match no. 2 pp. 37-43 1976

A NEW APPROACH TO THE GENERATION OF ISOMERS

Ödön Mestyanek[1]* and Péter Réti[2]

1. Budapest Technical University, Department of Inorganic Chemistry.
2. Computer Application R and D Centre of the Chemical Industry, Budapest, Hungary

(received: June 1976)

Abstract. The well-known problem of the generation of isomers are treated by the methods of mathematical linguistics. The main result is that all acyclic compounds can be generated by generalized grammars.

1. Introduction

There is a necessity to know all the structural isomers of a chemical compound given by its molecular formula. These problems are usually treated by graph-theoretical methods /see Balaban [1]/.

We suggest that the apparatus of mathematical linguistics be used for the solution of the generation of isomers. We feel that the advantage of this method lies in its simplicity and its easy and efficient application even on a small or medium-sized computer.

The subject of this work is only the generation of aliphatic compounds and we did not want to make our discussions complicated that is why the simplest such compound the methane and its derivatives are excluded from our research.

The next section will be dealing with coding the chemical compounds in a simple but adequate way. In the 3, 4, 5 sections alkanes, alkenes, substituated compounds are disserted respectively. The notations used in this paper are the same as in the standard textbooks /see for instance [2]/.

* To whom all correspondence shoud be addressed

2. Coding of aliphatic compounds

To solve the problem posed in the introduction one needs to code chemical compounds. In this section a linear code will be formulated which has a close relationship with the chemical nomenclature. Although there are several linear codes in the literature, however this one is very simple and sufficient to our goal.

There are essentially two parts of a chemical compound from the point of view of the generation of isomers: the chains consisting of carbon /C/ atoms and the functional groups. The internal structure of the latter we are not interested in.

The carbon atoms in the chain and the functional groups can be classified according to the number of bonds they are joined to non-H atoms. A C atom can be P /it is linked by one bond to non-H atom/, S /two bonds/, T /three bonds/, Q /four bonds/. It is possible to introduce similar notation to functional groups but since we restrict ourselves to monovalent groups we do not do that.
Let k/x/ denote the number of x type C atoms in one molecule. The following definitions are necessary to formulate the coding system.

Definition 1. The elements of

$$A = \left\{ \Psi = P^{\Psi} P \mid k(P) = 2 + k(T) + 2k(Q) \;, \Psi \in \left\{ P,S,T,Q \right\}^{*} \right\}$$

are called alkanes.

It can be seen that a bijection exists between A and the set of all structural formulae of saturated acyclic hydrocarbons.

The next definition generalizes definition 1 to compounds having double bonds as well.

Definition 2. The elements of

$$B = \left\{ \Psi \in \left\{ P,S,T,Q \right\}^{*} \mid k(P)+2k(S)+3k(T)+4k(Q) = 2(n+i-1) \right\}$$

are called n-i-alkenes.

/n, i are the number of C atoms and double bonds respectively, 1 triple bond counts 2 double bonds./

The coding as it is given in definition 2 easy to reconstruct that it can cope with such cases when one or more H atoms of the carbon chain are substituted by a monovalent functional group /nominated by F, their number in one molecule is k/F/ /. There are m different such groups

$$F = \left\{F_i\right\}_{i=1}^{m} .$$

Definition 3. The elements of

$$C=\left\{\psi \in \left\{P,S,T,Q,F\right\}^* \mid k(P)+2k(S)+3k(T)+4k(Q)-k(F)= 2(n+i-1)\right\}$$

are called n-i-k/F/ alkenes.

3. The generation of alkane isomers

The solution of this problem is interesting not only in itself since other aliphatic structures are formed from generated alkanes.

Let $G_1=(V,\Sigma,K,R_o)$ be a grammar where

$$\Sigma = \left\{P,S,T,Q\right\} ,$$

$$V = \Sigma \cup \left\{K\right\} ,$$

$$R_o = \begin{cases} r_o : K \to P\psi P \\ r_1 : S\eta\ S \to T\eta\ P \\ r_2 : S\varrho\ S\varphi S \to Q\varrho P\varphi P \end{cases}$$

$$\eta,\varrho,\varphi,\psi \in \left\{S,\Lambda\right\}^* .$$

The following theorem will give an insight to the relationship between grammars and generation of isomers.

Theorem 1.

$L/G_1/$ is equal to the set of all isomers of alkanes.

Proof.

1. $L/G_1/$ contains only alkanes because the product of the rule r_0 is alkane /see definition 1, all conditions are satisfied/, and r_1 and r_2 leave invariant the equation for the number of carbon atoms.

2. The whole set of alkanes is contained in $L/G_1/$. Let us take an arbitrary word $\Psi \in A$. It will be shown that $\Psi \in L/G_1/$. We are searching for such words from which Ψ can be derived by using either rule r_1 or rule r_2. If such a word exists then let us denote it by Ψ_1, if not then Ψ can be derived from K by applying the rule r_0. In Ψ_1 $k/S/$ must be larger and $k/T/$ or $k/Q/$ smaller then in Ψ but $\Psi_1 \in A$. This procedure can be continued up to that point when there is not any Q and T left in the word. Then this word can be derived from K by using the rule r_0, that is $\Psi \in L/G_1/$.

4. Generation of n-i-alkenes

To solve the problem mentioned in the title of this section we need some definitions about generalized grammars. There are plenty of them in the literature dealing with mathematical linguistics. Our notions follow the line proposed by Stotzki /see [3]/.

Let $G=(V,\Sigma,K,R)$ be a grammar where the rules in R are labelled by symbols from another alphabet Σ'. /We have done it up to now as well but only for the sake of convenience. Henceforth it belongs to the essence of the discussion./

Let $\alpha,\beta \in V^*$ be two words and $\alpha \overset{*}{\Rightarrow} \beta$.

