Universidad de las Ciencias Informáticas
Facultad 6 Bioinformática

Lenguaje Descriptor de Estructuras Químicas

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Autores: Raida Elinda Pérez Durán

Suanly Ortíz Tornín

Tutores: Dr. Ramón Carrasco Velar

Lic. Dannier Trinchet Almaguer

Ciudad de La Habana, Cuba

Julio 2007

Si buscas resultados distintos, no hagas siempre lo mismo.

Albert Einstein

DECLARACIÓN DE AUTORÍA

Nosotras, Raida Elinda Pérez Durán y Suanly Ortíz Tornín declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los _____ días del mes de _____ del _____.

_____	_____
Raida Elinda Pérez Durán (Autor)	Dr.Ramón Carrasco Velar (Tutor)
_____	_____
Suanly Ortíz Tornín (Autor)	Lic. Dannier Trinchet Almaguer (Tutor)

DATOS DE CONTACTO

Tutores:

Dr. Ramón Carrasco Velar.
Centro de Química Farmacéutica, Cuidad de La Habana, Cuba.
Email: rcarrasca@cqf.sld.cu

Lic. Dannier Trinchet Almaguer.
Universidad de las Ciencias Informáticas, Cuidad de La Habana, Cuba.
Email: trinchet@uci.cu

AGRADECIMIENTOS

Agradecemos a nuestros tutores Trinchet y Carrasco por toda la dedicación, apoyo, paciencia y esmero brindado.

A todas las personas que con su ayuda directa e indirecta han contribuido con la realización de este trabajo. En especial a nuestros padres y familiares que nos han ayudado y apoyado durante nuestra vida de estudiante, por la confianza depositada en nosotros y por su inseparable compañía.

A todos los profesores que pusieron su granito de arena para lograr nuestra formación profesional.

A Pla y a Jacque por dedicarnos parte de su valioso tiempo cuando los necesitamos y por su entera disposición para brindar su ayuda.

A nuestros amigos que nos han acompañado durante estos maravillosos 5 años, con los que hemos compartido nuestros logros y alegrías, pero también los momentos difíciles.

Un agradecimiento especial a nuestro Comandante por darnos la oportunidad de formarnos como ingenieros en una universidad de nueva creación, surgida al calor de la batalla de ideas, que ha marcado nuestro pensamiento revolucionario y comportamiento como verdaderos profesionales.

DEDICATORIA

A mis padres por depositar toda su confianza en mí, por acompañarme y apoyarme en todos los momentos de mi vida. *A mi hermana Thais por ser tan especial, y estar siempre a mi lado.* *A mi abuelo por ser sencillamente la persona más maravillosa que he conocido.* *A Yenisel por ser más que mi amiga, mi hermana de siempre.* *A Lola por demostrarme en tan poco tiempo que tengo una amiga más en quien confiar.* *A mis amigos por su sincera y sencilla amistad, por estar siempre cuando los necesito.* *Suanly Ortíz Tornín*	*A mis Tutos, por ser simplemente súper especiales, a ellos... mi vida entera.* *A mi amiga, o más que mí amiga mi hermana, que aunque lejos, siempre ha estado muy cerca: Denek.* *A todos mis amigos, especialmente a esos que siempre estuvieron ahí, cuando más lo necesité, cuando más los necesité.* *A ti...* *Raida Elinda Pérez Durán*

RESUMEN

El presente trabajo surge como parte del proyecto: ***"Plataforma Inteligente para la Predicción de Actividad Biológica de Compuestos Orgánicos"*** llevado a cabo entre la Facultad de Bioinformática de la Universidad de las Ciencias Informáticas y el Centro de Química Farmacéutica. El objetivo del mismo consiste en el desarrollo e implementación computacional de un lenguaje descriptor de estructuras químicas que permita realizar la descripción de compuestos orgánicos teniendo en cuenta, no solamente los aspectos topológicos de la estructura química sino una propiedad químico-física, la refractividad molecular, particionada sobre los átomos de la molécula que es muy utilizada en estudios de correlación entre la estructura química y la actividad biológica. Esto constituye una nueva forma de representación de la estructura molecular que responde a las necesidades particulares del proyecto. Se definieron ciertas agrupaciones atómicas como centros descriptores de la molécula y se determinaron los caminos de unión entre los mismos. Se presenta el conjunto de reglas gramaticales del lenguaje, el cual permitirá representar la molécula como un conjunto limitado de fragmentos notables asociables a la propiedad biológica.

ABSTRACT

This work arises as part of "*Plataforma Inteligente para la Predicción de Actividad Biológica de Compuestos Orgánicos" Project, taken to end between The University of Informatics Sciences and the* Centre of Pharmaceutical Chemistry. The objective of this consist in the development and computational implementation of a language describer of chemical structures that allows to realize a description of organic compounds having in bill, not only topological aspects of the chemical structure, but a property chemist – physics, the molecular refractivity, divided on the atoms of the molecule, that it is very used in studies of correlation between the chemical structure and the biological activity. This constitutes a new form of representation of the molecular structure that answers to the particular needs of the project. Certain atomic groups were defined as centers describers of the molecule and the ways of union decided between the same ones. There appears the set of grammatical rules of the language, which will allow representing the molecule as a limited set of notable fragments associable to the biological property.

ÍNDICE

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Desde hace años se han llevado a cabo investigaciones relacionadas con las características y propiedades biológicas de fármacos y tóxicos y sus interacciones con los organismos vivos. Este tema ha sido motivo de interés para bioquímicos, biólogos, químicos e investigadores de áreas afines, los cuales han llevado a cabo estudios al respecto con objetivos muy específicos. Estas investigaciones han inducido una serie de interrogantes como por ejemplo: ¿Cómo establecer correspondencia entre la estructura química de una sustancia y la respuesta que brinda en el medio biológico? ¿Cómo describir la estructura química de la manera más eficiente para lograr establecer estas correlaciones? ¿Existe algún modelo que permita realizar esta descripción de una forma eficiente? Para dar respuesta a estas interrogantes, se han desarrollado modelos matemáticos y diseñado lenguajes descriptores de la estructura química que permiten reflejar las características de las estructuras moleculares para situaciones específicas.

Los químicos de cualquier latitud se comunican entre sí mediante el lenguaje gráfico. Sin embargo, esta forma de comunicación, aunque práctica y eficiente, resulta inapropiada cuando se necesita manipular grandes volúmenes de información. Desde mucho antes del desarrollo de las ciencias de la computación y de la propia informática, los químicos podían representar la estructura de una sustancia mediante lo que se conoce como la fórmula empírica, según ésta, es posible conocer la composición química de una sustancia dada, aunque esto no resulta del todo suficiente pues existen múltiples compuestos con similar composición pero diferente forma de enlazarse los átomos entre sí. Eso condujo al desarrollo de otras formas de representar la estructura de una sustancia de manera más amplia. Ejemplos clásicos son los lenguajes SMILES (*Simplified Molecular Input Line Entry Specification*), [1] SMARTS (*SMiles ARbitrary Target Specification*), [2] e InChI (International Chemical Identifier) [3]. Estos lenguajes son muy útiles para el trabajo de búsqueda de fragmentos en grandes bases de datos. Sin embargo, todos poseen como limitante el hecho de que no contienen otra información estructural que la topológica. Si se desea incorporar información adicional a este tipo de descripción, no es posible hacerlo a partir de la sintaxis de dichos lenguajes. Un avance en este sentido se logró con la introducción del lenguaje DCAM (*Descriptor Centre Adjacency Matrix*) el cual es una versión ampliada del lenguaje SSFN (*Substructure Superposition Fragmental Notation*) que fue diseñado para el sistema experto OREX basado en la descomposición topológica de las moléculas en fragmentos estructurales y su asociación a las actividades biológicas. Este

sistema experto emplea la teoría lógico-combinatoria para el establecimiento de las reglas de inferencia [4].

Por otra parte, son múltiples los reportes acerca del empleo de las técnicas de teoría de grafos para describir moléculas. Una de las más populares es el desarrollo de los índices topológicos y topográficos. Todos ellos parten del cálculo de las matrices de adyacencia entre vértices y/o aristas o las matrices de distancias topológicas [5]. Más recientemente, se han introducido dos índices híbridos, llamados así por que en su definición se ha ponderado la matriz de conectividad entre los vértices del grafo químico por una propiedad químico-física particionada a nivel atómico como la refractividad molecular.

Los químicos han recurrido frecuentemente a la fragmentación de las estructuras químicas como procedimiento para determinar cuál o cuáles de sus partes son las responsables de determinada propiedad que ellas manifiesten. Esta concepción metodológica ha sido adoptada por especialistas en técnicas de inteligencia artificial para el desarrollo de diferentes sistemas computacionales de predicción de actividad biológica. [6]

Basado en estos mismos conceptos generales, dentro del proyecto que da origen a este trabajo (*Plataforma Inteligente para la Predicción de Actividad Biológica de Compuestos Orgánicos*) se planteó la necesidad de establecer correlaciones entre la presencia de fragmentos ponderados por el Índice del Estado Refractotopológico Total y la actividad biológica [7]. En esta primera aproximación se definieron como fragmentos los caminos de orden 1 hasta 10, clusters de orden 3 y 4, así como combinaciones de clusters con caminos hasta de orden 3. También se consideran fragmentos los ciclos desde orden 3 hasta 9. No obstante, no es posible afirmar que fragmentos pequeños seleccionados solamente con criterio matemático (subgrafos del grafo químico) sean los responsables directos de una propiedad tan compleja como la actividad biológica, la cual debía atribuirse a fragmentos mayores con características necesarias para dar la respuesta biológica. Esto planteó un ***problema*** importante, *¿cómo podemos representar en la Plataforma, fragmentos mayores dentro de las moléculas con un enfoque más químico-estructural?* El problema planteado, muestra la necesidad de disponer de una nueva forma de representar los fragmentos de las estructuras químicas.

Lo anterior indujo a plantear como ***hipótesis*** de trabajo que: *Es posible definir un lenguaje descriptor de estructuras químicas que incluya en su sintaxis información químico-física, para lograr una descripción más detallada de la estructura molecular para uso de la Plataforma.*

Este problema se enmarca en el ***objeto de estudio:*** *Estructura de las moléculas.*
Abarcando como ***campo de acción:*** *Codificación de estructuras químicas.*

Para dar solución a lo planteado se trazó el siguiente ***objetivo general:*** *Definir el lenguaje descriptor de estructuras químicas para la Plataforma, que incluya información químico-física y esté orientado a la descripción de la actividad biológica.*

Para cumplir este objetivo se plantean los siguientes ***objetivos específicos:***

- Diseñar e implementar un algoritmo para el reconocimiento de Centros Descriptores.
- Diseñar e implementar un algoritmo para la determinación de caminos de unión entre Centros Descriptores.
- Definir las reglas gramaticales del lenguaje.

