match no. 22 pp. 67-76 1987

ENUMERATION AND CONSTRUCTION OF MOLECULAR AND RATIONAL FORMULAS BY MEANS OF GENERATING FUNCTION

Tetsuo Morikawa
Department of Chemistry, Joetsu University of Education,
Yamayashiki, Joetsu, Niigata Prefecture 943, Japan

(Received: May 1987)

(Abstract) A new method is presented which enables us to find both the number and the explicit form, of all possible molecular (and/or rational) formulas with given positive integers. The method consists of three steps; first, we construct an algebraic function that generates a class of molecular formulas satisfying Senior's graph-theoretical conditions in question; second, each of the atomic symbols in the function is replaced with a parameter to a power that is decided by a given integer; last, the generating function just obtained is expanded into a power series whose coefficients answer the question.

Introduction. In the present note the number of each of atoms/elements in molecular formulas is expressed for convenience by an exponent although chemists usually use a subscript ; for example, the molecular formula C_2H_6 (ethane in chemistry) is

rewritten as C^2H^6, namely the power of the atomic symbols C and H with exponents 2 and 6. An algebraic summation of the alkanes on the basis of the power expression then becomes

$$CH^4 + C^2H^6 + C^3H^8 + \ldots \text{ (ad inf.)} = CH^4/(1 - CH^2).$$

Conversely each term in the series expansion of the function $CH^4/(1 - CH^2)$ can be interpreted as an alkane. In other words, the algebraic sum may be called "a generating function" for the alkanes. Such a concept of generating function introduced here will play an essential role for enumeration and construction of molecular formulas. Note that the term "molecular formula" in this paper means only a collection of the chemical elements and the number of atoms of each present in a molecule, and that it must be distinguished from the term "molecular structure (graph)".

As the above example shows, molecules are associated with positive integers, such as number of atoms, number of bonds, mass number, molecular number[1], and number of functional groups. In the identification of unknown chemical substances based on instrumental analysis chemists must determine the corresponding molecular (and/or rational[2]) formulas with integers of this kind. The object of this report is to set up a method for the determination.

Evidently we have to impose some restraints on a given set of atoms (a given molecular formula) in order to give it chemical meaning. In his paper on mathematical graph theory Senior[3] has established a necessary and sufficient theorem under which

a given set of atoms with valencies v_i is realizable as at least one molecular structure (graph). Here the valency of an atom is defined as the number of edges/bonds which meet at that atom. Senior's conditions: (i-1) The number of atoms with odd valency is even. (i-2) $\sum v_i \geq 2v_{max}$, where v_{max} stands for the maximal valency in the set. (ii) $\sum v_i \geq 2(n_A - 1)$, n_A being the number of atoms in the set. This note will deal with molecular formulas in such a way that they fulfill Senior's conditions.

Generating Function for Hydrocarbon Molecules

Let us derive a generating function for hydrocarbon molecules that are typical compounds in organic chemistry. Assume that carbon and hydrogen atoms have valency 4 and 1, respectively. Senior's conditions then become (i-1) n_H = even, (i-2) $4n_C + n_H \geq 8$, and (ii) $2n_C + 2 \geq n_H$, where n_C and n_H denote the number of carbons and hydrogens, respectively. Each set of C and H that satisfies Senior's conditions for n_C = 1, 2, 3, ..., can be listed step by step as follows:

$$
\begin{aligned}
& CH^4 \\
& + C^2H^6 + C^2H^4 + C^2H^2 \\
& + C^3H^8 + C^3H^6 + C^3H^4 + C^3H^2 \qquad \text{.....(1)} \\
& + C^4H^{10} + C^4H^8 + C^4H^6 + C^4H^4 + C^4H^2 \\
& + C^5H^{12} + C^5H^{10} + C^5H^8 + C^5H^6 + C^5H^4 + C^5H^2 \\
& + \ldots.
\end{aligned}
$$

The addition of these terms along each of the columns gives

$$\frac{CH^4}{1 - CH^2} + \frac{C^2H^4}{1 - CH^2} + \frac{C^2H^2}{1 - CH^2} + \frac{C^3H^2}{1 - CH^2} + \frac{C^4H^2}{1 - CH^2} + \ldots.$$

It can be seen easily that the first, the second, the third, ..., fractions represent the alkanes, the alkenes, the alkynes, etc. in the power expression. A straightforward calculation of this series gives $CH^2(H^2 + C - C^2H^2)(1 - C)^{-1}(1 - CH^2)^{-1}$.

To the above function two dummy molecules $H^2 + CH^2$ are added for simplification; then we arrive at the generating function required, and denote it as $G(C, H^2)$;

$$\frac{H^2}{(1 - C)(1 - CH^2)} = G(C,H^2).$$

The addition of CH^2 suggests the removal of the restraint (i-2). Each term in $G(C, H^2)$ is the multiplication of C to the power n_C by H^2 to the power $n_H/2$, where $n_C + 1 \geqq n_H/2$, $n_C \geqq 0$, $n_H/2 \geqq 1$. ($n_H/2$, integer). Note that $G(C, H^2)$ represents a convolution[4] because it consists of two generating functions $1/(1 - C)$ and $H^2/(1 - CH^2)$, and that the latter generating function equals the alkanes plus H^2.

