CHEMBIOMOL-07 : Chem. Biol & Med. Chem. Congress, Bilbao-Rostock, Germany, Galveston, USA, 2021.

SMILES Testing in Chemoinfortics software for prediction of intermolecular a-amidoalkylation reactions

Shan He [a], Sonia Arrasate [a], Humberto González-Díaz [a,b,] and Paula Carracedo[c]*

[a] Department of Organic and Inorganic Chemistry, Faculty of Science and Technology University of Basque Country UPV-EHU, 48940, Leioa, Basque Country, Spain.
[b] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.
[c]Department of Computer Science and Information Technologies, Faculty of Informatics, University of Coruña UDC, 15071, A Coruña, Spain.

Resumen

Los códigos SMILES son una especificación en forma de notación en línea para describir la estructura de las especies químicas empleando cadenas ASCll (American Standard Code for Information Interchang) cortas.[1]

Contiene la misma información que se puede encontrar en una tabla de conexión ampliada, pero presenta mayor utilidad ya que se trata de una construcción lingüística en lugar de una estructura de datos de ordenador. Otra propiedad importante del SMILES es que es bastante compacto en comparación con la mayoría de los otros métodos de representación de estructuras e implica menor espacio de archivo. Estas propiedades abren muchas puertas al programador de información química. Por ejemplo:

- Claves para el acceso a la base de datos
- Mecanismo para que los investigadores intercambien información química
- Sistema de entrada de datos químicos
- Parte de lenguajes para inteligencia artificial o sistemas de expertos en química.

Abstract

SMILES codes are a specification in the form of online notation to describe the structure of chemical species using short ASCll (American Standard Code for Information Interchang) strings.

It contains the same information that can be found in an extended connection table, but is more useful as it is a linguistic construct rather than a computer data structure. Another important property of SMILES is that it is quite compact compared to most other methods of representing structures and involves less file space. These properties open many doors to the programmer of chemical information. For instance:

- *Keys to access the database*
- *Mechanism for researchers to exchange chemical information*
- *Chemical data entry system*
- *Part of languages for artificial intelligence or chemistry expert systems.*

In this work, SMILES codes of all the compounds that participate in the intermolecular α-

[1]Weininger, D.; Weininger, A.; Weininger, Joseph L. "SMILES. 2. Algorithm for generation of unique SMILES notation". *Journal of Chemical Information and Modeling*. **1989**, 29 (2): 97–101.

En este trabajo, se utilizan los códigos SMILES de todos los compuestos que participan en la reacción de α-amidoalquilación intermolecular para el cálculo de los descriptores moleculares de la cadena de MARKOV, que posteriormente serán sustituidos en la ecuación $ee_R(\%)_{calc}$ proveniente del modelo de regresión implementado en el software MATEO para predecir el *ee*(%). Por lo tanto, en la prueba de software el reconocimiento y la identificación de los SMILES resultan de vital importancia. Además, durante el transcurso de la verificación y el chequeo del programa se han descubierto algunos aspectos deficientes relacionados con la programación del software.

En relación a los errores encontrados en el software MATEO para la identificación específica de algunos SMILES (Figura 1), se destaca para el caso 1 error en el cierre de anillo. Este defecto se produjo en el procedimiento de aprendizaje del Excel, al extender el mismo SMILES para el resto de las reacciones. Como la terminación del dicho SMILES es "1" el Excel lo reconoció como un número y se expandió este fallo para el resto de las reacciones. Aunque, fue un desliz ocasionado por el experimentalista, el programa no fue capaz de identificar particularmente el SMILES erróneo.

A pesar de que el SMILES de los alquenos en este trabajo no plantea serios problemas por la ausencia de los mismos en las reacciones estudiadas, pero si habría que tenerlos en cuenta en caso de que este modelo se extienda a otros tipos reacciones.

Para el caso 3, inicialmente el software no consiguió reconocer el SMILES debido a la presencia de la almohadilla, la cual es indicativo de un enlace triple. Este problema ha sido corregido y solucionado por Carracedo-Reboredo *et al*..

