659
3	0,13983	0,31039	0,17696	0,46372	0,54631	0,73699	0,52449	0,87387
4	0,14434	0,27720	0,22486	0,33731	0,52856	0,72426	0,42648	0,95670
5	0,13506	0,32405	0,22047	0,61996	0,51953	0,74473	0,53585	0,91316
6	0,14558	0,32209	0,26140	0,40450	0,52124	0,75093	0,53578	1,02836
7	0,15029	0,29948	0,19387	0,47899	0,53509	0,75410	0,50769	1,17760
8	0,14909	0,32494	0,28058	0,49937	0,53242	0,78292	0,55993	0,99182
9	0,15545	0,27064	0,21196	0,49787	0,52888	0,74008	0,47966	1,04189
10	0,13627	0,32837	0,23506	0,59767	0,52535	0,78521	0,47396	1,07254
p	–	–	0,00008	0,00006	–	–	≈ 0	0,00001

error obtenido en el conjunto H. Esta probabilidad es calculada para verificar si efectivamente existen diferencias en la separación de los compuestos en S_1 y S_2 en el error promedio de ambos subconjuntos con respecto al calculado en la columna H. Para realizar el cálculo de esta probabilidad, se realizó un test ANOVA con un diseño de bloques aleatorizado, en donde cada una de las réplicas fue asignada a un bloque distinto a fin de eliminar la varianza provocada por las diferentes particiones de los datos en T y H. En la Tabla 6.6 se reporta los porcentajes de los compuestos que fueron asignados a cada uno de los subconjuntos S_1, S_2 y S_3 para cada uno de los sets de datos.

Tabla 6.6: Porcentaje de los compuestos que fueron asignados a cada uno de los subconjuntos.

	DS1	**DS2**	**DS3**	**DS4**
S_1	63 %	64 %	50 %	79 %
S_2	8 %	17 %	13 %	15 %
S_3	29 %	19 %	37 %	6 %

Los principales desarrollos de esta metodología se encuentran publicados en Soto *et al.* (2009c).

6.4. Conclusiones

En el presente capítulo hemos presentado el resultado de nuestras investigaciones relacionadas con distintos análisis de similaridad aplicados en el dominio químico. La primera propuesta presentada mejoró la calidad predictiva de un modelo mediante la identificación de grupos de compuestos y su posterior entrenamiento individual. Esto permitió realizar modelos adaptados a las características propias de cada grupo, y así obtener mejores resultados con modelos más simples. Sin embargo, advertimos que la misma sufre de ciertas limitaciones. En primer lugar, uno de los objetivos del desarrollo de modelos QSPR es ganar conocimiento mediante la interpretación del modelo desarrollado. En este caso particular, los grupos obtenidos no tienen una interpretación física, por lo que la descripción del modelo final se vuelve más complicada.

La segunda propuesta, aún cuando no correspondería mencionarla como una metodología completa, consideramos que brindó resultados interesantes y nos permitió conocer las potencialidades de los SOMs como técnica de agrupamiento. El principal problema asociado de esta segunda estrategia, es que el modelo de corrección no es lo suficientemente general como para ser aplicado a cualquier método. Sin embargo, esta experiencia sirvió de base para desarrollar nuestra metodología de identificación de dominio de aplicación.

La mayor contribución de este capítulo corresponde a la metodología novedosa presentada al final del mismo, la cual apunta a la mejora de la calidad de los modelos de predicción basados en QSPR. Si bien los modelos de predicción no son

mejorados en sí mismos, la metodología ayuda a identificar compuestos cuyo error de predicción sea alto o bien que la precisión obtenida en la predicción sea incierta, cuando se le aplica un método de predicción determinado. Particularmente, se apuntó a separar los compuestos cuyas predicciones resultaran *confiables* (S_1) de los que fueran a resultar *no confiables* (S_2).

Resaltamos nuevamente que la metodología propuesta es más que una simple identificación de *outliers*. En otras palabras, no sólo se está buscando por problemas de extrapolación, sino también ciertas dificultades que pueden existir en la interpolación. Las tres conjeturas definidas apuntan a identificar las principales causas de error en la predicción de un compuesto, las cuales se pueden resumir como: falta de datos similares (C_1), problemas con los descriptores elegidos (C_2 y C_3) y problemas del modelo utilizado (C_3).

La manera en que cada una de las conjeturas es evaluada no es única, pero a razón de los resultados obtenidos, consideramos que representan una estrategia sensata. El uso de SOM para identificación de dominio de aplicación es de por sí, una contribución novedosa. El SOM tiene ciertas propiedades deseables para ser usado con este objetivo, ya que permite detectar regiones vacías dentro del volumen de los datos y no tiene problemas de cómputo ni de precisión con el aumento de los descriptores en el conjunto de entrenamiento. Asimismo, tampoco requiere del conocimiento de la probabilidad de distribución de los datos. Por otra parte, el estadístico de Hotelling fue elegido ya que es comúnmente usado para detectar observaciones anormales en el control y monitoreo de procesos (Qin (2003)). Este estadístico tiene la ventaja de que es invariante a la escala y considera la covarianza entre las variables. Esta distancia mantiene una relación de escala con la distancia de *leverage* que también fue utilizada en otras propuestas de dominio de aplicación, como por ejemplo en Gramatica & Papa (2003), Jaworska *et al.* (2005), Gramatica (2007).

Los errores obtenidos en los subconjuntos S_1 y S_2 en comparación con los obtenidos en H, en los cuatro sets de datos presentados, evidencian un desempeño muy interesante de nuestra técnica. Finalmente, nótese que las técnicas aquí enunciadas fueron concebidas para el dominio químico, pero su aplicación no escapa a otros posibles escenarios de predicción en un espacio altamente multivariado.

En términos generales, una de las limitaciones asociadas a la aplicación de aprendizaje no supervisado, es que las similitudes encontradas pueden no ser relevantes en términos de la variable que se está modelando. Esto hace que se formen asociaciones entre los datos que no guardan necesariamente una relación con la propiedad que se quiere predecir. En el capítulo siguiente, mostraremos una técnica que resuelve este problema, de manera que la similaridad entre los datos quede relacionada con la variable que se está modelando.

6. INFLUENCIA DEL APRENDIZAJE NO SUPERVISADO EN LA PREDICCIÓN E IDENTIFICACIÓN DE DOMINIO DE APLICACIÓN

Capítulo 7

Inferencias sobre espacios optimizados mediante matrices adaptivas

7.1. Introducción

En el capítulo 5 se presentaron técnicas de selección de variables, las cuáles consisten en hacer una búsqueda dentro del espacio multivariado previa a la tarea de predicción. En el caso de las técnicas presentadas en este capítulo, la selección de variables se realiza como parte del mismo proceso de construcción del modelo de predicción. A este tipo de métodos se los denomina como *embebidos*, los cuales difieren de los denominados *wrapper* presentados en el capítulo 5. Estas técnicas se encuadran también dentro de lo que se conoce con el nombre de *Ponderación de Variables* (Sun (2007); Xu *et al.* (2007)), que en lugar de realizar una búsqueda entre las distintas combinaciones de subconjuntos se aplica una reducción, escalado o proyección de los datos, de manera de obtener una transformación de los mismos en donde una cierta representación pueda ser optimizada. Usaremos el concepto de SARDUX (detección de relevancia de atributos en forma supervisada usando comparaciones cruzadas, del inglés *supervised attribute relevance detection using cross-comparisons*) descripto en el capítulo 3.4.1, y a partir de éste propondremos una serie de metodologías relacionadas, las cuales hemos aplicado en el campo de la quimioinformática. Básicamente, el algoritmo de SARDUX

plantea una transformación de los datos en donde se minimiza la separación entre compuestos de una misma clase y se maximiza la separación entre compuestos de clases diferentes.

Si bien al usar SARDUX la selección de variables se consigue de manera automática, el fin último de éstas técnicas consiste en encontrar un nuevo espacio matemático en donde las observaciones, o compuestos químicos en el caso de la quimioinformática, puedan ser mejor modelados. En la propuesta original de SARDUX, la técnica fue concebida para modelos de 2 clases; sin embargo, en este capítulo mostramos una extensión a múltiples clases y luego una estrategia más general la cual admite modelar tanto variables discretas como continuas.

Asimismo, si este nuevo espacio matemático es descripto mediante dos o tres dimensiones, la metodología permite un análisis visual que resulta de interés para analizar la disposición y la similaridad de los compuestos utilizados en relación con la variable experimental que se desea modelar. En comparación con otras técnicas de reducción, proyección o transformación de variables, como PCA (análisis de componentes principales), FA (análisis de factores) o MDS (escalado multi-dimensional), SARDUX, tal como su nombre indica, es una técnica supervisada. Esto significa que la variable de predicción es tenida en cuenta para la configuración del espacio de menor dimensionalidad.

7.2. 1º Propuesta: SARDUX utilizando dos clases

Como primera incursión en los métodos de transformación basados en matrices adaptivas, decidimos aplicar el algoritmo de SARDUX tal como se lo describe en la sección 3.4.1. Para ello utilizamos el set de datos descripto en el apéndice B.2 con el agregado de 61 descriptores tal como se hizo en la sección 5.3.5.2. El algoritmo de SARDUX requiere que los datos estén asignados a clases. Por tal motivo, aplicamos una separación del set de datos a utilizar de manera que logremos discretizar la variable a modelar en dos clases. La motivación de esta propuesta surge de la hipótesis que la ponderación de los descriptores para una tarea de clasificación entre compuestos con altos y bajos valores experimentales,

sería también de utilidad para ganar conocimiento sobre el modelo de la variable original descripta con números reales.

Para categorizar al set de compuestos en dos clases, aplicamos el algoritmo de las k-medias con $k = 2$. Se aplicaron 100 corridas de 1500 iteraciones cada una, de manera de asegurarnos resultados estables. De esta manera se determinó un umbral $\log P_* = 2{,}61$, según resultó el promedio entre el mínimo valor experimental de los compuestos vinculado al centroide de valor experimental más alto y el máximo valor experimental de los compuestos vinculado al centroide de valor experimental más bajo. De este modo, cerca del 57 % de los datos con valores experimentales menores o iguales a $\log P_*$, fueron asignados a la clase 0, y los restantes compuestos a la clase 1. Trabajar únicamente con dos clases tiene la ventaja que su posterior modelo de predicción puede establecerse con la determinación de un único punto sobre la recta a la cual se proyectan los datos.

Una vez categorizados todos los compuestos en una de las dos clases, separamos un subconjunto de los datos, los cuales utilizamos como conjunto de testeo. Tal como explicamos en la sección 5.4.4.1, el conjunto de testeo propuesto en Yaffe *et al.* (2002) no resulta representativo para mostrar la capacidad de generalización de un modelo. Por lo tanto en este caso, tomamos para el testeo 30 compuestos que cubrieran el rango de valores experimentales uniformemente y que no estuvieran confinados dentro del volumen descripto por los datos restantes. A todos los compuestos les aplicamos una normalización basada en los rangos de cada descriptor, de manera que ciertos descriptores no tuvieran un aporte más importante por el sólo hecho de tener una mayor varianza. Cabe destacar que los rangos son obtenidos de acuerdo a los compuestos no usados para testeo, por lo que ciertos compuestos usados para el testeo podrían estar afuera del hipercubo con longitud de lado de una unidad.

Teniendo en cuenta que la función a optimizar es derivable (Función 3.16), para realizar la minimización de la Función 3.12 de una manera eficiente hicimos uso de un método cuasi-Newton de memoria limitada (del tipo L-BFGS, Nocedal (1980)) .

A los compuestos no usados para el testeo les aplicamos el método SARDUX un total de 100 veces con incializaciones aleatorias de los λ_i y con particiones aleatorias en proporción 2:1 para el entrenamiento y la validación. La partición

de entrenamiento-validación se hizo de manera estratificada, es decir, manteniendo la proporción de 57:43 para cada clase. De manera de independizarse de la separación y la inicialización aplicadas, se obtuvo un modelo promediando los λ_i de las 10 mejores corridas. El desempeño de cada corrida se calcula de acuerdo al grado de solapamiento que existe al proyectar todos los puntos de la validación. Esta medida se obtiene aplicando para cada punto proyectado el cociente de la cantidad de compuestos pertenecientes a la otra clase que tienen un valor de proyección superior sobre aquellos que tienen un valor de proyección menor (para tener una uniformidad y para evitar divisiones por cero, se pone en el numerador el valor más chico de estos dos valores).

7.2.1. Resultados

Teniendo en cuenta que la matriz adaptiva Λ es de rango 1, la manera de cuantificar el grado de aporte individual de cada descriptor es analizando el valor absoluto de los λ_i: cuanto más grande es el valor, mayor es la influencia en la separación. La Figura 7.1 muestra el análisis de los valores absolutos de los λ_i. En el gráfico superior se muestran las 10 mejores corridas, donde a mayor numeración, mayor grado de solapamiento de esa corrida. Los colores azul, amarillo y rojo denotan un valor absoluto bajo, medio y alto respectivamente. A modo de conclusión puede apreciarse que existe cierta reproducibilidad considerando la homogeneidad de colores que, en general, existe en cada columna, considerando que se usaron distintas particiones y distintas inicializaciones. El gráfico inferior muestra los gráficos de caja, de acuerdo al ranking que obtiene cada descriptor en cada corrida: cuanto mayor es el ranking obtenido, mayor es el valor absoluto de ese descriptor en esa corrida. Vemos también que los desvíos en cada columna obedecen a la dificultad de cuantificar la relevancia de un descriptor de manera individual y la alta redundancia existente entre los descriptores.

Para analizar el grado de correlación de las variables realizamos un escalado multidimensional entre los 73 atributos del conjunto de entrenamiento utilizando una distancia basada en la correlación, esto es $\langle$*1-Correlación de Pearson*$\rangle$. Conservando las dos primeras dimensiones, en la Figura 7.2 se muestran las proyecciones de cada descriptor en dicho plano. Podemos apreciar que existe un alto

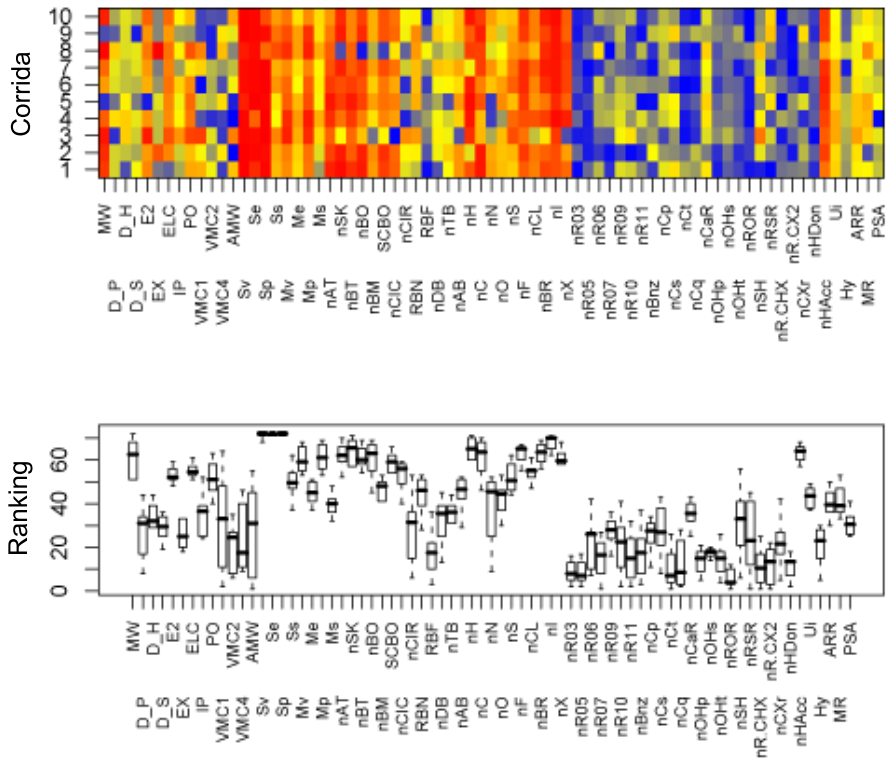


Figura 7.1: Arriba: Influencia de cada descriptor a partir de los valores de $|\lambda_i|$. Abajo: Posición en el ranking de cada descriptor.

grado de correlación entre la mayor parte de los descriptores considerados. Si bien el presente gráfico no indica relevancia de los descriptores, sí indica descriptores que son muy distintos al resto, como es el caso de nH, RBF, IP, EX y Ms.

