UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 6

TESIS EN OPCIÓN AL GRADO DE

MÁSTER EN BIOINFORMÁTICA.

"Desarrollo de modelos QSAR utilizando Programación Genética y Árboles de Regresión."

Autor: *Yania Molina Souto*.

Tutor: *Dr. Ramón Carrasco Velar*.

Ciudad de La Habana, Cuba.

Enero 2010

Tesis en opción al grado de Máster en Bioinformática

Título: Desarrollo de modelos QSAR utilizando Programación Genética y Árboles de Regresión.

Autor: Ing. Yania Molina Souto.

ymolinas@uci.cu

Tutor: Dr.C. Ramón Carrasco Velar

rcarrasco@uci.cu

Universidad de las Ciencias Informáticas.

Carretera a San Antonio de los Baños, Km 2½. Boyeros.

Ciudad de La Habana

Año: 2010

A mis padres…

A mis hermanos…

A toda mi familia…

Y especialmente a Jorge Iván…

AGRADECIMIENTOS.

Quiero agradecer primeramente a mi tutor, por brindar de forma desinteresada sus conocimientos y ayudarme en mi tesis de diploma y ahora en mi maestría. Agradecer a todos aquellos que tuvieron que ver de una forma u otra con mi formación profesional, a mis padres, a mis amistades que tanto apoyo me han dado y que a cada rato me preguntan “¿y la tesis?, ¿cómo vas?” A Jorge por su ayuda desinteresada, por su paciencia, por su amor y por las tantas veces que me consoló cuando sentía que nunca iba a terminar.

A mi mamá y a mi papá que estaban más preocupados que yo con el trabajo y a mis abuelos que siempre están al tanto de mi vida, apoyándome en todo lo que necesite.

RESUMEN.

Un área sumamente interesante dentro del modelado molecular es el diseño de nuevos compuestos. Los métodos QSAR han demostrado que las relaciones entre la actividad biológica y las propiedades físico-químicas se pueden cuantificar matemáticamente a partir de parámetros estructurales simples.

En los últimos años el interés por los modelos QSAR basados en técnicas de Minería de Datos ha crecido aceleradamente. La principal ventaja de estas técnicas es el hecho de poder construir un modelo sin especificar a priori la forma analítica según el comportamiento de los datos. La Programación Genética y los Árboles de Regresión pudieran ser muy útiles para establecer las complejas relaciones existentes entre la actividad biológica y los descriptores moleculares utilizados para describirla.

En el presente trabajo se proponen y estudian estas dos técnicas como opción tentativa para obtener las ecuaciones de regresión. Se estudian muestras diferentes de compuestos orgánicos y se analiza la competitividad de ambas técnicas en la búsqueda de soluciones.

Palabras claves: actividad biológica, descriptor molecular, QSAR.

ABSTRACT

The design of new compounds is an extremely interesting area inside the molecular modeling. The QSAR methods have demonstrated that the relationships between the biological activity and the physical-chemical properties of the compounds can be quantified mathematically starting from simple structural parameters. In the last years the interest for the model QSAR based in Data Mining techniques have grown quickly. The main advantage of these techniques is the fact of being able to build a model without specifying the analytic form a priori according to the behavior of the data. The genetic programming and the regression trees could be very useful to establish the complex existent relationships among the biological activity and the molecular descriptors used to describe them.

This work used Genetic Programming and Regression Trees as tentative option to obtain QSAR models. They are studied samples different from organic compound and the capacity of the artificial intelligence techniques is analyzed in the search of solutions.

Keywords: biological activity, molecular descriptor, QSAR.

GLOSARIO.

Descriptor: Número que describe la estructura química o una propiedad de la molécula o fragmento de esta.

Índices: Contienen información relacionada con la forma molecular, el grado de ramificación, tamaño molecular y la flexibilidad estructural.

Índice Topológico: Número que se calcula generalmente a partir de la matriz de adyacencia o de distancia de los elementos de un grafo molecular.

Compuestos orgánicos: Los compuestos o moléculas orgánicas son los compuestos químicos basados en Carbono, Hidrógeno y Oxígeno, y muchas veces con Nitrógeno, Azufre, Fósforo, Boro y halógenos.

Outlier: Medición atípica que altera fuertemente los resultados de un análisis de regresión.

Ligando: Iones o molécula que rodean a un metal en un complejo. Un ligando enlazado a un ion central se dice que está coordinado al ion.

Lipofilidad: Propiedad de las moléculas de disolverse en grasas.

Clusters: Se traduce como grupos. Crear clusters significa en ciencias de la computación agrupar por similitud.

Grafo molecular: Sistema que mediante punto y líneas representa la topología molecular.

Índice topológico: Número que se calcula generalmente a partir de la matriz de adyacencia o de distancias de los elementos de un grafo molecular.

Índice topográfico: Número que se calcula generalmente a partir de la matriz de adyacencia o de distancias entre los elementos de un grafo que han sido ponderados por un valor numérico que contiene información tridimensional del grafo molecular.

Modelo: Función de regresión que se obtiene para modelar una muestra de compuestos.

ÍNDICE

Introducción.

En la actualidad, tanto la cantidad como el tamaño de las bases de datos crecen aceleradamente. Este crecimiento de la información almacenada ha sido mayor que las potencialidades desarrolladas para procesarla; las capacidades para coleccionar y almacenar datos han sobrepasado la habilidad para analizarlos, resumirlos y extraer conocimiento a partir de ellos, conduciendo a un interés creciente por desarrollar aplicaciones que permitan extraer conocimiento útil de las mismas de forma automática.

En este sentido juegan un papel importante el uso de técnicas de Minería de Datos, a partir de la aplicación de procedimientos conocidos como modelos que, por su dependencia casi exclusiva de información histórica reportada, se conocen como modelos conducidos por datos. Dentro de estos modelos teóricos se pueden citar: las Redes Neuronales, los Árboles de Decisión, las Máquinas de Soporte Vectorial, la Programación Genética, los métodos de búsqueda heurística y metaheurística y las técnicas estadísticas, entre otras; incluyéndose estos algoritmos dentro del área de extracción del conocimiento en base de datos (Knowledge Discovery in Databases, KDD).

Cuando se utiliza una de estas técnicas para propósitos científicos se debe hallar un equilibrio entre la interpretabilidad y la potencia predictiva de las soluciones propuestas. En este sentido algunas técnicas hacen más énfasis en la primera mientras que otras son capaces de ajustar datos e inferir comportamientos a un grado de precisión alto, por ejemplo las Máquinas de Soporte Vectorial y la Programación Genética.

Cuando se investiga se deben balancear ambos aspectos, ya que soluciones con alta precisión y baja posibilidad para la interpretación, limitan el alcance de los resultados, mientras que soluciones fáciles de interpretar pero sin mucha precisión pueden concluir en ideas erróneas.

Un ejemplo en la búsqueda de conocimiento es la industria médico-farmacéutica. Estas instituciones invierten grandes períodos de tiempo para crear en el mercado un nuevo

medicamento, sin incluir el tiempo que demora el producto en ser aprobado por las entidades competentes, para lo cual se invierte una gran cantidad de dinero y recursos.

En el pasado, el descubrimiento de moléculas candidatas a fármacos con determinada acción o efecto terapéutico solía ser en la mayoría de los casos por azar, provocando en muchos casos reacciones adversas. Hoy día, el proceso de desarrollo de nuevos medicamentos tiene un enfoque mucho más seguro y diferente. Por lo general, se identifica primero la diana o blanco terapéutico y luego, a partir de técnicas disímiles se evalúan los compuestos candidatos.

Luego de múltiples pruebas se pasa a sintetizar la o las moléculas seleccionadas que seguidamente serán revisadas y aprobadas o no por entidades competentes después de pasar las evaluaciones preclínicas y clínicas que a nivel global controlan el proceso. Todo esto ha traído consigo grandes gastos y tiempo de espera, estimándose que el desarrollo completo de una molécula toma alrededor de 12-15 años y sólo 1 de cada 5.000 llega a ser un medicamento comercializable. A pesar de la gran cantidad de técnicas y tecnología moderna con la que cuenta la humanidad hoy, el tiempo de desarrollo de un fármaco sigue siendo elevado. Según un artículo publicado por la Revista Médica de Chile, en 1960 el tiempo promedio de desarrollo era de 8,1 años contra los 15,3 años que tomó en 1992, aumentando también los gastos asociados al proceso. En 1978 el costo total de desarrollo era de $54 millones de dólares; en 1990 era de $231 millones y en 1997 llegó a los $ 500 millones. Los controles cada vez más rigurosos por parte de las agencias regulatorias para asegurar que no se repitan los errores del pasado, contribuyen al aumento de los costos.

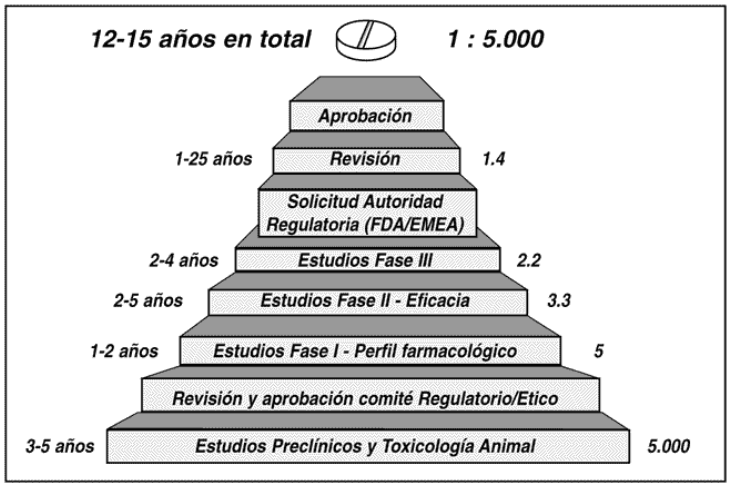


Ilustración 1. Diseño de Medicamentos, tomada de Revista Médica de Chile.

Por esta razón, en los últimos años la industria farmacéutica ha reorientado sus investigaciones y prestado más atención a aquellos métodos que permitan una selección racional o diseño de nuevos compuestos con propiedades deseadas. Uno de los nuevos enfoques es la utilización de métodos computacionales que relacionan la estructura química con la actividad biológica, que pueden dividirse en dos categorías: los Métodos de Modelación Molecular o SAR (Structure Activity-Relationships) y los Métodos QSAR (Quantitative Structure Activity-Relationships).

Los métodos SAR consideran las propiedades de las moléculas en tres dimensiones y son importantes en ellos aspectos como el análisis conformacional, la mecánica-cuántica, los campos de fuerzas y los gráficos moleculares interactivos (Carrasco, 2008).

Los métodos QSAR, a diferencia de los SAR, están basados en el empleo de técnicas de correlación entre la estructura química y la actividad biológica. Estas técnicas asumen que existe una relación implícita entre la actividad y la estructura molecular y tratan de establecer relaciones matemáticas simples para describir y luego extrapolar una o varias de esas actividades a un conjunto de compuestos que usualmente pertenecen a una misma familia. Dichas relaciones pueden determinarse a través de métodos matemáticos, ya sean, regresiones multilineales o métodos no lineales. Los

estudios QSAR utilizando técnicas estadísticas fueron aplicados exitosamente en diversos problemas de diseño de nuevos medicamentos (Hall, 1976) (Anzali, 1998).

A partir del impulso por la utilización de este tipo de técnicas, que permiten acortar los tiempos de investigación-desarrollo, con un gasto mínimo de recursos, se han desarrollado herramientas muy potentes que permiten modelar y predecir la actividad de un compuesto antes de que este sea sintetizado en un laboratorio, disminuyendo tiempo y esfuerzo por parte de los investigadores. Entre las que más se destacan concebidas para estos fines están: ADAPT, Accelerys y Trident que se basan en diferentes técnicas de procesamiento de la información, en métodos químico-cuánticos y en técnicas de Inteligencia Artificial. La gran mayoría de estas aplicaciones, sobre todo las de mayor prestigio, son muy costosas e inaccesibles para Cuba. En otros casos los investigadores necesitan varias herramientas para poder completar sus estudios, debido a que cada una está diseñada para fines específicos. Es por eso que surge la necesidad de desarrollar una plataforma que permita incluir todas las funcionalidades y permitir el diseño de un estudio estructura-actividad de forma completa, como contribución a la independencia tecnológica del país. Este proyecto surge con el nombre alasGRATO, un proyecto de investigación que se desarrolla en el grupo de Bioinformática de la Universidad de las Ciencias Informáticas (UCI) cuyo objetivo fundamental es permitir realizar estudios QSAR en compuestos orgánicos.

La plataforma está concebida para que el usuario pueda modelar sus propios compuestos o utilizar los ya existentes en la base de datos del proyecto. También se le brinda la facilidad de calcular los descriptores de las moléculas en estudio, y se pretende añadir la opción de generar los modelos QSAR o utilizar estos últimos para predecir actividad a partir de la descripción de la molécula por 2 vías fundamentales, fragmentos o descriptores.

Cuando el investigador realiza un estudio QSAR trata de cuantificar la relación que existe entre diferentes valores de descriptores, que representan de forma numérica características químicas o físicas, y la actividad biológica a partir de una función donde las variables independientes son el conjunto de descriptores que describen las moléculas o los fragmentos del ensayo y la variable dependiente es la actividad. Si se

pretende relacionar estos dos conjuntos de datos, siendo numéricas todas las variables, la vía más fácil para enfrentar el problema es utilizar un software estadístico que realice una regresión múltiple, sin embargo, esta variante no genera buenos resultados cuando en las muestras que se utilizan la relación entre las variables es no lineal. Ante este tipo de problema una vía de solución factible es la utilización de técnicas de Inteligencia Artificial (IA) para el análisis de los datos. Es por esto que se define como **problema científico**:

¿Cómo generar modelos QSAR en compuestos orgánicos a partir de descriptores moleculares utilizando técnicas de Inteligencia Artificial?

Teniendo como **objeto de estudio** la modelación de compuestos orgánicos empleando técnicas QSAR y como **campo de acción** la aplicación de técnicas de Inteligencia Artificial en la obtención de modelos QSAR.

El aporte práctico esperado con el desarrollo de este trabajo es brindar una propuesta de algoritmos que permitan y mejoren la generación de modelos de predicción a partir de descriptores moleculares, facilitando a los investigadores la realización de estudios QSAR de mejor calidad y con menos esfuerzo.

Para esto se trazó como objetivo general: Desarrollar modelos cuantitativos de relación estructura química-actividad biológica basados en Programación Genética y Árboles de Regresión que permitan predecir actividad en compuestos orgánicos.

A partir de un análisis del **objetivo general**, se derivan los siguientes objetivos específicos:

- ✓ Desarrollar modelos QSAR generados por Programación Genética.
- ✓ Desarrollar modelos QSAR empleando Árboles de Regresión.
- ✓ Validar los modelos QSAR.

Para lograr el cumplimiento de los objetivos propuestos se trazaron las siguientes tareas.

- ✓ Revisión del estado del arte respecto al uso de técnicas utilizadas en estudios QSAR.
- ✓ Análisis estadístico de las muestras utilizadas.
- ✓ Selección de variables utilizando algoritmos de selección.
- ✓ Generación de los modelos QSAR.
- ✓ Validación de los modelos desarrollados por Programación Genética y Árboles de Regresión a partir de análisis estadístico.