Enumeration and Construction of Hydrocarbon Molecules

In this section we solve five problems for counting hydrocarbon (molecular) formulas with a given integer.

1) Let the total number of atoms in a molecule be a given integer q. Each of the atomic symbols in a generating function for a given class of molecules in question should be replaced

with one parameter, say t; namely, for hydrocarbons, C is replaced with t^1, and H^2 with t^2 in $G(C, H^2)$ because $q = n_C + n_H$. Evaluating a power series of t obtained in this way is obviously equivalent to answering the problem: How many molecules totalling q atoms in set (1) are there ? Thus,

$$G(t, t^2) = \frac{t^2}{(1 - t)(1 - t^3)} = \sum_q g_{12}(q)t^q$$

$$= (t^2 + t^3) + t^4 + 2t^5 + 2t^6 + 2t^7 + 3t^8 + \ldots$$

The first two terms $(t^2 + t^3)$ result from the dummy molecules which were added to $G(C,H^2)$. The general answer $g_{12}(q)$ for a given integer q can be determined in terms of recurrence formulas:

$$1/(1 - t) = (1 - t^3)\sum g_{12}(q+2)t^q = \sum\left\{g_{12}(q+2) - g_{12}(q-1)\right\}t^q$$

which shows $g_{12}(q+2) - g_{12}(q-1) = 1$. If $q = 3k$, then $g_{12}(3k + 2) - g_{12}(3(k - 1) + 2) = 1$, which, by use of $g_{12}(3\cdot 0 + 2) = g_{12}(2) = 1$, give $g_{12}(3k + 2) = k + 1$. An argument similar to the case of $q = 3k$, if either $q = 3k + 1$ or $q = 3k + 2$, leads to $g_{12}(3k + 1 + 2) = k + 1$ or $g_{12}(3k + 2 + 2) = k + 1$. Then we have

$$g_{12}(q) = k + 1, \quad q - 2 = 3k + r, \quad 0 \leq r < 3$$

namely, the division of $q - 2$ by 3 leads to the quotient k and

the remainder r. The explicit forms of the k + 1 hydrocarbons are then

$$C^{n}H^{m},\quad C^{n-2}H^{m+2},\quad \ldots,\quad C^{n-2k}H^{m+2k},$$

where the parameters n and m in the starting molecule $C^{n}H^{m}$ for construction can be determined by $n = 3k + r$ and $m = 2$ because $3k + r + 2 = q$.

2) The number of molecular formulas with a given number q of bonds can be counted in a similar manner as above. It follows from (i-1) that $2q = \sum v_i$. For hydrocarbons in set (1), $2q = 4n_C + n_H$ or $q = 2n_C + n_H/2$, and therefore, $G(t^2, t)$ is the answer to the problem: Enumerate the number of hydrocarbon molecules with q bonds. Thus,

$$G(t^2, t) = t(1 - t^2)^{-1}(1 - t^3)^{-1} = \sum g_{21}(q)t^q$$

$$= (t + t^3) + t^4 + t^5 + t^6 + 2t^7 + t^8 + \ldots$$

The period of the generating function $1/(1 - t^3)$ is 3, and 3 terms in the period are 1, 0, 0; similar consideration is carried out for $1/(1 - t^2)$. It follows from this statement that

$$g_{21}(q) = \begin{cases} k & (r < s) \\ k + 1 & (r \geq s) \end{cases}$$

where $q - 1 = 2(3k + r) + s$, $0 \leq r < 3$, $0 \leq s < 2$. The parameters of the starting molecule $C^{n}H^{m}$ for construction are given

by $n = 3k + r$ and $m = 2(s + 1)$; $C^{n}H^{m}$, $C^{n-1}H^{m+4}$, ..., $C^{n-(k-1)}H^{m+4(k-1)}$ for $r < s$, and $C^{n-k}H^{m+4k}$ is added if $r \geq s$.

An algorithm for getting k, r, and s in $g_{21}(q)$ is simple and easy to perform: Divide q - 1 by 2, and get the quotient and the remainder s; divide the above quotient by 3, and get the quotient k and the remainder r; compare r with s, and have the choice of k or k + 1.

3) The total number of valence electrons in a molecule is selected as a given integer q. For the set (1) of hydrocarbon molecules, $q = 4n_C + n_H$ or $q/2 = 2n_C + n_H/2$; then $G(t^4, t^2)$ for q, or $G(t^2, t)$ for q/2, is the answer.

4) Counting the number of molecular formulas with a given molecular number q is also an easy calculation. For the set (1) of hydrocarbons, $q = 6n_C + n_H$, or $q/2 = 3n_C + n_H/2$; thus $G(t^3, t)$ is the answer.

5) The total mass number q of molecules is given. For hydrocarbons in set (1), $q = 12n_C + n_H$ or $q/2 = 6n_C + n_H/2$; then the generating function for the answer of q/2 is $G(t^6, t)$.