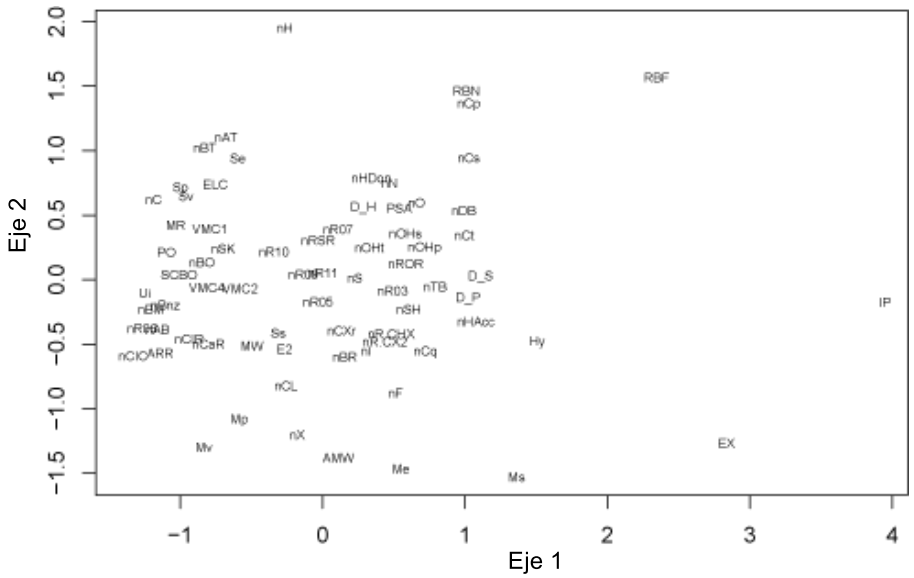


Figura 7.2: Escalado multidimensional usando la distancia de correlación sobre el espacio generado por los descriptores.

La Figura 7.3 muestra las proyecciones de los compuestos usados para entrenamiento (izquierda) y validación (derecha) de la corrida de SARDUX con menor grado de solapamiento en las proyecciones. El porcentaje de proyecciones solapadas, detallado entre paréntesis en la parte superior del gráfico, es relativamente bajo para el entrenamiento y el mismo no empeora demasiado para la validación. Un punto interesante de análisis surge de la linea de regresión, ya que las proyecciones encontradas contendrían información suficiente para armar un modelo de regresión para la propiedad experimental. Esta conclusión tiene dos puntos salientes:

1. Lo que se optimizó según la Función 3.12 es la separación entre dos clases y no un modelo de regresión. Esto nos lleva a creer que los descriptores relacionados con la separación discreta entre valores experimentales altos y bajos, también son de relevancia para confeccionar un modelo de regresión.

2. La cantidad de variables usadas para armar el modelo de regresión es sólo una. Todos los descriptores originales han sido codificados en una única variable mediante una transformación matricial. Desde el punto de vista de la selección de descriptores esto nos lleva a pensar que el método es apto para ser usado como método de selección de variables embebido (sección 5.1).

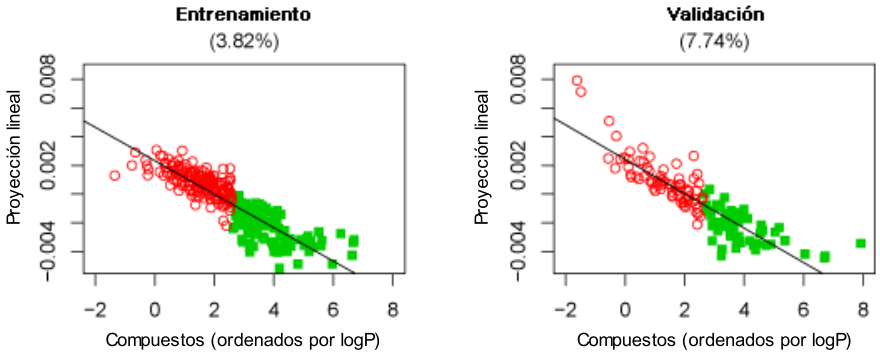


Figura 7.3: Proyección lineal de los compuestos. Los compuestos con valor experimental menor a 2.61 (círculos rojos) pertenecen a la clase 0, mientras que valores más grandes (cuadrados verdes) pertenecen a la clase 1.

Finalmente, se tomaron diferentes subconjuntos de descriptores y usando estos subconjuntos se entrenaron redes neuronales de retro-propagación elástica (*resilient back-propagation* Anastasiadis *et al.* (2005)). Las redes fueron evaluadas con los 30 compuestos separados para el testeo y sus resultados se resumen en la Tabla 7.1. En primer término, se tomó como referencia el subconjunto de 12 descriptores usados en el trabajo de Yaffe *et al.* (2002) y cuyos resultados se ilustran en el bloque A de la tabla anteriormente mencionada. Para el subconjunto del bloque B se eligieron los descriptores mejores rankeados de la mejor corrida de SARDUX, usando la misma cantidad de descriptores que en el caso anterior. Finalmente en el bloque C, se detallan los resultados promedios obtenidos al seleccionar al azar 12 de los 73 descriptores en 25 oportunidades diferentes.

Si bien los subconjuntos de descriptores reportados en los bloques A y B son disjuntos, la diferencia es bastante sutil en favor de la selección basada en

A						Descriptores de referencia					
MW	D_P	D_H	D_S	E2	EX	ELC	IP	PO	VMC1	VMC2	VMC4
ent-val:		$r^2 = 0{,}979$		MAE=0.153		testeo:		$r^2 = 0{,}814$		MAE=0.503	
B						Mejor corrida de SARDUX					
Sv	Se	Sp	nI	nF	nHAcc	nAT	nBR	Mp	nBT	nC	nH
ent-val:		$r^2 = 0{,}969$		MAE=0.177		testeo:		$r^2 = 0{,}830$		MAE=0.499	
D	Desempeño promedio de 25 selecciónes al azar de 12 descriptores										
ent-val:	$r^2 = 0{,}961 \pm 0{,}022$					testeo:		$r^2 = 0{,}681 \pm 0{,}143$			
	MAE=$0{,}200 \pm 0{,}0484$							MAE=$0{,}652 \pm 0{,}140$			

Tabla 7.1: Resultados de las predicciones de las redes neuronales usando distintas estrategias de selección de descriptores.

SARDUX. La selección aleatoria arroja un peor resultado en comparación a las estrategias de selección anteriores, aunque es importante mencionar que en dos oportunidades la selección aleatoria mejoró los resultados del bloque A y B. Esta probabilidad de selección por chance, la cual podríamos cuantificar como de 2/25, creemos que se debe al alto grado de redundancia existente entre el pool de descriptores. Esta conclusión se desprende también del gráfico de las proyecciones utilizando escalado multi-dimensional (Figura 7.2).

La metodología original está planteada como una técnica de clasificación sobre la cual observamos que poseería potencialidades para realizar inferencias en la predicción de valores continuos. Entre las principales limitaciones destacamos que un solo vector λ es optimizado y por lo tanto sólo se pueden realizar proyecciones en una única dirección. Asimismo, aún si se quiere usar para clasificación, advertimos que esta primera metodología carece de una especificación sobre cómo determinar los puntos usados como discriminantes para la predicción. Además, tampoco se considera el problema de la clasificación de datos en más de dos clases. Finalmente, la red neuronal de retro-propagación elástica no ofrece la misma capacidad de predicción que las redes neuronales con regularización bayesiana, por lo que futuras aproximaciones en el tema serán abordadas usando esta última técnica.

Los resultados de esta primera propuesta fueron publicados en Strickert *et al.*

(2009).

7.3. 2º Propuesta: SARDUX para múltiples clases (SARDUX-MC)

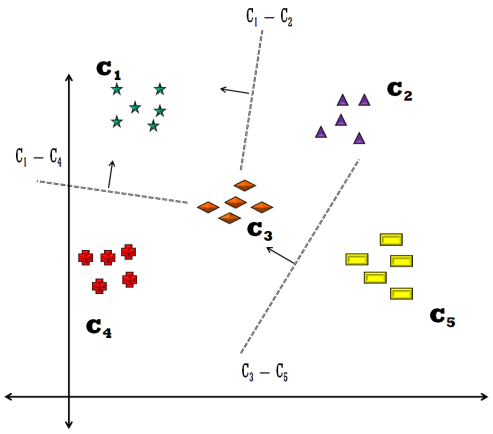
Tal como vimos en la sección anterior, SARDUX permite transformar el espacio original de los datos en un espacio alternativo en donde los datos pertenecientes a clases diferentes puedan ser más fácilmente separados y los pertenecientes a una misma clase queden lo más junto posible. Sin embargo, la matriz adaptiva planteada anteriormente es de rango 1, por lo que en conjuntos de datos con más de dos clases la transformación utilizando un único vector λ, no resultaría suficiente para que las proyecciones sobre una única dirección logren el propósito de compresión intra-clase y separación inter-clase de todas las clases entre sí.

La solución inmediata aparente sería usar 2 vectores adaptivos (o una matriz de rango 2) si se tienen 3 posibles clases, 3 vectores si se tienen 4 posibles clases y así sucesivamente; siempre usando un vector menos que la cantidad de clases a separar. Sin embargo, esto no resulta eficaz, ya que en ausencia de cualquier restricción todos los vectores estarían optimizando lo mismo, pudiendo llegar incluso al mismo resultado, y por lo tanto no se lograría el objetivo buscado. Una posible restricción sería plantear que los vectores adaptivos sean ortogonales entre sí (o lo menos correlacionados posibles), logrando de esta manera que los vectores tengan la menor redundancia posible. El problema aquí reside en que las direcciones ortogonales no necesariamente son las mejores para la separación de todas las clases.

Por lo tanto, nuestra propuesta para lograr la separación entre más de dos clases es considerar otro enfoque en donde tengamos en cuenta pares de clases y no considerar todas las clases a la vez. Por tal motivo, se plantearon dos estrategias para extender SARDUX: AGA y CC. Describiremos a continuación cada una de estas dos estrategias.

7.3.1. Todos contra todos - AGA

La primera de las estrategias para SARDUX-MC, llamada "*todos contra todos*" (AGA - "*all against all*") consiste en aplicar SARDUX para las observaciones de todos los posibles pares de clases. Esto implica que SARDUX debe ejecutarse $c \cdot (c-1)/2$ veces, obteniéndose así una dirección distinta por cada ejecución, las cuales se almacenan en una matriz de direcciones. Una vez obtenidas todas las direcciones, cuando se quiere determinar la clase de un compuesto no usado en el entrenamiento, el mismo debe ser proyectado de acuerdo a todas las proyecciones establecidas en la matriz de direcciones. Establecidos los discriminantes para cada dirección y de acuerdo a su comparación con el valor de la proyección, se arma una matriz de votos. Esta matriz posee la cantidad de veces que una observación es considerada como perteneciente a la clase correspondiente a su columna. Para cada compuesto (cada fila en la matriz de votos), la clase (o la columna) que más votos tenga es la clase final a la cual se asigna dicho compuesto. La Figura 7.4 muestra un esquema de la estrategia AGA.



Matriz de Direcciones

	C_1	C_2	C_3	C_4	C_5
C_1		C_1-C_2	C_1-C_3	C_1-C_4	C_1-C_5
C_2			C_2-C_3	C_2-C_4	C_2-C_5
C_3				C_3-C_4	C_3-C_5
C_4					C_4-C_5
C_5					

Matriz de Votos

	C_1	C_2	C_3	C_4	C_5
Testeo 1	2	3	4	1	0
Testeo 2	0	1	3	4	2
Testeo 3	4	3	2	1	0
...	...	...	...	...	...
Testeo *n*	2	4	3	1	0

Figura 7.4: Extensión de SARDUX para múltiples clases: esquema de la estrategia AGA para 5 clases.

La presente estrategia posee ciertas limitaciones. El primer problema surge de realizar proyecciones que carecen de significado. Por ejemplo, supongamos que un compuesto no usado para el entrenamiento pertenecería a la clase 4. Dado

que este compuesto va a ser proyectado en las direcciones discriminatorias de todos los posibles pares de clases, proyecciones entre clases distintas a las de su pertenencia, como por ejemplo la proyección entre C_1 y C_2, carece de significado y, más aún, podría conducir a resultados erróneos. Además, si bien la clase real del compuesto a testear debería siempre "ganar" en cantidad de votos, ya que debería ganar los votos cuando se compara su clase con el resto de las clases, las separaciones perfectas pueden no obtenerse en problemas reales, y esta situación podría dar lugar a desaciertos o empates en la matriz de votos. Los empates deberían resolverse usando la proyección sobre la dirección que separa las clases empatadas.

Nótese también que una estrategia similar a la presente consistiría en optimizar la separación entre una clase contra todas las restantes clases juntas. Esta estrategia, si bien evitaría el problema de las comparaciones espurias, también admite situaciones de empate en la matriz de votos.

7.3.2. Comparaciones en cascada - CC

La segunda estrategia, denominada "*comparaciones en cascada*" (CC - "*cascade comparisons*"), resuelve ciertas limitaciones de la estrategia anterior. Esta nueva estrategia fue ideada teniendo en mente que la clasificación proviene de la discretización de una variable continua según su percentil (tal como se hizo en la experimentación de la sección 7.2). La estrategia CC modifica la configuración original de las clases y las restructura en diferentes niveles. En el primer nivel, separa el conjunto de entrenamiento en dos clases, en el segundo nivel separa a los compuestos en tres clases, y así siguiendo hasta que en el último nivel se tiene el número de clases deseado. En cada nivel se optimiza usando SARDUX con los compuestos que pertenecen a clases contiguas. Nuevamente la cantidad de optimizaciones necesarias es $c \cdot (c-1)/2$. Sin embargo, cuando se testea un nuevo compuesto, el mismo no se lo proyecta en todas las direcciones generadas sino que se lo evalúa primero en la dirección del primer nivel y luego las direcciones sucesivas por las que se proyecta el compuesto depende del resultado obtenido en el nivel anterior.

En la Figura 7.5 se muestra un ejemplo de un compuesto evaluado de acuerdo a esta estrategia, en donde el conjunto de entrenamiento es divido en 5 clases, cada una conteniendo el 20 % de los datos. De acuerdo a la figura, en el primer nivel, el compuesto se proyecta en la única dirección que divide las dos mitades y es asignado a la clase que corresponde a los datos con valor experimental inferior al percentil 50. En el segundo nivel, el compuesto se proyecta únicamente en una de las direcciones de ese nivel (la dirección que divide a los compuestos del percentil 0-33 con los del percentil 33-66) y se determina que el mismo se encuentra entre el percentil 33 y el 66. El compuesto sigue siendo proyectado en cada nivel según las direcciones correspondientes hasta que se lo clasifica como miembro de la clase 3 (entre el percentil 40 y 60). Por lo tanto, cada compuesto es proyectado $c - 1$ veces, es decir, una vez por nivel.