La aplicación de técnicas de IA en el modelado molecular es cada vez más aceptada y utilizada, quedando en el pasado los métodos de prueba y error y el análisis estadístico simple. Una de las técnicas de IA más utilizadas son las Redes Neuronales Artificiales, sin embargo surgen otros algoritmos que no son los suficientemente explotados a pesar de sus ventajas. En esta tesis se trabaja con técnicas que, a pesar de no ser de amplia difusión en este campo de la Química Medicinal comienzan a brindar resultados alentadores, lo que pudiera introducir cambios y mejoras en las soluciones hasta ahora encontradas, como la Programación Genética y los Árboles de Regresión, de capacidad interpretativa y predictiva aceptable.

La estructura de la tesis cuenta con una *Introducción* donde se explica de forma detallada la necesidad y el por qué de la investigación. En el *Capítulo 1* se tratan temas que brindan una panorámica del estado del arte en los aspectos que serán abordados a lo largo de la tesis, relacionados con el diseño de fármacos y con algoritmos de Minería de Datos para la extracción de conocimiento. El *Capítulo 2* describe los algoritmos, aplicaciones y muestras que se emplearon en el trabajo y finalmente, en el *Capítulo 3* expone la discusión de los resultados que se obtienen en cada uno de los experimentos computacionales realizados.

Como novedad científica el trabajo presenta nuevos modelos de relación estructura química-actividad biológica de cefalosporinas y de inhibidores del Factor Esteroidogénico 1 utilizando dos técnicas de Inteligencia Artificial, Programación Genética y Árboles de Regresión, útiles para la modelación de nuevos agentes antibacteriales y antitumorales.

Capítulo 1: Revisión Bibliográfica.

1.1. Los estudios QSAR.

Durante muchos años los químicos han tratado de encontrar una relación entre la estructura molecular y determinadas propiedades de las sustancias. A finales del siglo XIX Richet formula la idea de que es posible relacionar las variaciones estructurales en una serie de ligandos, con variaciones en la actividad de forma cuantitativa, a través de la ecuación:

$$\Delta\phi = f(\Delta C)$$

Ecuación 1. Modelo de Richet.

siendo *C* la estructura química y *Φ* la medida de actividad biológica, comenzando así, de forma muy rudimentaria, la era de los estudios de relación entre la estructura química y la actividad biológica. Por otra parte, en 1933 Louis Hammet observa que la acidez de derivados del ácido benzoico es directamente proporcional a las constantes de ionización de los ácidos benzoicos de partida. Estos estudios dieron lugar al desarrollo de constantes de sustituyente que describen el efecto de los mismos en dichos sistemas aromáticos. Pero no fue hasta principio de los años 60 que se realizan las primeras aproximaciones QSAR realizadas con éxito en el diseño de nuevas moléculas; desarrollándose dos métodos, desde puntos de vista teóricos diferentes:

La aproximación de Hansch y Fujita (Hansh, 1964): que supone que la energía libre de unión ligando-receptor se puede aproximar mediante una combinación lineal de contribuciones lipofílica, electrónica y estérica, relacionando así la actividad biológica con parámetros físico-químicos relativos a la lipofilidad y la electronegatividad a través de ecuaciones del tipo:

$$\log\left(\frac{1}{C}\right) = k_1 \log P + k_2\sigma + k_3$$

Ecuación 2. Ecuación de Hansh.

donde el log (1/C) representa la actividad biológica en estudio P es el coeficiente de partición del compuesto en un sistema bifásico que generalmente es octanol/agua y σ es la constante electrónica de sustituyentes de Hammett.

La aproximación de Free-Wilson (Bort, 2001): considera los aportes que hacen a la actividad de un compuesto cada uno de sus sustituyentes X_i en la estructura química localizados en posiciones j de la estructura μ y que se podían calcular a partir de la Ecuación 3:

$$\log\left(\frac{1}{C}\right) = \sum x_i + \mu$$

Ecuación 3. Ecuación de Free-Wilson.

El método aditivo de Free-Wilson fue pensado para construir moléculas nuevas a partir de la unión de sus fragmentos pero solo pudo ser utilizado en estudios de relación estructura-actividad de series congenéricas.

Aunque Hansch y Fujita desarrollan durante la década de los 60 los primeros estudios QSAR, no es hasta los años 80 que se introduce el diseño computacional en el proceso de desarrollo de medicamentos. Esto se facilitó por el desarrollo teórico de técnicas de modelación molecular y la aparición de ordenadores personales. No obstante, la realización de estudios sistemáticos en cantidades apreciables de compuestos, aún de series congenéricas, se veía limitado por razones prácticas para la ejecución de los experimentos de laboratorio que aportan información químico-física. La necesidad de representar la estructura química mediante procedimientos sencillos para poder establecer dichas relaciones estructurales y de propiedad, permitió el resurgimiento de la teoría de grafos como herramienta para la descripción de la estructura química. Surgen entonces los índices topológicos que describen la estructura de la molécula mediante un número calculado a partir de la matriz de conectividad de los vértices del grafo molecular (Balaban, 1976). A partir del hecho de que las propiedades moleculares o actividades de un compuesto pueden representarse

mediante números, los modelos QSAR se reducen a una correlación entre dos conjuntos de números. Estas relaciones estructura-actividad permiten relacionar cuantitativamente los cambios estructurales de una serie de moléculas con los cambios en la actividad.

Actualmente, se utilizan múltiples descriptores de la estructura química combinados con la aplicación de técnicas de ayuda a la toma de decisiones, que incluyen, desde los métodos clasicos estadísticos hasta las más modernas técnicas de Inteligencia Artificial, como son los Algoritmos Genéticos (AG), las Redes Neuronales, las Máquinas de Soporte Vectorial, y otras.

1.2. Descriptores moleculares.

Los métodos QSAR han demostrado que las relaciones entre la estructura molecular y las propiedades físico-químicas de los compuestos se pueden cuantificar matemáticamente a partir de parámetros estructurales simples, conocidos como descriptores. La validez de estos métodos depende en gran medida de los descriptores utilizados para caracterizar la estructura química y de la calidad de los modelos matemáticos que describen los fenómenos biológicos. Un descriptor puede ser tanto teórico como experimental, resultado de la cuantificación de una propiedad o de un procedimiento matemático y lógico que caracterice al compuesto.

En estudios QSAR, a pesar de la investigación teórica y experimental, no existe acuerdo acerca de aquel conjunto de descriptores óptimo, y dado que diferentes descriptores codifican distinta información, la estrategia consiste en aplicar aquellos más relevantes según la particularidad del caso en estudio. Dentro de los más utilizados hasta el momento se encuentran los índices topológicos que se basan únicamente en la estructura 2D o topología de la molécula, los cuales se derivan generalmente de las matrices de conectividad o de distancia del grafo molecular. Se distinguen también los índices topográficos, los cuales incluyen además, otras propiedades estructurales de los átomos implicados y los índices basados en la teoría de la información. En general, estos índices contienen información relacionada con la forma molecular, el grado de ramificación, tamaño molecular y la flexibilidad estructural. Entre los más conocidos se

destacan los índices de conectividad, propuestos por Randic y desarrollados en profundidad por Hall y Kier. Son rápidos de calcular y se ha comprobado que correlacionan con diferentes propiedades químico-físicas y biológicas (Carrasco, 2008).

En el presente trabajo se utilizan varios descriptores moleculares, de los cuales se hablará de forma general en el próximo capítulo.

1.3. Técnicas estadísticas de análisis de datos.

Los datos de estructura química generados a partir de la teoría de grafos son voluminosos, pudiéndose definir una molécula a partir de múltiples forma que reflejan características distintas por poseer esos descriptores diferente contenido de información. Por esa razón, los descriptores calculados mediante los diferentes procedimientos no siempre resultan útiles para generar un modelo.

Para evaluar la calidad del conjunto de variables a utilizar se utilizan varias medidas estadísticas como la varianza y la correlación entre ellas. La varianza permite ver el grado de variación de un descriptor a lo largo del conjunto de datos, de manera que si esta es muy baja, el descriptor aporta muy poca información al conjunto. La correlación mutua orienta acerca del grado de redundancia interna. Si un par de descriptores independientes presentan un coeficiente de correlación nulo, se denominan ortogonales, y si la correlación es 1 se elimina una de las 2 variables del conjunto, ya que ambas aportan la misma información.

De forma general se podría resumir que, una vez que se calcula un conjunto de descriptores éstos no pueden utilizarse directamente para generar un modelo, ya que deben eliminarse tres tipos de problemas:

- ✓ La existencia de alta correlación entre las variables.
- ✓ Descriptores con información poco relevante.
- ✓ El número de descriptores es tan elevado que no es posible tratarlo computacionalmente.

De esta forma se realiza una primera depuración del conjunto de variables reduciéndose la dimensionalidad del problema y se obtiene un conjunto más reducido con una mayor densidad de información relacionada con la propiedad objetivo que en el caso que se analiza es la actividad biológica.

Incluso con esta primera depuración no es suficiente y se necesita encontrar un conjunto más reducido de variables para poder, de forma más factible y explicativa, modelar la muestra con respecto a la variable en estudio. Es aquí donde entran a jugar un papel importante los algoritmos de selección de variables que serán, algunos de ellos, detallados en el siguiente epígrafe.

1.4. Selección de variables.

Utilizando un criterio de selección de variables, la dimensionalidad de los datos puede reducirse sin perder información útil, minimizando al mismo tiempo la información compuesta por ruido. La estrategia para la selección más simple consiste en evaluar cada una de las variables de forma separada y seleccionar aquellas que aportan información de mayor calidad, pero esto no resulta eficiente, porque se ignora la redundancia y la sinergia entre los descriptores y difícilmente se llegará al conjunto óptimo. Ante este problema, una de las salidas es evaluar todos los posibles subconjuntos de variables que se puedan formar y escoger el óptimo por lo que se hace necesaria la utilización de un método de selección.

Cualquier procedimiento para seleccionar variables basa su funcionamiento en dos aspectos fundamentales:

- ✓ El criterio para seleccionar las variables.
- ✓ El método de búsqueda.

Por lo tanto, es preciso encontrar el criterio de selección y el procedimiento de búsqueda que encuentre resultados cercanos al óptimo global y una vez que se encuentre el conjunto de variables se pasará de entrada al algoritmo de reconocimiento de patrones para formar el modelo.

Para evitar una búsqueda exponencial exhaustiva, se han desarrollado diferentes métodos que exploran el espacio de una manera eficaz. Estas estrategias pueden clasificarse en: exponenciales, secuenciales y aleatorias (Guerrero, 2006).

Las técnicas exponenciales realizan búsquedas cuya complejidad crece exponencialmente a medida que aumenta el número de parámetros, haciendo un gran uso de los recursos computacionales. Los métodos secuenciales siguen estrategias que solo exploran partes del espacio de soluciones, aplicando búsqueda local. Entre ellos se pueden citar los ya mencionados Forward Selection, el Backward Elimination y el Stepwise Regression. Estos procedimientos tienden a quedar atrapados en mínimos locales con facilidad.

Por último, existen los algoritmos aleatorios que intentan minimizar los costos computacionales de los métodos exponenciales. Este tipo de estrategia realiza búsquedas locales alrededor de soluciones prometedoras pero posee una componente aleatoria que les permite explorar otras soluciones en el espacio de búsqueda minimizando el riesgo de quedar atrapados en mínimos locales.

Existen varios trabajos sobre selección de variables en estudios QSAR, entre ellos destacan los trabajos de Tang (Tang, 2002) y colaboradores que utilizan técnicas de selección de variables basadas en AG, demostrando la capacidad de estos métodos para describir la relación entre una serie de compuestos y su actividad biológica. Para construir los modelos, los autores emplean Partial Least Square (PLS) una vez hecha la selección lo que les permite evaluar la calidad de los mismos.

Lu Xu (Xu, 2001) y colaboradores, a diferencia de Tang, utilizan además de Algoritmos Genéticos otros métodos clásicos como Forward Selection, Backward Elimination, Stepwise Regression y otros más novedosos como Branch and Bound. Esto permitió establecer una comparación entre los distintos métodos a partir de modelar un conjunto de 35 nitrobenzenos correlacionando su actividad tóxica con descriptores calculados con MOPAC (Molecular Orbital Package), concluyendo que los resultados obtenidos con los AG eran superiores a los métodos estadísticos clásicos. Alexandridis (Alexandridis, 2005) utiliza, para seleccionar variables en modelos no lineales, AG y

Enfriamiento Simulado, demostrando las ventajas de estos dos algoritmos en distintos juegos de datos referenciados en varios trabajos y obtienen los modelos de regresión con una Red Neuronal RBF (Radial Basic Function).

Aunque ya existe tendencia al uso de métodos de Inteligencia Artificial en estudios QSAR, la mayoría de los investigadores se limitan a utilizar técnicas estadísticas de selección y regresión. Una parte de la literatura existente utiliza técnicas novedosas, y dentro de estas existe una tendencia al uso de Algoritmos Genéticos y Enfriamiento Simulado, incluso existen trabajos donde ambas técnicas son comparadas y se concluye que los resultados en ambos casos están por encima de la media de otros algoritmos de selección (Hernández, 2010)

1.4.1 Algoritmos Genéticos.

A lo largo del período de evolución de los seres vivos, se han ido seleccionando conductas que son adecuadas para la supervivencia de las especies y los individuos que sobreviven son aquellos que están mejor adaptados al ambiente que los rodea. Por esta razón pueden reproducirse y transferir a sus descendientes sus cualidades más beneficiosas y que le permitieron sobrevivir en el medio ambiente en que se desarrollan.

Este procedimiento aplicado en la naturaleza ha sido fuente de interés y de inspiración en el campo de las ciencias de la computación y la ingeniería para el desarrollo de métodos de optimización y búsqueda alternativa, conformando lo que se ha denominado en el contexto de aprendizaje de máquina un paradigma evolutivo (Holland, 1975) (Goldberg, 1989).

Los Algoritmos Genéticos utilizan operadores genéticos que son los encargados de dirigir la selección y la transformación de los individuos a lo largo del proceso simulado. Entre estos operadores se pueden mencionar: la selección, la reproducción, el cruce y la mutación.

Un AG simple, involucra los siguientes pasos:

1. Generar aleatoriamente una población inicial P de n individuos.

2. Evaluar los individuos de la población. Calcular la función de adaptación o ajuste asociada a cada individuo.

3. Mientras no se alcance la condición de parada, tiene lugar el ciclo evolutivo, este es:

3.1 Seleccionar individuos para tener descendencia.

3.2 Aplicar operadores genéticos a individuos seleccionados.

3.3 Evaluar descendencia.

4. Devolver mejor individuo o mejor conjunto de individuos de la población final según sea el caso.

Estos métodos son altamente paralelos y por esta razón pueden evaluar múltiples esquemas a la vez, a diferencia de otros métodos de búsqueda que solo exploran el espacio de soluciones hacia una solución en una única dirección y si esta no es la más óptima la asumen como tal.

La mayoría de los problemas prácticos tienen un espacio de soluciones enorme y es un reto encontrar cual de ellas es la mejor, corriendo el riesgo de quedar atrapados en aquellas que aunque prometedoras no son las mejores (Holland, 1975) (Koza, 1989).

1.4.2 Enfriamiento Simulado.

Este método de optimización fue desarrollado por Kirkpatrick y colaboradores y permite encontrar soluciones muy cercanas al óptimo global. El Enfriamiento Simulado simula un proceso en el cual un sólido es fundido aumentando su temperatura a un valor elevado y luego se va enfriando lentamente hasta llegar a un estado de mínima energía o equilibrio térmico (Zahavi, 2006).