Además, el programa no tiene en cuenta la quiralidad de las moléculas (caso 4), esto se ha conseguido remediar parcialmente mediante la multiplicación de los resultados de *ee*(%) por la

amidoalkylation reaction are used to calculate the molecular descriptors of the MARKOV chain, which will later be substituted in the equation $ee_R(\%)_{calc}$ *from the Regression model implemented in MATEO software to predict ee(%). Therefore, in software testing the recognition and identification of SMILES is of vital importance. Furthermore, during the course of the verification and testing of the program, some weak aspects related to the programming of the software have been discovered.*

In relation to the errors found in the MATEO software for the specific identification of some SMILES (Figure1), an error in the ring closure stands out for case 1. This defect occurred in the Excel learning procedure, when extending the same SMILES for the rest of the reactions. As the ending of the said SMILES is "1" Excel recognized it as a number and expanded this bug for the rest of the reactions. Although, it was a slippage caused by the experimentalist, the program was not able to particularly identify the wrong SMILES.

Although the SMILES of alkenes in this work does not pose serious problems due to their absence in the reactions studied, but they should be taken into account if this model is extended to other types of reactions.

For case 3, initially the software failed to recognize the SMILES due to the presence of the pad, which is indicative of a triple bond. This problem has been corrected and solved by Carracedo-Reboredo et al ..

Furthermore, the program does not take into account the chirality of the molecules (case 4), this has been partially remedied by multiplying the results of ee(%) by the chirality of the catalyst (+/-) 1. Although, there is no great significance in the prediction of ee(%) for the reactions studied in this work since no chiral substrates have been reported in the literature for

quiralidad del catalizador (+/-)1. Aunque, no hay gran trascendencia en la predicción de *ee*(%) para las reacción estudiadas en este trabajo ya que hasta ahora no se han reportado sustratos quirales en literatura para las reacciones de α-amidoalquilación intermolecular enantioselectivas. Sin embargo, se sugiere optimizar el software para ampliar su uso hacia futuras reacciones con reactivos quirales.

Por otro lado, se probaron diferentes alternativas de representaciones de SMILES para un mismo compuesto, concretamente el grupo nitro y los grupos aromáticos. Como resultado de dicho análisis, el software solo fue capaz de reconocer para el grupo nitro la primera opción mientras que para los grupos aromáticos ambas alternativas eran identificables.

Finalmente, se examinaron la posibilidad de reconocimiento de SMILES de compuestos enlazados no covalentemente (enlaces de hidrógeno (caso 7) y enlaces iónicos (caso 8), puesto que es común encontrarlos en la base de datos original bien sea la unión entre el resto de disolvente con el sustrato, en dicho caso es fácil de identificar por el gran tamaño del sustrato en comparación con el disolvente o la agrupación entre el disolvente y una impureza, en esta situación es más laborioso reconocer qué porción respecta al disolvente por la similitud del tamaño entre estas moléculas. En este estudio, se demostró que el programa no fue capaz de considerar este tipo de SMILES, por lo que en un trabajo previo se realizó una limpieza manual de los mismos y además resulta imposible identificar qué parte del complejo corresponde al disolvente y la porción referente a la impureza. A la vista de este inconveniente, se propone una limpieza automatizada por el software, dado que los códigos de SMILES no solo se pueden utilizar para las reacciones de α-amidoalquilación, sino que es posible extender hacia otros tipos de reacciones. Una manera de ampliar el uso del MATEO consiste en el desarrollo de nuevos modelos

enantioselective intermolecular α-amidoalkylation reactions. However, it is suggested to optimize the software to extend its use towards future reactions with chiral reagents.

On the other hand, different alternatives of SMILES representations were tested for the same compound, specifically the nitro group and the aromatic groups. As a result of this analysis, the software was only able to recognize the first option for the nitro group, while for the aromatic groups both alternatives were identifiable.