Para dar cumplimiento a los objetivos específicos se plantean las siguientes ***tareas:***

- Revisión bibliográfica y análisis crítico de la información sobre lenguajes descriptores de estructuras químicas.
- Determinación e implementación de las estructuras de datos que requiera el lenguaje para su desarrollo.
- Identificación de los Centros Descriptores presentes en las moléculas.
- Identificación de los caminos de unión entre Centros Descriptores.

Como resultado, se pretende brindar a la Plataforma un lenguaje capaz de describir estructuras moleculares a partir del análisis de fragmentos ponderados por el Índice del Estado Refractotopológico Total[1] [8] orientado a la descripción de la actividad biológica.

La tesis se dividió en Resumen, Introducción, dos capítulos que contienen la información referente a la investigación realizada, Conclusiones, Recomendaciones, Referencias Bibliográficas y Bibliografía. Los capítulos que constituyen el cuerpo de la tesis abordan los temas tratados de la siguiente forma:

Capítulo 1: Fundamentación teórica, se muestra un estudio detallado sobre los lenguajes utilizados en la actualidad a nivel mundial, sus principales características, aspectos comunes y diferentes. También se realiza un análisis de las tendencias y tecnologías actuales con el objetivo de seleccionar cuáles son las indicadas para desarrollar el lenguaje descriptor que garantice una descripción más detallada de las estructuras moleculares en la Plataforma.

Capítulo 2: Descripción del lenguaje, se aborda todo lo referente a las características y estructura que va a presentar el lenguaje descriptor, se hace un análisis crítico de los algoritmos necesarios implementados, con el objetivo de verificar los resultados teóricos esperados y se presentan las reglas gramaticales del lenguaje así como la solución alcanzada para la representación del mismo.

[1] **Índice del Estado Refractotopológico Total:** se desarrolla a partir de la teoría del grafo químico y de la partición de la refractividad atómica definida por Ghose y Crippen. El índice se basa en la influencia de las fuerzas de dispersión de cada átomo sobre cada uno de los restantes en la molécula, modificado por la topología molecular.

1

FUNDAMENTACIÓN TEÓRICA.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

1.1 Introducción

El lenguaje articulado, como medio de comunicación entre los seres humanos, es resultado del desarrollo del cerebro, y constituye un avance principal en la vida y desarrollo de la humanidad. Como parte de ese desarrollo, han surgido ciertos tipos de lenguajes que se emplean en la química y especialidades afines, para la comunicación e intercambio de información estructural. Sin embargo, hablar de lenguajes en general no constituye el objetivo de esta tesis. El interés de este trabajo se centra en el estado de desarrollo de algunos de los lenguajes que han sido diseñados para facilitar la comprensión y descripción de las estructuras químicas y que de algún modo han servido de guía para cumplir con el objetivo de esta tesis.

1.2 Estado del Arte

Los lenguajes descriptores de estructuras moleculares resultan poco conocidos para muchas personas, sin embargo son de gran importancia en el campo de la química, ya que permiten el reconocimiento y análisis de las estructuras y sus partes, brindan información sobre la composición química de la molécula y sus características principales y se utilizan para la recuperación de información en grandes bases de datos de productos químicos. Los más conocidos internacionalmente son el SMILES, SMARTS e InChI.

1.2.1_SMILES: **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem.

SMILES es un lenguaje empleado en la descripción de estructuras químicas que utiliza símbolos y el código ASCII de los caracteres para representar los átomos y enlaces presentes entre ellos. Las cadenas SMILES pueden ser importadas a través de editores gráficos de moléculas que realizan la conversión de estos formatos SMILES a modelos moleculares bi- o tridimensionales.
Las especificaciones del código SMILES fueron desarrolladas por David Weininger en la década del 80, usando el concepto de grafo molecular como base para la representación de la estructura química, tomando como nodos a los átomos y como aristas los caminos de enlace entre ellos. Este lenguaje ha sido modificado subsecuentemente y extendido notablemente por Daylight Chemical Information Systems Inc.

SMILES es un verdadero idioma, aunque con un vocabulario simple y sólo unas reglas gramaticales.
Es preciso aclarar que pueden existir varios códigos SMILES, válidos para una misma estructura, un ejemplo de ello se evidencia en la notación de estructuras lineales. [9]

1.2.2_SMARTS: **SM**iles **AR**bitrary **T**arget **S**pecification

SMARTS es un lenguaje que permite especificar subestructuras moleculares y utiliza reglas que son extensiones de las del SMILES. De hecho, casi todas las características técnicas de SMILES son válidas para el lenguaje SMARTS; utilizando este último pueden hacerse búsquedas eficaces y flexibles en grandes bases de datos y ofrecer información significativa para los usuarios.
SMARTS proporciona varios símbolos primarios que describen las propiedades atómicas más allá de aquellos usados en SMILES (el símbolo atómico, la carga, y las especificaciones isotópicas). Además, dentro de las etiquetas que se utilizan para especificar los nodos y enlaces del grafo, este lenguaje adiciona los operadores lógicos, lo que hace que presente una sintaxis más general.
Todas las expresiones de SMILES, son también expresiones SMARTS, pero la semántica cambia pues las expresiones de SMILES describen moléculas, mientras que las SMARTS describen fragmentos. Una molécula puede representarse por la cadena SMILES o SMARTS, pero el proceso inverso de representar un fragmento no ocurre. [10]
El surgimiento de varias versiones del lenguaje SMARTS, provocó desacuerdos entre los químicos y aparecieron incoherencias a la hora de representar las estructuras moleculares, constituyendo esta la desventaja más acentuada de dicho lenguaje. Para ello se tomaron acuerdos, para reunir todas las versiones existentes de SMARTS y se desarrolló un nuevo lenguaje que solucionó los problemas existentes en los anteriores, conocido con el nombre de InChI.

1.2.3_InChI: **In**ternational **Ch**emical **I**dentifier

InChI fue desarrollado con la cooperación de IUPAC (*International Union of Pure and Applied Chemistry*) y NIST (*National Institute of Standards and Technology, USA*) con el objetivo fundamental de establecer una única etiqueta, o identificador químico para utilizarlo en bases de datos impresas y electrónicas permitiendo así una recopilación más fácil de los datos existentes. Este lenguaje constituye la forma más nueva de descripción de las estructuras químicas y gana popularidad continuamente en la comunidad de la química informática por presentar rasgos muy interesantes. [11].

InChI resuelve muchas de las ambigüedades químicas no tratadas por SMILES, fue desarrollado por Dimitri Tchekhovskoi, Steve Stein y Steve Heller en el Instituto Nacional Americano de Normas y Tecnología (NIST).

El Identificador Químico Internacional (InChI) mantiene muy bien definidas las etiquetas para las sustancias químicas puras. Estas etiquetas se generan convirtiendo una entrada de la estructura química, en el formulario de una tabla de conexión, a una única y predecible serie de caracteres del código ASCII. Los identificadores de las estructuras químicas, no presentan un registro numérico y tampoco requieren del acceso a una base de datos. Esta facilidad se debe al desarrollo en primer lugar de un medio para "nombrar" un compuesto en los medios de comunicación digitales, aunque el identificador se expresa como texto simple que puede interpretarse de forma manual. [12]

El procedimiento para la conversión al lenguaje InchI genera un identificador diferente para cada compuesto, pero siempre da el mismo identificador para un compuesto particular sin tener en cuenta la forma de entrada de la estructura. El procedimiento se aplica igualmente a todo tipo de compuestos. [13]

Aunque los lenguajes anteriormente mencionados brindan solamente información relacionada con la composición química de las estructuras moleculares la forma de codificar la molécula es completamente diferente, es decir la sintaxis varía en dependencia del lenguaje que se utilice, como se observa en los siguientes ejemplos.

Ejemplos:

Conversión:

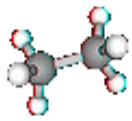
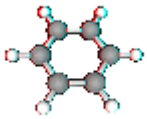
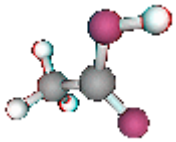
Nombre	SMILES	InChI	Estructura
Etano	CC	InChI=1/C2H6/c1-2/h1-2H3	
Dióxido de carbono	O=C=O	InChI=1/CO2/c2-1-3	
Ciclohexano	C1CCCCC1	InChI=1/C6H12/c1-2-4-6-5-3-1/h1-6H2	
Benceno	c1ccccc1	InChI=1/C6H6/c1-2-4-6-5-3-1/h1-6H	
Ácido acético	CC(=O)O	InChI=1/C2H4O2/c1-2(3)4/h1H3,(H,3,4)	

Tabla 1: Compuestos codificados según SMILES e InChI.

SMILES, SMARTS e InChI constituyen una herramienta muy útil en el trabajo de búsqueda de fragmentos en grandes bases de datos, aunque todos presentan como limitante el hecho de que sólo brindan información estructural topológica. En caso que se desee adicionar información a esta descripción, es imposible hacerlo mediante la sintaxis que presentan estos lenguajes.
Un paso de avance en este sentido se logró con la introducción del lenguaje DCAM (*Descriptor Centre Adjacency Matrix*) que constituye una versión ampliada del lenguaje SSFN (*Substructure Superposition Fragmental Notation*).

1.2.4_SSFN: **S**ubstructure **S**uperposition **F**ragmental **N**otation

SSFN es el lenguaje más simple de representación de la estructura, basado en el concepto de centros descriptores (CD) y las distancias entre ellos.
Todo los heteroátomos (átomos distintos de Carbono e Hidrógeno) incluidos N,O,S,P, halógenos, metales, etc., se toman como centros descriptores en el lenguaje SSFN, así como los sistemas aromáticos cíclicos en conjunto y los pares de átomos de carbono conectados por enlaces múltiples (doble o triple, pero no aromáticos). Cada CD del lenguaje SSFN se representa por un código formado por una cadena de caracteres.

Ejemplo:

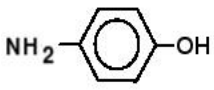
Compuesto	Código
	**0264111**

Tabla 2: Compuesto codificado según SSFN.

De esta forma queda representada una estructura molecular en el lenguaje SSFN, en este caso particular es preciso aclarar que en la figura existen 3 centros descriptores, cada uno de ellos identificado por un código único:

1. **NH_2:** 02
2. **Anillo aromático:** 64
3. **OH:** 11

4. El último valor (**1**) representa la conjugación en la molécula.

La conjugación se describe por un dígito simple, siendo igual a 1 cuando hay una unión y equivalente a 0 en ausencia de la misma. Se considera que existe conjugación cuando la longitud de la cadena es 00, o si todos los átomos de carbono presentes en ella tienen hibridación[2] sp o sp^2.