Este algoritmo explora el espacio de búsqueda formado por el universo de todas las posibles combinaciones de variables que puedan ser utilizadas para resolver el problema. El método parte de un conjunto de variables A, y calcula el error que se comente al usar este conjunto, luego elimina aleatoriamente de A un número de descriptores y forma un nuevo conjunto B. Se calcula el error de predicción que se comente al usar la nueva combinación y si el error disminuye al formar el conjunto B significa que la predicción ha mejorado eliminando esas variables. Si el error aumenta no se rechaza automáticamente la nueva selección, sino que se pasa a calcular la probabilidad p de aceptación:

$$P_i = \exp\left(- \Delta E / T_i \right)$$

Ecuación 4. Probabilidad de aceptación en Enfriamiento Simulado.

donde *Ti* es la temperatura inicial y *–ΔE* es el error de predicción del conjunto A menos el error del conjunto B. Si esta probabilidad de aceptación es menor que un número R escogido aleatoriamente entre 0 y 1 la nueva solución sigue resultando escogida y se siguen eliminando variables del conjunto B. En caso contrario el algoritmo prosigue desde A. El proceso se ejecuta un número N de iteraciones igual a la cantidad de variables menos 1(Cambell, 1995).

1.5. Algoritmos de Agrupamiento.

Las técnicas de agrupamiento o clustering se utilizan para encontrar conglomerados de casos que son relativamente homogéneos entre si pero existe un grado de heterogeneidad, sobre la base de un conjunto definido de variables, pero que por simple inspección no pueden ser agrupados. Estos métodos parten de una matriz inicial de datos, la cual está formada por las variables que se utilicen y el número de casos del ensayo. Para crear los conglomerados se utiliza una medida que evalúa las diferencias entre un caso y otro de la muestra.

La medida de similitud más utilizada es midiendo las distancias entre los objetos, si esta es pequeña los objetos se parecerán más entre si y serán ubicados en los mismos

grupos, en caso contrario se asignan a grupos distintos. También se puede utilizar la correlación, en el caso de variables métricas y el grado de asociación cuando estas son categóricas.

Cuando se determina cómo medir las diferencias entre los objetos el paso siguiente es estandarizar los valores de las variables si estas están en escalas diferentes. Luego se elige el método de agrupamiento que se utilizará y se pasa a formar los *clusters*. Estos métodos de agrupamiento se dividen en dos categorías:

Los métodos jerárquicos: Comienza creando para cada objeto un *cluster*. En cada iteración del método, el criterio por el que los objetos son separados se relaja y se van uniendo los conglomerados más cercanos a partir de la medida de similitud que se escogió. Los métodos jerárquicos determinan por si solos la cantidad óptima de grupos que se deben formar, aunque es válido aclarar que cuando los datos presentan mucho ruido puede que el procedimiento tenga problemas para llegar a una respuesta satisfactoria.

En estos algoritmos el agrupamiento se da a partir de un árbol jerárquico de *clusters* conocidos como dendogramas que le permite al investigador, en caso de que el número final de grupos sea muy grande, dividir el dendograma y seleccionar la cantidad de conjuntos que quiere para su estudio o escoger desde la raíz del árbol el número clusters que desee. En los dendogramas los casos se representan como nodos y las ramas indican los casos que se han ido uniendo en un mismo conglomerado.

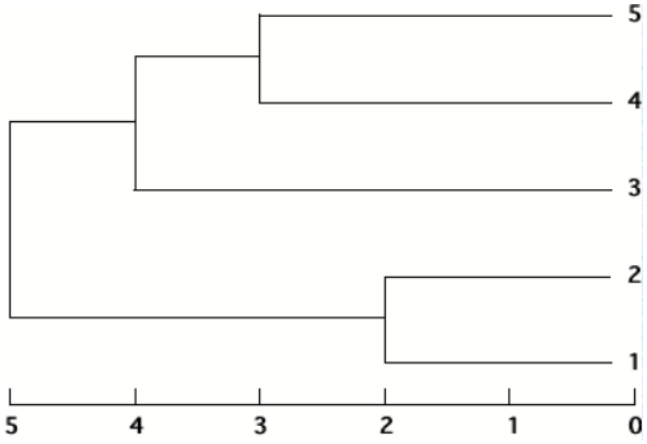


Ilustración 2. Dendograma de 4 *clusters*.

Los métodos no jerárquicos: Estos métodos a diferencia de los jerárquicos necesitan que se les defina un número de clusters. Normalmente consta de los siguientes pasos:

1.- Seleccionar K centroides iniciales, siendo K el número de clusters deseados.

2.- Asignar cada observación al cluster que le sea más cercano.

3.- Reasignar o relocalizar cada observación a uno de los K cluster de acuerdo con alguna regla de parada.

4.- Parar si no hay reasignación de los puntos o si la reasignación satisface la regla de parada. En otro caso se vuelve al paso dos.

Los algoritmos no jerárquicos son, en la mayoría de los casos, más robustos que los jerárquicos y más rápidos pero tienen como desventaja el hecho de tener que fijar *a priori* el número de *clusters*. Por esta razón no se considera que un tipo sea mejor que otro sino que ambos se compensan, es por eso que al hacer un análisis de agrupamiento primero se recomienda utilizar un método jerárquico que explore y defina la cantidad de conglomerados a formar y después de definir el número, utilizar un no jerárquico para construir los conjuntos.

Dentro de los jerárquicos los más utilizados son el Método del Centroide. Con este algoritmo, los grupos una vez formados se representan por sus valores medios para cada variable, es decir su vector de medias y las distancias entre los grupos se definen en términos de las distancia entre vectores de medias. Esto trae como consecuencia que cuando se crea un nuevo grupo esté más cercano a los grupos grandes y más alejado de los pequeños. Otro método es el Vecino más Cercano, su principal característica es que la distancia entre grupos se define como la del par de individuos que está más cerca.

Dentro de los jerárquicos un algoritmo que ha sido muy estudiado es el Método de Ward y en trabajos que han sido publicados se han hecho comparaciones y se considera como un procedimiento que se comporta muy estable y genera agrupamientos de calidad comportándose en la mayoría de los casos por encima de la

media de los de su especie (García, 2005) (Ruiz, 2006). En el caso de los no jerárquicos entre los más utilizados están K-Means, Fuzzy CMeans y los mapas autoorganizados. El Método de Ward y K-Means se utilizaron en una parte de la investigación y serán detallados en el Capítulo 2.

1.6. Algoritmos de optimización.

La búsqueda cuantitativa de la relación entre la estructura de un compuesto y su actividad biológica es, en esencia, un problema de regresión. Históricamente, la Regresión Lineal Múltiple (RLM), los Mínimos Cuadrados Parciales y el Análisis de Componentes Principales han sido los métodos más utilizados por los científicos para generar modelos QSAR y QSPR (Quantitive Structure Property Relationships). Sin embargo, en el caso de RLM se tiene el inconveniente que se debe fijar *a priori* el comportamiento de los datos, y ya esto representa una dificultad aun en aquellos con conocimientos bastos en el tema. Incluyendo el hecho de que estas técnicas no son capaces, por su naturaleza, de explorar todo el espacio de soluciones y encontrar la combinación óptima de variables que deben conformar el modelo. Esto presupone que el uso de técnicas de ayuda a la toma de decisiones puede introducir mejoras significativas en los modelos QSAR.

Actualmente existen un gran número de estas técnicas disponibles (Gutiérrez, 2009), el problema radica en saber distinguir cual de ellas se adapta mejor a las condiciones del problema a resolver. En el caso de los estudios QSAR no solo se trata de predecir con exactitud el valor de los datos sino también de llegar a ecuaciones de regresión de las cuales se pueda inferir algún tipo de conocimiento.

En los últimos años se han realizado varios estudios QSAR con técnicas novedosas donde se evidencia una tendencia al uso de las Redes Neuronales y las Máquinas de Soporte Vectorial para predecir. Entre ellos se pueden citar las investigaciones de Fatemi (Fatemi, 2009) donde se construyen modelos para predecir el factor de biomagnificación de algunos de los contaminantes organoclorados a partir de Regresión Lineal Múltiple y Redes Neuronales, para la selección de los descriptores utiliza AG. El autor establece una comparación entre las dos técnicas de regresión y obtiene las

mejores predicciones con RN con capacidad predictiva (Q^2) de 0.97. Goodarzi (Goodarzi, 2009) y colaboradores construyen modelos de regresión para la actividad inhibitoria de la glucógeno sintasa quinasa-3beta, para ello utilizan un algoritmo difuso basado en Máquinas de Soporte Vectorial (MSV), y establecen una comparación con RLM y RN obteniendo buenos resultados con MSV. Caballero (Caballero, 2008) utiliza un algoritmo que combina Redes Neuronales y AG y predice la actividad antagonista del receptor plaquetario humano contra la trombina, compara los resultados que obtiene con los que calcula con CoMFA(Comparative Molecular Field Analysis) y CoMSIA (Comparative Molecular Similarity Indices Analysis) y estos son superiores. Worachartcheewan (Worachartcheewan, 2009) y colaboradores utilizan una RN para modelar la actividad de los inhibidores de la furina donde los coeficientes de correlación del modelo que generan es de 0,92 y el error cuadrático medio de 0,3.

Todos estos trabajos muestran que el uso de técnicas de ayuda a la toma de decisiones pueden mejorar los resultados de los estudios QSAR. Sin embargo, existen algoritmos de este tipo que no han sido muy explotadas en problemas de Química Medicinal y que su uso en otros campos de investigación arroja resultados prometedores.

Dentro de la Computación Evolutiva, la Programación Genética es una técnica que promete evolucionar y desarrollar la investigación por su estructura flexible que permite afrontar problemas de clasificación y regresión. Sin embargo se emplea poco en problemas QSAR. Entre los trabajos que utilizan esta técnica está el de Nicolotti (Nicolotti, 2004) para analizar un conjunto de datos de 58 nicotinoides de gran actividad caracterizado por 56 descriptores y obtiene modelos de regresión con R^2=0,79 que en este tipo de estudios se consideran resultados buenos. Otros trabajos están dirigidos a la obtención de modelos SAR para clasificar o para enfrentar problemas en otros campos de la Bioinformática (Poli, 2008). En el caso de los Árboles de Regresión las aplicaciones en la Bioinformática son más significativas.

En Cuba, los grupos que investigan sobre Química Medicinal no han investigado mucho sobre estas técnicas. Por otra parte a nivel mundial se utilizan para construir modelos en varias áreas de la ingeniería.

Por esta razón en este trabajo se proponen, para generar los modelos QSAR, los Árboles de Regresión y la Programación Genética. En el caso de los Árboles de Regresión la principal ventaja radica en poder construir un árbol con varios puntos de ruptura y varios modelos lineales. Esta forma de representación en Química Medicinal le permite al investigador observar como puede comportarse el sistema para ciertos valores de una o varias variables a través de un modelo lineal y como se comporta de forma diferente para valores distintos. La idea de tener un árbol de modelos permite obtener mejores aproximaciones que otros métodos como RLM.

En el caso de la PG es una técnica que en principio fue creada para resolver problemas de regresión. Es un algoritmo muy creativo que permite explorar el conjunto de funciones, desde las más simples hasta las más complejas y entre ellas elegir la mejor para ajustar los datos. No necesita que se le especifique la estructura del modelo, ni se necesitan conocimientos de estadística solo se entran tuplas de entrada-salida para construir la ecuación de regresión. Al igual que los Algoritmos Genéticos este paradigma evalúa varias soluciones a la vez y posee una componente aleatoria que le permite buscar en una gran parte del espacio de posibles soluciones.

Por esta razón este trabajo tiene interés en utilizar estas técnicas en estudios QSAR y analizar las posibilidades de ambas de construir modelos de regresión de calidad.

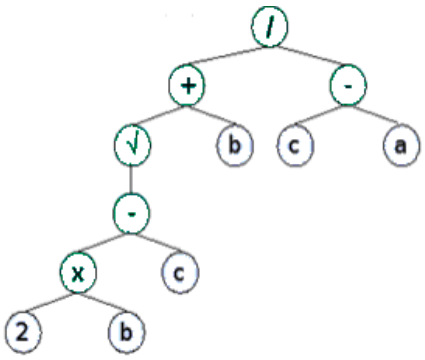
1.6.1. Programación Genética.

La Programación Genética (PG) es un algoritmo evolutivo concebido en un inicio para lograr la evolución automática de programas de computadoras usando las ideas basadas de la selección natural de Darwin.

Esta técnica fue creada por John Koza a finales de los 80's. Koza propuso, por medio de esta extensión, de lo que originalmente es un algoritmo genético, un método para la evolución de estructuras más complejas como pueden ser estructuras de programas de computadora o funciones matemáticas (Koza, 1992).

En la aplicación de la PG a un problema específico, existen cinco pasos preparatorios importantes, que deben ser definidos:

1. El conjunto de símbolos terminales: Los individuos, para su codificación, se representan con estructuras no lineales, como árboles. Por ejemplo para la ecuación siguiente la representación sería la que se muestra en la Ilustración 3:



$$x = \frac{\sqrt{2b - c} + b}{c - a}$$

Ilustración 3. Representación de un individuo en Programación Genética.

Los terminales son las hojas de los árboles que corresponden a variables o a valores constantes, en el caso específico de un estudio QSAR serían los descriptores y las constantes de ajuste que se generan en los modelos de regresión.

2. El conjunto de funciones permitidas: Contiene las funciones que se utilizarán para generar los nodos internos del árbol y pueden ser:

- ✓ operaciones aritméticas (+, -, *, etc.).
- ✓ funciones matemáticas (seno, coseno, log, etc.).
- ✓ operadores boléanos (and, or, not, etc.).
- ✓ operadores condicionales (if-then-else).
- ✓ funciones que causen iteración (do-until).
- ✓ funciones que causen recursión.

Si se buscan ecuaciones de regresión entonces se deben utilizar del conjunto de funciones las operaciones aritméticas y matemáticas que se consideren suficientes para modelar la actividad biológica. A partir de las funciones y los terminales que se definan se generarán árboles sintácticamente correctos que representan ecuaciones matemáticas como la que se puede observar en la Ilustración 3. Cada una de las funciones del conjunto F debe ser capaz de aceptar, como sus argumentos, cualquier valor y tipo de dato que pueda ser retornado por cualquier función del conjunto de

funciones, y cualquier valor y tipo de dato que pueda tomar por cualquier terminal del conjunto T. A partir de las funciones que se definan y los terminales se construyen los individuos de la población, o lo que es lo mismo, un modelo para solucionar total, o parcialmente, un problema.

3. La medida de aptitud: Cada individuo en la población, se mide en términos de qué tan bien se comporta en el ambiente del problema en particular. En muchos casos, la aptitud de un individuo se mide ponderando el error que se produce al utilizarlo para dar una respuesta. En regresión puede calcularse a través del error que se comete al usar la función de regresión para modelar el sistema.

4. Los parámetros para controlar la corrida y el criterio de terminación: Normalmente, los parámetros que controlan la ejecución en PG son el tamaño de la población, el tamaño de los árboles, el número de generaciones y las probabilidades de aplicación de los operadores genéticos de cruce y mutación o cuando algún predicado de éxito específico del problema sea alcanzado (como por ejemplo, encontrar una solución 100% correcta).

5. El método para designar el resultado: Consiste en seleccionar al mejor individuo que haya aparecido en cualquier generación, para lo cual es necesario colocarlo en memoria durante la ejecución. Otro método alternativo es el de designar como resultado el mejor individuo de la población al concluir el algoritmo.