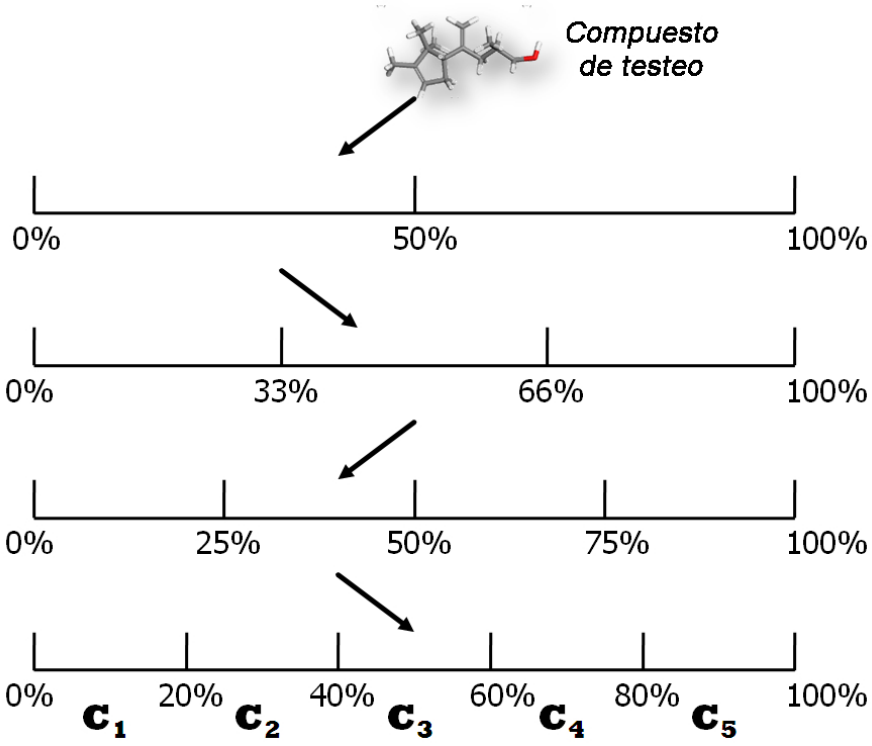


Figura 7.5: Extensión de SARDUX para múltiples clases: esquema de la estrategia CC para 5 clases.

Como rápidamente se puede observar, esta estrategia no sólo no realiza proyecciones carentes de sentido, sino que además evita las posibilidades de ambigüedad por empates. Vale destacar que este procedimiento se podría haber hecho de una manera no solapada, es decir, donde la cantidad de proyecciones se duplican por nivel. Esto sería, tener dos clases en el primer nivel, cuatro en el segundo, y así siguiendo. Sin embargo, mantener esta estrategia en la forma presentada posee una ventaja interesante. La ventaja es que para un compuesto a testear se puede tener una asignación equivocada de clase en un nivel superior, y aún así ser capaz de finalizar en la clase correcta en el último nivel.

La única desventaja de CC es que es únicamente aplicable para situaciones en donde las clases fueron generadas desde una variable continua. En una situación donde la variable es originalmente discreta, puede aplicarse un esquema de comparación en cascada de manera no solapada, únicamente si la variable discreta en cuestión admite una relación de orden entre los posibles distintos valores.

7.3.3. Determinación del punto de umbral para el discriminante

En forma independiente a como se definen las proyecciones, una vez obtenida una dirección como resultado de la optimización estipulada, se debe definir un punto sobre dicha dirección para determinar cuando un compuesto pertenece a una clase o a la otra. La estrategia más simple consistiría en buscar el punto medio entre los centroides de las proyecciones de una clase y otra.

Sin embargo, esta estrategia no necesariamente resulta la mejor en cuanto a la posibilidad de que un punto sea incorrectamente clasificado. Por tal motivo, proponemos una estrategia que apunte a la minimización de la *confusión*[1]en la predicción de clases. Para llevar a cabo esta idea, una vez proyectados todos los puntos sobre la dirección optimizada, se prueba en forma iterativa el punto medio entre todos los pares de puntos contiguos. Estos puntos medios son tratados como umbral tentativo, en donde el umbral definitivo se definirá como aquel que minimiza la confusión de clases al aplicarse sobre un conjunto de validación o, en su defecto, sobre el mismo conjunto de testeo.

[1]El término *confusión* es usado aquí como la proporción o cantidad de clases incorrectamente

7.3.4. Resultados

Para evaluar SARDUX-MC utilizamos nuevamente el mismo set de datos de la sección anterior. Se decidió realizar distintos experimentos, en donde en cada uno de ellos la variable experimental se dividió en 2, 3, 4, 5 y 6 clases, teniendo cada grupo la misma probabilidad *a priori* (es decir separando por iguales percentiles). 88 de los 440 compuestos, lo que corresponde al 20 % de los datos, fueron utilizados para testeo. Para cada experimento se realizaron 10 corridas de SARDUX-MC, en donde la inicialización de los pesos iniciales y la separación de los datos se reiniciaba para cada iteración. A fin de proveer un marco de comparación con otro método de clasificación, se realizaron los mismos experimentos utilizando análisis de discriminante lineal (LDA, *linear discriminant analysis*), el cual resulta un método de confiada referencia por su popularidad y su aceptable capacidad.

En primer lugar nos interesa mostrar algunos detalles sobre el entrenamiento realizado con las dos alternativas. En la Figura 7.6 se muestran las proyecciones de todos los compuestos del entrenamiento usando el esquema AGA, al optimizar la primera clase versus las clases restantes para una dada corrida. Podemos apreciar que la separación obtenida es aceptable, haciéndose cada vez mejor a medida que se compara con una clase más lejana. Esta característica resalta como la relación de orden de las clases queda reflejada en el espacio proyectado. Dicho de otra manera, hay una correlación entre las distancias del espacio de la variable de salida y el de los descriptores.

Otro análisis interesante proviene de observar los λ_i de los distintos vectores de proyección que se generan utilizando SARDUX-MC. En la Figura 7.7 se muestran los gráficos de caja para los 10 descriptores con valor de λ_i más alto para cada una de las proyecciones de la Figura 7.6. En forma similar, en la Figura 7.8 se muestran los 10 descriptores más influyentes considerando todas las proyecciones anteriores. El gráfico permite apreciar la incidencia de un descriptor en una proyección y en otra. Este análisis resulta de interés, ya que se puede identificar la influencia de los distintos descriptores dependiendo de la clase a la cual pertenece. Por ejemplo, el descriptor 'Sp' resulta muy influyente en la primera proyección (separación de las

clasificadas por un clasificador.

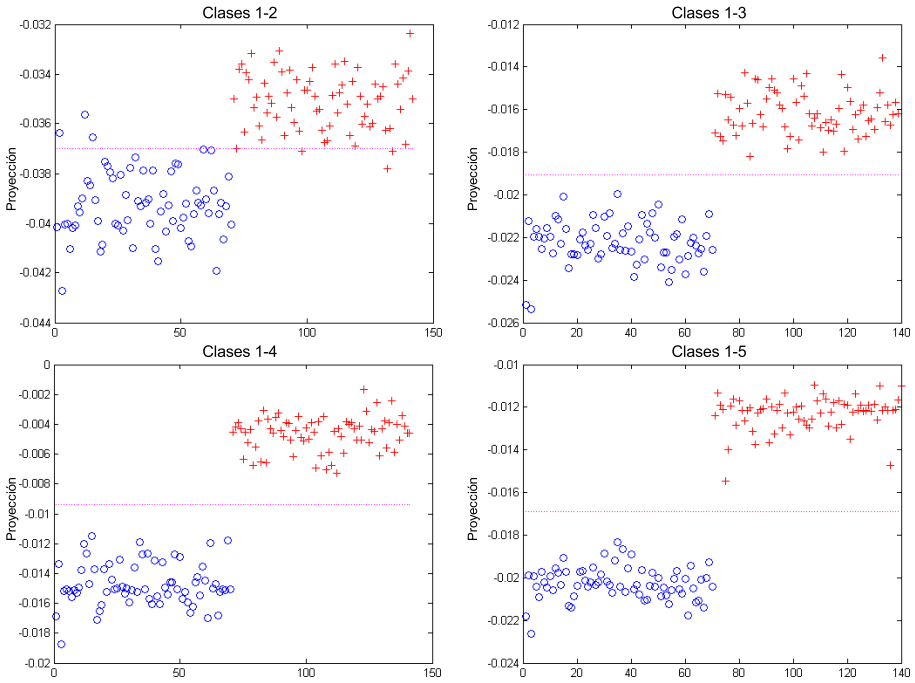


Figura 7.6: Proyecciones según el enfoque AGA de la clase 1 (en azul) versus las otras clases (en rojo).

clases 1 y 2) pero no tanto en la cuarta (separación de las clases 4 y 5), mientras que sucede exactamente lo contrario con el descriptor 'Se'.

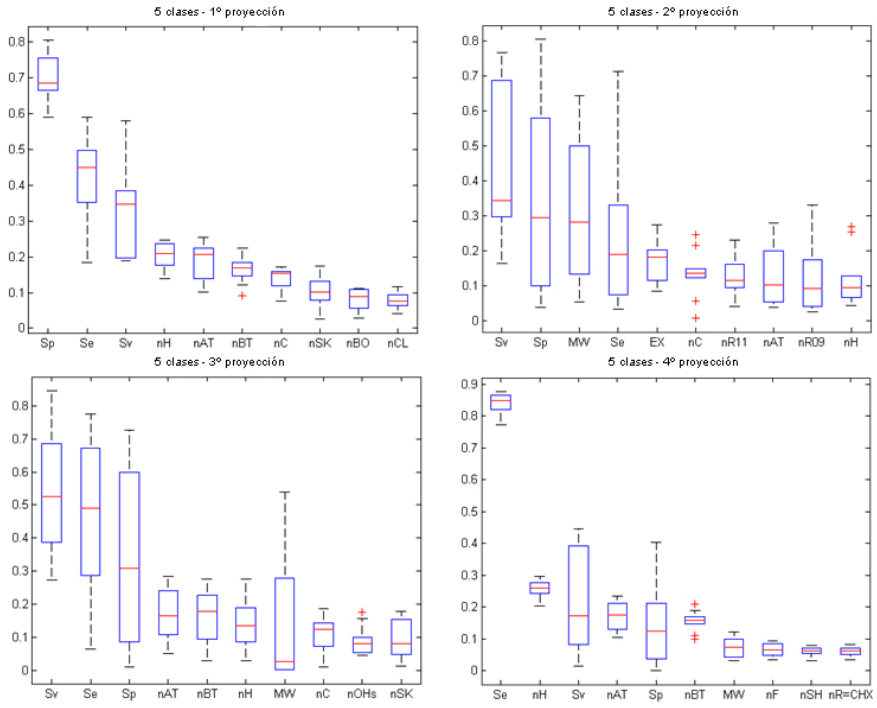


Figura 7.7: λ_i de las distintas direcciones obtenidas en el último nivel usando CC.

La Tabla 7.2 muestra los resultados de cada experimento para las dos estrategias de SARDUX-MC y para LDA. De cada metodología se muestran dos resultados que provienen de analizar la matriz de confusión: el porcentaje de compuestos mal clasificados ("*% Mal clasif.*") y el error ponderado ("*Error Pond*"). El primero claramente surge de dividir la cantidad de elementos mal clasificados por el método por la cantidad de compuestos testeados. El error ponderado posee el mismo denominador, pero en el numerador, al contar la cantidad de elementos mal clasificados se multiplica por la distancia a la clase a la cual dicho elemento debiera haber sido clasificado. Por ejemplo, si un compuesto que pertenece a la clase 1, se clasificara como clase 4, el error contaría por tres $(4-1)$. Esta métrica, en definitiva, intenta analizar, en el caso de errores, cuán lejos se predijo su clase. Esto es únicamente de interés cuando existe una relación de orden entre las

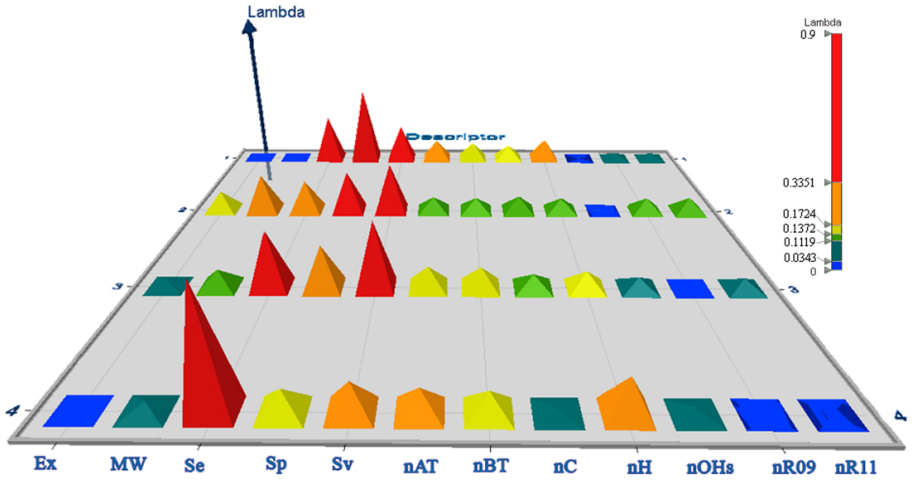


Figura 7.8: Análisis de los 10 λ_i más destacados en todas las proyecciones del último nivel usando CC.

Tabla 7.2: Resultados de clasificación para SARDUX (AGA y CC) y LDA.

Clases	SARDUX				LDA	
	AGA		CC			
	% Mal clasif.	Error pond.	% Mal clasif.	Error pond.	% Mal clasif.	Error pond.
2	0,2193	0,2193	0,2170	0,2170	0,2159	0,2159
3	0,2398	0,2511	0,2420	0,2511	0,3341	0,3386
4	0,3580	0,3977	0,3500	0,3738	0,4602	0,5091
5	0,3398	0,4352	0,3352	0,3943	0,4898	0,6227
6	0,4590	0,6215	0,4273	0,5511	0,5455	0,7784

clases. De la comparación entre las dos alternativas de SARDUX-MC vemos que no existen diferencias importantes entre ellas, aunque el error se vuelve un poco más alto en el caso de AGA cuando se compara usando el error ponderado. Por otra parte, al comparar con LDA, notamos que los resultados son muy similares cuando se trabajan con 2 clases, sin embargo LDA desmejora drásticamente su desempeño a medida que se incorporan más clases al problema. Esto no sucede para SARDUX-MC.

En conclusión, destacamos que la extensión de SARDUX para múltiples clases, resulta un método al menos competitivo con LDA y, a diferencia de éste, pensado para problemas de clasificación donde las clases tienen un valor ordinal asociado. Más aún, LDA no es del todo robusto, en el sentido que requiere ciertas asunciones sobre las matrices de covarianza de cada clase que, si no se cumplen, no brindan

resultados confiables.

7.4. 3º Propuesta: generalización de SARDUX

Finalmente, un último enfoque aplicado en el marco de esta tesis, consistió en desarrollar un esquema más general para aplicar mapeos a espacios de menor dimensionalidad usando distancias basadas en matrices adaptivas. La motivación primaria de esta generalización parte de la necesidad de contar con una metodología que sea independiente de la naturaleza de la variable a predecir. Esto significa que la variable destino puede ser un tipo de dato continuo o discreto, ya sea éste ordinal o categórico. El objetivo central de esta propuesta generalizada, a la cual llamaremos regresión de subespacios multivariados (MSR - *multivariate subspace regression*) es mejorar la precisión de los métodos de predicción, permitiendo ahora modelos de regresión.

Al igual que en las propuestas anteriores de este capítulo, el método MSR busca optimizar la representación de los compuestos dentro del espacio determinado por los descriptores, de manera que se mantenga un vínculo entre esta representación y la variable a modelar de los compuestos. Más específicamente, lo que se quiere hacer es que en el espacio de baja dimensionalidad (o espacio proyectado), las distancias entre compuestos se asemeje lo más posible a las distancias de los compuestos en la variable de modelado. Para medir las distancias de los compuestos, las cuales se miden en el espacio determinado por los descriptores, se utiliza una medida de distancia basada en matrices adaptivas similar a la usada en 3.15. Para las distancias medidas en el espacio de la variable a modelar, se utilizó la distancia Euclídea.