El lenguaje SSFN proporciona las bases para:

- La representación adecuada de farmacóforos[3] [14] pero solamente con carácter topológico.
- El diseño de procedimientos eficaces para la identificación de farmacóforos durante el proceso de obtención de la estructura y datos de la actividad para los compuestos químicos.
- El desarrollo de algoritmos fiables para la predicción de actividad biológica de compuestos basadas en grupos farmacofóricos encontrados.

Aportes y limitaciones del lenguaje SSFN:

- La ventaja principal del lenguaje SSFN reside en la simplicidad de representación de las características estructurales esenciales de los compuestos para la manifestación de la actividad biológica, por medio de la sucesión de descriptores.
- Puede usarse en algoritmos de reconocimiento de modelos estadísticos y para la evaluación cuantitativa de estructuras similares entre agentes biológicamente activos.
- El lenguaje SSFN resulta fácil de aprender por los químicos. La codificación de un compuesto requiere, un tiempo promedio de 2 minutos y su código contiene alrededor de 14 -15 descriptores.

[2] **Hibridación:** Consiste en una mezcla de orbitales puros en un estado excitado para formar orbitales híbridos equivalentes con orientaciones determinadas en el espacio.

[3] **Farmacóforo:**

Emil Fischer (1984) y Paul Ehrlich (1909)

Término, introducido por Ehrlich:

“Esqueleto" molecular que "transporta" (*phoros*) los elementos esenciales responsables de que un compuesto tenga una actividad biológica determinada (*pharmacon*).

La desventaja principal del lenguaje SSFN y otros sistemas de notación de fragmentos es la desintegración de la estructura. La información solo almacena los fragmentos estructurales de la molécula, pero no proporciona la información sobre su interconexión. Es más, la manifestación de la actividad biológica frecuentemente depende de la distancia global entre dos centros activos que no sólo pueden separarse por una cadena de carbono, sino también por un heteroátomo. Esta cadena en el lenguaje SSFN está rota en varios descriptores y no puede encontrarse de forma íntegra.

Más allá del desarrollo de los principios fundamentales del lenguaje SSFN se ha producido un nuevo método de representación estructural que, satisfaciendo los requisitos principales de este lenguaje, retiene la topología de una molécula en conjunto. La nueva forma de representación del lenguaje se llama Matriz de Adyacencia de Centros Descriptores **(DCAM).**

1.2.5_DCAM: **D**escriptor **C**entre **A**djacency **M**atrix

La estructura se describe por una matriz simétrica:

$$DCAM = \begin{pmatrix} E_{11} & E_{12} & E_{n} \\ & E_{22} & E_{2n} \\ & & E_{nn} \end{pmatrix}$$

Figura 1: Estructura de la Matriz DCAM.

La diagonal principal (**E_{11}, E_{22},...E_{nn}**) de la matriz de adyacencia caracteriza los centros descriptores de la estructura, y las entradas no diagonales (E_{12}, E_{n}, $E_{2n,...}$) contienen los parámetros para describir las características de los enlaces presentes entre estos centros descriptores.

Ejemplo:

Representación de una estructura usando el lenguaje DCAM

	1	2	3
1	**66**	-2	17
			37
2		**66**	-2
3			**66**

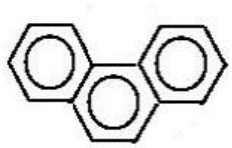


Figura 2: Compuesto representado en la matriz DCAM.

Donde la cantidad de centros descriptores presentes en el compuesto químico: (**1, 2, 3**) representan las filas y columnas de la matriz.

En la diagonal principal se representa el código de cada CD presente en la estructura molecular, en este caso (**66**) porque aparece 3 veces en el compuesto.

Los valores (**17, 37**) corresponden a características del enlace entre los centros descriptores (simples, múltiples, aromáticos).

El valor (**-2**) representa que dos anillos comparten la misma arista o enlace.

En el lenguaje DCAM las dimensiones de la matriz dependen de la cantidad de CD que forme la estructura del compuesto.

Los lenguajes como DCAM y similares se orientan a problemas. Esto significa que esa información adicional que falta en las fórmulas estructurales se introduce en ellos. Así, se introduce el postulado de los centros activos y su clasificación en SSFN y DCAM.

Naturalmente, esta información puede tenerse en cuenta cuando se diseñan los algoritmos para la selección de rasgos de actividad y análisis cualitativos de relación estructura-actividad (SAR por sus siglas en inglés *Structure Activity Relationships***)**. Sin embargo, esto puede llevar a una falta de universalidad, a la complicación de los algoritmos, y puede aumentar el costo computacional.

Al mismo tiempo los lenguajes orientados a problemas, como el DCAM, no pueden reemplazar las fórmulas estructurales en todos los casos porque en ellos no existe una correspondencia 1 a 1 entre la descripción de dos estructuras, en cambio presenta una sintaxis más abarcadora y brinda información más detallada sobre las propiedades, características y composición química de las estructuras moleculares, alcanzando un punto de avance respecto a los restantes lenguajes analizados, puesto que logra hacer la descripción íntegra de una estructura molecular, partiendo del análisis de centros descriptores activos dentro de la estructura, asociados con la actividad biológica. No obstante, aún no soluciona del todo los problemas que se plantean para lograr una adecuada descripción molecular que responda a las necesidades del proyecto, por lo que se decidió desarrollar un lenguaje similar al DCAM pero con nuevas particularidades.

1.2.6_Nuevo lenguaje descriptor de estructuras químicas

El desarrollo del lenguaje descriptor de estructuras químicas tiene gran similitud con el lenguaje DCAM pues está basado también en la descripción a partir de centros descriptores presentes en la estructura molecular y caminos de unión existentes entre ellos, pues de esta forma se puede obtener la representación de los fragmentos notables dentro de una molécula, los que constituyen las partes fundamentales de la estructura y a su vez los responsables potenciales de manifestar la actividad biológica.

Para definir cuáles de estos fragmentos presentan mayores posibilidades de interactuar o reaccionar ante determinadas situaciones en el medio en que se encuentran, es preciso expresar características, mediante valores numéricos, que permitan distinguir fragmentos topológicamente iguales dentro de una molécula o moléculas desiguales que presentan diferente grado de responsabilidad en la mayor o menor actividad. Para esto, el nuevo lenguaje introduce los valores del Índice del Estado Refractotopológico Total **(R-state, R_i)** (Ver epígrafe 2.2.1).

Este lenguaje almacena la información necesaria para realizar posteriores análisis en los restantes módulos del proyecto, permite hacer también un estudio más detallado de un compuesto químico, ya que no solo brinda información de la composición química de una estructura molecular, sino su capacidad para interactuar con otras moléculas. Esta información está contenida en los valores del *Índice del Estado Refractotopológico Total* que se calcula para cada fragmento. Además, se utiliza una nueva forma de representación de la información que permite incluir los datos que anteriormente eran imposibles de añadir en los lenguajes anteriores por las restricciones y características particulares en la sintaxis.

1.3 Tendencias y tecnologías a considerar

Existen un conjunto de tendencias y tecnologías que pueden ser adecuadas para llevar a cabo el lenguaje descriptor de estructuras químicas que se pretende desarrollar. Para la selección del lenguaje de programación, así como la utilización de los diversos componentes; se consultaron diferentes características como portabilidad, compatibilidad y facilidad de uso, resultando los seleccionados los que más se acercan al cumplimiento del objetivo general planteado en el trabajo.
Según el análisis efectuado se determinó que para la definición del lenguaje descriptor no es preciso regirse por una metodología de desarrollo de software como Rational Unified Process (**RUP**) o Extreme Programing (**XP**) por citar algunos ejemplos, puesto que en la presente investigación no se lleva a cabo un proceso de software con las fases, procedimientos, reglas, técnicas y documentación requerida por una metodología porque no constituye el principal objetivo del trabajo, pero si es preciso el uso del Lenguaje Unificado de Modelado (**UML**) para realizar la representación de los diagramas o modelos que contribuyen a lograr el propósito planteado.

1.3.1_Lenguaje representativo UML

El Lenguaje de Modelado Unificado (UML - Unified Modeling Language) fue creado por un grupo de estudiosos de la Ingeniería de Software formado por: Ivar Jacobson, Grady Booch y James Rumbaugh en el año 1995. Desde entonces, se ha convertido en el estándar internacional para definir, organizar y visualizar los elementos que configuran la arquitectura de una aplicación orientada a objetos. Tiene como objetivo brindar un material de apoyo que le permita al lector poder definir diagramas propios como también entender diagramas ya existentes.

“El UML es un lenguaje para la construcción de modelos; no guía al desarrollador en la forma de realizar el análisis y diseño orientados a objetos ni le indica cual proceso de desarrollo adoptar” [15].

UML prescribe un conjunto de notaciones y diagramas estándar para modelar sistemas orientados a objetos, y describe la semántica esencial de lo que estos diagramas y símbolos significan. Mientras que ha habido muchas notaciones y métodos usados para el diseño orientado a objetos, ahora los modeladores sólo tienen que aprender una única notación.

1.3.2_Herramientas CASE (Computer Aided Software Engineering)

Visual Paradigm: Es una herramienta CASE que utiliza "UML" como lenguaje de modelado. VP-UML soporta los últimos estándares de anotaciones de JAVA y UML, provee soporte para la generación de código y la ingeniería inversa para Java, además se integra con Eclipse, JBuilder, NetBeans IDE, Oracle JDeveloper, BEA Weblogic, Borland®, para soportar las fases de implementación en el desarrollo de un software.
Está disponible en varias ediciones, cada una destinada a diferentes necesidades: Enterprise, Professional, Community, Standard, Modeler y Personal. [16]

Visual Paradigm ofrece:

- Entorno de creación de diagramas para UML 2.0.
- Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- Modelo y código que permanece sincronizado en todo el ciclo de desarrollo.
- Disponibilidad de múltiples versiones, para cada necesidad.
- Disponibilidad de integrarse en los principales IDEs.
- Disponibilidad en múltiples plataformas.

1.3.3_Lenguaje de programación utilizado (Java)

Características del lenguaje:
El lenguaje de programación Java es un lenguaje de propósito general, concurrente, basado en clases y orientado a objetos, eso implica que su concepción es muy próxima a la forma de pensar humana. Ofrece toda la funcionalidad de un lenguaje potente, pero sin las características menos usadas y más confusas de éstos. C++ es un lenguaje que adolece de falta de seguridad, pero C y C++ son lenguajes más difundidos, por ello Java se diseñó para ser parecido a C++ y así facilitar un rápido y fácil aprendizaje, además el mismo elimina muchas de las características de otros lenguajes como C++, para mantener reducidas las especificaciones del lenguaje y añadir características muy útiles.

Es un lenguaje libre por lo que brinda la posibilidad de utilizar el compilador y la máquina virtual de forma gratuita, es un lenguaje de programación que es compilado, y genera ficheros de clases compiladas, pero estas clases, son en realidad interpretadas por la máquina virtual de java, la cual se encarga de mantener el control sobre las clases que se estén ejecutando.

