V.3 SISTEMA EVOLUTIVO GRAFICADOR DE MOLÉCULAS ORGÁNICAS

*Jesús Manuel Olivares Ceja***

Resumen

Existen diversas propuestas para modelar moléculas orgánicas, se presenta una alternativa en donde se obtiene la representación gráfica a partir del nombre de la molécula. Los nombres se procesan utilizando una gramática con atributos obtenida de los ejemplos dados al sistema en un proceso evolutivo, esto es, al ir recibiendo más ejemplos, la gramática junto con su semántica se reestructura en caso necesario.

Palabras clave: Sistemas Evolutivos, Gramáticas, Moléculas Orgánicas, Graficación

INTRODUCCIÓN

Actualmente existen diferentes propuestas para graficar moléculas orgánicas, en algunos casos se especifican las coordenadas que conforman los átomos, en otros, se bosqueja la estructura mediante una interfase gráfica. Dos ejemplos de entre decenas de ellos son: Geom [10], MOGRA [9].

La propuesta presentada es una extensión a [2], en la cual se obtiene la representación gráfica de las moléculas orgánicas mediante una gramática obtenida a partir de las reglas de la IUPAC [1]. En este artículo, el mismo sistema ayuda para obtener la gramática y asociarle su semántica de acuerdo con el enfoque de los Sistemas Evolutivos [3][7].

---

* Jesús Manuel Olivares Ceja realizó este trabajo dentro del curso de Autómatas Celulares en el Centro de Investigación y de Estudios Avanzados del IPN Sección de Computación del Departamento de Ingeniería Eléctrica en mayo 1994

La modelación de moléculas es una herramienta útil para el análisis y diseño de experimentos, por ejemplo, en la industria farmacéutica cada día crecen las aplicaciones para diseñar o modificar medicamentos [9]. También es útil en cuestiones de simulación y cristalografía.

II El Lenguaje de los Nombres de Moléculas Orgánicas

La nomenclatura utilizada se basa en propuesta por la IUPAC [1]. Algunos nombres deben reexpresarse para explicitar las cadenas y subcadenas que lo componen para aplicar los mismos procesos lingüísticos descritos más adelante, esto es, en el caso de nombres comerciales o nomenclaturas distintas o especiales de la IUPAC. Para esto se utiliza una tabla de reescritura (figura 1) en donde se tiene el nombre reexpresable y su asociado. Conforme se ingresan datos que corresponden a casos nuevos la tabla crece o se reduce al eliminar algunos obsoletos. La oración a procesar es la resultante de la reescritura.

ORACION	SE REESCRIBE COMO
ISOBUTANO	2-METILPROPANO
NEOPENTANO	2,2-DIMETILPROPANO

Figura 1 Tabla de reescritura

Cada nombre de molécula después de reescribirse, se divide en unidades léxicas, las que no se conocen se dan de alta. A cada unidad léxica se le asocia un tipo (categoría léxica). Al conjunto de tipos de una oración se le nombra oración canónica. Los tipos utilizados se muestran en la figura 2.

TIPO	DESCRIPCIÓN
prefijo	MET, ET, PROP, etc.
a	ANO
e	ENO
l	IL
n	número
c	cuenta
,	guión
i	ignorar, se utiliza en unidades no significativas

Figura 2. Tabla de unidades léxicas

Dado que las familias de nombres orgánicos convergen en una terminación común de acuerdo con cada una: alcanos (ano), alquenos (eno), alcoholes (ol), etc., se ve la conveniencia de tomar la oración canónica en sentido inverso (figura 3), de tal forma que las producciones inician por el identificador de cada familia.

ORACIÓN	ORACIÓN CANÓNICA INVERTIDA
2,3-DIMETILHEXANO	a hex l met c - n
PINTA EL 2 -METILPROPANO	a prop l met - n
PENTANO	a pent
DIBUJA UN HEXANO	a hex

Figura 3. Oraciones canónicas invertidas

Se aplica distribución lingüística [5] en las oraciones para explicitar los atributos de subcadenas en una oración, como en el ejemplo del 2,3-dimetilhexano, el cual reexpresamos introduciendo paréntesis y los operadores algebraicos de unión (+) y concatenación (*) como sigue:

a hex ((l met) * (n1 + n2)) = a hex (l met n1 + l met n2)

se realizan las operaciones indicadas y se eliminan paréntesis, resultando:

a hex l met n1 l met n2

lo que equivale a:

2 MET IL 3 MET IL HEX ANO

Con las oraciones canónicas invertidas y distribuidas se construye una gramática inicial de la que se muestran sus producciones:

S → a hex l met n l met n
S →a prop l met n
S →a pent
S →a hex

En la obtención de la gramática, se debe detectar la recursividad [4] en producciones que tienen repeticiones internas para simplificarlas como en el caso de 2,3-dimetilhexano y similares en donde se tiene el factor de repetición "l met n". En este caso se coloca un punto (.) como la convergencia de la repetición para evitar la cadena vacía.