7.4.1. Descripción de la metodología

Sea $\mathbf{x_i}$ el i-ésimo compuesto del conjunto de datos X, donde cada dato $\mathbf{x} \in \mathbb{R}^n$ y $|X| = m$. Sea también $l_i \in L \subset \mathbb{R}^q$ la variable a modelar de $\mathbf{x_i}$. Nótese que el valor a modelar para cada compuesto no es un vector unidimensional, sino que también puede ser un vector multivariado de q componentes. En forma

general, usaremos $q = 1$ para problemas de regresión y $q = k$ para problemas de clasificación categórica, donde k es la cantidad de clases posibles.

La transformación del espacio de descriptores se da a partir de la optimización de la correlación entre las distancias, tal como se describe en la siguiente ecuación:

$$\underset{\lambda}{\operatorname{argmax}}\, S_d = r(D_L, D_X^{\boldsymbol{\lambda}}) \tag{7.1}$$

En la Ecuación 7.1, D_L son las distancias entre todos los pares de compuestos en su variable a modelar, en donde para este caso se usó la distancia Euclídea. $D_X^{\boldsymbol{\lambda}}$ es la distancia entre todos los pares de compuestos medidas en el espacio de los descriptores moleculares, la cual se calcula de acuerdo a la Ecuación 7.2, y la misma depende de la matriz adaptiva $\boldsymbol{\lambda}$. Finalmente, r hace referencia al cálculo de la correlación de Pearson.

La distancia usada para esta generalización de SARDUX es similar a la propuesta en la Ecuación 3.15, aunque posee ciertas diferencias tal como se muestra a continuación:

$$(D_X^{\boldsymbol{\lambda}})_{i,j} = d(\mathbf{x_i}, \mathbf{x_j}|\boldsymbol{\lambda}) = \sqrt{(\mathbf{x_i} - \mathbf{x_j})^\intercal \cdot \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^\intercal \cdot (\mathbf{x_i} - \mathbf{x_j})} \tag{7.2}$$

En primer lugar, en este caso se aplica la raíz cuadrada a toda la distancia de manera de conciliar con la distancia Euclídea aplicada en la variable a predecir. Por otra parte, el parámetro $\boldsymbol{\lambda}$ no es un vector como en el caso anterior sino que $\boldsymbol{\lambda} \in \mathbb{R}^{n \times p}$. Esto admite una mayor flexibilidad por parte del método, dado que el subespacio proyectado $X \cdot \boldsymbol{\lambda}$ posee una dimensionalidad igual a p, el cual se establece de acuerdo al grado de optimización requerido de la Ecuación 7.1 o de la conveniencia en la dimensionalidad de la representación buscada. De este modo, si por ejemplo se requiere visualizar el subespacio proyectado, se usará un valor de $p \leq 3$.

La derivada parcial de la función a optimizar respecto del parámetro $\boldsymbol{\lambda}$ (Ecuación 7.3) puede ser resuelta analíticamente haciendo uso de los resultados del trabajo de Strickert *et al.* (2008a) y resolviendo la derivada parcial de de la Ecuación 7.2. Al poder obtener esta derivada analítica, nuevamente hicimos uso del método de optimización L-BFGS (Nocedal (1980)).

$$\frac{\partial S_d}{\partial \boldsymbol{\lambda}} = \frac{\partial r(D_L, D_X^{\boldsymbol{\lambda}})}{\partial D_X^{\boldsymbol{\lambda}}} \cdot \frac{\partial D_X^{\boldsymbol{\lambda}}}{\partial \boldsymbol{\lambda}}. \tag{7.3}$$

$$\frac{\partial d(\mathbf{x_i}, \mathbf{x_j}|\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \frac{(\mathbf{x_i} - \mathbf{x_j}) \cdot ((\mathbf{x_i} - \mathbf{x_j})^\intercal \cdot \boldsymbol{\lambda})}{d(\mathbf{x_i}, \mathbf{x_j}|\boldsymbol{\lambda})}. \tag{7.4}$$

7.4.2. Evaluación de la calidad del mapeo

Uno de los problemas más importantes de los métodos de visualización, agrupamiento y reducción de dimensionalidad, surge porque suelen ser métodos de naturaleza no supervisada y con esto se genera la controversia sobre su capacidad de evaluación más allá de los criterios subjetivos. En cambio, los métodos presentados en este capítulo poseen una naturaleza supervisada, dada por las funciones de optimización planteadas. Esto admite una comparación entre los diferentes resultados finales alcanzados.

Sin embargo, para evaluar nuestra metodología, no utilizaremos el resultado de la optimización alcanzada, lo que correspondería en este caso a analizar el índice de correlación, sino que usaremos la matriz del subespacio proyectado $X \cdot \boldsymbol{\lambda}$ con algún método de regresión/clasificación. Esto significa que a esta última matriz la usaremos como nuestra nueva matriz de compuestos, de p dimensiones, la cual corresponde a una transformación lineal de la matriz con los n descriptores originales. El error en la predicción alcanzado será utilizado como factor de evaluación de las distintas proyecciones.

A fin de no producir resultados sobreestimados, separamos los conjuntos de compuestos utilizados en subconjuntos de entrenamiento, validación y testeo. El conjunto de validación es utilizado para provocar una terminación temprana del proceso de optimización, dependiendo del monitoreo aplicado sobre alguna medida de desempeño. Experimentamos con dos medidas de desempeño para la validación, donde una era la correlación de las distancias de la proyección del conjunto de validación con las distancias en la variable a modelar, mientras que la otra era analizar el error de predicción utilizando RL, en donde se usa el subespacio proyectado por el conjunto de validación como las variables independientes.

El conjunto de testeo es usado para validar la metodología y en donde como método de predicción se utilizó RL y CRN. La metodología se repite una cantidad r de réplicas, en donde para cada réplica se reinicializa la separación de los compuestos y los valores iniciales utilizados para la matriz $\boldsymbol{\lambda}$.

Como se hizo en las metodologías de este capítulo antes presentadas, los descriptores deben ser normalizados a fin de evitar que descriptores con un rango más amplio de valores afecten por demás el cálculo de las distancias entre los descriptores. En este sentido se aplicaron dos estrategias de normalización: una basada en la transformación en una variable con distribución normal con media 0 y desvío 1 (transformación z) y otra basada en una transformación por rankings *ad hoc*.

Finalmente, y tal como se hizo en las secciones anteriores, se puede obtener un perfil de cuáles fueron los descriptores que más contribuyeron en la separación de los compuestos en el subespacio proyectado, y por ende los que resultan más relevantes para la variable a modelar. Este perfil surge de analizar las componentes de la matriz $\boldsymbol{\lambda}$. Es importante destacar que, en esta variante, la matriz $\boldsymbol{\lambda}$ puede tener más de una columna, por lo que el impacto efectivo surge de sumar los pesos de cada descriptor en sus p dimensiones.

A modo de resumen de la metodología, detallamos en un pseudo-código los pasos que seguimos en esta nueva metodología. Los pasos 2-7 pueden repetirse una cierta cantidad de veces, a fin de reducir y determinar la variabilidad de los resultados obtenidos.

1. Se elige el número de dimensiones del espacio a proyectar.
2. Se definen valores aleatorios para las componentes de la matriz $\boldsymbol{\lambda}$.
3. Se divide el set de datos en subconjuntos de entrenamiento, validación y testeo.
4. Aplicamos L-BFGS para optimizar la Ecuación 7.3 para los datos del conjunto de entrenamiento. El algoritmo se detiene de acuerdo a un monitoreo sobre los resultados obtenidos en el conjunto de validación.
5. Los datos de testeo son proyectados a un subespacio definido por $X \cdot \boldsymbol{\lambda}$.

6. Se aplica un método de predicción sobre los datos de testeo proyectados, a fin de evaluar la precisión en la predicción.

7. Como información adicional puede realizarse un análisis de la relevancia de los descriptores a partir de la inspección de la matriz $\boldsymbol{\lambda}$.

7.4.3. Resultados

Los resultados mostrados en esta sección corresponden a un subconjunto de la totalidad de los experimentos realizados, en donde por cuestiones de espacio se restringió a los resultados más interesantes y representativos. Para cada variante del método se aplicaron 10 réplicas de la metodología. En primer lugar destacamos que no se encontraron diferencias significativas en la aplicación de las normalizaciones basadas en desviación estándar y rankings, como así tampoco se hallaron diferencias entre los dos métodos mencionados de detención temprana de la optimización.

Evaluamos nuestro método MSR utilizando el set de datos basado en logP, también usado en las secciones anteriores, con el agregado de los sets de datos descriptos en B.3 y B.4. Se aplicó una separación a los conjuntos de datos, de manera que 64 % se usen para el entrenamiento, 16 % para la validación y 20 % para el testeo. Para cada uno de los sets de datos se evaluó su comportamiento utilizando 1, 2 y 3 dimensiones para el subespacio proyectado, es decir seteando $u = 1$, $u = 2$ y $u = 3$ en cada experimento.

Los resultados de las predicciones se resumen en las Tablas 7.3, 7.4 y 7.5. En cada una de estas tablas, la primera columna indica el método de regresión que se utilizó para evaluar la calidad de la proyección encontrada y en la columna siguiente se muestra la métrica con la que se evaluó la calidad de la regresión. En la tercera columna se muestran los promedios de los resultados obtenidos cuando no se utilizó ninguna proyección, es decir, utilizando todo el conjunto de descriptores original. Las últimas tres columnas detallan los resultados obtenidos cuando se aplica MSR, diferenciando en la cantidad de dimensiones utilizadas para el subespacio proyectado.

La Tabla 7.3 muestra los resultados de las experimentaciones para el conjunto de datos que tiene a logP como variable a modelar. Este conjunto de datos parte

Tabla 7.3: Resultados en la predicción para el set de datos descripto en B.2.

		Sin proyección	MSR		
			1D	2D	3D
RL	MAE	0,384974	0,356693	0,353844	0,360319
	MSE	0,332484	0,231318	0,22789	0,22903
	r	0,916236	0,942381	0,943225	0,942676
	q^2	0,840655	0,888405	0,890008	0,888911
CRN	MAE	0,320318	0,353011	0,348459	0,355348
	MSE	0,225588	0,22438	0,223071	0,228558
	r	0,944357	0,943746	0,944508	0,942478
	q^2	0,892192	0,890977	0,892437	0,888551

de una considerable menor cantidad de descriptores que los otros dos conjuntos de datos utilizados, y recordando lo mencionado en capítulos anteriores, todos los descriptores iniciales son potencialmente relevantes. Podemos ver en este caso que los resultados de la RL mejoran notablemente cuando se aplica una proyección basada en MSR. Además estos resultados son bastantes similares a los obtenidos cuando se aplica CRN sobre el conjunto inicial de descriptores. Este resultado es interesante teniendo en cuenta que una combinación de métodos lineales, como lo son MSR y RL, logran ser competitivos con un método no lineal como las redes neuronales. Destacamos acá también que no advertimos diferencias significativas entre el uso de distintas dimensiones para el subespacio proyectado.

En el caso de los resultados obtenidos para el set de datos relacionado con la penetración en la barrera hemato-encefálica (Tabla 7.4) vemos que se obtienen beneficios más grandes por aplicar MSR. En primer lugar advertimos que la combinación de MSR y RL es incluso mejor que el uso de redes neuronales sin reducción de la dimensionalidad. Vemos también en esta tabla cómo el uso de dimensiones más grandes permite un incremento gradual de la calidad predictiva. Las predicciones obtenidas usando el CRN y la proyección en 3 dimensiones logran los mejores resultados.

Al analizar los resultados obtenidos sobre el conjunto de datos que contiene información experimental del logHIA (Tabla 7.5), observamos que se da una situación análoga al caso del conjunto de datos anterior. Esto es, buen desempeño de MSR en conjunción con RL y mejor aún cuando MSR se lo combina con CRN. Esta similitud de los resultados es natural de imaginar teniendo en cuenta que ambos sets de datos poseen características similares, en cuanto a la cantidad de

Tabla 7.4: Resultados en la predicción para el set de datos descripto en B.3.

		Sin proyección	MSR		
			1D	2D	3D
RL	MAE	0,378431	0,314052	0,308478	0,307171
	MSE	0,292641	0,166621	0,163895	0,162229
	r	0,600260	0,75623	0,764861	0,771124
	q^2	0,365451	0,572346	0,585997	0,596036
CRN	MAE	0,375536	0,311881	0,301672	0,29988
	MSE	0,262865	0,166003	0,161526	0,156409
	r	0,623113	0,755732	0,767215	0,778426
	q^2	0,394464	0,571678	0,58934	0,607367

Tabla 7.5: Resultados en la predicción para el set de datos descripto en B.4.

		Sin proyección	MSR		
			1D	2D	3D
RL	MAE	0,871241	0,320915	0,32311	0,328171
	MSE	1,309979	0,221933	0,247464	0,227273
	r	0,306270	0,709907	0,739158	0,702077
	q^2	0,145344	0,520064	0,581718	0,511655
CRN	MAE	0,361462	0,281496	0,304981	0,27693
	MSE	0,264432	0,195896	0,227438	0,175602
	r	0,660222	0,731243	0,735626	0,754319
	q^2	0,459259	0,560959	0,577137	0,586624

información que poseen.

Finalmente ilustramos cómo a partir de los componentes de la matriz $\boldsymbol{\lambda}$ se puede analizar la incidencia de cada uno de los descriptores. En la Figura 7.9 se muestra un diagrama de caja de las variaciones en cada una de las 10 corridas para los componentes de $\boldsymbol{\lambda}$ cuando se aplica la proyección a un subespacio de dimensión 1 sobre el primer conjunto de datos. Se puede observar como los descriptores que surgen como más relevantes guardan una relación con los descriptores elegidos en el capítulo 5 cuando el mismo set de datos fue utilizado.

7.4.3.1. Resultados adicionales

En este último apartado expondremos ciertas experimentaciones y análisis sobre la metodología que no son centrales dentro del marco de esta tesis, aunque permiten resaltar otros aspectos sobre la potencialidad del método.

Una de las bondades de MSR que no ha sido analizada anteriormente es su potencialidad para la visualización. Esto permite que un ser humano pueda

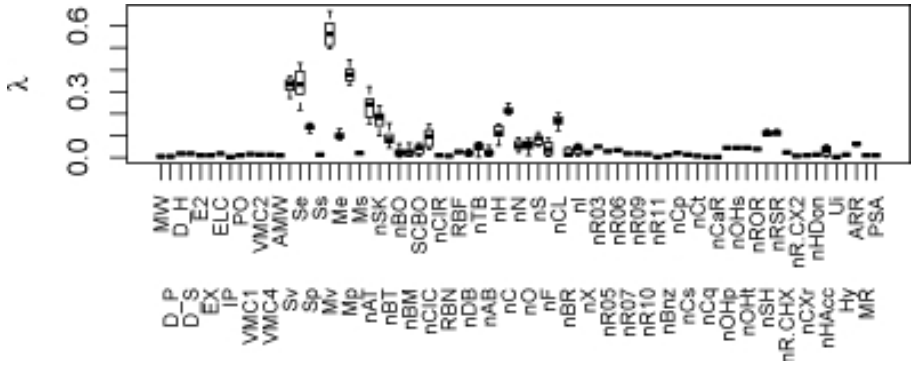


Figura 7.9: Diagrama de caja de la contribución de cada descriptor en las diferentes réplicas para el conjunto de datos descripto en B.2

vislumbrar en un espacio manejable, por ejemplo de 2 ó 3 dimensiones, la configuración espacial de sus compuestos. Tal como hemos reiterado, esta configuración espacial está en consonancia con la variable a modelar, lo que le otorga más significado a la configuración espacial que el que le puede otorgar un método no supervisado, como por ejemplo el análisis de componentes principales. Asimismo, un compuesto no usado para el entrenamiento puede ser proyectado en el subespacio en cuestión, y así anticipar en forma aproximada el posible valor de una propiedad o actividad experimental.