Java es multiplataforma lo que permite que un mismo código java funcione en cualquier otro sistema operativo que tenga instalada la máquina virtual java, la cual le confiere una gran seguridad a este lenguaje, ya que al ejecutar el código java, la máquina virtual realiza comprobaciones de seguridad, además el propio lenguaje carece de características inseguras, como por ejemplo los punteros.
Gracias al API de java podemos ampliar el lenguaje para que sea capaz de comunicarse con equipos mediante red, acceder a bases de datos, crear páginas HTML dinámicas, crear aplicaciones visuales al estilo Windows, y realizar diversas operaciones.[17]

1.3.4_Entorno de desarrollo (Eclipse)

Eclipse es una poderosa herramienta que permite integrar diferentes aplicaciones para construir un entorno integrado de desarrollo (IDE). Es un potente entorno de desarrollo de Java, usa java como lenguaje de programación ya que soporta la programación orientada a objetos (POO) y la implementación de aplicaciones resulta mucho más sencilla. Mediante Eclipse se puede crear diversas aplicaciones como son sitios Web, programas Java, C++ y Enterprise Java Beans. Su principal aplicación es JDT (Java Development Tool), herramienta para crear aplicaciones en Java. Otras aplicaciones pueden ser integradas a eclipse en forma de plugins, que son reconocidos automáticamente por Eclipse al iniciar el mismo. [18].

Beneficios:

- Es una herramienta open-source.
- Soporta la construcción de una variedad de herramientas para el desarrollo de aplicaciones.
- Soporta herramientas que manipulan diferentes tipos de archivos como por ejemplo Java, C, C++, EJB, HTML, GIF, etc.
- Corre en una gran cantidad de sistemas operativos incluyendo Windows y Linux.
- Provee a los desarrolladores, herramientas (ej.- PDE) que facilitan la creación de plugins.

1.3.5_Sistemas gestores de bases de datos (SGBD)

Se pueden definir como un paquete generalizado de software, que se ejecuta en un sistema computacional anfitrión, centralizando los accesos a los datos y actuando de interfaz entre los datos físicos y el usuario. Las principales funciones que debe cumplir un SGBD se relacionan con la creación y mantenimiento de la base de datos, el control de accesos, la manipulación de datos de acuerdo con las necesidades del usuario, el cumplimiento de las normas de tratamiento de datos, evitar redundancias e inconsistencias y mantener la integridad.[19]

MySQL:

Es un sistema de gestión de bases de datos relacional, licenciado bajo la GPL de la GNU. Su diseño multihilo le permite soportar una gran carga de forma muy eficiente.

Este gestor de bases de datos es, probablemente, el gestor más usado en el mundo del software libre, debido a su gran rapidez y facilidad de uso. Esta gran aceptación es debida, en parte, a que existen infinidad de librerías y otras herramientas que permiten su uso a través de gran cantidad de lenguajes de programación, además de su fácil instalación y configuración.[20]

Las principales características de este gestor de bases de datos son las siguientes:

- Aprovecha la potencia de sistemas multiprocesador, gracias a su implementación multihilo.
- Soporta gran cantidad de tipos de datos para las columnas.
- Dispone de API's en gran cantidad de lenguajes (C, C++, Java, PHP, etc.)
- Gran portabilidad entre sistemas.
- Soporta hasta 32 índices por tabla.
- Gestión de usuarios y contraseñas, manteniendo un muy buen nivel de seguridad en los datos.

1.4 Conclusiones

Teniendo en cuenta que el lenguaje descriptor constituye un módulo de la Plataforma, se decidió mantener una uniformidad con las herramientas a utilizar y se escogió:

- El lenguaje de programación Java por ser un lenguaje de propósito general, por su portabilidad y uso gratuito.
- El Visual Paradigm como herramienta CASE para realizar el modelado siguiendo el estándar UML.
- El Eclipse como entorno para llevar a cabo la programación por sus grandes potencialidades, además de ser una herramienta de código abierto.
- Como gestor de base de datos MySQL por sus facilidades de uso, rapidez y compatibilidad con otras versiones, además de ser libre y de bajo costo.

2

DESCRIPCIÓN DEL LENGUAJE.

CAPÍTULO 2: DESCRIPCIÓN DEL LENGUAJE

2.1 Introducción

La necesidad de obtener una vía para predecir la actividad biológica de las moléculas constituye la pauta fundamental que dio origen al desarrollo de una nueva forma de representación de la estructura molecular que satisfaga las necesidades particulares del proyecto:***"Plataforma Inteligente para la Predicción de Actividad Biológica de Compuestos Orgánicos"*** , por lo que el lenguaje que a continuación se presenta está encaminado a dar solución a la problemática existente con respecto a la descripción molecular de los compuestos orgánicos que el proyecto necesita obtener.

2.2 Lenguaje Descriptor. Definiciones.

Se define como

- Centro Descriptor (**CD**): Las agrupaciones de átomos que constituyen o no grupos funcionales dentro de una molécula.

- Camino de unión (**CU**): Fragmento molecular enlazado por cada uno de sus extremos a un **CD**.

El lenguaje consiste pues, en: La representación de una molécula por un conjunto de fragmentos que le pertenecen. Esos fragmentos están construidos por dos CD's unidos entre si por un camino de unión.

Tanto los **CD** como los **CU** están ponderados por el **(R-state, R_i)**.

Como los valores del **(R-state, R_i)** de los fragmentos dependen del entorno molecular en el que están insertados, cada uno tendrá un valor propio que lo identifica, por lo tanto, ese valor, junto con la topología del fragmento es lo que debe distinguir el grado de participación de dicho fragmento en la respuesta biológica.

2.2.1_Índice del Estado Refractotopológico Total

El índice que se utiliza en el lenguaje es el Índice del Estado Refractotopológico el cual se desarrolla a partir de la teoría del grafo químico y de la partición de la refractividad atómica definida por Ghose y Crippen. Se basa en la influencia de las fuerzas de dispersión de cada átomo sobre cada uno de los restantes en la molécula, modificado por la topología molecular.

Definición de Ri

El **R-state R_i**, para un átomo **i** se define por la ecuación:

$$R_i = AR_i + \Delta AR_i$$

Donde **ARi** es el valor de refractividad intrínseco del átomo **i** y **ΔARi** es un término de perturbación definida por la ecuación:

$$\Delta AR_i = (AR_i - AR_j)/r^2_{ij}$$

Donde se suman todos los vértices **j** adyacentes en el grafo, **ARi** y **ARj** son los valores intrínsecos de la refractividad de los átomos **i** y **j**, respectivamente, y r^2_{ij} es el número de átomos del camino más corto entre los átomos **i** y **j**, incluyendo tanto a **i** como a **j**. Al igual que en el Estate y en el S-state, la distancia topológica cuadrática indica que debe haber una disminución de la interacción, con el aumento de la distancia de separación entre los átomos.

Características de Ri

A diferencia de otros índices topológicos, los cuales no consideran la influencia de los átomos de hidrógeno, este índice si incluye sus contribuciones al valor intrínseco de los átomos pesados. Esta inclusión refleja la capacidad potencial de interacción total del grupo con una supuesta proteína ligando y esto es importante no solamente para el valor intrínseco del átomo al cual esta directamente enlazado sino para los otros grupos presentes en la molécula. En resumen, el Índice del Estado Refractotopológico Total, refleja el papel de las fuerzas de dispersión de London a escala de cada átomo en la molécula, considerando la influencia de su entorno molecular.

2.2.2_Reglas gramaticales del lenguaje

Las reglas gramaticales constituyen la guía fundamental para la definición del lenguaje, tanto para la especificación de los CD como los caminos de unión entre estos. Estas reglas están estrechamente relacionadas con características y propiedades químicas de compuestos orgánicos y para la definición de las mismas se tuvo en cuenta el análisis y criterio de especialistas químicos.

R1. Se definen como centros descriptores las siguientes agrupaciones de átomos:

- Los ciclos o anillos, aromáticos y/o alicíclicos, de tres hasta nueve miembros.
- Los halógenos: F, Cl, Br, I.
- Cluster de orden 3.
- Cluster de orden 4.

R2. Los centros descriptores están unidos por caminos (fragmentos) que no incluyen otros CD's.
R3. Los caminos de unión entre CD's pueden ser más de uno.
R4. Tanto los CD's como los caminos entre ellos, se ponderan con el correspondiente valor calculado del Índice del Estado Refractotopológico Total.

2.2.3_Representación

La información se representa mediante una matriz conformada por una lista de elementos que representan los centros descriptores de la estructura molecular que se esté analizando, el código que identifica cada centro descriptor, la información referente a los caminos de unión entre ellos y el valor del índice del estado refractotopológico total correspondiente a cada uno. La longitud de la lista depende de la cantidad de centros descriptores que sean encontrados en la estructura molecular a modelar.

En la figura 3 se muestra la representación conceptual del compuesto como la matriz antes mencionada. Nótese la similitud con el lenguaje DCAM.

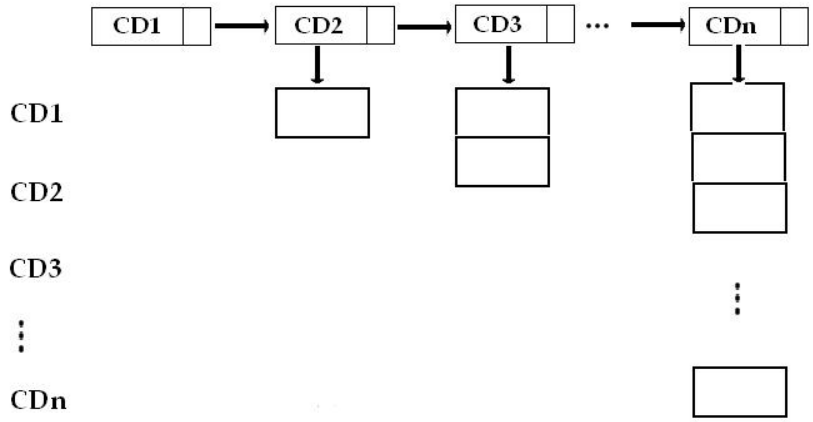


Figura 3: Representación de la matriz de adyacencia de CD.

Con esta nueva forma de representación se obtiene una solución que permite en gran medida una descripción más acertada de la estructura molecular atendiendo a parámetros específicos, que brindan una información más precisa y permiten obtener una respuesta concreta al problema científico planteado en el trabajo. A diferencia de la matriz DCAM esta forma de representación permite almacenar varios caminos de unión entre CD's, sin tener que encontrarse los mismos solapados entre sí.

2.3 Descripción de la solución propuesta

La solución propuesta a continuación, utiliza los componentes del modelo de dominio, atendiendo a las características del lenguaje UML para la representación del mismo. De este modelo se tendrá en cuenta la definición de los conceptos principales utilizados en el lenguaje así como su representación gráfica. También se presentan otros diagramas que contribuyeron a lograr la solución planteada.