La producción S → a hex l met n l met n , al sustituirla por una recursiva se expresa como:

S → a hex X
X → l met n X | .

En la gramática se aplica la factorización para simplificar las repeticiones entre producciones (externas), ejemplo:

Sean las producciones:

S → a hex X
S → a prop X
S → a pent
S → a hex
X → l met n X | .

al factorizar "a" resulta

S → a Y
Y → hex X
Y → prop X
Y → pent
Y → hex
X → l met n X | .

En algunos casos se requiere que se aplique la sustitución de una producción por su símbolo no terminal que lo representa antes de aplicar alguna operación lingüística. Sean por ejemplo las producciones:

S → A | B
A → a X

X → met
X → prop
X → dec
B → l Y
Y → et
Y → but

Si llega la oración canónica "a dec l et n l et n l et n" aplicando sustitución queda como A B n B n B n, en donde es posible aplicar la operación de recursividad, dando lugar a las producciones:

S → A Z
Z → B n Z | .

con las mismas producciones A y B, porque sólo cambio S y se introdujo Z.

III ASOCIACIÓN DE SEMÁNTICA

A cada elemento de la gramática se le asocia un objeto gráfico que puede consistir de un átomo u otra molécula. Esto se indica delante del final de cada producción terminal en la gramática, dando lugar a lo que se nombra como gramática con atributos.

La semántica se expresa siguiendo la notación que consiste de alguna molécula o átomo existente y una lista de enlaces con otras moléculas que la complementan

(MOLECULA_1 (átomo_1 enlace_1
MOLECULA_2 átomo_2 enlace_2)
. . .
(átomo_1 enlace_1 MOLECULA_n átomo_n enlace_n))

en donde

MOLECULA_i: puede ser un átomo o una molécula;

átomo_i: es el número del átomo con que se enlaza a la MOLECULA_j. Su valor es relativo a la molécula a la que pertenece, si se une a otra molécula su valor puede cambiar
enlace_i: es el número del enlace utilizado, siempre es menor ó igual a la valencia del átomo enlazante. El enlace puede ser sencillo, doble o triple. Su valor se conserva en cualquier molécula en la que se coloque.

Ejemplos:

Sea el caso del metano construido a partir de átomos, uno

de carbono con hibridación sp3 y tres hidrógenos:

S → a met : (Csp3 (1 1 H 1 1) (1 2 H 1 1) (1 3 H 1 1))

Sea el caso del 2-METILHEXANO reconocido por la gramática con producciones:

S → A B
A → a hex : (hexano)
B → l met : (metil 1 3)

su interpretación es la composición de las cadenas que la forman, como sigue:

(hexano (2 1 metil 1 3)).

el radical metil ya incluye el átomo y enlace disponible para unirse a una cadena principal, en este caso en el carbono 2 que es la posición donde se indicó que va el radical usando su primer enlace del carbono.

Los átomos constituyentes de las moléculas se describen mediante vectores en el espacio, descritos por sus coordenadas cartesianas (x, y, z) teniendo el origen (0, 0, 0) como el centro del átomo. Al concatenarse con otros se rota y su centro se traslada según se requiera. Cada enlace puede ser sencillo, doble ó triple; se unen únicamente con su correspondiente: sencillo-sencillo, doble-doble ó triple-triple.

La gramática con atributos obtenida se expresa formalmente como:

G = <N, T, P, S>

donde

S es el símbolo inicial

T = {a, hex, prop, pent, met, n, ...} es el conjunto de terminales, identificados por iniciar con minúscula o ser signo de puntuación

N = {S, X, Y, ...} es el conjunto de variables, identificados por iniciar con mayúscula

P es el conjunto de producciones que incluye la interpretación en término de elementos gráficos

como se ha visto el número de terminales, no terminales y producciones está sujeto a cambios constantes, conforme se dan de alta nombres de moléculas.