1.6.2. Generación de la población inicial.

La población inicial en PG está compuesta por individuos representados por expresiones. Cada expresión se comienza mediante la generación al azar de un árbol.

Se comienza seleccionando una de las funciones en el conjunto F al azar, para ello puede utilizarse una distribución uniforme de probabilidad. La Ilustración 4 muestra el comienzo de la creación de un árbol aleatorio. Se seleccionó la función suma del conjunto de funciones, la cual lleva dos argumentos.

Ilustración 4. Raíz de un árbol.

Cuando se selecciona un nodo a continuación se generarán n líneas desde ese punto, donde n es la aridad de la función que se escogió. Por cada línea, un elemento del conjunto de funciones y de terminales es seleccionado al azar para ser el punto final de esta. Mientras se seleccionan funciones como rótulo, el proceso de generación sigue recursivamente, pero si se selecciona un terminal ese punto se convierte en una hoja del árbol y el proceso de generación es terminado para ese nodo.

El tamaño de los árboles va a ser menor que un tope permitido, este tope puede ser en profundidad o en el número de nodos máximo por cada árbol, discriminando todos aquellos individuos que lo excedan. Para introducir diversidad entre la población de estructuras se puede generar un porcentaje de árboles de cada tamaño válido o generar estructuras puramente aleatorias que hagan uso de todas las entradas.

Las tres formas de inicialización de la población más utilizadas son el método *full* en la cual se genera una población de árboles de un mismo tamaño escogiéndose siempre nodos del conjunto de funciones hasta tener casi todo el árbol formado y luego completar las hojas con los terminales del problema. El segundo método es el método *grow* donde se genera aleatoriamente el árbol hasta que se cumpla el tamaño determinado por el usuario o hasta que todas las hojas estén cubiertas por terminales del problema. El tercer método fue propuesto por Koza en el año 1992, *half to half*, que propone la generación de la población usando el método *grow* y el método *full* para generar las dos mitades de la población, obteniéndose diversidad de individuos (Baños, 2008).

1.6.3. Operadores genéticos.

Aquí se describirán los operadores genéticos reproducción, cruce y mutación utilizados en la Programación Genética.

Reproducción: Esta operación, la cual es asexual porque actúa sobre un solo individuo a la vez, consiste en dos pasos muy simples. Primero, un único individuo es seleccionado utilizando algún método de selección, luego es copiado sin alteración, desde la población actual hacia la nueva población (nueva generación).

Cruzamiento: El operador de cruzamiento comienza con dos individuos padres y produce dos descendientes. Es una operación sexual, los padres son seleccionados según algún método de selección. Para cada padre se selecciona un punto de cruce y el árbol que se encuentra debajo de este nodo se intercambia con otro subárbol de cruce de otro padre. Dicho subárbol puede consistir en un solo terminal (si el punto de cruce es un terminal) o incluso en el árbol completo que representa a la expresión (si el punto de cruce es la raíz de dicho árbol). El primer descendiente se producirá borrando el fragmento del primer padre e ingresando el fragmento de cruce del segundo padre en el punto del primero. El segundo descendiente se obtiene con igual procedimiento, borrando el fragmento de cruce del segundo padre e ingresando el fragmento de cruce del primer padre en el punto de cruce del segundo (Poli, 2001). En la Ilustración 5 se muestra el procedimiento.

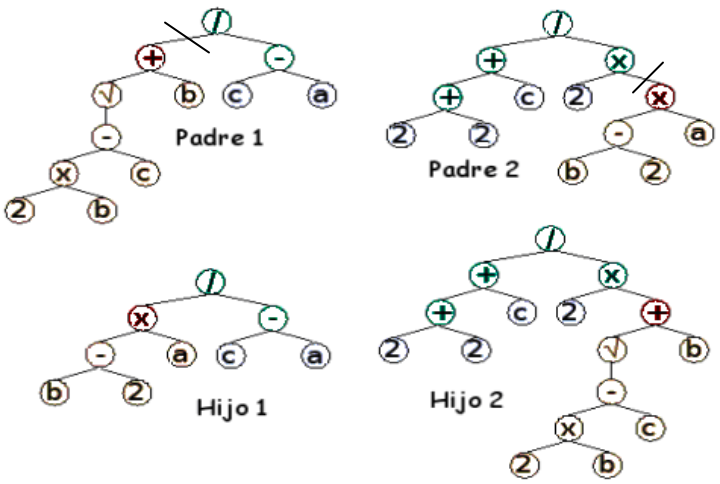


Ilustración 5. Operador de cruce en PG.

Mutación: La mutación es asexual y opera sobre una única expresión, la cual se selecciona en base a una probabilidad proporcional a su aptitud. La operación comienza seleccionando un punto aleatorio dentro del árbol, dicho punto puede ser interno o externo. Se remueve el nodo que se seleccionó y todo aquello que esté por

debajo de este. Luego inserta un árbol creado aleatoriamente en ese punto (Ilustración 6). Esta operación es controlada por un parámetro que especifica el máximo tamaño (medido por profundidad) de cada individuo.

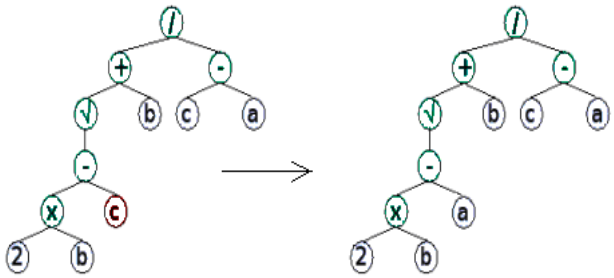


Ilustración 6. Operador de mutación.

Existe otra variante de mutación que es también muy utilizada, conocida como mutación puntual, donde se selecciona a partir de un método de selección un nodo del árbol y seguidamente este es sustituido por un nodo del mismo conjunto del nodo seleccionado.

1.6.4. Métodos de selección.

Existen diversos métodos de selección que pueden ser utilizados durante la selección de progenitores y de sobrevivientes. Algunos de ellos son:

Selección elitista: se garantiza la selección de los k miembros más aptos de cada generación.

Selección proporcional a la aptitud: los individuos más aptos tienen más probabilidad de ser seleccionados aunque no siempre se selecciona el mejor.

Selección por ruleta: una forma de selección proporcional a la aptitud en la que la probabilidad de que un individuo sea seleccionado es proporcional a la diferencia entre su aptitud y la de sus competidores. Tiene semejanza con un juego de ruleta donde el individuo obtiene una sección de la ruleta, pero los más aptos obtienen secciones

mayores que las de los menos aptos. Luego la ruleta se hace girar, y en cada ocasión se elige al que “tenga” la sección en la que se asiente la bola.

1.7. Árboles de regresión.

Un árbol de regresión puede ser interpretado como una ecuación de regresión construida a trozos, ya que las hojas de los árboles contienen valores numéricos o rectas de regresión (Quinlan, 1993).

Entre los algoritmos que generan este tipo de árboles es muy popular el CART (Classification and Regression Trees) (Wang, 1997), en este método el valor de predicción de cada hoja es el valor medio de todos los conjuntos de entrenamiento alcanzados por esa hoja. Basándose en la idea de los algoritmos CART surge una nueva variante conocida como M5, esta última construye árboles cuyas hojas tienen funciones lineales, extendiendo la aplicación de los CART que es un caso particular de árboles donde las funciones que se consideran son constantes.

Ambas estrategias construyen un primer árbol mediante un algoritmo de inducción de árboles de decisión, la diferencia radica en que en los árboles de decisión los puntos de corte se determinan maximizando en todo momento la ganancia de información, mientras que en los de regresión el corte se realiza maximizando la reducción de la varianza.

En el caso del algoritmo M5 se construye un árbol dividiendo las instancias basándose en los valores de los atributos predictivos. Una vez que el árbol ha sido construido, el método genera un modelo lineal para cada nodo. Seguidamente las hojas del árbol comienzan a podarse mientras que el error disminuya, calculándose el error de cada nodo como el valor absoluto de la diferencia entre el valor predicho y el valor actual de cada ejemplo del conjunto de entrenamiento que alcanza dicho nodo. Este error es ponderado con un peso que representa el número de ejemplos que alcanza ese nodo, repitiéndose el proceso hasta que todos los ejemplos son cubiertos por una o más reglas (Breiman, 1984). Finalmente se obtendrán varios puntos de ruptura y una ecuación de regresión para cada ruptura.

1.8. Conclusiones Parciales.

A lo largo del capítulo se habló sobre la importancia de los algoritmos de selección de variables para disminuir la dimensionalidad del problema y luego generar los modelos predictivos de actividad, facilitando el descubrimiento de grupos de variables óptimos para la generación de los modelos, basados en las relaciones entre estas. También se habló sobre el beneficio que puede traer la aplicación de técnicas de optimización o de reconocimiento de patrones en la obtención de modelos QSAR, por sus capacidades para buscar en todo el espacio de posibles soluciones y proponer entre ellas las mejores. Por esta razón en esta tesis se utilizará, para seleccionar el conjunto de descriptores, los Algoritmos Genéticos y el Enfriamiento Simulado por su capacidad de escapar de mínimos y máximos locales en lugares donde pueden existir muchos picos como el caso en estudio donde, por la complejidad del problema, no se ha podido encontrar un patrón de comportamiento estable para todos los casos. Para construir los modelos se utilizarán los Árboles de Regresión y dentro de los paradigmas de Computación Evolutiva se seleccionó un híbrido entre PG y AG, teniendo en cuenta que permiten enfrentar problemas de regresión y por su capacidad de precisión en la obtención de soluciones incluyendo que los modelos son ecuaciones de regresión que pueden interpretarse.

Capítulo 2: Programas y métodos.

2.1. Programas empleados.

Los softwares que se explican a continuación fueron utilizados para la selección de variables y la obtención de los modelos de regresión.

2.1.1. Keel 1.0.

KEEL (Knowledge Extraction based on Evolutionary Learning) es una herramienta software desarrollada dentro del proyecto KEEL para utilizar y construir diferentes modelos de Minería de Datos. Esta herramienta tiene implementadas una gran cantidad de técnicas de Inteligencia Artificial con código abierto en Java. Dentro de las características que hacen de esta aplicación una herramienta importante se pueden destacar (Alcalá, 2000):

Implementación de una librería de algoritmos de extracción del conocimiento, supervisado y no supervisado, entre los que se pueden encontrar Redes Neuronales, Lógica Difusa, Máquinas de Soporte Vectorial y diferentes paradigmas evolutivos. Incluye una librería de programación de algoritmos evolutivos en Java, JCLEC (Java Class Library for Evolution Computation). Dispone de test paramétricos y no paramétricos para el análisis de los resultados obtenidos. De este software se utilizó el algoritmo híbrido de Programación Genética y Algoritmos Genéticos GAP, los Árboles de Regresión M5 y la Regresión Lineal Múltiple.

2.1.2. Weka.

Weka es un software programado en Java que está orientado a la extracción de conocimientos desde bases de datos con grandes cantidades de información. Desarrollado bajo licencia GPL lo que ha hecho de esta aplicación una alternativa muy interesante. Contiene una colección de herramientas de visualización y algoritmos para el análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades.

Una de las propiedades más interesantes de este software, es su facilidad para añadir extensiones y modificar métodos. Su versión original se diseñó para analizar datos de la agricultura pero la versión más reciente basada en Java se utiliza en diversas áreas, con fines docentes y de investigación.

En este trabajo se utilizan los AG implementados en esta aplicación y el algoritmo SimpleK-Means para el agrupamiento.

2.1.3. SPSS(Statistical Package for the Social Sciences).

Este paquete estadístico es muy utilizado en investigaciones que requieran de análisis de datos debido a su facilidad para manejar grandes volúmenes de información. Las versiones más recientes están desarrolladas para manipular 2 000 000 de instancias y un cuarto de millón de variables. Cuenta con varios módulos que incluyen varias funcionalidades al módulo central.

Dentro de estas funcionalidades están la construcción de modelos de regresión, reducción de muestras, clasificación, pruebas paramétricas y no paramétricas, análisis cluster y validación de datos, entre otras.

2.2. Algoritmos de Clustering.

En los dos algoritmos de clustering que se utilizaron se definió como medida de similitud la Distancia Euclidiana que se define como la distancia ordinaria entre dos puntos de un espacio euclídeo y se calcula a partir de la Ecuación 5:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ki} - x_{kj})^2}$$

Ecuación 5. Distancia Euclidiana.

D_{ij} es la distancia que hay entre la observación i y la observación j , X_{ki} es el valor de la variable X_k en el caso i.

Después de que se eligiera la medida de similitud se utilizó el Método de Ward para determinar la cantidad de clusters y el SimpleK-Medias para formar el número de clusters que devuelve el algoritmo anterior. Ambos métodos se explican a continuación.

2.2.1. Simple K-Medias.

Inicialmente se tienen n características en vectores x1, x2,..., xn, donde cada x está representado en un espacio de m dimensiones y agrupados en k conjuntos definidos (k<n). Se define mj como la media del j-ésimo conjunto. Si los conjuntos están bien separados, se puede decir que xi está en el j-ésimo conjunto si la norma || xi – mj || es el mínimo con respecto a los k conjuntos. Esto sugiere el siguiente algoritmo para formar los grupos:

1. El número *k* de *clusters* es fijo.
2. Se proporciona un conjunto inicial de *k* "semillas" (grupos).
    - K primeros elementos.
    - Otras semillas.
3. Dado un cierto umbral, las unidades son asignadas a la semilla más cercana del grupo.
4. Se calculan nuevas semillas.
5. Volver a la etapa 3 hasta que no sea necesaria una reclasificación.

2.2.2. Método de Ward.

Este método es jerárquico y no necesita que se le defina la cantidad de conjuntos a formar. El criterio de agrupación que utiliza se basa en los estudios de Ward que demuestra que la información que se pierde al integrar varios individuos en un mismo *cluster* puede medirse sumando todos los cuadrados de las desviaciones entre cada individuo y la media del *cluster* en el que se asigna este. Por esta razón el algoritmo en cada iteración considera la posibilidad de unión de cada par de grupos y selecciona aquella que menos incrementa la suma de los cuadrados de las desviaciones al unirse.

2.3. Algoritmos de reconocimiento de patrones utilizados.

2.3.1 Árbol de regresión M5.

Los árboles de regresión M5 son construidos por el método divide y vencerás. Inicialmente el atributo A es asociado a un nodo del árbol. Para determinar qué atributo es el mejor para dividir el conjunto de entrenamiento en un nodo determinado se usa la desviación estándar como medida de error.

Primero se calcula el error que se comete al utilizar cada uno de los atributos como punto de corte y el que minimiza el error es elegido como punto de corte en dicho nodo. La reducción del error viene dada por la Ecuación 6:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \cdot sd(T_i)$$

Ecuación 6. Fórmula para calcular el error en M5.

donde Ti son los conjuntos resultantes de dividir el nodo usando el atributo elegido A, T es el conjunto de entrenamiento y sd() es la desviación estándar de un conjunto.

El proceso de cálculo de los puntos de corte termina cuando las predicciones de los ejemplos del conjunto de entrenamiento que alcanzan el nodo varían ligeramente, es decir la desviación estándar de este subconjunto es pequeña con respecto a la desviación estándar del conjunto total o cuando ese nodo es alcanzado por muy pocos ejemplos. Luego se calcula un modelo lineal para cada nodo. Los modelos tienen la forma de la Ecuación 7:

$$a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

Ecuación 7. Estructura de las funciones en M5.

donde ai son valores de los atributos y xi son los pesos o coeficientes que se calculan mediante una regresión. En cada función de regresión sólo intervienen los atributos del subárbol que corresponden al nodo que se está analizando.