2.3.1_Modelo de Dominio

Teniendo en cuenta que la definición de procesos y roles del proceso del negocio no se pueden definir en esta investigación porque no existen, se ve a simple vista la necesidad de describir el funcionamiento del lenguaje mediante una serie de conceptos, entidades y sus relaciones, agrupándolos en un modelo del dominio con el fin de facilitar el entendimiento del mismo.
Los objetos del dominio representan los entes o eventos que suceden en el entorno en el que se trabaja, el modelo de dominio captura los tipos más importantes de objetos en el contexto de este entorno. La modelación del dominio tiene como objetivo fundamental la comprensión y descripción de estos objetos, conceptos o eventos.

2.3.2_Definición de los principales conceptos

Con la aplicación del modelo del dominio se detectaron los conceptos siguientes:

- **Matriz:** Conjunto de elementos de cualquier naturaleza aunque, en general, suelen ser números ordenados en filas y columnas.
- **Átomo:** Es la entidad química mas pequeña (a representar), el mismo está compuesto de electrones, protones y neutrones.
- **Molécula:** Es la más pequeña cantidad de materia que retiene todas sus propiedades químicas. Está formada por un conjunto de átomos unidos por enlaces covalentes.
- **Fragmento:** Un fragmento es un conjunto de átomos interconectados formando una estructura determinada que es a su vez, parte de una molécula. Por ejemplo, caminos, ciclos, clusteres.
- **Ciclo:** Es la unión de 3 o más átomos formando un ciclo dentro de la molécula.
- **Cluster:** No es más que un átomo conectado a 3 o 4 átomos al mismo tiempo. Este número determina su grado (cluster grado 3 y cluster de grado 4)

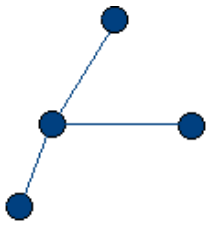


Figura 4: Representación de un cluster.

- **Halógeno:** Elemento químico perteneciente al grupo siete de la tabla periódica de Mendeleiev. Son halógenos el flúor F, cloro Cl, bromo Br y yodo I.
- **Centro descriptor:** Agrupación de átomos que determinan propiedades relevantes dentro de la molécula. No son necesariamente grupos funcionales.
- **Caminos de unión:** Conjunto de elementos químicos de una molécula que unen dos centros descriptores sin incluir un tercero.
- **Índice refractotopológico total:** Representa un valor numérico, está basado en la influencia de las fuerzas de dispersión de cada átomo sobre cada uno de los restantes en la molécula, modificado por la topología molecular.

2.3.3_Descripción Textual del Modelo del Dominio.

En el modelo se representa una matriz de adyacencia que contiene información de varios centros descriptores y los caminos de unión existente entre ellos, esta matriz solo puede estar asociada o pertenecer a una única molécula, la cual tiene asociado un conjunto de fragmentos.
Un fragmento puede representar un centro descriptor o un camino de unión según sus características particulares. Tanto los centros descriptores como los caminos de unión tienen asociado un valor de Índice del Estado Refractotopológico Total, en el caso particular de los CD, este valor puede variar en dependencia de la posición que ocupen dentro de la molécula, en cambio a los caminos de unión les corresponde un único valor de índice. Un centro descriptor es representado por un conjunto de átomos que forman una determinada estructura. Ejemplo (Ciclo, Cluster, Halógeno).

Representación del modelo del dominio.

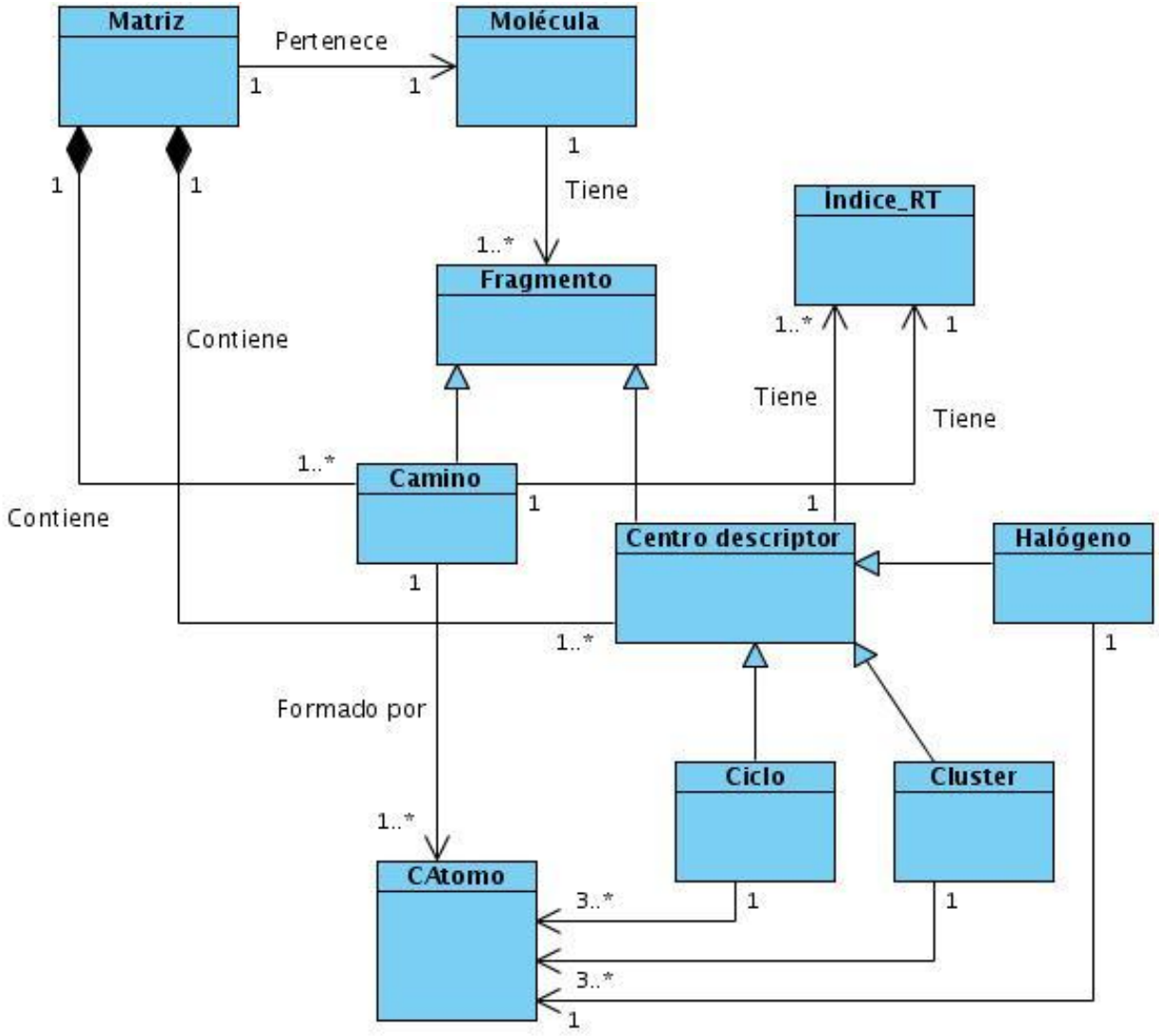


Figura 5: Modelo de Dominio.

2.3.4_Implementación.

Partiendo de las reglas gramaticales definidas para el lenguaje se implementaron un conjunto de clases y métodos que permiten realizar el proceso de reconocimiento de CD y los caminos de unión presentes en una estructura química, en general la traducción al lenguaje propuesto.

En primer lugar se definió una clase llamada **parameters** que es la encargada de la configuración del lenguaje, en ella se definen un conjunto de parámetros constantes tales como:

- La cantidad de caminos máximos a considerar entre dos CD.
- El nombre del fichero donde se encuentra almacenado el código correspondiente a cada CD.
- El tipo de fragmento que se esté analizando, dígase un ciclo, cluster, halógeno, etc.
- Parámetros de configuración de posibles bases de datos a usar.

Fueron implementadas un conjunto de clases vinculadas con la definición del lenguaje entre ellas se encuentran:

- **Class TAtom:** Representa un átomo y contiene la información referente al mismo.
- **Class TFragment:** Permite obtener toda la información relacionada con un fragmento molecular hibridación, tipo de fragmento, valor del Índice del Estado Refractotopológico Total, la cadena SMILE correspondiente, el conjunto de átomos que conforman el fragmento.
- **Class TBounds:** Es la clase que permite adicionar los caminos que existen entre los centros descriptores encontrados en una molécula.
- **Class TDCNode:** Su función principal es representar los nodos que conforman la matriz de adyacencia, es la que contiene la información relacionada con los CD.
- **Class TDescriptorCenter:** Permite representar los CD definidos en una molécula y el código que lo identifica.
- **Class TDCAMCP:** A través de esta clase es posible realizar la representación de la matriz de adyacencia, aquí es donde se almacena toda la información de los centros descriptores pertenecientes a una misma molécula y los caminos que los unen.

- **Class TDCAMControl:** Es la clase principal de traducción, permite obtener toda la información referente a una molécula de una base de datos, comprueba que un fragmento cumpla las reglas gramaticales definidas para el lenguaje, traduce o convierte la información obtenida en la matriz de CD.
- **Class TCoord:** Permite obtener la posición o coordenadas que ocupa un determinado CD dentro de la matriz.

Clases para establecer conexión con la base de datos:

- **Class DBControl**: Es la encargada de obtener la información de los fragmentos en la base de datos.
- **Class TConnection:** Permite realizar la conexión a la base de datos.

Clases para almacenar la información de los CD:

- **Class TTableDCInfo:** Permite representar un fragmento con su código correspondiente en una tabla, siempre que el fragmento constituya un CD.
- **Class TTableOfCDDefinition:** Es utilizada para leer todos los CD almacenados en un archivo.

Representación del diagrama de clases del diseño.