IV SISTEMA EVOLUTIVO GRAFICADOR DE MOLÉCULAS ORGÁNICAS

El sistema se compone de los módulos siguientes:

1) Tabla de reescritura, para adecuar los nombres comerciales y aquellos aceptados por la IUPAC para simplificar otros.

2) Constructor Léxico, el cual le asocia su tipo a cada unidad léxica del nombre de la molécula. También da de alta aquellos que no se tengan registrados y formen parte del lenguaje. Esto permite que inicie su operación sin conocer las palabras del lenguaje. El resultado es la oración canónica invertida de la oración de entrada.

3) Constructor sintáctico-semántico, realiza el reconocimiento sintáctico del nombre de la molécula al tiempo que ubica su interpretación semántica. Si la oración canónica NO es reconocida y es válida de acuerdo con el usuario, se integra a la gramática aplicando alguna o varias de las operaciones

lingüísticas: sustitución, distribución, factorización, recursividad. Si le hace falta algún componente para su interpretación se le solicita al usuario que lo proporcione, una vez integrada continua el proceso de interpretación.

4) Presentación visual, una vez reconocido e interpretado el nombre de la molécula, a sus componentes gráficos se le aplican operaciones de rotación y traslación para examinar la molécula.

V OBTENCIÓN DEL NOMBRE A PARTIR DE SU REPRESENTACIÓN GRÁFICA

Dado el vínculo que se tuvo con el Departamento de Química del CINVESTAV-IPN, surgió el requerimiento de un sistema para obtener el nombre de una molécula dada su representación gráfica, en este sentido, Oscar Zarza Villavicencio, cursando el segundo semestre de la licenciatura en Informática en IPN-UPIICSA, realizó un sistema en donde la estructura de la molécula se representa con un grafo. Sobre el grafo se encuentra el camino más largo del grupo funcional que caracteriza a la molécula, mismo que establece el nombre principal. Aplicando recursivamente la búsqueda de caminos más largos se encuentran las subcadenas. Concatenando los nombres y posiciones encontrados se obtiene el nombre de la molécula.

CONCLUSIONES

Se han mostrado las características del lenguaje de las moléculas orgánicas junto con los módulos de un sistema para su validación gramatical y su interpretación gráfica. Es una alternativa en la modelación que las industrias requieren para analizar y proponer nuevas moléculas Se mencionó también una alternativa para la generación automática de nombres de moléculas.

REFERENCIAS BIBLIOGRÁFICAS

[1] IUPAC, International Union of Pure an Applied Chemistry, "Nomenclature of Organic Chemistry", Editorial Advisory Board H. W. Thompson, London, 1971
[2] J. M. Olivares C., H. V. McIntosh, "Graficación Tridimensional de Moléculas Orgánicas a Partir de su Análisis Lingüístico" (por editarse), CINVESTAV-IPN, México, D.F., 1994
[3] F. Galindo S., "Sistemas Evolutivos: Nuevo Paradigma de la Informática" en Memorias de la XVII Conferencia Latinoamericana de Informática, Caracas, Venezuela, julio 1991
[4] E. Berruecos R., "Sistema Evolutivo Generador de Esquemas Lógicos de Base de Datos", IPN-UPIICSA, México, D. F., 1990
[5] J. M. Olivares C., "Sistema Evolutivo para Representación del Conocimiento", IPN-UPIICSA, México, D. F., 1991
[6] C. Olicón N., "Sistema Evolutivo Generador de Paisajes", IPN-UPIICSA, México, D. F., 1992
[7] F. Galindo S., "Sistemas Evolutivos" en Boletín de Política Informática, INEGI-SPP, México, D. F., septiembre 1986
[8] N. Chomsky, "Estructuras Sintácticas", 9na edición, introducción, notas, apéndice y traducción de C. P. Otero, Editorial Siglo XXI, México, 1987
[9] N. V. Ramana, I. Gosh, "Software report: Molecular graphics software MOGRA" en Computer & Graphics Vol. 17 No. 4, Great Britain, 1993
[10] H. V. McIntosh, "Geom for drawing spheres and things", Departamento de Aplicación de Microcomputadoras, Instituto de Ciencias de la UAP, Puebla, Puebla, México, noviembre 1993