Mostramos entonces las proyecciones del conjunto de datos de logP, las cuales han sido aplicadas a un subespacio de 1, 2 y 3 dimensiones según se muestra en la Figura 7.10. Cada punto representa un compuesto y su color es el valor experimental de la propiedad que se está modelando. De estas figuras puede apreciarse que la disposición de los compuestos adquiere más libertad a medida que se incrementa la cantidad de dimensiones. Asimismo destacamos que los conjuntos de testeo se adaptan a la disposición obtenida en el entrenamiento, lo que nos lleva a afirmar que el modelo es generalizable.

Otra de las bondades de MSR sobre la cual no nos hemos explayado aún, es la capacidad de manejarse para sets de datos para problemas de clasificación. El caso discreto ordinal no presenta ningún rasgo distintivo con respecto al caso continuo, en donde simplemente usaremos un número entero en la variable a modelar para expresar el orden de la clase de cada compuesto. Sin embargo, el

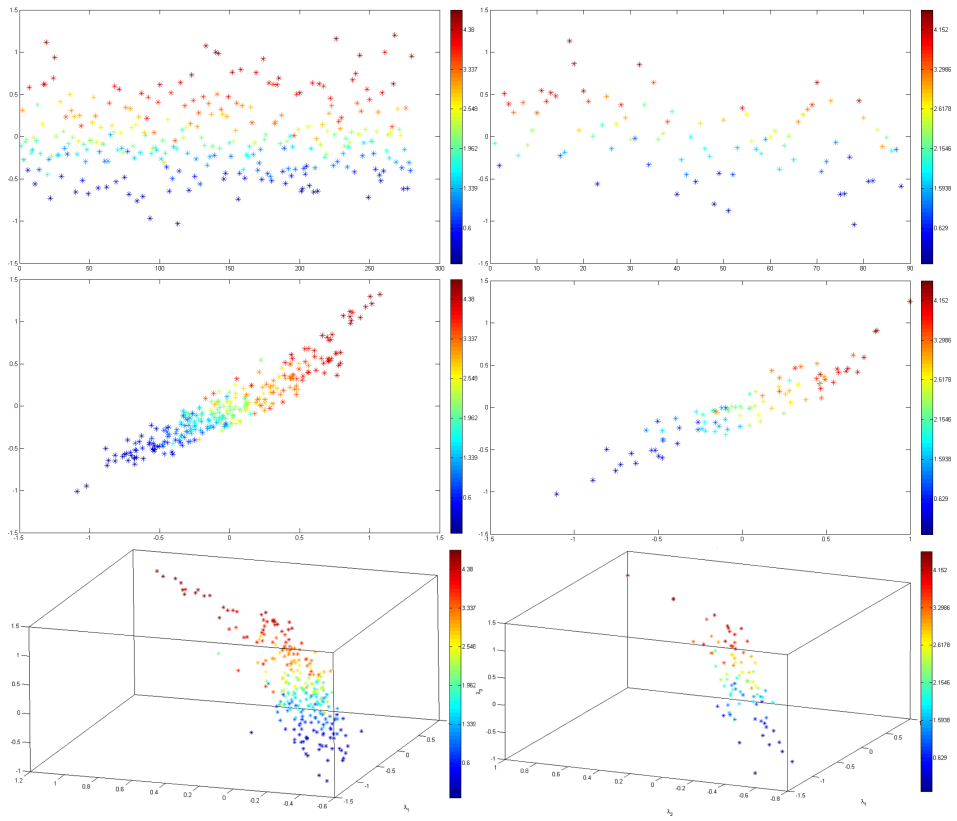


Figura 7.10: Proyección de los compuestos de entrenamiento (izquierda) y de los compuestos de testeo (derecha) a subespacios de 1 dimensión (arriba), 2 dimensiones (medio) y 3 dimensiones (abajo).

caso discreto categórico no admite esta representación, dado que se consideraría el número asignado como un valor ordinal y, por ejemplo, se encontrarían distancias más grandes entre la primera clase y la última, que entre dos clases con valores más cercanos. Por lo tanto la manera de lidiar con esta dificultad es asignando un vector de tantas componentes como clases posibles haya a cada uno de los compuestos. Cada vector tiene norma 1, y cada par de vectores diferentes deben ser ortonormales. De esta manera, si se tienen 3 clases por ejemplo, a una clase se le asigna el vector $\langle 1, 0, 0\rangle$, a otra el $\langle 0, 1, 0\rangle$ y a la última el $\langle 0, 0, 1\rangle$. Nótese que en la sección 7.4.1 se indica que $l_i \in L \subset \mathbb{R}^q$ haciendo referencia a que $q > 1$ en los casos de clasificación categórica.

Para recrear esta situación tomamos el percentil 33,3 % más bajo, medio y más alto del conjunto de compuestos basado en logP, y en donde asignamos la clase $\langle 1, 0, 0\rangle$ para los compuestos con logP $< 1{,}78$, $\langle 0, 1, 0\rangle$ para compuestos en el rango $1{,}78 \leq$ logP $< 3{,}0132$ y $\langle 0, 0, 1\rangle$ para valores de logP $\geq 3{,}0132$. Previa a la separación en clases, separamos un conjunto de testeo en forma aleatoria. Vemos en la Figura 7.11 la proyección en 2 dimensiones de los sets de datos de entrenamiento y testeo para las diferentes clases. Si bien la separación entre las clases no está completamente definida, el grado de solapamiento es menor y la lógica del modelo se captura. Otra observación interesante surge de analizar que la disposición de las clases en el plano proyectado no es a lo largo de una dirección como sucede en los casos de la Figura 7.10, sino que los datos se disponen de una manera más homogénea. Esto se da debido a la equidistancia asignada entre todas las clases.

Los resultados mostrados en esta sección fueron discriminados del resto debido que a que no corresponden con una problemática central que se deseaba desarrollar en esta tesis y por lo tanto no hemos hecho un análisis en profundidad al respecto. Por otra parte, las técnicas de visualización por sí solas, sólo pueden ser evaluadas desde un punto de vista subjetivo, lo que conlleva a otro tipo de análisis y al planteo de establecidas controversias al respecto. Por último, el análisis de MSR aplicado a la clasificación categórica no se hace en profundidad dado que no contamos con datos de este tipo para QSAR/QSPR. El uso de esta estrategia para los casos discretos ordinales queda cubierto con las metodologías descriptas en la sección 7.3.

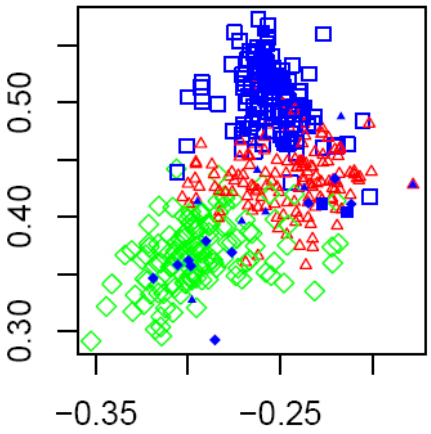


Figura 7.11: Proyección de los compuestos de entrenamiento (símbolos sin pintar) y testeo (símbolos pintados) pertenecientes a tres variables categóricas (discriminadas por color).

Los resultados obtenidos en esta sección fueron presentados en Strickert *et al.* (2010) y Soto *et al.* (2010).

7.5. Conclusiones

Este capítulo se centró en la aplicación de métodos de mapeo a espacios de menor dimensionalidad para la aplicación de modelos QSAR/QSPR. Los métodos de mapeo empleados tienen por característica común que se aplican de una manera supervisada, es decir teniendo en cuenta la variable a modelar. La primera de las aplicaciones mostradas fue tomada de la literatura (sección 7.2) y su utilización sirvió como punto de partida para subsecuentes variaciones desarrolladas en el marco de esta tesis doctoral (secciones 7.3 y 7.4).

Enfocándonos en los métodos propuestos en la sección 7.3 realizamos una extensión de la metodología original de manera que pudiera utilizarse para casos de múltiples clases. El objetivo de esta modificación se centró en la posibilidad de transformar una variable continua en k clases distintas, y así aproximarse a una situación de regresión. De las dos alternativas desarrolladas, AGA y CC, destacamos a CC como la alternativa más confiable, aún cuando no existen grandes

diferencias entre los resultados obtenidos con cada una de ellas. Asimismo, enfatizamos que esta alternativa es altamente competitiva en comparación a LDA, el cual es uno de los métodos de discriminantes lineal más establecido en la literatura.

La propuesta presentada en la sección 7.3, denominada MSR, realiza una modificación más acentuada respecto de la metodología anterior, en donde la metodología de mapeo permite trabajar sobre datos donde la variable a modelar es continua, es decir datos para ser modelados mediante una regresión. Por corresponder a una técnica de mapeo o de reducción de la dimensionalidad, la misma permite una visualización en un espacio de 1, 2 ó 3 dimensiones. Sin embargo el objetivo central de MSR, es el de mapear los datos de manera que usando el subespacio proyectado se logre mejorar la predicción del modelo, tal como se muestra en los resultados obtenidos en la sección 7.4.3.

Asimismo, también se realiza una selección implícita de los descriptores más influyentes. Por tal motivo, MSR corresponde a un método de selección de variables embebido, aunque en términos de la calidad de la selección de los descriptores, esta metodología no llega a ser tan potente como los métodos presentados en el capítulo 5. Por otra parte, también se mostró cómo la propuesta permite ser igualmente utilizada para problemas de clasificación ordinal y categórica. En el primer caso, no se requiere de ninguna modificación con respecto al modo de utilizarse para el caso de regresión, mientras que para el segundo caso, se requiere de la utilización de vectores ortonormales para la asignación de cada clase.

Una ventaja interesante de MSR es que sólo un parámetro debe ser establecido antes de ejecutar el método, el cual corresponde a la dimensionalidad del subespacio (p). Esto es altamente deseable, dado que no se requiere una alta cantidad de corridas para ajustar los parámetros óptimos para cada set de datos. Finalmente mencionamos que el conjunto de técnicas expuestas en este capítulo son igualmente aplicables a otras áreas de considerable interés en la ciencia de hoy en día, como los estudios en proteómica, metabolómica y transcriptómica en donde el gran número de variables hace que un mapeo a bajas dimensiones permita extraer información relevante sobre las variables y las relaciones existentes entre los datos.

7. INFERENCIAS SOBRE ESPACIOS OPTIMIZADOS MEDIANTE MATRICES ADAPTIVAS

Capítulo 8

Conclusiones

El objetivo global de esta tesis consistió en desarrollar conocimiento en el campo de la ciencia de la computación aplicado en el área de Quimioinformática. Más específicamente, se planteó el desarrollo de un conjunto de herramientas diseñadas para mejorar las metodologías existentes de predicción de propiedades biológicas o fisicoquímicas, las cuales son de alta utilidad durante el proceso de descubrimiento de nuevos fármacos.

El mecanismo de acción de las drogas medicinales es un área de gran complejidad e incertidumbre. Por esta razón, el proceso de descubrimiento de nuevas drogas es un proceso lento y cargado de resultados inesperados que obligan a abandonar un proyecto sobre una droga candidata tras varios años de estudio. El uso de herramientas computacionales que asisten en el desarrollo de nuevos medicamentos resulta de especial interés, no sólo desde el punto de vista económico, sino también desde un aspecto social teniendo en cuenta la fuerte implicancia que origina la posibilidad de generar nuevos medicamentos sobre la calidad de vida de los seres humanos.

Esta tesis se centró en los métodos basados en el enfoque QSAR/QSPR. Dichos métodos apuntan a analizar los parámetros químicos y estructurales de un compuesto para que sean relacionados con propiedades biológicas o físicoquímicas, en donde dichas propiedades requieren de costoso tiempo de trabajo experimental para poder ser medidas. El objetivo central de QSAR es, por lo tanto, capturar la relación estructura-propiedad y desarrollar un modelo que permita predecir la propiedad a partir de su composición molecular. Los parámetros químico-

estructurales usados se denominan descriptores, y su correcta selección resulta un punto crucial para que el modelo predictivo capture correctas relaciones del modelo. Asimismo, el desarrollo de metodologías de predicción basadas en algoritmos de aprendizaje automático requieren de un correcto diseño y aplicación. En este sentido, resulta importante proveer de metodologías de validación que permitan estimar la capacidad predictiva del modelo y su dominio de aplicación, de manera que la estimación de la capacidad predictiva se mantenga dentro de los valores esperados. Este punto resulta de considerable relevancia para que estos métodos de cálculo *in silico* puedan ser usados confiadamente sobre librerías de compuestos virtuales.

En este capítulo enunciaremos a modo de resumen las principales contribuciones de esta tesis, analizando las implicancias de las distintas propuestas. Además, se presentan recomendaciones y lineamientos generales para futuras investigaciones en el tema, así como también sugerencias sobre extensiones naturales del trabajo desarrollado en esta tesis.

8.1. Resumen de las contribuciones

En primer instancia, nuestro enfoque estuvo centrado en encontrar una forma de identificar los descriptores que resultan relevantes para la predicción de una determinada propiedad. La idea de nuestra metodología es seleccionar en forma automática ante un *pool* de descriptores de diferentes familias, uno o más subconjuntos de descriptores cuya información codificada sea de relevancia para capturar las relaciones que definen el comportamiento de una determinada propiedad.

En virtud del análisis bibliográfico y como consecuencia de diferentes experimentaciones, se propusieron dos modelos de selección de descriptores basados en el uso de técnicas de computación evolutiva. La diferencia esencial entre estos dos enfoques es que uno utiliza un método de búsqueda mono-objetivo, mientras que el otro realiza la búsqueda teniendo en cuenta dos objetivos simultáneos. La estrategia multi-objetivo representa una propuesta novedosa dentro de la literatura de selección de variables para problemas de tipo QSAR/QSPR. Más aún, esta estrategia proporciona ciertas ventajas por sobre la propuesta mono-objetivo. En

primer lugar, el criterio de búsqueda multi-objetivo permite concentrarse dentro de un área más acotada del posible espacio de combinaciones de variables y, por ende, encontrar mejores subconjuntos de variables. En forma simultánea, se demostró que el carácter multi-objetivo de la búsqueda hace que los conjuntos de descriptores encontrados sean menos propensos a generar modelos sobreajustados o con posibilidades de correlaciones por chance. Además, la búsqueda de subconjuntos de menor cardinalidad permite una mejor interpretación teórica de las relaciones estructura-propiedad.

No obstante las diferencias entre una metodología y la otra, ambas propuestas aportan interesantes ideas desde el punto de vista metodológico. En primer lugar se propuso que la metodología de selección de descriptores esté desdoblada en dos fases bien diferenciadas, en donde la primera se encarga de realizar una rápida exploración dentro del espacio de búsqueda, y la segunda fase realiza una selección final utilizando la información de la búsqueda de la primera fase y métodos de regresión de mayor precisión. Más aún, el estudio de distintas combinaciones de métodos para la primera fase aportó conocimiento sobre los diferentes resultados esperados en la aplicación de distintas estrategias de diseño. Finalmente, destacamos que este enfoque de dos fases, particularmente en su versión multi-objetivo, proporcionó la construcción de mejores modelos que los propuestos en la literatura para la predicción de la hidrofobicidad, la absorción intestinal y la penetración sobre la barrera hemato-encefálica. Vale destacar también que en esta misma línea de investigación se estudió la aplicación de métodos de selección de descriptores embebidos. Estos enfoques se basan en técnicas de búsquedas de proyecciones, los cuales permiten hacer una ponderación de las variables en forma automática.

Por otra parte, nuestras investigaciones estuvieron centradas en la mejora de las propuestas de predicción existentes. Como consecuencia de este estudio, se propuso una metodología que combina aprendizaje supervisado y no supervisado, para desarrollar un modelo de *mezcla de expertos*. El objetivo de esta propuesta es encontrar agrupaciones naturales de los compuestos y, de esta manera, aplicar modelos de predicción específicos para cada grupo de compuestos. En consecuencia, cuando un nuevo compuesto no observado previamente se presenta para

predecir su valor experimental, se evalúa su similaridad con los grupos conformados y luego se le aplica el modelo específico. Esta técnica nos ha proporcionado mejoras significativas en situaciones donde existen compuestos de composición heterogénea, en comparación a la construcción de un único modelo de predicción.