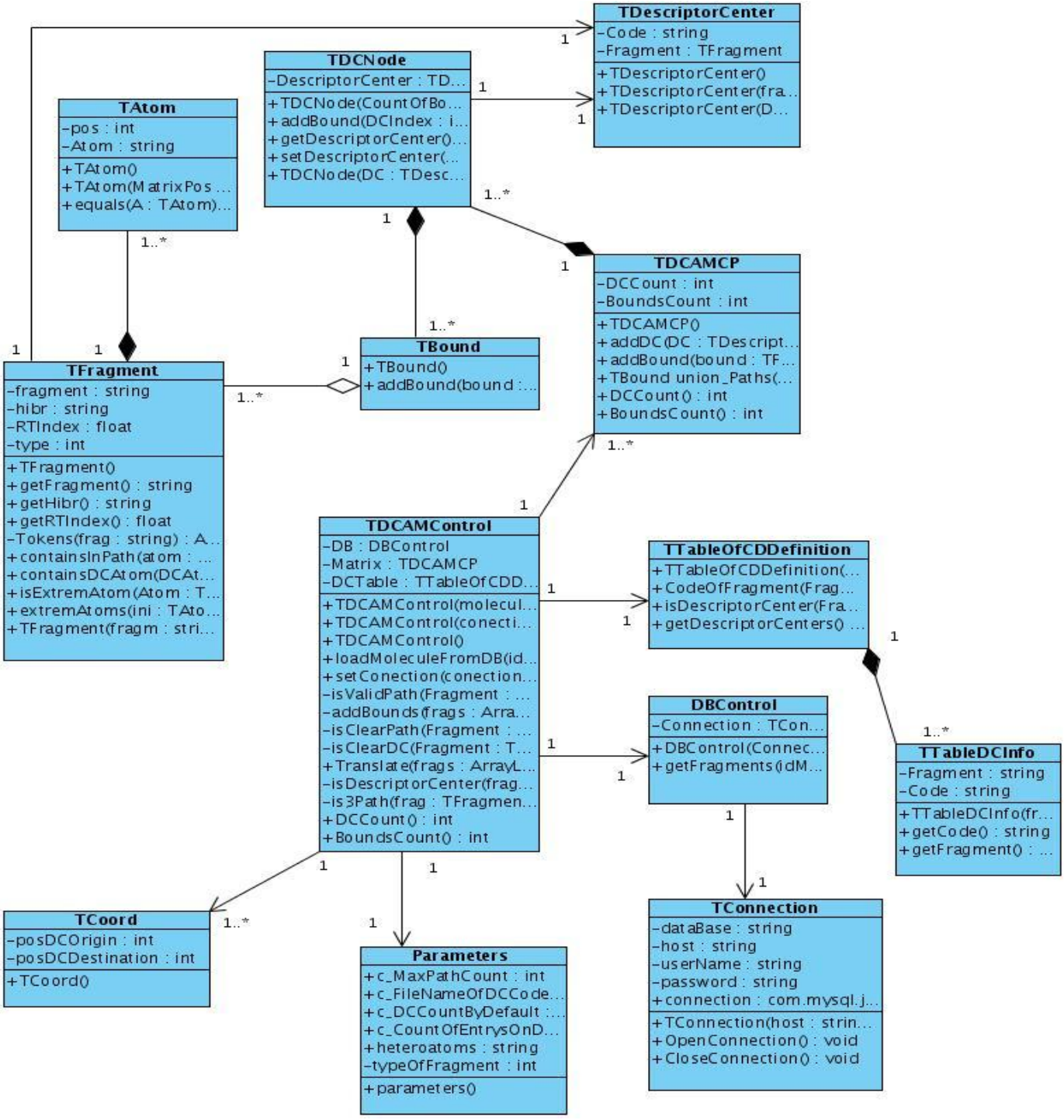


Figura 6: Diagrama de clases del diseño.

2.3.5_Ejemplo de Identificación de CD y Caminos de Unión.

Para realizar la conversión de una estructura química al lenguaje se tuvo en cuenta además de las reglas gramaticales definidas, la matriz de conectividad[4] de los átomos que forman la molécula, pues mediante la misma puede construirse el grafo molecular correspondiente al compuesto analizado.

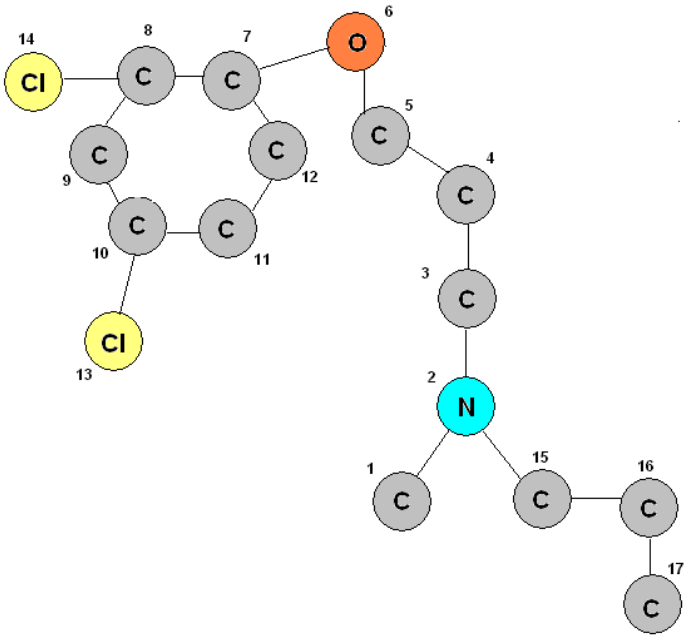


Figura 7: Grafo molecular.

Mediante los algoritmos implementados para el reconocimiento de centros descriptores y caminos de unión, que están basados directamente en las reglas gramaticales definidas para el lenguaje, pueden encontrarse centros descriptores representados estructuralmente de forma muy similar, pero esto no quiere decir que sean iguales, puesto que la posición que ocupa cada átomo en la estructura que lo forma influye de forma significativa en los valores del Índice del Estado Refractotopológico Total correspondiente a cada CD encontrado. Para la molécula representada en la figura 7 se obtienen los siguientes resultados.

[4] Los índices en la figura expresan la posición de cada átomo dentro de la matriz.

Centros Descriptores:

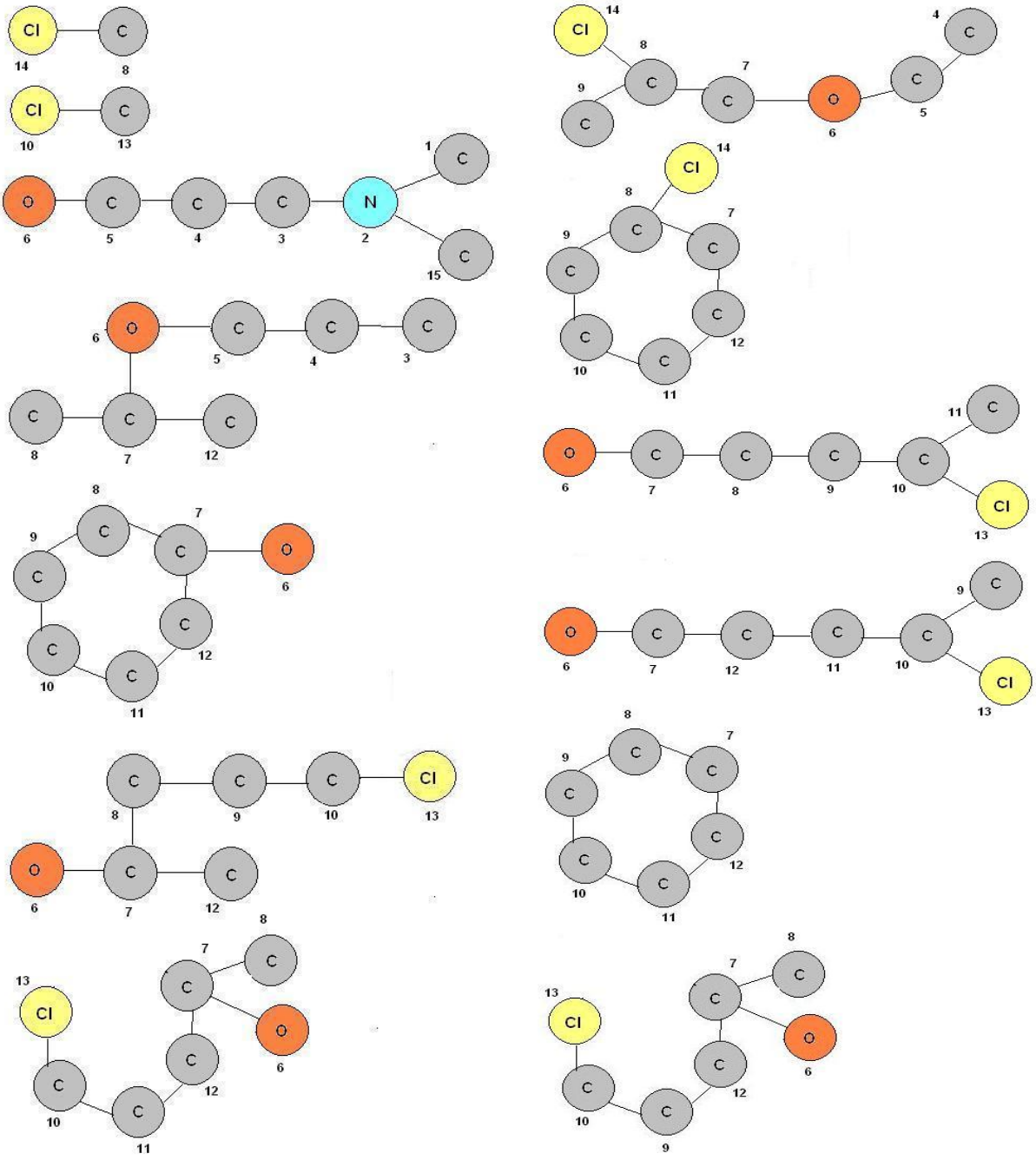
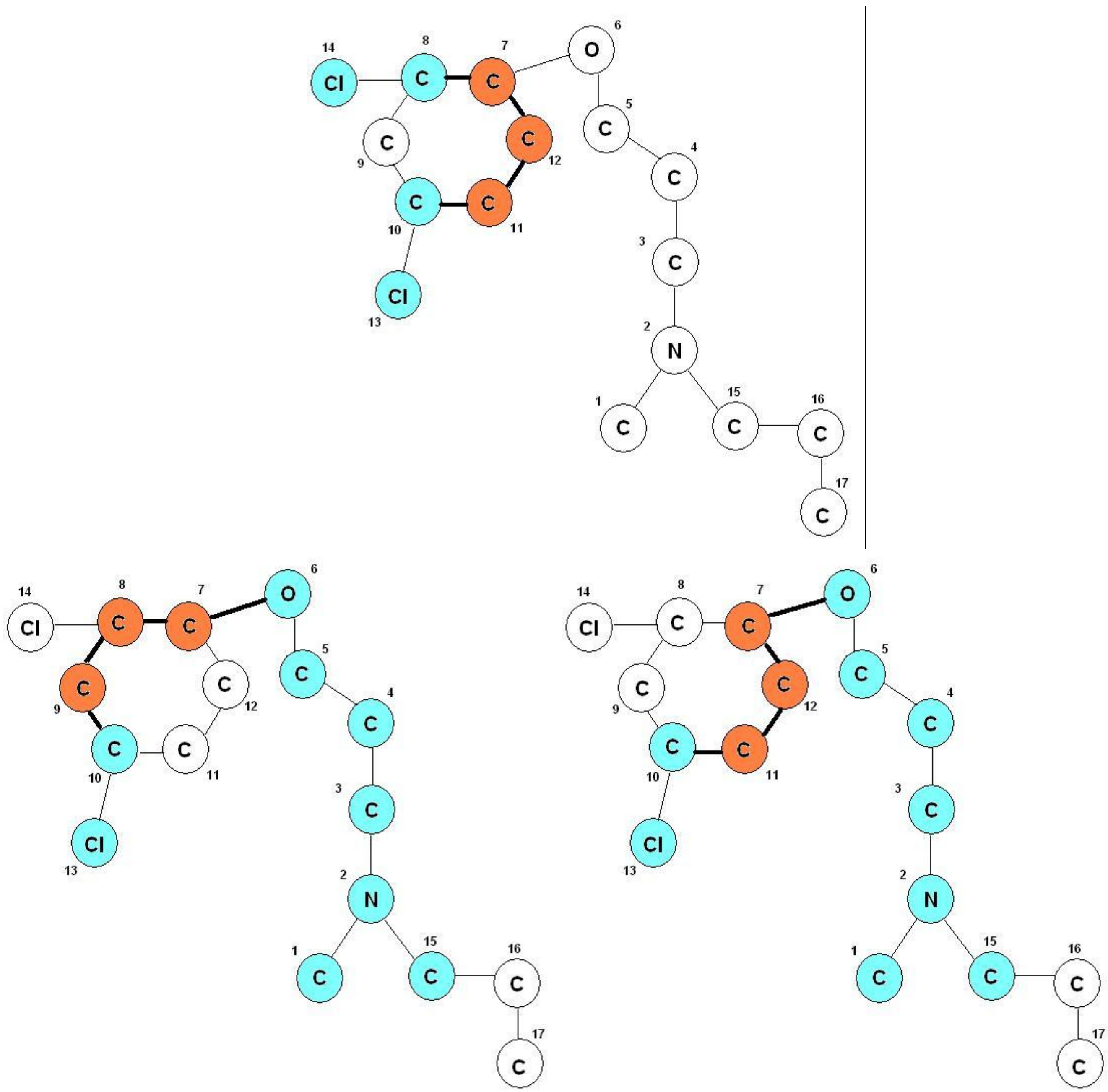