Definition 4.

The characteristics of a derivation $\alpha \overset{*}{\Rightarrow} \beta$ is a word $\varphi_1 \varphi_2 \cdots \varphi_n \in (\Sigma')^*$ where φ_i is the label of the derivation $x_{i-1} \rightarrow x_i$ $(x_0=\alpha\ ,\ x_n=\beta)$.

Let $G=(V,\Sigma,K,R)$ and $G'=(V',\Sigma',K',R')$ be two grammars.

Definition 5.

The ordered pair of grammars $GG=/G,G'/$ is called generalized grammar if Σ' contains the labels of the rules in R and in G only such derivations are permitted of which at least one characteristics is contained in $L/G'/$.

Let $G^{n,1}=(G, G')$ be a generalized grammar where $G=(V,\Sigma,K,R)$, Σ, V, K are the same as in the previous section,

$R = R_0 \cup \{r_3, r_4, r_5, r_6, r_7\}$,

r_3 : SP→TS
r_4 : SS→TT
r_5 : TP→QS
r_6 : TS→QT
r_7 : TT→QQ

It is worth mentioning that r_3-r_7 are such context-free rules that on the left-hand side terminal symbols stand only. According to a well-known theorem of mathematical linguistics these rules can be altered that only grammatical symbols remain on the left-hand side. However we do not do this alteration for the sake of simplicity.

$G'=(V',\Sigma',K',R')$ where $\Sigma'=\{r_0, r_1, r_2, \ldots, r_7\}$,

$V'=\Sigma' \cup \{K', A, B_1, B_2, B_3, \ldots, B_i\}$,

$$R'=\begin{cases} r_0' : K' \to r_0 A \\ r_1' : A \to r_1 A \\ r_2' : A \to r_2 A \\ r_3' : r_{i\in I}{}^{A} \to r_{i\in I} r_{j\in J} B_1 \quad I=\{1,2\},\ J=\{3,4,5,6,7\} \\ r_4' : B_1 \to r_{j\in J} B_2 \\ r_5' : B_2 \to r_{j\in J} B_3 \\ \quad \vdots \\ r_{i+2}' : B_{i-1} \to r_{j\in J} \end{cases}$$

/For instance, r'_4 means that rewriting can be made for all $j \in J$./

Theorem 2.

Supposing that in r_0 $|\Psi| = n-2$, then $L/G^{n,i}/$ is equal to the set of all n-i-alkenes.

Proof.

/It is only the skeleton of the proof. The detailed one is very similar to that of the last theorem./

There are only n-i-alkenes in $L/G^{n,i}/$ because of the structure of $G^{n,i}$. It is true that r_3 through r_7 give the whole set of possible transformations /this statement is of chemical nature/ therefore $L/G^{n,i}/$ contains all the n-i-alkenes.

5. Generation of n-i-l-alkenes

Here the notations used previously are not defined only the new ones are explained.

Let $G = (V, \Sigma, K, R)$ be a grammar where

$$\Sigma = \{P, S, T, Q, F_1, F_2, \ldots, F_m\},$$

$$V = \Sigma \cup \{K\},$$

$$R = \{r_0, r_1, r_2, \ldots, r_7, r_8, r_9, r_{10}\},$$

/R is understood to contain not only the labels but the rules themselves as well/, the new rules are:

$$r_8 : P \rightarrow SF_{j \in J'}$$

$$r_9 : S \rightarrow TF_{j \in J'}$$

$$r_{10} : T \rightarrow QF_{j \in J'}$$

$$J' = \{1, 2, 3, \ldots, m\}.$$

Let $G' = (\Sigma', V', K', R')$ be another grammar where $\Sigma' = \{r_i\}_{i=1}^{10}$ is the set of characteristics for G,

$$V' = \Sigma' \cup \{K', A_1, A_2, A_3, \ldots, A_l, C_1, C_2, \ldots, C_\ell\},$$

$$R' = \{r'_0, r'_1, r'_2, \ldots, r'_{i+1}, r'_{i+2}, r'_{i+3}, \ldots, r'_{i+\ell+1}\}$$

and the definition of the new rules is the following

$$r'_{i+2} : B_{i-1} \longrightarrow r_{j\in J} C_1$$
$$r'_{i+3} : C_1 \longrightarrow r_{i\in I'} C_2 \qquad I' = \{8,9,10\}$$
$$r'_{i+4} : C_2 \longrightarrow r_{i\in I'} C_3$$
$$\vdots$$
$$r'_{i+\ell+2} : C_\ell \longrightarrow r_{i\in I'}$$

Let $G^{n,i,\ell} = /G,G'/$ be a generalized grammar.

Theorem 3.

Supposing that in r_o $|\psi| = n-2$ then $L/G^{n,i,\ell}/$ is equal to the set of n-i-ℓ-alkenes.

The proof of the theorem 3 is completely analogous to that of the theorems 1 and 2.

Remark.

The elements of $L/G^{n,i,\ell}/$ are not isomers of proper chemical sense, here F_i-s are regarded as one substance. A slight modification of the $G^{n,i,\ell}$ grammar gives the isomers corresponding to the usual chemical convention.

Bibliography

1 A.T. Balaban, Match 1, 123 /1975/

2 Barron Brainerd: Introduction to the Mathematics of Language Study. American Elsevier, New York, 1971.

3 E.D. Stotzki: Generalized Grammars, in "Research in the Field of Mathematical Linguistics" /in Russian/ Nauka, Moscow, 1972.