Después de construido el árbol inicial se pasa a las siguientes etapas:

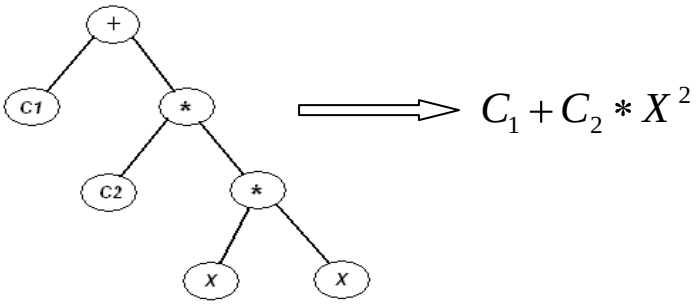
Poda: Una vez obtenido el modelo lineal para un nodo este se simplifica eliminando términos mientras que el error estimado disminuya. Cuando se terminan de simplificar los modelos lineales el árbol se poda desde las hojas. Se continuará podando mientras el error siga disminuyendo.

Suavizado: En esta etapa el algoritmo eliminará las discontinuidades que existen en los modelos lineales podados, sobretodo los modelos que abarcan un conjunto pequeño de ejemplos.

2.3.2 Algoritmos GAP.

El principal problema de la Programación Genética en la resolución de problemas de regresión es el uso de constantes para ajustar las funciones. Para enfrentar este problema en esta tesis se utiliza una técnica evolutiva conocida como GAP, donde se evoluciona la estructura de las funciones matemáticas mediante Programación Genética y las constantes reales o coeficientes de ajuste evolucionan por Algoritmos Genéticos (Sánchez, 2000).

En este híbrido, entre ambas técnicas de computación evolutiva, cada individuo consta de dos partes, un árbol sintácticamente correcto que al ser leído se obtiene la función de ajuste que representa una posible solución del algoritmo y una cadena de coeficientes de ajuste, evolucionando simultáneamente pero de forma independiente (Howard, 1995). Como resultado se obtiene un individuo representado de la siguiente forma de la Ilustración 6.



$$C_1 + C_2 * X^2$$

C1: 3,8521	C2: 1,256

Ilustración 6. Individuo: Árbol que representa la función por PG y cadena de constantes por AG.

En este híbrido los operadores de cruce y mutación para cada tipo de algoritmo son diferentes, pudiéndose definir distintos valores de probabilidad en cada caso por separado. Por una parte evolucionan los árboles mientras que por otra evolucionan los coeficientes de ajuste.

Las poblaciones en GAP convergen hacia subpoblaciones de funciones de estructuras semejantes y distintos valores de constantes. En las primeras corridas del algoritmo se tienen poblaciones con diversidad de estructuras y a medida que avanzan las generaciones se puede observar como van quedando las estructuras funcionales de mejor ajuste al problema.

Al inicio la Programación Genética tiene más peso en el algoritmo pues se buscan aquellas formas funcionales más robustas y luego que la población comienza a converger hacia ellas juegan un papel importante los Algoritmos Genéticos evolucionando los coeficiente y ajustando cada vez más las salidas del problema. Aunque es válido aclarar que ambos algoritmos trabajan paralelamente a lo largo de la ejecución del programa.

2.4. Muestras utilizadas.

2.4.1 Cefalosporinas.

Para la construcción de los modelos de regresión se utilizó una muestra de cefalosporinas, con 104 compuestos reportados en la literatura con actividad biológica

conocida. Las cefalosporinas pertenecen al la familia de los ß-lactámicos con actividad antibacterial reportada.

Según la opinión de varios especialistas esté tipo de compuestos eran prácticamente imposibles de modelar debido a la naturaleza compleja y específica de su mecanismo de acción. A partir de estas ideas el Dr. Ramón en su tesis doctoral (Carrasco, 2003) se plantea demostrar que si es posible construir modelos QSAR sobre estos ß-lactámicos a partir de conocimientos y herramientas estadísticas. Teniendo en cuenta lo planteado acerca de las dificultades intrínsecas para desarrollar modelos cuantitativos estructura-actividad en β-lactámicos, preparó una muestra con diversidad de descriptores que describen características topológicas, topográficas y electrónicas de cada uno de los compuestos. En esta tesis se trabajará con la misma muestra utilizada en la tesis doctoral y se harán algunas comparaciones entre los resultados obtenidos en ambos trabajos con el fin de arribar a conclusiones. En la Tabla 1 se muestran los compuestos con los respectivos valores de actividad biológica.

Tabla 1. Muestra de cefalosporinas.

	Valor Observado		Valor Observado		Valor Observado		Valor Observado
Cefaclor	5.962	hp16e	5.877	ki16o	5.791	SN16f	6.404
Cefdinir	6.596	hp16f	5.524	ki16n	6.092	SN16g	6.406
Cefixime	4.258	hp16g	5.573	ki16o	5.791	SN16h	6.394
Cefotaxima	5.464	hp16i	5.183	kKI22a	5.749	SN16i	6.416
Cefpodoxima	5.437	hp25a	5.214	kKI22b	5.750	SN16ia	6.645
Ceftibuten	3.612	hp25e	5.216	kKI22c	6.695	SN16ib	6.658
CY1a	6.094	hp25f	5.226	kKI23a	6.064	SN16j	6.429
CY1b	5.505	hp25g	5.226	KI23b	6.065	SN16m	5.779
CY1c	5.833	hp25l	5.537	KI23d	5.424	SN16n	5.792
CY1d	6.129	hp25n	4.915	KI23e	5.716	SN16p	6.105
CY1f	5.541	hp25p	4.615	KI23f	5.430	SN29d	6.658
CY1h	5.295	ki16a	4.926	KI23g	5.735	SN29e	6.672
CY1k	6.185	ki16b	5.235	KI23h	6.101	SN29f	6.383
CY2b	5.269	ki16c	5.703	KI23i	6.114	SN29g	6.387
CY2c	5.191	ki16d	5.402	KI23j	6.418	SN29h	6.117
CY2d	5.758	ki16e	4.797	KI23k	5.785	SN29j	6.383
CY2e	5.769	ki16f	6.017	KI24a	6.403	SN29k	6.699
CY2f	6.069	ki16g	5.731	KI24b	6.404	SN29l	6.382
CY2g	6.100	ki16h	6.637	KI25b	6.094	SN29m	6.695
CY2h	5.725	ki16i	6.651	KI25c	6.734	SN29n	6.406
CY3a	4.637	ki16j	5.438	KI26b	6.410	SN29o	6.107
hp16a	5.487	ki16k	6.067	KI26i	6.756	SN29p	6.419
hp16b	5.838	ki16l	6.645	SN16c	6.670	SN29q	5.465
hp16c	5.789	ki16m	5.722	SN16d	6.683	SN29r	5.769
hp16d	4.925	ki16n	6.092	SN16e	6.695	SN29s	5.781
						SN29t	5.797

2.4.2 Contraensayo para inhibidores del Factor 1 del receptor nuclear esteroidogénico (SF-1).

Un ensayo de dosis-respuesta basado en células para la inhibición del receptor huérfano. El SF-1 se expresa en las glándulas adrenal, pituitaria, testículos, y ovarios y regula la producción de la hormona esteroidea a diferentes niveles, incluyendo la expresión directa de la enzima P-450 principal involucrada en la síntesis de la hormona esteroidea. Se predice, en particular, que el diseño apropiado de antagonistas del SF-1 tienen utilidad terapéutica en el tratamiento del cáncer de próstata metastásico a través de la supresión, tanto de la síntesis de testosterona gonadal como de andrógeno adrenal. Otro beneficio potencial de este esfuerzo puede ser la identificación de ligandos del SF-1 que pudieran convertirse en una nueva clase de pequeñas moléculas reguladoras del metabolismo energético y la obesidad.

Este ensayo contiene 315 moléculas entre activas e inactivas. Los descriptores con los que se modelará la actividad de estos compuestos fueron calculados por el módulo de cálculo de descriptores de la plataforma alasGRATO.

Se emplearon para describir las moléculas del ensayo los índices de Rándic (Randic, 1975), Valencia, Partición de la Refractividad Molecular, Conectividad entre Aristas, Momentos Espectrales, Randic Topográfico, Valencia Topográfico, Conectividad entre Aristas Topográfico, Momentos Espectrales Topográfico. Los interesados en una visión más amplia y profunda acerca de la definición y aplicaciones de cada uno de estos índices pueden visitar las siguientes referencias bibliográficas (Balaban, 1976) (Bonchev, 1983). En todos los descriptores se tuvo en cuenta los caminos, clusters y cluster caminos hasta orden 4.

Capítulo 3: Resultados y Discusión.

3.1. Introducción.

Para la generación de los modelos QSAR se plantea la utilización, como ya se explicó en el Capítulo 2, de los Árboles de Regresión y la Programación Genética combinada con Algoritmos Genéticos. El uso de estos potentes algoritmos de regresión podría introducir mejoras en los estudios cuantitativos de relación estructura-actividad. Por esta razón a lo largo de este capítulo se analizarán los resultados obtenidos con el empleo de estas dos técnicas.

3.2. Generación de modelos QSAR en cefalosporinas.

La muestra de cefalosporinas contiene 104 compuestos y 181 descriptores. El paso inicial consistió en la eliminación de todas las variables que presentan varianza cercana a cero quedando finalmente 161 descriptores. Con esta primera reducción el espacio dimensional es aún demasiado grande, pues se estima que la relación óptima debe ser de alrededor de un descriptor por cada 5 instancias para desarrollar un modelo.

Por esta razón, se realiza una selección de variables para disminuir la dimensión del problema. Como no se tiene conocimiento del método que proveerá la mejor combinación de parámetros se utilizan dos algoritmos conocidos de selección de variables y se predice con cada una de las selecciones hechas con vistas a buscar la mejor combinación. Los algoritmos seleccionados para la reducción son Algoritmos Genéticos implementado en Weka y el Enfriamiento Simulado implementado en alasGrato[1].

Para la selección de variables a partir de AG se estableció un tamaño de población de 300 individuos y 100 generaciones, tratando de cubrir un espacio amplio de posibles respuestas. La probabilidad de mutación se fijó en 0,01 tal como recomienda la literatura para evitar que exista mucha diversidad y porque se considera que este

[1] Implementado por Yaikiel Hernández Díaz en su Tesis de Maestría. UCI, 2010. Dirección electrónica yhernandezd@uci.cu.

operador debe ser utilizado con baja probabilidad, preferentemente por debajo de 0,2. La probabilidad de cruce se varió desde 0,6 hasta 0,9 que según la bibliografía consultada es el rango donde se encuentran las mejores probabilidades de este operador (Goldberg, 1989).

Para encontrar los modelos de regresión se dividió la muestra, el 90 % se empleó para el entrenamiento y el 10% restante para la prueba, sin realizar análisis de outliers.

Tabla 2. Selección de variables con algoritmos genéticos.

No.	Pc	# Variables	% red	# Final de Variables	R
1	0,6	161	80,74	31	0,9462
2	0,7	161	78,26	35	0,9421
3	0,8	161	76,39	38	0,9567
4	0,9	161	78,26	35	0,9567

En la Tabla 2 se puede observar el por ciento de reducción obtenido en la selección de variables con Algoritmos Genéticos y en la última columna la calidad de cada selección utilizando una regresión estándar con todas las variables. Se puede observar que las mejores combinaciones se obtienen con probabilidad de cruce de 0,8 y 0,9 para un coeficiente de regresión múltiple estándar de 0,9567.

Los resultados de la selección de variables con Enfriamiento Simulado se resumen en la Tabla 3.

Tabla 3. Selección de variables con Enfriamiento Simulado.

No.	alfa	# Variables.	% red.	# Final de Variables.	R
5	0,6	161	85,09	24	0,9186
6	0,7	161	94,40	9	0,9421
7	0,8	161	84,47	25	0,9462
8	0,9	161	91,30	14	0,9380

Para esta selección de variables se modificó el parámetro alfa desde 0,6 hasta 0,9 y se evaluó la calidad de la selección con las mismas condiciones que en la selección anterior. De forma general se puede observar que las combinaciones obtenidas con Algoritmos Genéticos, a pesar de que el conjunto de variables es significativamente mayor en la mayoría de los casos, la calidad es superior por lo que se decide trabajar con la selección que brinda estos.

En la Tabla 2, se muestran dos combinaciones de variables con igual valor de R lo que significa que en principio se puede trabajar con cualquiera de las dos, sin embargo, se escogió la selección hecha con probabilidad 0,9 por tener un menor número de descriptores.

Para generar los modelos QSAR a partir de la selección 4 (Tabla 2) se utilizaron los Árboles de Regresión (M5) y el híbrido entre Programación Genética y Algoritmos Genéticos (GAP) que se proponen en la tesis, así como la Regresión Lineal Múltiple (RLM) como criterio de comparación.

En el caso del algoritmo GAP la configuración de los parámetros fue la siguiente: Se utilizó una población de 300 individuos, y se ejecutaron 1000 generaciones. Las probabilidades de cruce del AG y LA PG que conforman el híbrido se varió desde 0,6 hasta 0,9 y la de mutación para ambos de 0,01. El tamaño de las cadenas de constantes en Algoritmos Genéticos se determinó que sería de 35 y en el caso de Programación Genética los árboles se definieron con tamaño 35. Las funciones permitidas para construir los modelos de regresión fueron: exponencial, raíz cuadrada, suma, resta, multiplicación y división para no obligar a la técnica a construir funciones solamente lineales sino que brindara la función de mejor ajuste, aunque esta no fuera lineal.

En la Tabla 4 se muestra el resumen de los mejores modelos encontrados por los 3 algoritmos utilizados y los coeficientes de correlación correspondientes a cada uno de ellos. Se puede observar también la cantidad de variables inicial con que se generan las funciones de regresión y la cantidad de variables final que conforman los modelos.

Tabla 4. Datos obtenidos de la regresión con la selección de 35 variables.

Algoritmo	# Final de Variables	Coeficiente R	Error Relativo Medio (%)
RLM	8	0,81	5,5
GAP	4	0,84	5,1
M5	6	0,95	4,9

Analizando los datos de la Tabla 4 se puede observar que los modelos encontrados por M5 son los de mayor calidad. En el caso de la Regresión Lineal Múltiple (RLM) y el algoritmo híbrido, los resultados son semejantes pero teniendo en cuenta que la cantidad de instancias de la muestra es de 104 la proporción de casos por descriptor es mejor en el algoritmo GAP. La baja calidad del algoritmo de regresión con respecto al M5 se puede explicar a partir de que la regresión utiliza el método Forward Stepwise para incluir variables en el modelo y a medida que avanza la selección de las variables a incluir, pueden perderse combinaciones de descriptores con aquellos que fueron desechados en los pasos previos y con ello, este método clásico encuentra dificultades para escoger los descriptores que conformarán la mejor ecuación de regresión. En el caso del M5, que aplica un método de búsqueda exhaustiva, los coeficientes de regresión aumentan, esto manifiesta una buena selección de variables por parte de los algoritmos genéticos y un buen desempeño de M5 con un mayor número de parámetros con respecto a RLM.

Los modelos formados por ambos algoritmos son los que se muestras en las Ecuaciones 8 y 9.