Finally, the possibility of SMILES recognition of non-covalently linked compounds (hydrogen bonds (case 7) and ionic bonds (case 8) was examined, since it is common to find them in the original database, either the union between the rest of solvent with the substrate, in this case it is easy to identify by the large size of the substrate compared to the solvent or the grouping between the solvent and an impurity, in this situation it is more laborious to recognize which portion refers to the solvent by the similarity of the size between these molecules. In this study, it was shown that the program was not able to consider this type of SMILES, so in a previous work a manual cleaning of them was carried out and it is also impossible to identify which part of the complex corresponds to the solvent and the reference portion to impurity. In view of this drawback, an automated cleaning by the software is proposed, since the SMILES codes can not only be used for α-amidoalkylation reactions, but it is possible to extend to other types of reactions. One way of expanding the use of MATEO consists of the development of new chemoinformatic models for the prediction of the chemical reactivity of other reactions and in this sense, although the error in the SMILES code in this master's thesis is not so important due to the scarcity of cases of non-covalently linked compounds, it is necessary to

quimioinformáticos para la predicción de la reactividad química de otras reacciones y en este sentido, aunque el error del código de SMILES en este trabajo de fin de master no tiene tanta importancia por la escasez de casos de compuestos enlazados no covalentemente, es necesario solventarlo como fuente de código SMILES que permanece para nuestro grupo de investigación.

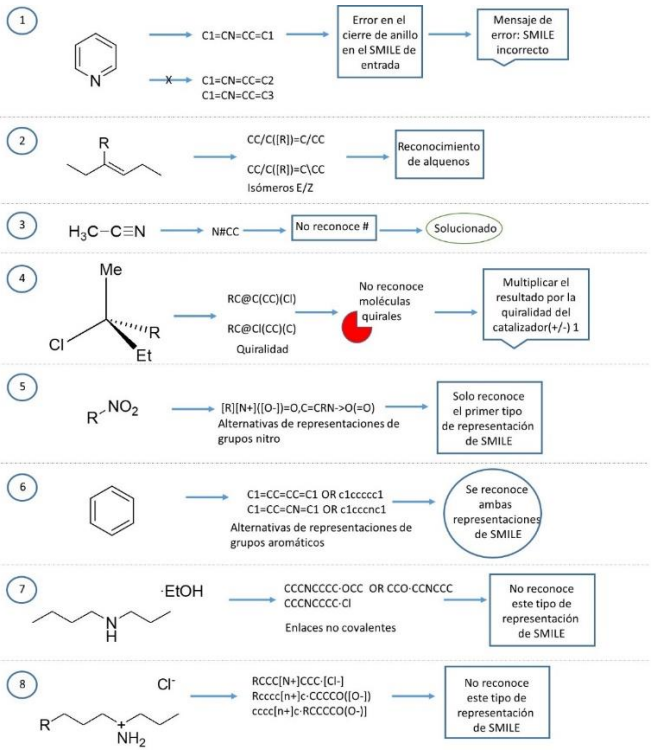

Figura 1: Problema de reconocimiento de algunos SMILES por el software MATEO.

solve it as a source of SMILES code that remains for our research group.

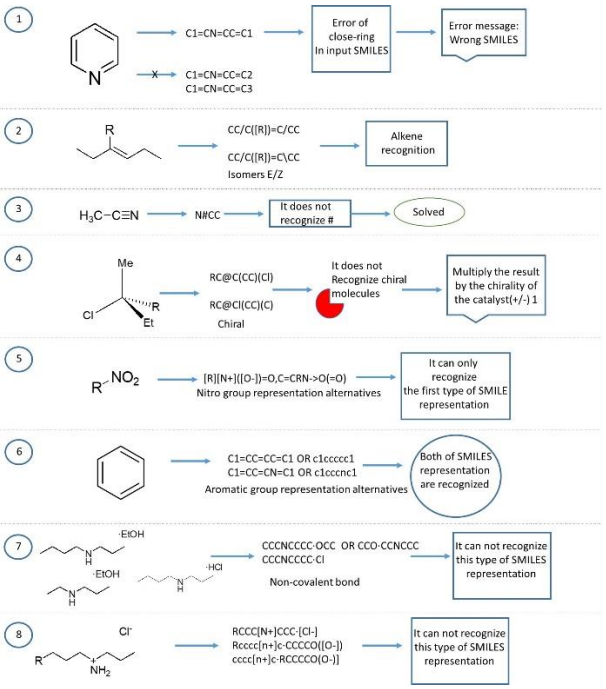

Figure 1: Problem of recognition of some SMILES by the MATEO software.