Otra importante contribución en el tema consistió en la proposición de una estrategia para evaluar la confiabilidad en la predicción de una propiedad biológica o fisicoquímica. La identificación de predicciones de compuestos no confiables resulta vital para el proceso actual de descubrimiento de nuevas drogas. Esta metodología utiliza un mapa auto-organizativo que actúan en combinación con el método de predicción que se desea evaluar. De esta manera, el método permite discriminar cada predicción en tres clases distintas, en función de si la predicción va a resultar *Confiable*, *No confiable* o de *Confiabilidad no categorizada*. La propuesta presentada se destaca ante otras existentes en el sentido que no sólo se identifican situaciones de extrapolación, sino que también se identifican ciertos problemas dentro del dominio de aplicación, los cuales pueden provenir de los datos utilizados o del método de predicción aplicado. La propuesta fue aplicada en cuatro sets de datos diferentes, obteniéndose resultados interesantes y significativos desde el punto de vista estadístico para todos los sets de datos.

Finalmente, la técnica de selección de descriptores embebida posee otra arista que resulta novedosa y al mismo tiempo interesante dentro de la literatura de modelos QSAR. En primer lugar, la técnica permite una reducción del espacio original a un subespacio de menor cardinalidad en donde las distancias representadas guardan relación con la distancia en el espacio de la variable a modelar. Esta idea se ha logrado generalizar para problemas de clasificación multi-clase y para problemas de regresión. El espacio reducido permite ser usado directamente para construir un modelo de predicción, obteniéndose resultados competitivos con técnicas ampliamente usadas como LDA. Además, ésta propuesta es robusta y eficiente en cuanto a su cálculo computacional. Por último, este tipo de técnicas permite también analizar las relaciones entre los datos por inspección visual.

8.2. Investigaciones futuras

Como extensión natural a las investigaciones realizadas en esta tesis, se proyecta ver la posibilidad de continuar mejorando los métodos desarrollados, su aplicación a otros campos de estudio y su mejora en términos de eficiencia computacional. En particular, la técnica de identificación de dominio de aplicación requiere ciertas mejoras en pos del establecimiento de umbrales en forma automática. Asimismo, la metodología de proyección a subespacios también resulta interesante de ser analizada como técnica de identificación de confiabilidad en las predicciones.

Es importante destacar que la naturaleza de las técnicas desarrolladas durante el transcurso de esta tesis doctoral, son también aplicables y atractivas para ser empleadas en otras aplicaciones del área de Bioinformática que presentan dificultades similares en la predicción de modelos. Un caso de estudio interesante, resultaría evaluar la metodologías presentadas a la selección de genes representativos que definen los patrones de expresión génica.

Por otra parte, los métodos evolutivos de selección de descriptores pueden mejorarse haciendo uso de técnicas de computación paralela. Esta mejora resulta de aplicación directa, teniendo en cuenta la naturaleza paralela de búsqueda de los métodos basados en algoritmos evolutivos.

Apéndice A

Lista de abreviaciones y notación matemática

A.1. Lista de abreviaciones

AD: árbol de decisión.

ADMET / ADME-Tox: absorción, distribución, metabolismo, excreción y toxicidad.

AGA: del inglés *All Against All*, "todos contra todos".

ANOVA: del inglés *ANalysis Of VAriance*, análisis de la varianza.

BBB: del inglés *blood-brain barrier*, barrera hemato-encefálica.

CC: comparaciones en cascada.

CRN: comité de redes neuronales.

DMS: diferencia mínima significativa.

kVC: k-vecinos más cercanos.

HIA: del inglés *human intestinal absorption*; absorción intestinal humana.

HTS: del inglés *high-throughput screening*; tamizaje de alto rendimiento.

LDA: del inglés *linear discriminant analysis*; análisis de discriminantes lineal.

LFER: del inglés *linear free-energy relationship*; relaciones de energía libre lineal.

LVQ: del inglés *learning vector quantization*; cuantización del vector de aprendizaje.

MAE: del inglés *mean absolute error*; error absoluto medio.

MSE: del inglés *mean square error*; error cuadrado medio.

NG: nodo ganador de un mapa auto-asociativo.

OECD: del inglés *Organisation for Economic Co-Operation and Development*; Organización para el Desarrollo y la Cooperación económica.

PCA: del inglés *principal component analysis*; análisis de componentes principales.

PLS: del inglés *partial least squares*; cuadrados mínimos parciales.

QSAR: del inglés *quantitative structure-activity relationship*; modelos cuantitativos de relación estructura-actividad.

QSPR: del inglés *quantitative structure-property relationship*; modelos cuantitativos de relación estructura-propiedad.

REACH: del inglés *Registration, Evaluation, Authorisation and restriction of Chemical Substances*; Registro, Evaluación, Autorización y restricción de Sustancias Químicas.

RL: regresión lineal.

RNL: regresión no lineal.

SARDUX: del inglés *supervised attribute relevance detection using cross-comparisons*; detección de relevancia de atributos supervisada usando comparaciones cruzadas.

SARDUX-MC: SARDUX para múltiples clases.

SOM: del inglés *Self Organizing Maps*; mapas auto-oganizativos.

A.2. Notación matemática

El significado de cada letra puede variar de un capítulo a otro, indicando explícitamente su significado según corresponda. En general consideramos que un conjunto de datos posee m observaciones y n variables en total.

Los vectores se muestran con una letra romana minúscula, como por ejemplo, $\mathbf{x}$ y se asume que es un vector columna. El correspondiente vector fila se indicará como $\mathbf{x}^T$ o como $(x_1, x_2, ..., x_n)$. Las letras mayúsculas serán reservadas para matrices, como en el caso de M.

Por lo tanto, si se tienen m observaciones de un espacio $\mathcal{X} \subset \mathbb{R}^n$, donde cada observación $\mathbf{x_1}, ..., \mathbf{x_m}$ son vectores n-dimensionales $\mathbf{x} = (x_1, x_2, ..., x_n)^T$, estos datos se pueden combinar en una matriz $X \in \mathbb{R}^{m \times n}$ en donde la i-ésima fila de X corresponde a la observación $\mathbf{x_i}^T$. Luego, el elemento i, j de X corresponde a la j-ésima variable de la i-ésima observación $\mathbf{x_i}$.

A. LISTA DE ABREVIACIONES Y NOTACIÓN MATEMÁTICA

Apéndice B

Sets de compuestos utilizados en nuestras metodologías

En el presente apéndice se resumen las principales características de los diferentes sets de datos usados en nuestras experimentaciones.

B.1.

Este set de datos contiene información de los compuestos extraídos de la base de datos PHYSPROP (*Physical Properties*) (PHYSPROP (1999)). Al momento de la extracción de los datos la base de datos contenía información de poco más de 13000 entidades químicas. De cada compuesto se dispone de su número de CAS (*Chemical Abstract Service* - un identificador único de entidades químicas), su notación SMILES (*Simplified Molecular Input Line Entry System* - una notación usada para representar estructuras moleculares por una cadena de caracteres) y del valor experimental del logaritmo del coeficiente de partición octanol-agua (logP).

Sobre cada entidad química se realizó la optimización de su estructura molecular, en pos de conocer su disposición geométrica y así poder calcular todos sus

descriptores, especialmente los descriptores 2D y 3D. Los descriptores fueron calculados usando el software DRAGON v3.0 (DRAGON (2003)). De esta manera se calcularon 1497 descriptores los cuales están divididos en 20 familias, según se detalla en Todeschini & Consonni (2000).

B.2.

El segundo set de compuestos consiste de 442 compuestos orgánicos, los cuales fueron compilados del trabajo de Yaffe *et al.* (2002). La variable a predecir es nuevamente el valor de logP a 25°C. Entre las 440 entidades químicas se tienen: hidrocarburos, halógenos, súlfidos, anilinas, alcoholes, ácidos carboxílicos entre otros. Para el presente artículo se usaron 12 descriptores, los cuales fueron obtenidos con un método de selección de variables. Los descriptores elegidos son: peso molecular, potencial de ionización, momento dipolar molecular promedio, dipolo total (carga puntual), dipolo total (hibridación), dipolo total (suma total), energía de doble-centro total, energía de cambio, energía de inacción electroestática total y los índices de conectividad molecular de valencia uno, dos y cuatro.

B.3.

Este conjunto de entidades químicas fue extraído del artículo de Konovalov *et al.* (2008) y nombrado en el mismo como 'KS289-logBB'. La variable experimental es logBB, la cual es comúnmente usada para medir el grado de penetración sobre la barrera hemato-encefálica. De este trabajo se reportan 289 compuestos, a los cuales se les calcularon los 1666 descriptores que permite la versión *on-line* de Dragon (E-DRAGON (2007)). Una vez removidos los descriptores con valores constantes o nulos para todos las entidades químicas, cada compuesto quedó especificado por 1501 descriptores. Este set de datos contiene un descriptor adicional 'Iv' que distingue si el logBB fue medido desde un ensayo *in vivo* o *in vitro*.

B.4.

Los compuestos y descriptores de este conjunto también fueron extraídos del trabajo de Konovalov *et al.* (2008), donde se lo refiere con el nombre de 'KS172-HIA'. En este caso la variable experimental de este conjunto de datos es el logHIA. Esta variable es una transformación no lineal de la absorción intestinal, expresada como porcentaje absorbido (%HIA), es decir, es el porcentaje de la dosis que se observa en la vena porta. Este set de datos contiene 127 compuestos y 1499 descriptores. Es importante mencionar que en el trabajo de Konovalov *et al.* fueron removidos los compuestos que presentaban un %HIA de 0 % ó 100 %.

B. SETS DE COMPUESTOS UTILIZADOS EN NUESTRAS METODOLOGÍAS

Referencias

Abraham, M., Zhao, Y., Le, J., Hersey, A., Luscombe, C., Reynolds, D., Beck, G., Sherbone, B. & Copper, I. (2002). On the mechanism of human intestinal absorption. *European Journal of Medicinal Chemistry*, **37**, 595–605. 26

Abraham, M., Ibrahim, A., Zhao, Y. & Acree, W. (2006). A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *Journal of Pharmaceutical Sciences*, **95**, 2091–2100. 61

Agrafiotis, D., Cedeño, W. & Lobanov, V. (2002). On the use of neural network ensembles in QSAR and QSPR. *Journal of Chemical Information and Computer Sciences*, **42**, 903–911. 63

Anastasiadis, A., Magoulas, G. & Vrahatis, M. (2005). New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing*, **64**, 253–270. 82, 159

Anderberg, M. (1973). *Cluster Analysis for Applications*. Academic Press, New York. 41

Aoyama, T., Suzuki, Y. & Ichikawa, H. (1989). Neural networks applied to pharmaceutical problems. I. method and application to decision making. *Chemical & Pharmaceutical Bulletin*, **37**, 2558–2560. 62

Arodź, T., Yuen, D. & Dudek, A. (2006). Ensemble of linear models for predicting drug properties. *Journal of Chemical Information and Modeling*, **46**, 416–423. 63, 119

Banik, G. (2004). In silico ADME-Tox prediction: The more, the merrier. *Current Drug Discovery*, **4**, 31–34. 1

Baumann, K. (2005). Chance correlation in variable subset regression: Influence of the objective function, the selection mechanism, and ensemble averaging. *QSAR & Combinatorial Science*, **24**, 1033–1046. 76, 111, 119

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, New Jersey. 44

Beresford, A., Segall, M. & Tarbit, M. (2004). In silico prediction of ADME properties: Are we making progress? *Current Opinion in Drug Discovery & Development*, **7**, 36–42. 124

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press. 34

Bishop, C. (2006). *Pattern Recognition and Machine Learning*, vol. I. Springer Science + Business Media, New York. 1, 27

Bodor, N., Gabanyi, Z. & Wong, C. (1989). A new method for the estimation of partition coefficient. *Journal of American Chemical Society*, **111**, 3783–3786. 60

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123–140. 37

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall/Crc, Boca Raton, Florida. 31, 78

Bremermann, H. (1958). The evolution of intelligence. The nervous system as a model of its environment. Tech. rep., Dept. Mathematics, Univ. Washington,. 46

Brown, R. & Martin, Y. (1996). Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences*, **36**, 572–584. 56

Bruneau, P. & McElroy, N. (2006). logD7.4 modeling using bayesian regularized neural networks. Assessment and correction of the errors of prediction. *Journal of Chemical Information and Modeling*, **46**, 1379–1387. 63, 66, 67

Bryson, A. & Ho, Y.C. (1969). *Applied Optimal Control*. Blaisdell, New York. 34

Burden, F. & Winkler, D. (1999). Robust QSAR models using bayesian regularized neural networks. *Journal of Medicinal Chemistry*, **42**, 3183–3187. 62, 100, 115

Burden, F. & Winkler, D. (2009). Optimal sparse descriptor selection for QSAR using bayesian methods. *QSAR & Combinatorial Science*, **28**. 64, 100

Camenisch, G., Folkers, G. & van de Waterbeemd, H. (1998). Shapes of membrane permeability-lipophilicity curves: Extension of theoretical models with an aqueous pore pathway. *European Journal of Pharmaceutical Sciences*, **6**, 321–329. 62

Caron, G. & Ermondi, G. (2008). Lipophilicity: Chemical nature and biological relevance. In R. Mannhold, ed., *Molecular Drug Properties: Measurement and Prediction*, vol. 37 of *Methods and Principles in Medicinal Chemistry*, 315–330, Wiley-VCH Velag GmbH & Co. 24

Caruana, R., Lawrence, S. & Giles, C. (2000). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Neural Information Processing Systems*, 402–408, MIT Press, Denver, USA. 138, 141

Clark, D. (1999). Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *Journal of Pharmaceutical Sciences*, **88**, 807–814. 63

Coello Coello, C. (2006). Evolutionary multiobjective optimization: a historical view of the field. *IEEE Computational Intelligence Magazine*, **1**, 28–36. 95

Coello Coello, C., Lamont, G. & Van Veldhuizen, D. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer Science+Business Media, LLC, New York. 49, 96

Collins, F., Morgan, M. & Patrinos, A. (2003). The human genome project: Lessons from large-scale biology. *Science*, **300**, 286–290. 8

Cronin, M. & Schultz, T. (2003). Pitfalls in QSAR. *Journal of Molecular Structure (Theochem)*, **622**, 39–51. 56, 62

Davies, S. & Rusell, S. (1994). NP-completeness of searches for smallest possible feature sets. 76

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. (2002). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, **50**, 1–18. 46, 143

Deb, K. (2004). *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons Ltd., Chichester. 95, 112

Deb, K. & Raji Reddy, A. (2003). Reliable classification of two-class cancer data using evolutionary algorithms. *Biosystems*, **72**, 111–129. 74

Deb, K., Pratap, A., Agrawal, S. & Meyrivan, T. (2002). A fast and elitist multiobjective genetic algorithm : NSGA-II. *IEEE Transactions on Evolutionary Computation*, **6**, 182–197. 50, 96

DiMasi, J., Hansen, R. & Grabowski, H. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, **22**, 151–185. 18

Dimitrov, S., Mekenyan, O., Sinks, G. & Schultz, T. (2003). Global modeling of narcotic chemicals: Ciliate and fish toxicity. *Journal of Molecular Structure*, **622**, 63–70. 66

Doweyko, A. (2008). QSAR: dead or alive? *Journal of Computer-Aided Molecular Design*, **22**, 81–89. 70, 124

Downs, G. (2004). Molecular descriptors. In P. Bultinck, H. De Winter, W. Langenaeker & J.P. Tollenaere, eds., *Computational medicinal chemistry for drug discovery*, Marcel Dekker Inc. 3, 58

DRAGON (2003). V3.0. Milano Chemometrics and QSAR Research Group. Department of Environmental Sciences, University of Milano-Bicocca, Italy. 58, 194