Por RLM:

$$Pot\_SA = 8{,}32 + 0{,}64(OMP01) - 7{,}5(E2ROP1) - 0{,}74(E2ROPC4) - 0{,}33(OQPCL04) + 4{,}09(E2P1) + 0{,}201(ETOT) + 0{,}718(O2QP6) + 3{,}52(E2ROP2)$$

R= 0,81 error relativo medio= 5,5% desviación estándar=0,5

Ecuación 8. Modelo utilizando Regresión Lineal Múltiple.

Por M5:

$O2C5 > 0{,}0695$ entonces *LM1*

$O2C5 \leq 0{,}0695$ entonces *LM2*

LM1: $7{,}23 + 6{,}56(SE\_NETA3)$

LM2: $5{,}31 + 0{,}175(RANP01) - 0{,}536(OQCPCL04) - 0{,}448(O2P2) + 3{,}91(O2QP6)$

R= 0,95 error relativo medio= 4,9% desviación estándar=0,6

Ecuación 9. Modelo utilizando Árboles de Regresión.

La calidad de un modelo de regresión no solo se mide por la precisión con que ajustan los datos sino también por la simplicidad del mismo. Un modelo de pocas variables da por lo general, una mejor idea del comportamiento de cada una de las variables y permite llegar, de forma más sencilla, a la interpretación de los resultados. Según la regla LM1, valores menores o iguales a 0,0695 de la variable O2C5, es decir, a menor ramificación en el anillo cefalosporánico, el peso de la actividad es más dependiente del carácter electrofílico de la posición 3 en dicho anillo y por lo tanto de la naturaleza del grupo saliente en dicha posición. Como la variable O2C5 es de carácter topográfico estrechamente relacionado con el grado de ramificación, debe por lo tanto estar condicionada a un menor tamaño en el grupo sustituyente. Cuando se sobrepasa ese umbral (LM2), es decir, cuando la ramificación en el anillo cefalosporánico es elevada, la responsabilidad de la actividad se reparte en toda la molécula según se infiere de la aparición de variables topológicas vinculadas a la estructura del sustituyente en la posición 7 del anillo cefalosporánico (RANP01 y OQCPCL04) y en menor grado a la ramificación y tamaño de una estructura más polar de dicho anillo (O2P2 y O2QP6 respectivamente).

Al predecir con este modelo en el conjunto de datos de prueba de 10 cefalosporinas se obtienen los valores predichos que se observan en la Tabla 5.

Tabla 5. Valores predichos por la Ecuación 9.

Comp.	Observado	Predicho	Error	Comp.	Observado	Predicho	Error
CEFDIMIR	6,596	6,3737	0,2223	KI25I	6,756	6,2354	*0,5206*
SN29S	5,781	5,4797	0,3013	SN16H	6,394	6,4262	0,0322
SN16IA	6,645	6,5453	0,0997	SN16J	6,429	6,4262	0,0028
SN16N	5,792	5,9151	0,1231	KI16I	6,650	6,5048	0,1452
KI25B	6,094	6,300	0,206	SN29T	5,797	5,5967	0,2003

El mayor error absoluto cometido en los compuestos de prueba fue de 0,52. El resto se comporta por debajo de 0,3 obteniéndose una media de 0,18 de error absoluto cometido al predecir con el modelo. El error relativo de la predicción es del 2,9 %. Esto presupone que el modelo es general y que logra modelar todos los compuestos del ensayo.

En la tesis de doctorado citada anteriormente, se logró modelar la misma muestra de cefalosporinas a partir de 9 descriptores para un coeficiente de correlación R =0,8777 y un error relativo del 4%, para ello se tuvo que realizar un análisis de outliers donde fueron eliminados 5 compuestos. Sin embargo esta vez la muestra se modela completa con solamente seis variables. En el test de validación con el 10% de los compuestos que se utilizan se obtiene un coeficiente de correlación de 0,87 lo que demuestra que el modelo es estable.

Otro de los modelos generados por M5 pero esta vez con la reducción de dimensionalidad por algoritmos genéticos con probabilidad 0,8 que brinda una selección de 38 descriptores en el conjunto inicial, se genera un modelo con nueve variables independientes (Ecuación 10).

$$\begin{aligned} Pot\_SA = {} & 5,84 + 0,21(EP\_P\_01) - 0,81(EPPCL06) - 3,58(SE\_SIG3) + \\ & 4,87(SE\_NETA3) + 2,96(SE\_SIG4) + 0,0096(ETOT) + 0,036(ETTOT) \\ & - 0,22(HOMO) + 2,11(LUMO2) \end{aligned}$$

R=0,90 error relativo medio= 5,9% desviación estándar=0,6

Ecuación 10. Modelo generado por Árboles de Regresión con una muestra de 38 variables.

3.3 Análisis de los algoritmos.

La muestra de cefalosporinas se dividió por validación cruzada de 10x1. Luego se hizo un análisis del comportamiento de los algoritmos en cada una de las muestras por separado. En la Tabla 6 se pueden ver los coeficientes de correlación obtenidos por cada uno de los modelos que se construyeron.

Tabla 6. Coeficiente de regresión de cada modelo en las 10 muestras de entrenamiento.

Algoritmo	1	2	3	4	5	6	7	8	9	10
RLM	0,82	0,83	0,83	0,82	0,84	0,81	0,83	0,84	0,82	0,83
M5	0,94	0,92	0,93	0,91	0,95	0,92	0,93	0,92	0,92	0,91
GAP	0,87	0,87	0,89	0,87	0,84	0,82	0,81	0,86	0,85	0,87

Los mejores coeficientes de regresión se obtienen con M5 en todos los casos. Los valores predichos en cada partición formada se pueden observar en el Anexo 1. En el caso de GAP y RLM los resultados son muy cercanos aunque se puede ver que en algunas muestras los valores de GAP son mayores que en RLM. Para poder establecer una diferencia entre los resultados, se aplicaron diferentes técnicas estadísticas.

Inicialmente se realiza un test de Friedman que asigna en este caso el orden mayor al mejor de los algoritmos y el menor al peor (Tabla 7). Esta asignación se efectúa bajo el criterio de la hipótesis nula, que se forma a partir de suponer que los resultados de los algoritmos son equivalentes y, por tanto, sus rankings son similares.

Tabla 7. Resultados del test de Wilconxon.

Algoritmo	**Ranking**
M5	3,00
GAP	1,85
RLM	1,15
Significación	0.00

De acuerdo al valor que se muestra en la Tabla 7 el mejor ranking corresponde a M5. La prueba de Friedman resulta significativa (0.000) y por lo tanto se acepta que existen diferencias globales entre los tres algoritmos siendo M5 el de mejor resultado según el lugar que ocupa en el ranking.

Para observar cuán significativa es la diferencia se hicieron comparaciones de algoritmos dos a dos. Para ello se utilizó el test de Wilconxon (Tabla 8) donde se demuestra que las comparaciones entre los algoritmos son significativas y resalta el algoritmo M5 como el de mejor comportamiento, confirmándose el resultado del test de Friedman.

Tabla 8. Resultados del test de Wilconxon.

Comparación	R+	R-	p-Value	Hipótesis (α = 0.05)
GAP vs RLM	42,50	2,50	0.018	Rechaza a favor de GAP
M5 vs GAP	55,0	0,00	0.005	Rechaza a favor de M5
M5 vs RLM	55,00	0,00	0.005	Rechaza a favor de M5

Cuando se utilizan los 10 modelos que se forman en cada una de las muestras para predecir sobre la muestra completa se detectan 4 compuestos como *outliers* de ellos 3 (CY1G, CY1L, HP16H) coinciden con los identificados por (Carrasco, 2003) en su investigación (Anexo 1).

3.4 Análisis de Validación Cruzada.

Después del análisis de los algoritmos se evaluó la calidad de los modelos a partir del análisis de validación cruzada. El método que se utilizó elimina una muestra cada vez y calcula un modelo, la restante la utiliza para predecir la actividad biológica. Este proceso se repite hasta que cada una de las muestras ha sido excluida. En este análisis se utilizaron los siguientes descriptores estadísticos:

✓ Suma de los cuadrados de los errores residuales de predicción:

$$PRESS = \sum (y_{CV} - y_{obs})^2$$

✓ Error estándar de predicción de la validación cruzada:

$$SPRESS = \sqrt{PRESS / n}$$

✓ Coeficiente de correlación de la validación cruzada:

$$RPRESS = \sum (y_{CV} - y_{med})^2 / \sum (y_{CV} - y_{med})^2$$

Donde:

y_{CV} : es el valor predicho por validación cruzada.

y_{obs} : es el valor de referencia. y_{med} : el valor medio.

Los resultados de la validación cruzada se muestran en la Tabla 9.

Tabla 9. Resultados de la Validación Cruzada.

Muestras empleadas en la validación.	Error estándar de la validación.	Coeficiente de correlación. (R)
10	0,7	0,84

Se obtiene un coeficiente de correlación de 0,84 (Tabla 9) en la validación cruzada, por encima de 0,6 se considera al sistema como predictivo por lo que los resultados se consideran aceptables.

3.5. Modelos QSAR para Inhibidores del Factor Esteroidogénico-1.

Como se detalló en el Capítulo 2, el ensayo 599 de la base de datos del National Cancer Institute (NCBI) es de interés para la generación de modelos matemáticos que contribuyan a la caracterización teórica de los rasgos estructurales de los compuestos evaluados en el ensayo que permitan modelar y/o explicar la relación existente entre la estructura química y la actividad biológica mostrada por los mismos.

Cuando se trabajó en este ensayo se eliminaron de inicio, todas las variables con valores de varianza cercanos a cero ya que en esos casos no aportan información estructural relevante al estudio. La muestra que se obtuvo se utilizó para la generación y validación de modelos y los coeficientes de correlación resultaron muy bajos. El mejor

modelo presentó un R = 0,59 con un 30% de error relativo medio. Se variaron las probabilidades de cruzamiento del Algoritmo Genético y de la Programación Genética de GAP, pero los resultados continuaron bajos. Lo mismo ocurrió cuando se utilizó M5. Al analizar la dispersión de los datos del ensayo de forma gráfica se pudo evidenciar la dispersión del conjunto de instancias por lo que se utilizaron técnicas de agrupamiento para dividir los compuestos.

Para formar los clusters se utilizó el algoritmo SimpleK-Means implementado en WEKA. Este algoritmo es no jerárquico y se necesita para poder hacer la agrupación, especificar a priori la cantidad de grupos que se desean. Por esta razón, se exploró primeramente el conjunto completo con un procedimiento jerárquico, el Método de Ward implementado en SPSS y el resultado que se obtuvo fue de 4 conjuntos. Esta cantidad se pasó como número de *clusters* al no jerárquico y se formaron las particiones que se observan en el Gráfico 1.

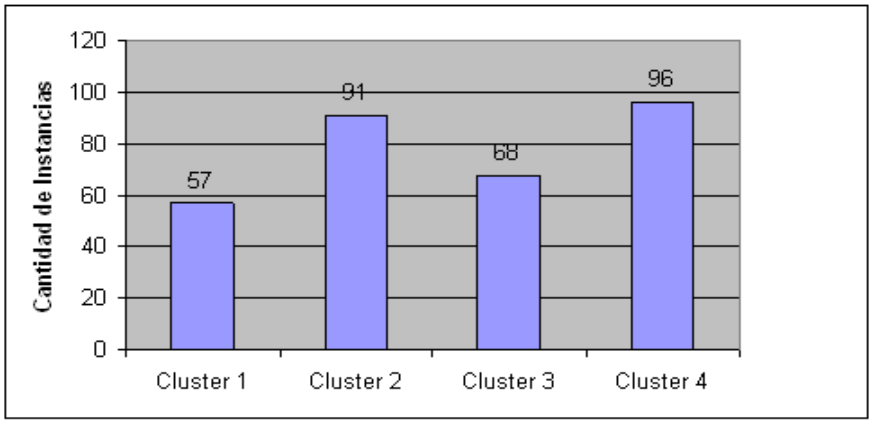


Gráfico 1. Agrupamiento de la muestra 599.

Después de crear los 4 *clusters* se eliminaron los descriptores de varianza cero y se pasó a seleccionar variables con algoritmos genéticos. En la Gráfico 2 puede verse el comportamiento de la reducción en cada uno de los *clusters* variando las probabilidades de cruzamiento.

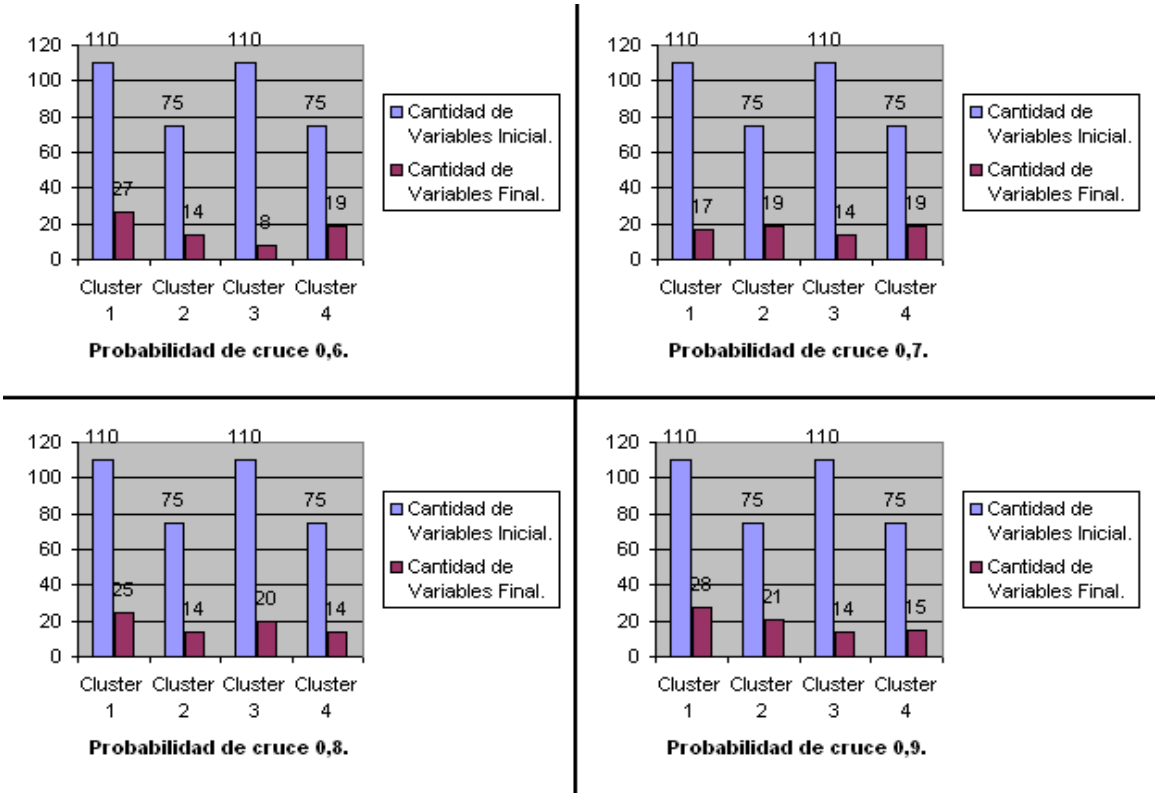


Gráfico 2. Selecciones con diferentes probabilidades de cruce.

Después de reducir la dimensionalidad de las muestras se estimó la calidad de cada una de las selecciones hechas utilizando RLM estándar. Los resultados se muestran en la Tabla 10.