Figura 8: Centros Descriptores.

Caminos de Unión:



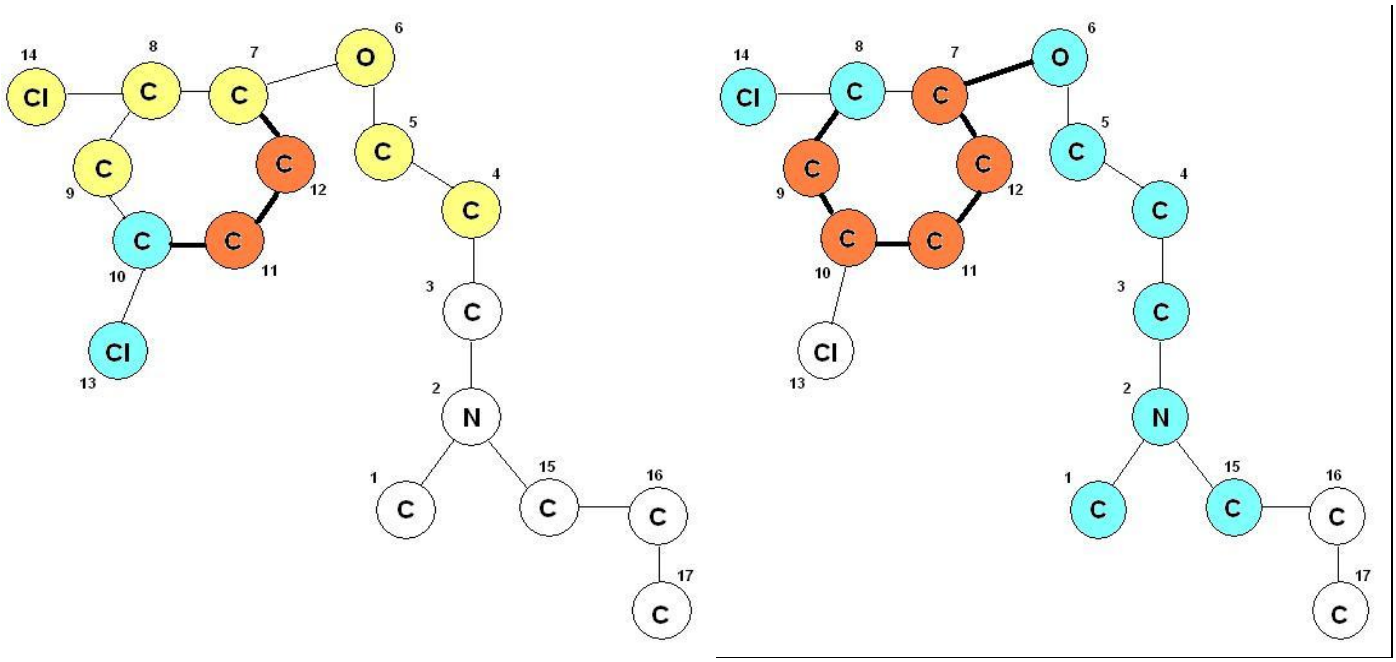


Figura 9: Caminos de unión válidos.

2.4 Conclusiones

En este capítulo se introduce el estudio de una propuesta basada en el modelo del dominio, donde se definen 10 de los conceptos, que se estima puedan tener alguna ambigüedad en la comprensión de dicho modelo. Se presentan las reglas gramaticales para la obtención del lenguaje, así como las características específicas del Índice del Estado Refractotopológico Total que constituye un elemento fundamental en la definición del lenguaje, se muestra una vía de solución para la implementación del mismo, así como las clases del diseño definidas, representadas con sus relaciones correspondientes en el diagrama de clases del diseño.

CONCLUSIONES

- Se desarrolló un lenguaje descriptor de estructuras químicas, para realizar la descripción de compuestos orgánicos, el cual permite representar la molécula como un conjunto limitado de fragmentos notables asociables a la propiedad biológica.
- Se definieron las reglas gramaticales para la definición de centros descriptores y caminos de unión, y se incorporaron los valores del Índice del Estado Refractotopológico Total correspondiente a cada fragmento molecular.
- Se implementaron los algoritmos necesarios para la traducción al lenguaje planteado.

RECOMENDACIONES

- Incorporar nuevos índices que contribuyan al enriquecimiento de la descripción de compuestos orgánicos como el ***Índice del Estado Refractotopográfico total*.**

- Crear una tabla de centros descriptores que almacene toda la información referente a estos, donde se identifiquen por un código único, sin importar la molécula a la que pertenecen.

- Hacer la traducción a partir de la identificación de la molécula y no a partir de los fragmentos de esta.

- Implementar una herramienta que permita evaluar visualmente, la traducción de una molécula expresada en el lenguaje propuesto.

REFERENCIAS BIBLIOGRÁFICAS

1. ALISO VIEJO, C." *SMILES - A Simplified Chemical Language"* [Página web]. Daylight Chemical Information Systems, Última actualización: 2007. [Consultado el: 20 de noviembre de 2006]. Disponible en: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

2. ---. *"SMARTS"- A Language for Describing Molecular Patterns* [Página web]. Daylight Chemical Information Systems, Última actualización: 2007. [Consultado el: 20 de noviembre de 2006]. Disponible en: http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html.

3. KOSATA, B. *"www.InChI.info"* [Página web]. Última actualización: 15.2.2007. [Consultado el: 20 de noviembre de 2006]. Disponible en: http://inchi.info/.

4. SYNTHESIS, I. O. O. *OREX, Expert System for Drug Design.* [Sistema experto]. Letonia: 1990. En Latvian Academy of Science.

5. ESCALONA, J. C.; CARRASCO, R.*, et al. Introducción al diseño de Fármacos.* [Folleto para la docencia de la asignatura de Farmacia]. Universidad de Oriente: 39 p.

6. ROZENBLIT, A. B. y GOLENDER, V. E. *Logico-Combinatorial Algorithm in Drug Design.* Research Studies Press Ltd., 1989. 289 p. ISBN 0 86380 006 8.

7. CARRASCO, R.; PADRÓN, J. A.*, et al. Definition of a novel atomic index for QSAR: the refractotopological state* [Página web]. J Pharm Pharmaceut Sci (http://www.cspscanada.org/), Última actualización: April 16, 2007. [Consultado el: 10 de enero de 2007]. Disponible en: http://www.ualberta.ca/~csps/JPPS7(1)/R.Carrasco/QSAR.htm

8. HERNÁNDEZ, N. R. F. *Sistema "GRATO (GRAph-TOol)" para la Visualización Molecular y el Cálculo de Descriptores.* Tesis de Diploma, Universidad de las Ciencias Informáticas, 2006.

9. JAMES, C. A. *An introduction to the Computer Science and Chemistry of Chemical Information Systems* [Página web]. eMolecules, Última actualización: 2007. [Consultado el: 10 de febrero de 2007]. Disponible en: http://www.emolecules.com/doc/cheminformatics-101.htm.

10. ALISO VIEJO, C. *SMARTS - A Language for Describing Molecular Patterns* [Página web]. Daylight Chemical Information Systems, Última actualización: 2007. [Consultado el: 12 de febrero de 2007]. Disponible en: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

11. IUPAC. *Chemical Nomenclature and Structure Representation Division (VIII)* [Página web]. International Union of Pure and Applied Chemistry, Última actualización: 19 April 2007. [Consultado el: 13 de febrero de 2007]. Disponible en: http://www.iupac.org/projects/2000/2000-025-1-800.html.

12. *The IUPAC International Chemical Identifier* [Página web]. [Consultado el: 13 de febrero de 2007]. Disponible en: http://inchi.sourceforge.net/.

13. CHEMISTRY, R. S. O. *16 May 2005: International chemical identifier goes online* [Revista]. Royal Society of Chemistry 2007, Disponible en: http://www.rsc.org/chemistryworld/News/2005/May/16May2005Internationalchemicalidentifiergoesonline.asp.

14. FONT., M. *El diseño de fármacos racional asistido por computadoras* [PDF]. Universidad de Navarra, [Consultado el: 18 de mayo de 2007]. Dpto de Química Orgánica y Farmacéutica. Sección de Modelización Molecular.

15. LARMAN, C. *UML y Patrones.Introducción al Análisis y Diseño Orientado a Objetos.* Prentice Hall Hispanoamérica, SA, 1999. vol. I, 291 p.

16. *Visual Paradigm for UML* Programación en castellano, Última actualización: 5 de julio de 2005. [Consultado el: 13 de mayo de 2007].
Disponible en: http://www.programacion.com/noticia/1363/.