Du, Y., Liang, Y., Li, B. & Xu, C. (2002a). Orthogonalization of block variables by subspace-projection for quantitative structure property relationship (QSPR) research. *Journal of Chemical Information and Computer Sciences*, **42**, 993–1003. 68

Du, Y., Liang, Y. & Yun, D. (2002b). Data mining for seeking an accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. *Journal of Chemical Information and Computer Sciences*, **42**, 1283–1292. 68

Duch, W. (2006). Filter methods. In I. Guyon, S. Gunn, M. Nikravesh & L.A. Zadeh, eds., *Feature Extraction: Foundations and Applications*, Studies in Fuzziness and Soft Computing, 89–117, Springer- Verlag Berlin Heidelberg. 74

Dunn, W., Wold, S., Edlund, U., Hellberg, S. & Gasteiger, J. (1984). Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: The PLS method. *Quantitative Structure-Activity Relationships*, **3**, 131–137. 62

Dunnet, C. (1955). A multiple comparisons procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096–1121. 85

Duprat, A., Huynh, T. & Dreyfus, G. (1998). Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of logP. *Journal of Chemical Information and Computer Sciences*, **38**, 586–594. 60, 88

Dutta, D., Guha, R., Wild, D. & Chen, T. (2007). Ensemble feature selection: Consistent-descriptor subsets for multiple QSAR models. *Journal of Chemical Information and Modeling*, **47**, 989–997. 64, 78, 83, 119

Dy, J. (2008). Unsupervised feature selection. In H. Liu & H. Motoda, eds., *Computational Methods of Feature Selection*, Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC, Boca Raton. 74

E-DRAGON (2007). Dragon v5.4; http://www.vcclab.org/lab/edragon/. 106, 194

Emmanouilidis, C., Hunter, A. & Macintyre, J. (2000). A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. In *Evolutionary Computation*, vol. 1, 309–316. 117

Eriksson, L., Jaworska, J., Worth, A., Cronin, M., McDowell, R. & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspective*, **111**, 1361–1375. 62, 66

Erös, D., Kovesdi, I., Örfi, L., Takács-Novák, K., Acsády, G. & Kéri, G. (2002). Reliability of logP predictions based on calculated molecular descriptors: A critical review. *Current Medicinal Chemistry*, **9**, 1819–1829. 59, 60

Ertl, P. (2008). Polar surface area. In R. Mannhold, ed., *Molecular Drug Properties: Measurement and Prediction*, vol. 37 of *Methods and Principles in Medicinal Chemistry*, 111–126, Wiley-VCH Velag GmbH & Co, Weinheim. 26, 110

Fechner, N., Jahn, A., Hinselmann, G. & Zell, A. (2010). Estimation of the applicability domain of kernel-based machine learning models for virtual screening. *Journal of Cheminformatics*, **2**. 68

Figueiredo, M. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1150–1159. 64, 100, 101, 119, 120

Fisher, R. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics*, **8**, 376–386. 44

Fogel, L., Owens, A. & Walsh, M. (1966). *Artificial Intelligence through Simulated Evolution*. Wiley, New York. 46

Fontaine, F. (2005). *Development and applications of new 3D molecular descriptors*. Ph.D. thesis, Pompeu Fabra University. 58

Fox, T. & Kriegl, J. (2006). Machine learning techniques for in silico modeling of drug metabolism. *Current Topics in Medicinal Chemistry*, **6**, 1579–1591. 2

Freund, Y. & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computational System Sciences*, **55**, 119–139. 37

Fröhlich, H., Wegner, J. & Zell, A. (2004). Towards optimal descriptor subset selection with support vector machines in classification and regression. *QSAR & Combinatorial Science*, **23**, 311–318. 64, 119

Fujita, T., Iwasa, J. & Hansch, C. (1964). A new substituent constant, π, derived from partition coefficients. *Journal of American Chemical Society*, **86**, 5175–5180. 59

Gama, J. & Brazdil, P. (2000). Cascade generalization. *Machine Learning*, **41**, 315–343. 119

Gan, G., Ma, C. & Wu, J. (2007). *Data Clustering: Theory, Algorithms and Applications*. SIAM, Philadelphia, Pennsylavania. 27

Garg, P. & Verma, J. (2006). In silico prediction of blood brain barrier permeability: An artificial neural network model. *Journal of Chemical Information and Modeling*, **46**, 289–297. 61

Geman, S., Bienenstock, E. & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, **4**, 1–58. 37

Gharagheizi, F. (2008). QSPR studies for solubility parameter by means of genetic algorithm-based multivariate linear regression and generalized-regression neural network. *QSAR & Combinatorial Science*, **27**, 165–170. 77

Ghasemi, J. & Saaidpour, S. (2007). Quantitative structure-property relationship study of n-octanol-water partition coefficients of some of diverse drugs using multiple linear regression. *Analytica Chimica Acta*, **604**, 99–106. 60

Göller, A., Hennemann, M., Keldenich, J. & Clark, T. (2006). In silico prediction of buffer solubility based on quantum-mechanical and HQSAR- and topology-based descriptors. *Journal of Chemical Information and Modeling*, **46**, 648–658. 67

Godden, J. & Bajorath, J. (2003). An information-theoretic approach to descriptor selection for database profiling and QSAR modeling. *QSAR & Combinatorial Science*, **22**, 487–497. 64

Gola, J., Obrezanova, O., Champness, E. & Segall, M. (2006). ADMET property prediction: The state of the art and current challenges. *QSAR & Combinatorial Science*, **25**, 1172–1180. 1, 2, 19, 124

Goldberg, D. (1989). *Genetic Algorthms in Search, Optimization and Machine Learning*. Addison-Wesley, New York. 47, 48

Goldberg, D. & Deb, K. (1991). A comparative analysis of selection schemes used in genetic algorithms. In G.J.E. Rawlins, ed., *Foundations of genetic algorithms*, 69–93, Morgan Kaufmann Publishers, San Mateo, California. 81

Goldman, B. & Walters, W. (2006). Machine learning in computational chemistry. In D. Spellmeyer, ed., *Annual Reports in Computational Chemistry 2, 2*, vol. 2, 127–140, Elsevier. 62

Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science*, **26**, 694–701. 39, 56, 70, 150

Gramatica, P. & Papa, E. (2003). QSAR modeling of bioconcentration factor by theoretical molecular descriptors. *QSAR & Combinatorial Science*, **22**, 374–385. 150

Guerra, A., Paéz, J. & Campillo, N. (2008). Artificial neural networks in ADMET modeling: Prediction of blood-brain barrier permeation. *QSAR & Combinatorial Science*, **27**, 586–594. 61

Guha, R. & Jurs, P. (2004). Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors. *Journal of Chemical Information and Computer Sciences*, **44**, 2179–2189. 79

Gunturi, S. & Narayanan, R. (2007). In silico ADME modeling 3: Computational models to predict human intestinal absorption using sphere exclusion and kNN QSAR methods. *QSAR & Combinatorial Science*, **26**, 653–668. 62

Gutman, I., Ruščić, B., Trinajstić, N. & Wilcox, C. (1975). Graph theory and molecular orbitals. XII. Acyclic polyenes. *Journal of Chemical Physics*, **62**, 3399–3405. 57

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182. 74, 75, 119

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422. 74

Hammer, B., Strickert, M. & Villmann, T. (2004). Relevance LVQ versus SVM. In L. Rutkowski, J. Siekmann, R. Tadeusiewicz & L. Zadeh, eds., *Artificial Intelligence and Softcomputing*, vol. 3070, 592–597, Springer. 45

Handl, J. & Knowles, J. (2006). Feature subset selection in unsupervised learning via multiobjective optimization. *International Journal of Computational Intelligence Research*, **2**, 217–238. 77, 78

Hansch, C. & Fujita, T. (1964). *rho-σ-π* analysis. A method for the correlation of biological activity and chemical structure. *Journal of American Chemical Society*, **86**, 1616–1626. 59

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning. Data Mining, inference and predictions*. Springer. 27, 33, 38

Hawkins, D. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, **44**, 1–12. 123

Hawkins, D., Basak, S. & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, **43**, 579–586. 119

Hoffman, B., Cho, S., Zheng, W., Wyrick, S., Nichols, D., Mailman, R. & Tropsha, A. (1999). Quantitative structure-activity relationship modeling of dopamine D1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and k-nearest neighbor methods. *Journal of Medicinal Chemistry*, **42**, 3217–3226. 63

Holland, J. (1975). Adaptation in natural and artificial systems. Tech. rep., University of Michigan Press. 46

Hopfinger, A., Wang, S., Tokarski, J., Jin, B., Albuquerque, M., Madhav, P. & Duraiswami, C. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of American Chemical Society*, **119**, 10509–10524. 58

Horvath, D., Bonachera, F., Solov'ev, V., Gaudin, C. & Varnek, A. (2007). Stochastic versus stepwise strategies for quantitative structure-activity relationship generation-How much effort may the mining for successful QSAR models take? *Journal of Chemical Information and Computer Sciences*, **47**, 927–939. 65, 97, 109

Hou, T. & Wang, J. (2008). Structure - ADME relationship: still a long way to go? *Expert Opinion on Drug Metabolism & Toxicology*, **4**, 759–770. 2, 70, 124

Hou, T. & Xu, X. (2002). ADME evaluation in drug discovery: 1. Applications of genetic algorithms to the prediction of blood-brain partitioning of a large set of drugs. *Journal of Molecular Modeling*, **8**, 337–349. 61

Hou, T. & Xu, X. (2003). ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *Journal of Chemical Information and Computer Sciences*, **43**, 2137–2152. 61

Hou, T., Wang, J. & Li, Y. (2007). ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *Journal of Chemical Information and Modeling*, **47**, 2408–2415. 62

Iwase, K., Komatau, K., Hirono, S., Nakagawa, S. & Moriguchi, I. (1985). Estimation of hydrophobicity based on the solvent-accessible surface area of molecules. *Chemical & Pharmaceutical Bulletin*, **33**, 2114–2121. 60

Izrailev, S. & Agrafiotis, D. (2001). A novel method for building regression tree models for QSAR based on artificial ant colony systems. *Journal of Chemical Information and Computer Sciences*, **41**, 176–180. 63

Jacobs, R., Jordan, M., Nowlan, S. & Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87. 37

Jaszkiewicz, A. (2004). Evaluation of multiple objective metaheuristics. In X. Gandibleux, M. Sevaux, K. Sörensen & V. T'kindt, eds., *Metaheuristics for Multiobjective Optimisation*, vol. 535 of *Lecture Notes in Economics and Mathematical Systems*, 65–89, Springer, Berlin. 96

Jaworska, J., Comber, M., Auer, C. & Van Leeuwen, C. (2003). Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environmental Health Perspectives*, **111**, 1358–1360. 66, 69

Jaworska, J., Nikolova-Jeliazkova, N. & Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Alternatives to Laborative Animals*, **33**, 445–459. 66, 150

Jónsdóttir, S., Jørgensen, F. & Brunak, S. (2005). Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics*, **21**, 2145–2160. 4, 124

Johnson, R. & Wichern, D. (1992). *Applied Multivariate Statistical Analysis*, vol. III. Prentice Hall. 33

Johnson, S. (2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *Journal of Chemical Information and Modeling*, **48**, 25–26. 70, 119, 124

Jung, E., Kim, J., Kim, M., Jung, D., Rhee, H., Shin, J.M., Choi, K., Kang, S.K., Kim, M.K., Yun, C.H., Choi, Y.J. & Choi, S.H. (2007). Artificial neural network models for prediction of intestinal permeability of oligopeptides. *BMC Bioinformatics*, **8**, 245–253. 62

Kühne, R., Ebert, R.U. & Schüürmann, G. (2006). Model selection based on structural similarity-method descrption and application to water solubility prediction. *Journal of Chemical Information and Modeling*, **46**, 636–641. 67, 125, 130

Kier, L. & Hall, L. (1976). *Molecular connectivity in chemistry and drug research*. Academic Press, New York. 57

Klamt, A. & Smith, B. (2008). Challenge of drug solubility prediction. In R. Mannhold, ed., *Molecular Drug Properties: Measurement and Prediction*, vol. 37 of *Methods and Principles in Medicinal Chemistry*, 283–314, Wiley-VCH Velag GmbH & Co. 26

Kohavi, R. & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, **97**, 273–324. 78

Kohonen, T. (1997). *Self-Organizing Maps*, vol. II. Springer-Verlag, Heidelberg. 41

Konovalov, D., Coomans, D., Deconinck, E. & Heyden, Y. (2007). Benchmarking of QSAR models for blood-brain barrier permeation. *Journal of Chemical Information and Modeling*, **47**, 1648–1656. 61

Konovalov, D., Sim, N., Deconinck, E., Heyden, Y. & Coomans, D. (2008). Statistical confidence for variable selection in QSAR models via Monte Carlo cross-validation. *Journal of Chemical Information and Modeling*, **48**, 370–383. 61, 64, 100, 101, 103, 119, 120, 194, 195

Kuhn, H. & Tucker, A. (1951). Nonlinear programming. In J. Neyman, ed., *Second Berkeley Symposium on Mathematical Statistics and Probability*, 481–492, University of California Press, California. 95

Leach, A. (2001). *Molecular Modelling: Principles and Applications*. Pearson Education EMA. 22

Leach, A. & Gillet, V. (2007). *An Introduction to Chemoinformatics*. Springer, Dordrecht, The Netherlands. 54

Leo, A., Hansch, C. & Elkins, D. (1971). Partition coefficients and their uses. *Chemical Reviews*, **71**, 525–616. 25

Leonard, J. & Roy, K. (2005). On selection of training and test sets for the development of predictive QSAR models. *QSAR & Combinatorial Science*, **25**, 235–251. 140

Li, L., Weinberg, C., Darden, T. & Pedersen, L. (2001). Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142. 79

Liang, Y., Yuan, D., Xu, Q. & Kvalheim, O. (2007). Modeling based on subspace orthogonal projections for QSAR and QSPR research. *Journal of Chemometrics*, **22**, 23–35. 68

Liao, Q., Yao, J. & Yuan, S. (2006). SVM approach for predicting logP. *Molecular Diversity*, **10**, 301–309. 60

Lin, T.H., Chiu, S.H. & Tsai, K.C. (2006). Supervised feature ranking using a genetic algorithm optimized artificial neural network. *Journal of Chemical Information and Modeling*, **46**, 1604–1614. 65

Lipinski, C., Lombardo, F., Dominy, B. & Feeney, P. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **46**, 3–26. 25, 88

Liu, H. & Motoda, H. (2008). Less is more. In H. Liu & H. Motoda, eds., *Computational Methods of Feature Selection*, Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC, Boca Raton. 64, 94, 118

Liu, H. & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowlewdge and Data Engineering*, **17**, 1–12. 74

Liu, Y. (2004). A comparative study on feature selection methods for drug discovery. *Journal of Chemical Information and Computer Sciences*, **44**, 1823–1828. 64, 74

Livingstone, D. (2000). The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Sciences*, **40**, 195–209. 58

Livingstone, D. & Salt, D. (2005). Judging the significance of multiple linear regression models. *Journal of Medicinal Chemistry*, **48**, 661–663. 110

Loughrey, J. & Cunningham, P. (2005). Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search. Tech. rep., Trinity College Dublin. 78

MacKay, D. (1992a). Bayesian interpolation. *Neural Computation*, **4**, 415–447. 36

MacKay, D. (1992b). A practical bayesian framework fo back-propagation networks. *Neural Computation*, **4**, 448–472. 36

Madsen, K., Nielsen, H. & Tingleff, O. (2004). Methods for non-linear least squares problems. Tech. rep., Technical University of Denmark. 79

Manallack, D., Tehan, B., Gancia, E., Hudson, B., Ford, M., Livingstone, D., Whitley, D. & Pitt, W. (2003). A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *Journal of Chemical Information and Computer Sciences*, **43**, 674–679. 64, 67

Mannhold, R., Poda, G., Ostermann, C. & Tetko, I. (2008). Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *Journal of Pharmaceutical Sciences*, **98**, 861–893. 23, 60