Tabla 10. Calidad de la selección de variables por algoritmos genéticos.

Pc	***Cluster 1***	***Cluster 2***	***Cluster 3***	***Cluster 4***
0,6	0,9918	0,9614	0,9900	0,9421
0,7	0,9918	0,9500	0,9900	0,9972
0,8	0,9921	0,9614	0,9971	0,9972
0,9	0,9921	0,9614	0,9971	0,9972

Se puede observar que las correlaciones son muy altas en todos los casos, y la mejor combinación se obtiene con probabilidad 0,8. Esta selección será la que se utilice para obtener los modelos. Al modelar cada uno de los clusters con Programación Genética (GAP) se tuvo en cuenta nuevamente la necesidad de ajustar la probabilidad de

cruzamiento, y se utilizó la mutación en 0,01. En la Tabla 11 se resumen los datos de los modelos generados para el *Cluster 1*.

Tabla 11. Coeficiente de correlación de los modelos en el *Cluster 1*.

Pc	# Variables en el Modelo	Coeficiente R	Error Relativo Medio %
0,6	5	0,9895	8,2
0,7	5	0,9918	8,6
0,8	5	0,9894	9,0
0,9	4	0,9892	8,9

El modelo de más alta correlación se muestra en la Ecuación 11.

$$LogIC50 = \sqrt{(X23 * \sqrt{2e^{(X10*X20)}} + X15) + X25}$$

r = 0,9918 desviación estándar = 0,5253 error relativo medio= 8%.

Ecuación 11. Modelo que se genera por PG para el *Cluster 1*.

En donde,

X10→ ConectividadAristasCamino_4

X20→ ValenciaTopClusterOrden3_Camino_3

X15→RandicTopClusterOrden3_Camino_2

X23→ ValenciaTopClusterOrden4_Camino_1

X25→ ConectividadAristasTopClusterOrden4_Camino_4

La Ecuación 11 es de naturaleza lo suficientemente compleja como para reducir considerablemente la posibilidad de realizar un análisis fenomenológico o pretender dar una explicación de la influencia de cada una de las variables con respecto a la actividad. La calidad del modelo está dada por sus elevados coeficientes de correlación, bajo valor de la desviación estándar y bajo error relativo en la predicción. Esto indica que el sistema

que se pretende modelar tiene necesariamente un comportamiento no-lineal. Con esto se puede afirmar que el modelo constituye una herramienta útil para el tamizaje de compuestos potenciales como agentes quimioterapéuticos frente al cáncer de próstata y la obesidad. Las correlaciones en el resto de los clusters se pueden ver en los Anexos 2, 3 y 4 respectivamente.

3.6. Consideraciones Importantes.

Se ha mencionado que algunos plantean que la modelación cuantitativa de β-lactámicos era un sueño imposible (Frere, 1989). Sin embargo, se demostró que si podían ser modelados y se reportaron modelos de RLM significativos (Carrasco, 2003). La reevaluación de la muestra permitió generar nuevos modelos QSAR demostrándose a lo largo del Capítulo 3 que el uso de algoritmos de selección de variables y los Árboles de Regresión pueden mejorar dichos modelos, encontrando funciones que relacionan la actividad del compuesto con los descriptores seleccionados, encargándose los algoritmos de selección de encontrar las mejores combinaciones de variables sin la necesidad de análisis de *outliers*.

La utilizacion de la Programación Genética y los Árboles de Regresión, técnica novedosa, demostró la capacidad de las mismas para enfrentar problemas de regresión en muestras estructuralmente diversas. Los errores relativos de los modelos de cefalosporinas, oscilan entre el 4,9 y el 5,5 por ciento para las muestras completas sin análisis de *outliers*. Las correlaciones que se encuentran en la muestra modelada están por encima de R = 0,80 y hasta R = 0,95 en la muestra de entrenamiento y de 0,87 en la muestra de prueba, lo que para este tipo de problema constituye un buen resultado.

Se generaron modelos para SF-1 con coeficientes de regresión de 0,92 y hasta 0,9945 con errores relativos medios entre 6 y 11 por ciento en todos los casos, para los 4 *clusters* analizados.

Se demostró que el uso correcto de técnicas de selección de variables para la reducción de dimensionalidad puede mejorar considerablemente los resultados en la obtención de modelos QSAR, permitiendo encontrar combinaciones de descriptores para generar modelos que mejoran el desempeño del algoritmo de regresión que se utilice.

Se observó que resulta difícil encontrar un único algoritmo para generar los modelos óptimos en toda clase de compuestos. En los experimentos realizados, los resultados obtenidos con Árboles de Regresión resultaron superiores a los de programación genética en uno de los ensayos, mientras para el ensayo 599 se observó que ambas técnicas se comportan de forma similar.

Se evidencia que con el uso de técnicas de optimización o de reconocimiento de patrones se pueden obtener mejores respuestas que con los clásicos algoritmos de regresión para la representación de las relaciones estructura-actividad. Los modelos que se muestran en este capítulo, así lo demuestran.

Conclusiones.

- ✓ Se desarrollaron modelos de regresión de cefalosporinas empleando Árboles de Regresión y Programación Genética con R = 0,95 y error relativo medio del 5% mejores que los reportados en la literatura lo que confirmó que es posible hacer modelos QSAR de ß-lactámicos.

- ✓ Se proponen modelos de regresión de la actividad de inhibidores del Factor 1 del receptor nuclear esteroidogénico (SF-1) con coeficientes de regresión superiores a 0,92 y error relativo medio por debajo del 9%.

Recomendaciones.

- ✓ Utilizar las técnicas propuestas en la tesis en otros estudios QSAR y analizar otras posibles variantes, principalmente en el caso del algoritmo GAP que permitan mejorar la precisión de los modelos.

- ✓ Utilizar en la muestra SF-1 un espectro más amplio de descriptores que permitan encontrar modelos más descriptivos y que puedan aportar mayor conocimiento sobre el comportamiento de los datos.

Bibliografía.

1. **Alcalá, Jesús.** Proyecto KEEL: Desarrollo de una herramienta para el análisis e implementación. *KEEL.* [Online] [Citado em: 3 de septiembre de 2009.] www.keel.es.
2. **Alexandridis, Alex.** *A two-stage evolutionary algorithm for variable selection in the development of RBF Neural Network models.* [ed.] Chemometrics and Intelligents Laboratory Systems. 2005. pp. 114-121.
3. **Anzali, S.** *The use of Self-Organizing Neural Networks in Drug Design.* NL : Academic Press, 1998. 273.
4. **Balaban, A.** *Chemical Applications of Graph Theory.* s.l. : Academic Press Inc, 1976. 978-0120760503.
5. **Baños, Angel.** *Programación Genética, evolución gramatical y programación por expresión genética* [ed.] Escuela de Ingeniería de Sistemas y Computación. s.l. : Universidad del Valle, 2008.
6. **Bonchev, Danail.** *Information theoretic indices for characterization of chemical structures.* Nueva York : Research Studies Press, 1983. 0471900877.
7. **Bort, Juan Manuel Andrés.** *Química teórica y computacional.* España : s.n., 2001. p. 542.
8. **Breiman, L.** *Classification and Regression Tree.* Belmont : s.n., 1984.
9. **Caballero, J.** *Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularization genetic neural networks (BRGNN): application to the prediction of the antagonistic activity.* Talca : MedLine, 2008, Vol. VIII. 1873-4294.
10. **Cambell, Matthew.** *Electronic component placement using Simulated Annealing under thermal constraint.* Departament the engineering design research center. San Francisco : Universidad de Pitsburg, 1995. International Mechanical Engineering Congress and Exposition.
11. **Camps, Joan.** *Algunas mejoras en la evalución de los individuos en Programación Genética para la Regresión Simbólica.* 2004.
12. **Carrasco, Ramón.** *Nuevos descriptores atómicos y moleculares para estudios estructura-actividad. Aplicaciones.* Ciudad de La Habana : Editorial Universitaria, 2003. pp. 35-36.

13. **Carrasco, Ramón .** *Introducción al diseño racional de fármacos.* Ciudad de La Habana : Editorial Universitaria, 2008. p. 15. Vol. 1. 959-16-0647-1.
14. **Chozua, Mariano.** *Técnicas de optimización para el diseño de circuitos analógicos.* Buenos Aires : s.n., 2008. Tesis Maestría.
15. **Fatemi, M.** *A novel quantitative structure-activity relationship model for prediction of biomagnification factor of some organochlorine pollutants.* Babolsar : s.n., agosto de 2009, MedLine. 1573-501X.
16. **Frere, J. M.** *Molecular Pharmacology.* 1989. pp. 125-132.
17. **García, Reynaldo Gil.** Algoritmos de agrupamiento sobre grafos y su paralelización. *www.uji.es.* [Online] 2005 de julio. [Citado em: 2 de octubre de 2009.] http://www.cerpamid.co.cu/index.php?option=com_content&task=view&id=21&Itemid=51.
18. **Goldberg, D.** *Genetic Algorithm in Search.* Massachusetts : Addison-Wesley, 1989.
19. **Goldberg, D.** *Genetic Algorithms in Searching, Optimization and Machine Learning.* Boston : Addison-Wesley, 1989. 0201157675 .
20. **Goodarzi, M.** *New hibrid genetic based Support Vector Regression as QSAR approach for analyz flavonoids-GABA(A) complexes.* La Plata : MedLine, junio de 2009, MedLine, Vol. VI. 1549-9596.
21. **Guerrero, Oscar.** *Desarrollo de diferentes métodos de selección de variables para sistemas multisensoriales.* Tarragona : s.n., 2006. p. 82. Tesis Doctoral.
22. **Hall, L.** *Molecular Connectivity in Chemistry and Drug Research.* Nueva York : s.n., 1976.
23. **Hansh, C.** *Method for correlation of biological activity of chemical.* s.l. : Journal of American Chemical Society, 1964.
24. **Holland, John.** *Adaptation in Natural and Artificial Systems.* [ed.] Universidad de Michigan. s.l. : Ann Arbor, 1975.
25. **Howard, L.** *The GA-P: A Genetic Algorithm and Genetic Programming Hybrid.* s.l. : Universidad de Georgia, 1995.
26. **Hernández, Yaikiel.** *Desarrollo de modelos de clasificación de actividad biológica empleando Máquinas de Soporte Vectorial.* [ed.] Universidad de las Ciencias Informáticas. Ciudad de La Habana : s.n., 2010. Tesis de Maestría.

27. **Koza, John.** *Genetic Programming on the programming of computers by means of natural selection.* Massachusetts : The MIT Press, 1992.
28. **Koza, John.** *Hierarchical Genetic Algorithms operating on populations of computer programs.* [ed.] Morgan Kauffman. 1989.
29. **Troncoso, Alicia .** *Técnicas avanzadas de predicción y optimización aplicadas a sistemas de potencia.* Sevilla : s.n., 2005. Tesis Doctoral.
30. **Nicolotti, O.** *Neuronal nicotinic acetylcholine receptor agonist: pharmacophores, evolutionary QSAR and 3D-QSAR models* Bari : s.n., Italia, MEDLINE, Vol. IV. 1568-0266.
31. **Poli, Ricardo.** *Exact schema theorems for GP one-point crossover and standar crossover operating on linear structures and their application to the study of the evolution of size.* San Francisco, California : s.n., 2001. Genetic Programming, Proceedings of the Genetic and Evolutionary Computation Conference(GECCO).
32. **Quinlan, J.** *Induction of Decision Trees.* Singapore : World Scientific, 1993. Proceedings of the 5th Australian Joint Conference on Artificial Intelligent. Vol. I, pp. 81-106.
33. **Quinlan, J. 2006.** *Learning with continuous classes.* Sydney : Universidad de Sydney, 2006. www.keel.es.
34. **Randic, M.** *Indice de Randic.* s.l. : Journal American Society of Chemical , 1975.
35. **Ruiz, Roberto.** *Comparación entre métodos de agrupamiento en la selección de características.* [ed.] Universidad de Oriente. [prod.] Departamento de lenguajes y sistemas informáticos. Santiago de Cuba, Cuba : s.n., 2006.
36. **Sánchez, Luciano.** *Fuzzy random variables-based modeling with GA-P algorithms.* [ed.] Proyecto KEEL. Oviedo : s.n., 2000.
37. **Tonghua, Li.** *Combining PLS with GA-PLS for QSAR.* [ed.] Chemometrics and Intelligent Laboratory Systems. 2002. pp. 55-64.
38. **Osuna, Ricardo.** *Retrieval in Sensor Networks.* Las Vegas : IEEE Sensors Journal, 13 de enero de 2009, Vol. 3, pp. 235-246.
39. **Wang, Y.** *Induction of model trees for predicting continuos classes.* 1997. European Conference on Machine Learning.
40. **Worachartcheewan, A.** *Modeling the activity of furing inhibitors using artificial neural network.* s.l. : MEDLINE, 2009. 1664-73.

41. **Xu, Lu.** *Comparison of differents methods for variables selection.* 2001. pp. 447-483.
42. **Zahavi, Jacob.** *Using Simulated Annealing to optimize the feature selection problem in marketing applications.* 3 de junio de 2006, European Journal of Operational Research, Vol. 171, pp. 842-858.

Anexos.

Anexo 1. Valores predichos para cefalosporinas con 10 muestras diferentes con M5.