The procedure described above for enumeration and construction of molecular formulas can be summarized, in the case of two parameters, as follows. Let us consider the problem of finding the number of solutions for a linear indeterminate (Diophantine) equation

$$ax + by = p \quad (a, b, p, \text{positive integers})$$

with two unknowns x and y under the inequality restraints x

$+ 1 \geq y$. The summation of the parameters X and Y in the form X^xY^y's, in which the exponents satisfy $x + 1 \geq y$, $x \geq 0$, and $y \geq 1$, is equal to $G(X,Y) = Y(1 - X)^{-1}(1 - XY)^{-1}$ because the inequality is obtained from the condition (ii) for hydrocarbons after setting $n_C = x$ and $n_H = 2y$. The technique for the solution of the problem is that of working with $G(t^a,t^b)$, (t, any parameter), rather than with the linear indeterminate equation itself; this is based on the observation that each term on the left-hand side of the linear indeterminate equation, say ax, means "X to the power a" in the power expression, and that the replacement of X by t means "counting of the total number of X ". That is to say, the coefficient of t^p in the power expansion of $G(t^a,t^b)$ is just the solution.

Theoretical treatment similar to that for two unknowns is possible for three (or more) unknowns.

Enumeration of Rational Formulas

We proceed to showing that the generating function method for molecular formulas is applicable to count the number of rational formulas (in a given molecular formula) with given functional groups (or with subgroups). This process is illustrated by use of two examples. The first selects >CH-CH< (valency 4) and $-CH^3$ (methyl, valency 1) as subgroups, which are substituted for C and H in $G(C, H^2)$; then

$$G(C^2H^2, C^2H^6) = C^2H^6(1 - C^2H^2)^{-1}(1 - C^4H^8)^{-1}.$$

The coefficient of C^nH^m in the series expansion of this function

represents the number of rational formulas with the two subgroups $>C^2H^2<$ and $-CH^3$ which constitute the hydrocarbon molecule C^nH^m.

If $>N-N<$ (valency 4) and $-NH^2$ (valency 1) are chosen as functional groups in N^nH^m, then $G(N^2, N^2H^4)$ becomes the required function.

Types of Generating Function for Molecular Formulas

The generating function $G(C, H^2)$ in the previous sections has been constructed in terms of C (valency 4) and H (valency 1). There are however many types of generating functions corresponding to many classes of molecules (or functional groups) containing atoms with several kinds of valencies.

None of the atoms with valency 2 make any contribution to Senior's conditions (i-1) and (ii), apparently. A generating function for such a kind of atom, say oxygen O, is hence written as the product of $G(C, H^2)$ by O; therefore, summing up $G(C,H^2)$, $O^1G(C, H^2)$, $O^2G(C, H^2)$, $O^3G(C, H^2)$, ..., and so on, leads to the function $G(C,H^2)/(1 - O)$ required.

For molecular formulas containing only nitrogen N (valency 3) and hydrogen H (valency 1), Senior's conditions (i-1) and (ii) are expressed as $n_N + n_H$ = even and $n_N + 2 \geqq n_H$. If $n_N = 2m_N$ and $n_H = 2m_H$, then $m_N + 1 \geqq m_H$; therefore, by the same argument as in hydrocarbons, the generating function $G(N^2,H^2)$ is obtained. If $n_N = 2m_N - 1$ and $n_H = 2m_H - 1$, then also $m_N + 1 \geqq m_H$; therefore, $N^{-1}H^{-1}G(N^2,H^2)$. Note: $m_N \geqq 0$, $m_H \geqq 1$.

When molecular formulas contain only C, H, and N, the following Senior's conditions are derived: (i-1) $n_H + n_N$ = even and

(ii) $2n_C + n_N + 2 \geq n_H$. If $n_H = 2m_H$ and $n_N = 2m_N$, then $n_C + m_N + 1 \geq m_H$. Thus the generating function for fixed m_N is

$$N^{2m_N} H^2(1 - CH^2)^{-1}\{(1 - C)^{-1} + H^2 + H^4 + \ldots + H^{2m_N}\}.$$

The summation of these functions over all m_N (= 0, 1, 2, ...) takes the form

$$G(C,H^2)(1 - CH^2N^2)H^{-2}G(N^2,H^2).$$

In a similar fashion if $n_H = 2m_H - 1$ and $n_N = 2m_N - 1$, then $n_C + m_N + 1 \geq m_H$; then, the generating function is the multiplication of the above function by $N^{-1}H^{-1}$.

Notes and References

[1]About the term "molecular number", refer to Brewster, R. Q.; McEwen, W. E. "Organic Chemistry"; Prentice-Hall : Englewood Cliffs, 1961, chap. 1, sec. 14.

[2]A rational formula for a given molecule is defined as a set of functional groups whose atoms constitute that molecule and whose valencies satisfy Senior's conditions.

[3]Senior, J. K., Amer. J. Math. 73, 663(1951).

[4]Liu, C. L. "Elements of Discrete Mathematics" ; McGraw-Hill: New York, 1977, chap. 6.