Martin, Y. (2007). We can't predict logP, so why should we expect to predict binding affinity? In *CUP VIII*, Santa Fe, USA. 60

Martin, Y., Kofron, J. & Traphagen, L. (2002). Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, **45**, 4350–4358. 125

Mattioni, B. & Jurs, P. (2002). Development of quantitative structure-activity relationship and classification models for a set of carbonic anhydrase inhibitors. *Journal of Chemical Information and Computer Sciences*, **42**, 94–102. 64

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley. 45

Michalewicz, Z. & Schoenauer, M. (1996). Evolutionary algorithms for constrained parameter optimization problems. *Evolutionary Computation*, **4**, 1–31. 97

Molina, E., Díaz, H., González, M., Rodríguez, E. & Uriarte, E. (2004). Designing antibacterial compounds through a topological substructural approach. *Journal of Chemical Information and Computer Sciences*, **44**, 515–521. 64

Narayanan, R. & Gunturi, S. (2005). In silico ADME modelling: prediction models for blood-brain barrier permeation using a systematic variable selection method. *Bioorganic & Medicinal Chemistry*, **13**, 3017–3028. 61

Netzeva, T., Worth, A., Aldenberg, T., Benigni, R., Cronin, M., Gramatica, P., Jaworska, J., Kahn, S., Klopman, G., Marchant, C., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G., Perkins, R., Roberts, D., Schultz, T., Stanton, D., Sandt, J., Tong, W., Veith, G. & Yang, C. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *Alternatives to Laborative Animals*, **33**, 155–173. 66, 67

Nicolotti, O. & Carotti, A. (2006). QSAR and QSPR studies of a highly structured physicochemical domain. *Journal of Chemical Information and Modeling*, **46**, 264–276. 65

Nikolova-Jeliazkova, N. & Jaworska, J. (2003). Approaches to measure chemical similarity - a review. *QSAR & Combinatorial Science*, **22**, 1006–1026. 66

Nikolova-Jeliazkova, N. & Jaworska, J. (2005). An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN. *Alternatives to Laborative Animals*, **33**, 461–470. 66

Nilsson, N. (1965). *Learning machines; foundations of trainable pattern-classifying systems*. McGraw-Hill, New York. 38

Nocedal, J. (1980). Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, **35**, 773–782. 155, 171

OECD (2004). Principios OECD para la validación, para propósitos regulatorios, de modelos (Q)SAR www.oecd.org/dataoecd/33/37/37849783.pdf. 69

Oliveira, L., Sabourin, R., Bortolozzi, F. & Suen, C. (2003). A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition. *International Journal of Pattern Recognition*, **17**, 903–929. 77, 117

Oprea, T. (2005). *Chemoinformatics in Drug Discovery*, vol. 23 of *Methods and Principles in Medicinal Chemistry*. WILEY-VCH Verlag. 2

Osborne, M. & Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press, Cambridge. 48

Papa, E., Villa, F. & Gramatica, P. (2005). Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in pimephales promelas (fathead minnow). *Journal of Chemical Information and Modeling*, **45**, 1256–1266. 56

Papa, E., Kovarich, S. & Gramatica, P. (2009). Development, validation and inspection of the applicability domain of QSPR models for physicochemical properties of polybrominated diphenyl ethers. *QSAR & Combinatorial Science*, **28**, 790–796. 4

Paris, G. (1999). Meeting of the American Chemical Society. 54

PhRMA (2005). The Pharmaceutical Research and Manufacturers of America (PhRMA); What goes into the cost of prescription drugs? www.phrma.org. 18

PHYSPROP (1999). The Physical Properties Database; Syracuse Research Corporation (SRC); http://www.syrres.com/esc/physdemo.htm. 193

Platts, J., Abraham, M., Zhao, Y., Hersey, A., Ijaz, L., Butina, D. & D. Jouan-Rimbaud, D. (2001). Correlation and prediction of a large blood-brain distribution data set-an LFER study. *European Journal of Medicinal Chemistry*, **36**, 719–730. 61

Polley, M., Burden, F. & Winkler, D. (2005). Predictive human intestinal absorption QSAR models using bayesian regularized neural networks. *Australian Journal of Chemistry*, **58**, 859–863. 62, 63

Qin, S. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, **17**, 480–502. 150

Randic, M. (1975). On characterization of molecular branching. *Journal of American Chemical Society*, **97**, 6609–6615. 57

Rücker, C., Rücker, G. & Meringer, M. (2007). y-randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Computer Sciences*, **47**, 2345–2357. 76, 111

REACH (2007). Regulación (EC) No 1907/2006 del Parlamento Europeo y del Consejo (18/12/2006); http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=oj:l:2006:396:0001:0849:en:pdf . 69

Rechenberg, R. (1973). Evolutionstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Tech. rep., Frommann-Holzboog. 46

Rekker, R. & Manhold, R. (1992). The hydrophobic fragmental constant approach. In *Calculation of Drug Lipophilicity*, 112, VCH, Weinheim. 60

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408. 33

Rumelhart, D., Widrow, B. & Lehr, M. (1986). Learning representations by back-propagation errors. *Nature*, **323**, 533–536. 34

Safe, M., Carballido, J., Ponzoni, I. & Brignole, N. (2004). On stopping criteria for genetic algorithms. In A. Bazzan & S. Labidi, eds., *Advances in Artificial Intelligence - SBIA 2004*, vol. 3171/2004 of *Lecture Notes in Computer Science*, 405–413, Springer Berlin / Heidelberg. 82

Schneider, P., Biehl, M. & Hammer, B. (2008). Matrix adaptation in discriminative vector quantization. Tech. rep., Clausthal University of Technology. 45

Schroeter, T., Schwaighofer, A., Mika, S., Laak, A., Suelzle, D., Ganzer, U., Heinrich, N. & Müller, K.R. (2007). Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *Journal of Computer-Aided Molecular Design*, **21**, 651–664. 66, 68

Scsibrany, H., Karlovits, M., Demuth, W., Müller, F. & Varmuza, K. (2003). Clustering and similarity of chemical structures represented by binary substructure descriptors. *Chemometrics and Intelligent Laboratory Systems*, **67**, 95–108. 56

Segall, M., Beresford, A., Gola, J., Hawksley, D. & Tarbit, M. (2006). Focus on success: using a probabilistic approach to achieve an optimal balance of compound properties in drug discovery. *Expert Opinion on Drug Metabolism & Toxicology*, **2**, 325–337. 1

Selick, H., Beresford, A. & Tarbit, M. (2002). The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today*, **7**, 109–116. 1, 19

Shen, J., Du, Y., Zhao, Y., Liu, G. & Tang, Y. (2008). In silico prediction of blood - brain partitioning using a chemometric method called genetic algorithm based variable selection. *QSAR & Combinatorial Science*, **27**, 704–717. 65

Sheridan, R., Feuston, B., Maiorov, V. & Kearsley, S. (2004). Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences*, **44**, 1912–1928. 66, 125

Simmons, K., Kinney, J., Owens, A., Kleier, D., Bloch, K., Argentar, D., Walsh, A. & Vaidyanathan, G. (2008). Practical outcomes of applying ensemble machine learning classifiers to high-throughput screening (HTS) data analysis and screening. *Journal of Chemical Information and Modeling*, **48**, 2196–2206. 64

So, S.S. & Karplus, M. (1996). Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. *Journal of Medicinal Chemistry*, **39**, 1521–1530. 65

Soto, A., Cecchini, R., Vazquez, G. & Ponzoni, I. (2008a). An evolutionary approach for feature selection applied to ADMET prediction. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, **12**, 55–63. 88, 117

Soto, A., Cecchini, R., Vazquez, G. & Ponzoni, I. (2008b). A wrapper-based feature selection method for ADMET prediction using evolutionary computing. In E. Marchiori & J.H. Moore, eds., *Sixth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics - EvoBIO 2008*, vol. 4973 of *Lecture Notes in Computer Science*, 188–199, Springer / Heidelberg. 93, 117

Soto, A., Cecchini, R., Vazquez, G. & Ponzoni, I. (2009a). Supporting Information http://www3.interscience.wiley.com/journal/123212859/suppinfo. 109

Soto, A., Cecchini, R., Vazquez, G. & Ponzoni, I. (2009b). Multi-objective feature selection in QSAR using a machine learning approach. *QSAR & Combinatorial Science*, **28**, 1509–1523. 93, 117

Soto, A., Ponzoni, I. & Vazquez, G. (2009c). Segregating confident predictions of chemicals' properties for virtual screening of drugs. In S. Omatu, ed., *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, vol. 5518/2009, 1005–1012, Springer Berlin / Heidelberg. 149

Soto, A., Strickert, M. & Vazquez, G. (2010). A mapping method for linking chemical compounds to biological and physicochemical properties in drug discovery. In *ISCB Latin America*, Montevideo, Uruguay. 180

Strickert, M., Schleif, F.M., Seiffert, U. & Villmann, T. (2008a). Derivatives of Pearson correlation for gradient-based analysis of biomedical data. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, **12**, 37–44. 171

Strickert, M., Schneider, P., Keilwagen, J., Villmann, T., Biehl, M. & Hammer, B. (2008b). Discriminatory data mapping by matrix-based supervised learning metrics. In L. Prevost, S. Marinai & F. Schwenker, eds., *Artificial Neural Netwroks in Pattern Recognition*, vol. 5065 of *Lecture Notes in Computer Science*, 78–89, Springer. 45

Strickert, M., Soto, A., Keilwagen, J. & Vazquez, G. (2009). Towards matrix-based selection of feature pairs for efficient ADMET prediction. In *10^{th} Argentine Symposium on Artificial Intelligence, ASAI 2009*, 83–94, Mar del Plata. 160

Strickert, M., Soto, A. & Vazquez, G. (2010). Adaptive matrix distances aiming at optimum regression subspaces. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2010*, Bruges, Belgium. 180

Sun, Y. (2007). Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 1035–1051. 45, 153

Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R. & Feuston, B. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, **43**, 1947–1958. 63

Talete (2007). Dragon; http://www.talete.mi.it/help/dragon_help/ index.html?DRAGONDescriptorBlockList. 58

Tarbit, M. & Bermann, J. (1998). High-throughput approaches for evaluating absorption, distribution, metabolism and excretion properties of lead compounds. *Current Opinion in Chemical Biology*, **2**, 411–416. 21

Taskinen, J. & Yliruusi, J. (2003). Prediction of physicochemical properties based on neural network modelling. *Advanced Drug Delivery Reviews*, **55**, 1163–1183. 62

Testa, B., Carrupt, P.A., Gaillard, P., Billois, F. & Weber, P. (1996). Lipophilicity in molecular modeling. *Pharmaceutical Research*, **13**, 335–343. 62

Tetko, I. (2002). Associative neural network. *Neural Processing Letters*, **16**, 187–199. 130

Tetko, I. & Bruneau, P. (2004). Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *Journal of Pharmaceutical Sciences*, **93**, 3103–3110. 23

Tetko, I., Livingstone, D. & Luik, A. (1995). Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences*, **35**, 826–833. 123

Tetko, I., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V., Radchenko, E., Zefirov, N., Makarenko, A., Tanchuk, V. & Prokopenko, V. (2005). Virtual computational chemistry laboratory - design and description. *Journal of Computer-Aided Molecular Design*, **19**, 453–463. 88

Tetko, I., Bruneau, P., Mewes, H.W., Rohrer, D. & Poda, G. (2006). Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today*, **11**, 700–707. 3, 65, 67, 124

Tiño, P., Nabney, I., Williams, B., Lösel, J. & Sun, Y. (2004). Nonlinear prediction of quantitative structure-activity relationships. *Journal of Chemical Information and Computer Sciences*, **44**, 1647–1653. 60, 62

Todeschini, R. & Consonni, V. (2000). *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, Germany. 58, 83, 127, 194

Todeschini, R. & Gramatica, P. (2006). New 3D molecular descriptors: The WHIM theory and QSAR applications. In H. Kubinyi, G. Folkers & Y.C. Martin, eds., *3D QSAR in Drug Design*, vol. 2, 355–380, Springer Netherlands. 58

Tong, W., Hong, H., Fang, H., Xie, Q. & Perkins, R. (2003). Decision forest: Combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences*, **43**, 525–531. 63

Topliss, J. (1979). Chance factors in studies of quantitative structure-activity relationships. *Journal of Medicinal Chemistry*, **22**, 1238–1244. 62, 76, 110

Topliss, J. & Costello, R. (1972). Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, **15**, 1066–1068. 76, 110

Trevino, V. & Falciani, F. (2006). GALGO: An R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, **22**, 1154–1156. 79, 89

Tropsha, A. (2006). Variable selection QSAR modeling, model validation, and virtual screening. In D. Spellmeyer, ed., *Annual Reports in Computational Chemistry*, vol. 2, 113–126, Elsevier. 66

Tropsha, A., Gramatica, P. & Gombar, V. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, **22**, 69–77. 3, 65, 67, 70, 119

Varnek, A. & Tropsha, A. (2008). *Chemoinformatics Approaches to Virtual Screening*. Royal Society of Chemistry, Cambridge, UK. 54

Wang, R., Fu, Y. & Lai, L. (1997). A new atom-additive method for calculating partition coefficients. *Journal of Chemical Information and Computer Sciences*, **37**, 615–621. 88

Weinberger, K. & Saul, L. (2008). Fast solvers and efficient implementations for distance metric learning. In A. McCallum & S. Roweis, eds., *International Conference on Machine Learning*, vol. 307, 1160–1167, Helsinki, Finlandia. 45

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, **28**, 31–36. 57

Werbos, P. (1974). *Beyond Regression: New tools for prediction and analysis in the behavioral sciences*. Ph.D. thesis, Harvard. 34

Whitley, D., Ford, M. & Livingstone, D. (2000). Unsupervised forward selection: A method for eliminating redundant variables. *Journal of Chemical Information and Computer Sciences*, **40**, 1160–1168. 64

Wiener, H. (1947). Structural determination of paraffin boiling point. *Journal of American Chemical Society*, **69**, 17–20. 57

Winer, B., Brown, D. & Michels, K. (1991). *Statistical Principles in Experimental Design*. McGraw-Hill, New York. 85

Winkler, D. (2002). The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery. *Briefings in Bioinformatics*, **3**, 73–86. 1

Winkler, D. (2004). Neural networks in ADME and toxicity prediction. *Drugs of the Future*, **29**, 1043–1057. 2, 62

Winkler, D. & Burden, F. (2004). Modelling blood-brain barrier partitioning using bayesian neural nets. *Journal of Molecular Graphics and Modelling*, **22**, 499–505. 63

Wolpert, D. & Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, **1**, 67–82. 118

Xu, L., Wua, J.H.L., Shena, G.L. & Yu, R.Q. (2007). Variable-weighted PLS. *Chemometrics and Intelligent Laboratory Systems*, **85**, 140–143. 153

Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A. & Giralt, F. (2002). Fuzzy ARTMAP and back-propagation neural networks based quantitative structure - property relationships (QSPRs) for octanol-water partition coefficient of organic compounds. *Journal of Chemical Information and Computer Sciences*, **42**, 162–183. 88, 89, 100, 102, 104, 137, 155, 159, 194

Yang, L. & Jin, R. (2006). Distance metric learning: A comprehensive survey. Tech. rep., Michigan State University. 45

Yang, S.S., Lu, W.C., Gu, T.H., Yan, L.M. & Li, G.Z. (2009). QSPR study of n-octanol/water partition coefficient of some aromatic compounds using support vector regression. *QSAR & Combinatorial Science*, **28**, 175–182. 60, 77

Zheng, W. & Tropsha, A. (2000). Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Sciences*, **40**, 185–194. 63

Zhu, Z., Ong, Y.S. & Dash, M. (2007). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, **40**, 3236–3248. 74

Zitzler, E., Laumanns, M. & Thiele, L. (2002). SPEA2: Improving the strength pareto evolutionary algorithm. 50, 96