Comp.	V. Obs	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6	Mod 7	Mod 8	Mod 9	Mod 10	V. Pred	ERM
cefaclor	5,962	4,888	5,3058	4,9235	4,9704	4,8654	5,1069	5,1132	4,9524	5,0382	4,9734	5,014	0,159
CEFDIMIR	6,596	6,3737	5,4645	6,6001	5,5869	6,6096	5,6645	5,9754	5,5222	5,5775	5,8165	5,919	0,103
CEFIXIME	4,258	4,9756	4,9336	4,9337	5,0918	4,8247	5,2985	5,2581	5,0637	4,9947	5,0664	5,044	0,185
CEFOTAX	5,464	5,576	5,6353	5,6483	5,8089	5,533	5,8275	5,5807	5,3928	5,7371	5,7734	5,651	0,034
CEFPOD	5,437	5,3299	5,6729	5,3032	5,6108	5,2979	5,557	5,349	5,9786	5,5853	5,6104	5,530	0,017
CEFTIBUTE	3,612	4,6611	4,3343	4,5992	4,6289	4,5813	4,381	4,7269	4,3467	4,5622	4,7384	4,556	0,261
CY1A	6,094	5,8624	5,7333	6,098	6,096	6,098	5,9957	6,2718	6,1889	5,9734	6,2082	6,053	0,007
CY1B	5,505	5,7116	5,7694	5,8438	6,1199	5,6997	5,8703	5,6338	6,0638	5,895	6,0021	5,861	0,065
CY1C	5,833	5,5345	5,8071	5,6468	5,8274	5,5146	5,1193	5,2696	6,101	5,449	5,5784	5,585	0,042
CY1D	6,129	5,8624	5,7811	6,098	5,9865	6,098	5,4694	6,1226	6,2973	5,6552	5,7369	5,911	0,036
CY1E	5,864	6,26	4,957	6,9	6,751	6,801	4,23	7,304	7,8	3,75	7,72	6,247	0,065
CY1F	5,541	5,3868	5,3102	5,4883	5,6321	5,359	5,663	5,5788	5,78	5,4885	5,51	5,520	0,004
CY1G	4,358	6,26	5,01	7,12	6,821	7,103	4,02	7,6	4,14	3,57	4,93	5,657	*0,298*
CY1H	5,295	5,6117	5,1402	5,6807	5,513	5,6072	5,6582	5,312	5,1641	5,4932	6,0987	5,528	0,044
CY1J	4,957	6,26	4,55	7	5,921	6,9	4,58	4,81	4,81	3,89	4,95	5,367	0,083
CY1K	6,185	5,5349	6,2769	5,6165	5,5665	5,5221	5,5066	5,7126	6,3982	5,7618	6,356	5,825	0,058
CY1L	3,660	6,83	3,83	7,21	4,856	7,102	4,802	7,401	4,8	4,4	5,08	5,631	*0,539*
CY2B	5,269	5,8034	5,7643	5,2902	5,9189	5,23	5,8206	6,171	5,6903	5,6104	6,0737	5,737	0,089
CY2C	5,191	5,7903	5,7712	6,098	5,8296	5,23	5,2572	5,7156	5,7522	5,5296	6,0025	5,698	0,098
CY2D	5,758	5,7772	5,7781	5,796	5,951	5,7415	5,682	6,0467	5,9281	5,7841	6,0955	5,858	0,017
CY2E	5,769	5,7772	5,7781	5,7846	5,951	5,5101	5,6935	6,0541	5,8121	5,7276	6,0955	5,818	0,009
CY2F	6,069	5,7641	5,785	6,098	5,951	6,098	5,6496	6,0255	5,7459	5,6224	6,0955	5,884	0,031
CY2G	6,100	5,7641	5,785	6,098	6,1187	6,098	5,9966	6,3529	5,9759	5,8491	6,2284	6,027	0,012
CY2H	5,725	5,7706	5,7816	5,6502	5,5869	5,7415	5,3296	5,6684	5,6591	5,6297	5,8165	5,663	0,011
CY3A	4,637	5,3573	5,2024	5,4575	5,6009	5,3277	5,6298	5,5545	5,7444	5,3906	5,4581	5,472	0,180

HP16A	5,487	5,4642	5,6353	5,6207	5,6794	5,8573	5,6629	5,4646	5,6502	6,0422	5,6721	5,675	0,034
HP16B	5,838	5,5393	5,4524	5,5286	5,6507	5,9783	5,9698	5,6408	5,8568	6,1124	5,6604	5,739	0,017
HP16C	5,789	5,4797	5,4524	5,6076	5,7093	5,6173	5,7001	5,489	5,738	5,8306	5,6944	5,632	0,027
HP16D	4,925	5,2685	5,6353	5,5306	5,2155	5,5241	5,4988	5,3609	5,437	5,7286	5,3316	5,453	0,107
HP16E	5,877	5,6032	5,6353	5,7213	5,6225	5,9147	6,0777	5,7686	5,8518	6,0282	5,648	5,787	0,015
HP16F	5,524	5,595	5,449	5,5729	5,7578	6,0669	6,0073	5,6645	5,9107	6,1907	5,7402	5,796	0,049
HP16G	5,573	5,7156	5,6353	5,6925	5,7986	6,0124	6,1632	5,778	5,7309	6,1055	5,7813	5,841	0,048
HP16H	4,267	7	6,48	6,61	5,987	7,702	5,45	6,801	4,808	6	6,16	6,300	*0,476*
HP16I	5,183	5,4507	5,6353	5,5558	5,6164	5,4535	5,6968	5,4754	5,7249	5,6776	5,6273	5,591	0,079
HP25A	5,214	5,3403	5,3766	5,2603	5,3339	5,1339	5,1085	4,9158	5,2896	5,3743	5,4209	5,255	0,008
HP25E	5,216	5,3392	5,3697	5,2593	5,3319	5,0868	5,1071	5,2624	5,2935	5,3329	5,4194	5,280	0,012
HP25F	5,226	5,2209	5,3766	5,0779	5,0529	5,3943	5,029	5,3458	5,0756	5,5792	5,2145	5,237	0,002
HP25G	5,226	5,3962	5,3731	5,3373	5,3946	5,3791	5,2353	5,3595	5,4432	5,5936	5,4687	5,398	0,033
HP25L	5,537	5,3408	5,3731	5,2613	5,335	5,092	5,1072	4,9156	5,0749	5,3374	5,4217	5,226	0,056
HP25N	4,915	5,3392	5,3697	5,2593	5,3319	5,0868	5,1071	5,2624	4,9235	5,3329	5,4194	5,243	0,067
HP25P	4,615	5,3962	5,3731	5,3373	5,3946	5,3943	5,2353	5,3595	5,1806	5,5936	5,4687	5,373	0,164
KI16A	4,926	4,9932	5,6601	4,9283	4,8355	4,9375	4,9812	5,2531	5,349	5,3107	5,7943	5,204	0,056
KI16B	5,235	5,0194	5,1526	4,9859	5,1207	4,9583	4,6338	4,841	5,2134	5,019	5,0121	4,996	0,046
KI16C	5,703	4,8155	4,5856	4,7693	4,8699	4,743	5,1044	5,095	4,7228	4,9396	4,8964	4,854	0,149
KI16D	5,402	5,385	5,7314	5,3821	5,6364	5,3426	5,8208	5,4655	5,7614	5,7906	6,0066	5,632	0,043
KI16E	4,797	5,6476	5,661	5,6427	5,8026	5,6242	5,8811	5,6472	5,6388	5,8252	5,7357	5,711	0,191
KI16F	6,017	5,813	5,661	5,8031	5,8753	5,8025	5,7938	5,7192	5,6448	5,772	5,5167	5,740	0,046
KI16G	5,731	5,5735	5,9491	5,5493	5,5869	5,5493	6,3801	6,0309	6,3063	6,6169	5,8501	5,939	0,036
KI16H	6,637	5,8399	5,6997	5,8187	5,7984	5,8338	6,2165	6,0523	5,9049	6,1538	5,4915	5,881	0,114
KI16I	6,651	5,8407	5,85	5,8114	5,73	5,8366	6,0522	6,0204	6,0645	6,1287	5,4394	5,877	0,116
KI16J	5,438	6,2579	6,0362	6,2293	6,0276	6,283	6,1744	6,2947	5,8458	6,1262	6,0067	6,128	0,127
KI16K	6,067	6,2238	6,3429	6,1863	5,9276	6,2486	5,9691	6,2285	6,0874	6,0741	6,1422	6,143	0,012
KI16L	6,645	6,5453	6,3454	6,5066	6,1417	6,593	6,0842	6,4649	6,0868	6,1412	5,9384	6,285	0,054

KI16M	5,722	5,8292	5,6855	5,8052	5,767	5,823	5,8806	5,8795	5,5464	5,8438	5,726	5,779	0,010
KI16N	6,092	6,0145	5,8834	5,979	5,7989	6,024	5,6529	5,9303	5,8599	5,7434	5,6088	5,850	0,040
KI16O	5,791	5,9151	5,8759	5,8852	5,7767	5,9163	5,7895	5,9155	6,074	5,8258	5,756	5,873	0,014
KI22A	5,749	5,7551	5,692	5,7414	6,135	5,6749	6,2981	5,854	5,9942	6,0381	6,3898	5,957	0,036
KI22B	5,750	5,7648	5,692	5,7049	6,1535	5,6076	6,3523	5,889	5,9701	5,9789	6,4036	5,952	0,035
KI22C	6,695	5,6051	6,0415	5,835	5,7372	6,0621	6,0978	5,753	6,1315	6,3582	6,0997	5,972	0,108
KI23A	6,064	6,0177	6,0426	6,002	6,3012	5,9563	6,3583	6,0357	6,0092	6,0725	6,1189	6,091	0,005
KI23B	6,065	6,0274	6,0426	5,9655	6,3197	5,8891	6,4126	6,0707	5,9816	6,0135	6,1327	6,086	0,003
KI23D	5,424	5,7201	6,3811	5,7969	5,9031	5,7465	5,8837	5,6654	5,9923	5,9238	5,8127	5,883	0,085
KI23E	5,716	5,8076	6,009	5,8071	6,0245	5,7064	6,0753	5,8103	5,9532	5,8808	5,9057	5,898	0,032
KI23F	5,430	5,8952	6,0193	5,9315	6,1459	5,8246	6,1446	5,8754	5,8514	5,9747	5,9988	5,966	0,099
KI23G	5,735	5,8952	6,0426	5,9107	6,1459	5,964	6,1669	5,89	6,0381	6,0944	5,9988	6,015	0,049
KI23H	6,101	6,1017	6,0426	6,2263	6,3536	6,1455	6,3719	6,0728	6,2665	6,2182	6,164	6,196	0,016
KI23I	6,114	6,0537	6,0426	6,1354	6,2226	6,0297	6,3631	6,0537	6,123	6,1072	6,0687	6,120	0,001
KI23J	6,418	5,866	6,0426	6,1172	5,8631	5,9815	6,1133	5,896	5,989	6,065	5,8009	5,973	0,069
KI23K	5,785	6,14	6,0426	6,0806	6,437	6,0823	6,5131	6,1295	6,4204	6,1643	6,2256	6,224	0,076
KI24A	6,403	6,594	6,5287	6,5457	6,4262	6,5807	6,4463	6,6169	6,5255	6,3214	6,5254	6,511	0,017
KI24B	6,404	6,6036	6,4781	6,5091	6,4447	6,5135	6,5006	6,6519	6,3887	6,2624	6,5393	6,489	0,013
KI25B	6,094	6,3943	6,4781	6,3018	6,316	6,2889	6,1844	6,3538	6,0757	5,9317	6,0058	6,233	0,023
KI25C	6,734	6,2346	6,8483	6,432	5,8996	6,744	5,9299	6,2178	6,1893	6,3114	5,702	6,251	0,072
KI26B	6,410	6,209	5,9956	6,128	6,2841	6,088	6,412	6,303	6,059	6,0321	6,123	6,163	0,039
KI26I	6,756	6,2354	6,6295	6,2982	6,1869	6,2282	6,3622	6,2858	6,3034	6,1255	6,059	6,271	0,072
SN16C	6,670	6,4327	6,4921	6,583	6,1248	6,5918	6,1367	6,3731	6,1788	5,7996	6,2295	6,294	0,056
SN16D	6,683	6,4262	6,4806	6,5699	6,2215	6,5781	6,2164	6,4443	6,0852	6,0757	6,3043	6,340	0,051
SN16E	6,695	6,4262	6,4999	6,5574	6,3532	6,5651	6,2583	6,4888	6,1101	6,0981	6,4049	6,376	0,048
SN16F	6,404	6,4262	6,4131	6,5546	6,4745	6,5621	6,4237	6,6067	6,5622	6,1062	6,498	6,463	0,009
SN16G	6,406	6,4327	6,5124	6,5072	6,3263	6,5126	6,2423	6,4878	5,9561	6,028	6,3868	6,339	0,010
SN16H	6,394	6,4262	6,4818	6,5201	6,1307	6,5261	6,2154	6,4534	6,3766	6,5625	6,2386	6,393	0,000

SN16I	6,416	6,4262	6,5227	6,3913	6,351	6,3916	6,2032	6,4798	6,6509	6,6406	6,4088	6,447	0,005
SN16IA	6,645	6,5179	6,4781	6,5862	5,8821	6,5952	6,0018	6,2358	6,2608	5,7153	6,0434	6,232	0,062
SN16IB	6,658	6,5048	6,4716	6,5304	5,8869	6,5369	6,0847	6,3166	5,9214	6,0753	6,0518	6,238	0,063
SN16J	6,429	6,4262	6,4366	6,3643	6,4714	6,3634	6,3402	6,5913	6,2751	6,7034	6,5011	6,447	0,003
SN16M	5,779	5,7248	5,6353	5,7291	6,1488	5,4616	6,0091	5,7337	5,5849	5,6835	6,0234	5,773	0,001
SN16N	5,792	5,8013	5,6353	5,8202	6,2454	5,7257	6,0878	5,8044	5,7542	5,9049	6,0982	5,888	0,017
SN16P	6,105	5,7718	5,6353	5,7774	6,1536	5,9093	6,0848	5,8126	5,8956	6,0588	6,0318	5,913	0,031
SN29D	6,658	6,5048	6,4674	6,5752	5,9787	6,5837	6,0831	6,2004	6,261	5,8127	6,1182	6,259	0,060
SN29E	6,672	6,4983	6,4852	6,5615	6,1104	6,5694	6,1253	6,3661	5,7672	5,7842	6,2189	6,249	0,063
SN29F	6,383	6,5114	6,4955	6,5123	6,0845	6,518	6,1127	6,3967	6,2391	6,0418	6,2015	6,311	0,011
SN29G	6,387	6,5114	6,3923	6,5489	6,2318	6,5562	6,286	6,4462	6,2923	6,2274	6,3119	6,380	0,001
SN29H	6,117	6,5245	6,2882	6,3986	6,3167	6,3992	6,3239	6,6387	4,9785	6,3107	6,3786	6,256	0,023
SN29J	6,383	6,4983	6,4814	6,4899	5,9372	6,4946	6,1105	6,2827	5,7066	6,5461	6,0921	6,264	0,019
SN29K	6,699	6,5048	6,4017	6,451	6,0895	6,4539	6,2273	6,4766	6,2345	6,5402	6,2081	6,359	0,051
SN29L	6,382	6,5048	6,5031	6,4455	5,9879	6,4482	6,0559	6,3159	6,3775	6,6365	6,1305	6,341	0,006
SN29M	6,695	6,5048	6,5085	6,3991	6,1082	6,3998	6,0738	6,3457	6,2609	6,5839	6,2228	6,341	0,053
SN29N	6,406	6,4983	6,5181	6,3739	6,2276	6,3734	6,1115	6,3911	5,996	6,2106	6,3142	6,301	0,016
SN29O	6,107	6,5048	6,4224	6,3721	6,2286	6,3716	6,2109	6,4572	6,3949	6,643	6,315	6,392	0,047
SN29P	6,419	6,5048	6,4498	6,2838	6,0665	6,2793	6,1413	6,4368	6,4527	6,2606	6,1964	6,307	0,017
SN29Q	5,465	5,5497	5,6353	5,4783	5,906	5,3858	5,8764	5,3643	5,7748	5,6375	5,8374	5,645	0,033
SN29R	5,769	5,6262	5,6353	5,5694	6,0026	5,7415	5,9577	5,6804	5,56	5,7356	5,9122	5,742	0,005
SN29S	5,781	5,5967	5,6353	5,5279	5,9109	5,8282	5,9596	5,7198	5,6021	6,0081	5,8458	5,763	0,003
SN29T	5,797	5,5967	5,6353	5,5282	5,9109	5,8102	5,9593	5,6786	6,0081	5,9922	5,8458	5,797	0,000
Error Relativo Medio (%)												6,10	

Anexo 2. Resultados para el *Cluster 2* con GAP.

Pc	# Variables en el Modelo	Coeficiente R	Error Relativo Medio %
0,6	3	0,9614	7,2
0,7	5	0,9247	6,3
0,8	4	0,9302	6,2
0,9	6	0,9770	8,4

Anexo 3. Resultados para el *Cluster 3* con GAP.

Pc	# Variables en el Modelo	Coeficiente R	Error Relativo Medio %
0,6	5	0,9914	7,4
0,7	4	0,9926	8,2
0,8	4	0,9943	8,1
0,9	4	0,9900	8,1

Anexo 4. Resultados para el *Cluster 4* con GAP.

Pc	# Variables en el Modelo	Coeficiente R	Error Relativo Medio %
0,6	5	0,9427	8,3
0,7	5	0,9326	9,3
0,8	6	0,9900	9,4
0,9	5	0,9425	11,0