17. EXES. *Características del lenguaje Java* [Sitio web]. MailxMail.com, [Consultado el: 14 de febrero de 2007]. Curso gratis
Disponible en: http://www.mailxmail.com/curso/informatica/java/capitulo2.htm

18. LEAD, P. P. *Introducción a la plataforma Eclipse* [PDF]. InCo, Última actualización: 16 noviembre de 2006. [Consultado el: 14 de febrero de 2007].

19. TRAMULLAS, J. y KRONOS, E. *Los sistemas de gestión de bases de datos* [Página web]. Última actualización: 1997-2000. [Consultado el: 15 de febrero de 2007]. Disponible en: http://tramullas.com/documatica/2-4.html.

20. PECOS, D. *PostGreSQL vs. MySQL* [Página web]. MySQL AB, [Consultado el: 15 de febrero de 2007]. [MySQL_Manual] Manual de MySQL,
http://www.mysql.com/documentation/index.html.
Disponible en: http://www.netpecos.org/docs/mysql_postgres/x57.html.

BIBLIOGRAFÍA

1. ALISO VIEJO, C. SMARTS - A Language for Describing Molecular Patterns [Página web]. Daylight Chemical Information Systems, Última actualización: 2007. [Consultado el: 12 de febrero de 2007]. Disponible en: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

2. ---. "SMARTS"- A Language for Describing Molecular Patterns [Página web]. Daylight Chemical Information Systems, Última actualización: 2007. [Consultado el: 20 de noviembre de 2006]. Disponible en: http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html.

3. ---. " SMILES - A Simplified Chemical Language" [Página web]. Daylight Chemical Information Systems, Última actualización: 2007. [Consultado el: 20 de noviembre de 2006]. Disponible en: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

4. AVIDON, V. V.; POMERANTSEV, I. A., et al. Structure-Activity Relationship Oriented Languages for Chemical Representation. 1982, vol. 22, 207-214 p.

5. CARRASCO, R.; PADRÓN, J. A., et al. Definition of a novel atomic index for QSAR: the refractotopological state [Página web]. J Pharm Pharmaceut Sci (http://www.cspscanada.org/), Última actualización: April 16, 2007. [Consultado el: 10 de enero de 2007]. Disponible en: http://www.ualberta.ca/~csps/JPPS7(1)/R.Carrasco/QSAR.htm

6. CHEMISTRY, R. S. O. 16 May 2005: International chemical identifier goes online [Revista]. 2007, Disponible en: http://www.rsc.org/chemistryworld/News/2005/May/16May2005Internationalchemicalidentifiergoesonline.asp.

7. D.WEININGER. "SMILES. 3. Depict: Graphical Depiction of Chemical Structures" [Página web]. JCICS, 30(3), 1990,237-243., [Consultado el: 22 de marzo de 2007].

8. ESCALONA, J. C.; CARRASCO, R., et al. Introducción al diseño de Fármacos. [Folleto para la docencia de la asignatura de Farmacia]. Universidad de Oriente: 39 p.

9. EXES. Características del lenguaje Java [Sitio web]. MailxMail.com, [Consultado el: 14 de febrero de 2007]. Curso gratis Disponible en: http://www.mailxmail.com/curso/informatica/java/capitulo2.htm

10. FONT., M. El diseño de fármacos racional asistido por computadoras [PDF]. Universidad de Navarra, [Consultado el: 18 de mayo de 2007]. Dpto de Química Orgánica y Farmacéutica. Sección de Modelización Molecular.

11. GUTMAN, D. I. y POLANSKY, D. O. E. Mathematical Concepts in Organic Chemistry Berlin: Akademic-Verlang, 1987. 200 p.

12. GUTMAN, I. y ESTRADA, E. Topological Indexes Based on the Line Graph of the Molecular Graph. 1996, vol. 36, 541-543 p.

13. HALL, L. H. y KIER, L. B. Topological Indexes Based on the Line Graph of the Molecular Graph 1995, vol. 35, 1039 - 1045 p.

14. HERNÁNDEZ, N. R. F. Sistema "GRATO (GRAph-TOol)" para la Visualización Molecular y el Cálculo de Descriptores. Tesis de Diploma, Universidad de las Ciencias Informáticas, 2006.

15. IUPAC. Chemical Nomenclature and Structure Representation Division (VIII) [Página web]. International Union of Pure and Applied Chemistry, Última actualización: 19 April 2007. [Consultado el: 13 de febrero de 2007]. Disponible en: http://www.iupac.org/projects/2000/2000-025-1-800.html.

16. JAMES, C. A. An introduction to the Computer Science and Chemistry of Chemical Information Systems [Página web]. eMolecules, Última actualización: 2007. [Consultado el: 10 de febrero de 2007]. Disponible en: http://www.emolecules.com/doc/cheminformatics-101.htm.

17. KEIR, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. 1986, vol. 5, 1-7 p.

18. KIER, L. B. Indexes of molecular shape from chemical graphs. 1986, vol. 36, 171-188 p.

19. KOSATA, B. "www.InChI.info" [Página web]. Última actualización: 15.2.2007. [Consultado el: 22 de marzo de 2007]. Disponible en: http://inchi.info/converter_en.html.

20. LARMAN, C. UML y Patrones.Introducción al Análisis y Diseño Orientado a Objetos. Prentice Hall Hispanoamérica, SA, 1999. vol. I, 291 p.

21. LEAD, P. P. Introducción a la plataforma Eclipse [PDF]. InCo, Última actualización: 16 noviembre de 2006. [Consultado el: 14 de febrero de 2007].

22. MENÉNDEZ, J. Á. El carbono, formas alotrópicas y estructuras de los carbones. [Página web]. [Consultado el: 30 de mayo de 2007]. Disponible en: http://www.oviedo.es/personales/carbon/estructuras/estructuras.htm.

23. PECOS, D. PostGreSQL vs. MySQL [Página web]. MySQL AB, [Consultado el: 15 de febrero de 2007]. [MySQL_Manual] Manual de MySQL, http://www.mysql.com/documentation/index.html Disponible en: http://www.netpecos.org/docs/mysql_postgres/x57.html.

24. ROMERO, A. V. G. SISTEMA PARA PREDICCIÓN ACTIVIDAD BIOLÓGICA DE COMPUESTOS ORGÁNICOS. Tesis de diploma, Instituto Superior Politécnico "José Antonio Echeverría", 2006.

25. ROZENBLIT, A. B. y GOLENDER, V. E. Logico-Combinatorial Algorithm in Drug Design. Research Studies Press Ltd., 1989. 289 p. ISBN 0 86380 006 8.

26. STEPHEN E. STEIN; STEPHEN R. HELLER, et al. The IUPAC Chemical Identifier – Technical Manual. [PDF]. Editado por: Division, P. A. C. P. Gaithersburg, Maryland, U.S. 20899-8380: National Institute of Standards and Technology,

27. SYNTHESIS, I. O. O. OREX, Expert System for Drug Design. [Sistema experto]. Letonia: 1990., En Latvian Academy of Science.

28. TEKNODA. Notas técnicas de JAVA - Tip en detalle Nro. 1 [PDF]. Teknoda S.A, [Consultado el: 3 de mayo de 2007].

29. TRAMULLAS, J. y KRONOS, E. Los sistemas de gestión de bases de datos [Página web]. Última actualización: 1997-2000. [Consultado el: 15 de febrero de 2007]. Disponible en: http://tramullas.com/documatica/2-4.html.

30. Visual Paradigm for UML Programación en castellano, Última actualización: 5 de julio de 2005. [Consultado el: 13 de mayo de 2007]. Disponible en: http://www.programacion.com/noticia/1363/.

31. VIZCAÍNO, A.; GARCÍA, F. Ó., et al. Prácticas Ingeniería del Software 3. Una Herramienta CASE para ADOO: Visual Paradigm Universidad de Castilla- La Mancha [Consultado el: 25 de enero de 2007].

GLOSARIO DE TÉRMINOS

Actividad biológica: Actividad que caracteriza el comportamiento biológico en compuestos químicos (Molécula o Fragmento).

API: El API Java es una Interface de Programación de Aplicaciones (API: por sus siglas en inglés) provista por los creadores del lenguaje Java, y que da a los programadores un ambiente de desarrollo completo así como una infraestructura. La API Java está organizada en paquetes, donde cada paquete contiene un conjunto de clases relacionadas semánticamente.

CASE: Acrónimo inglés de Computer Aided Software Engineering, que significa Ingeniería de Software Asistida por Ordenador.

C++: Es un lenguaje de programación, diseñado a mediados de los 80, por Bjarne Stroustrup, como extensión del lenguaje de programación C.

GNU: es un acrónimo recursivo para «GNU No es Unix» y se pronuncia fonéticamente en español. El proyecto GNU fue lanzado en 1984 para desarrollar un completo sistema operativo tipo Unix, bajo la filosofía del software libre: el sistema GNU.

GPL: Licencia Pública General de GNU, llamada comúnmente GNU GPL, la usan la mayoría de los programas de GNU y más de la mitad de las aplicaciones de software libre.

Heteroátomo: Átomos distintos del Carbono e Hidrógeno.

IDE: Un "Integrate Development Enviroment" es una herramienta de soporte al proceso de desarrollo de software que integra las funciones básicas de edición de código, compilación y ejecución de programas, entre otras.

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

Java: Es un lenguaje de programación, de alto nivel, orientado a objetos y desarrollado por Sun Microsystems.

Matriz de conectividad: Matriz que se construye a partir de la conexión de cada átomo en la molécula con los adyacentes.

Open Source: Cualidad de algunos softwares de incluir el código fuente en la distribución del programa. En general se usa para referirse al software libre.

OREX: Expert System for Drug Design. Sistema Experto desarrollado en 1990 basado en la descomposición topológica de las moléculas en fragmentos estructurales y su asociación a las actividades biológicas.

PDE: Acrónimo inglés de Plugin Development Enviroment.

Plugins: Función o utilidad generalmente muy específica. Se adiciona a algún programa para ser ejecutado. Los plugins típicos tienen la función de reproducir determinados formatos de gráficos, reproducir datos multimedia, etc.

POO: Programación Orientada a Objeto**,** estilo de programación muy difundido en la actualidad por las ventajas que tiene.

RUP: Proceso Unificado (Rational Unified Process) metodología para el desarrollo de sistemas informáticos, dirigidos por casos de uso.

Topología Molecular: Es toda la información (y la única) que puede obtenerse de la conectividad mutua entre todos los pares de átomos en una molécula.

UML: Lenguaje de Modelado Unificado (Unified Model Languaje), lenguaje gráfico que brinda un vocabulario y reglas para especificar, construir, visualizar y documentar los artefactos de un sistema utilizando el enfoque orientado a objetos.

Windows: Microsoft Windows es el nombre de una familia de sistemas operativos no libres desarrollados por la empresa de software Microsoft Corporation.

XP: Metodología para el desarrollo de sistemas informáticos que se basa en el intercambio constante con el cliente.