THE TEACHING OF MATHEMATICS

2008, Vol. XI, 2, pp. 53–61

THE CHEMICAL FORMULA C_nH_{2n+2} AND ITS MATHEMATICAL BACKGROUND

Ivan Gutman

Abstract. Already in the elementary school, on the chemistry classes, students are told that the general formula of alkanes is C_nH_{2n+2}. No proof of this claim is offered, either then or at any higher level. We now show how this formula can be proven in a mathematically satisfactory manner. To do this we have to establish a number of elementary properties of the mathematical objects called *trees*.

ZDM Subject Classification: K34; *AMS Subject Classification*: Primary: 05C05; Secondary 05C90.

Key words and phrases: Graph Theory, Trees, Chemistry, Alkanes.

1. Introduction

It is usually believed that chemistry is a science in which no knowledge of mathematics and no mathematical skills are needed, and that mathematics is a science in which no knowledge of chemistry and no chemical skills are needed. As a consequence, students talented for mathematics use to ignore and despise chemistry ("the stinky science") whereas students interested in chemistry are usually recruited among those for whom mathematics is "to hard to be understood".

The author of this article does not want to claim that there is a lot of interesting mathematics in chemistry and that students who love mathematics should focus their attention to chemical matters. Yet, there exist mathematically non-trivial themes in chemistry. One of these, chosen to be suitable for secondary-school students, is outlined in the present article.

2. The alkane formula

The simplest organic chemical compounds are the so-called *alkanes* (in older times also called *paraffins*). The chemical formulas of the eight smallest alkanes are found in Fig. 1.

Alkanes are "hydrocarbons" which means that their molecules consist only of carbon and hydrogen atoms. Alkanes are "saturated hydrocarbons" which means that the number of hydrogen atoms (for a given number of carbon atoms) is maximum possible. Alkanes are "acyclic", which means that their molecules contain no cycles.

In chemistry it is known that the general formula of alkanes is C_nH_{2n+2}, where, of course, n is a positive integer. This formula has to be understood as follows:

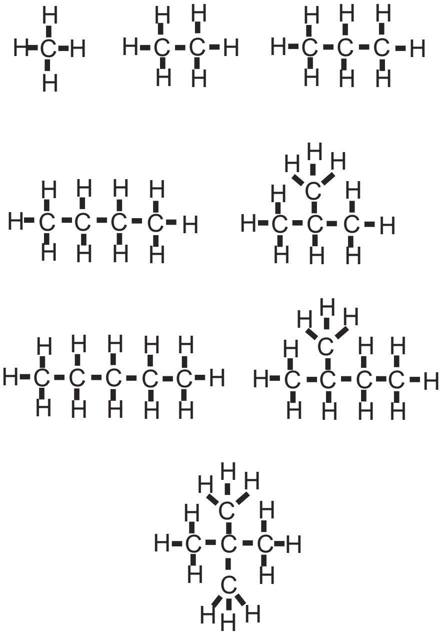

Fig. 1. The structural formulas of all alkanes with $n = 1, 2, 3, 4, 5$ carbon atoms. Note that for $n = 4$ there are two such alkanes (two isomers), whereas for $n = 5$ the number of possible isomers is already three. The number of isomers rapidly increases with n, as seen from Table 1.

Whenever in a molecule of an alkane there are n carbon atoms, then in this molecule there are exactly $2n+2$ *hydrogen atoms.* There are very many different alkanes with a given number n of carbon atoms (see below), but in *all* such molecules the number of hydrogen atoms is $2n + 2$. This "*all*" gives to the formula C_nH_{2n+2} a certain mathematical flavor.

Namely, a sceptical student may ask: How we know that *all* alkanes with n carbon atoms contain $2n+2$ hydrogen atoms? Wouldn't it be possible that some of the zillion possible alkanes possesses more than $2n+2$ or less than $2n+2$ hydrogen atoms?

If someone wants to embarrass his chemistry teacher, he may ask him this question. It is unlikely that the teacher will be able of offer a satisfactory answer. In the best case (typical for chemists) he will provide a number of examples, each consistent with the C_nH_{2n+2} formula.

Now, the problem with alkanes is that for a given value of n there exist very many *isomers* – various arrangements of n carbon and $2n + 2$ hydrogen atoms. For $n = 1$, $n = 2$, and $n = 3$, these arrangements are unique, as shown in Fig. 1. However, already for $n = 4$ there exists two distinct isomers, and for $n = 5$ three isomers. (Ask your chemistry teacher for the names of these compounds; this he will know.)

For $n > 5$ the number of isomeric alkanes rapidly increases, as seen from the following table:

n	$NI(n)$		n	$NI(n)$		n	$NI(n)$
1	1		11	159		21	910 726
2	1		12	355		22	2 278 658
3	1		13	802		23	5 731 580
4	2		14	1 858		24	14 490 245
5	3		15	4 347		25	36 797 588
6	5		16	10 359		26	93 839 412
7	9		17	24 894		27	240 215 803
8	18		18	60 523		28	617 105 614
9	35		19	148 284		29	1 590 507 121
10	75		20	366 319		30	4 111 846 763

Table 1. The number $NI(n)$ of distinct structural isomers of the C_nH_{2n+2} alkanes. In reality the number of isomers is still greater because of the so-called stereoisomerism.

Students who are curious, may try to check the above numbers for $n = 6$ and, perhaps, $n = 7$ and $n = 8$, but then should stop. Counting the number of isomeric alkanes is a very difficult mathematical problem, requiring the usage of advanced combinatorial techniques, that even university students do not master. Details of the history of this problem are found in the last section of the present article.

3. Graphs and trees

In this article we cannot give a full account of *Graph Theory*, a part of modern mathematics. Interested students are referred to some of the numerous existing textbooks (see, for example, [1–4]).

In nutshell, a *graph* is an object consisting of two sorts of elements, called *vertices* and *edges*. Graphs are usually (but not necessarily) depicted as diagrams in which the vertices are represented by small circles or big dots, and the edges by lines connecting some pairs of vertices. It is assumed that the number of vertices is finite, and that the lines pertaining to the edges are not directed.

In Fig. 2 is depicted a graph with 8 vertices and 9 edges.

The graph shown in Fig. 2 is *cyclic*, because it possesses cyclic arrangements of vertices. In particular, it possesses a five-membered cycle $(4, 5, 6, 7, 8, 4)$, a four-membered cycle $(4, 5, 6, 7, 4)$ and a three-membered cycle $(4, 7, 8, 4)$. This graph is *connected*, because it is possible to go (via the edges) from any vertex to any other vertex. The graph shown in Fig. 3 is also cyclic, but not connected. The graphs shown in Fig. 4 are acyclic and connected.

Definition 1. A connected acyclic graph is called a *tree*.

In fact, in Fig. 4 are depicted all trees with 4, 5, and 6 vertices.

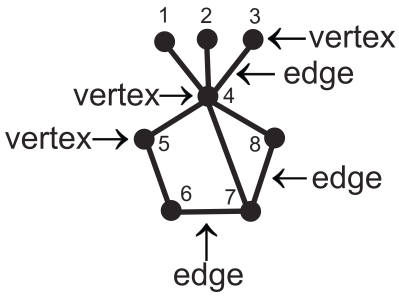

Fig. 2. A graph with 8 vertices and 9 edges. This graph is cyclic and connected.

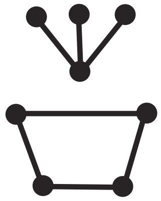

Fig. 3. A graph that is not connected.

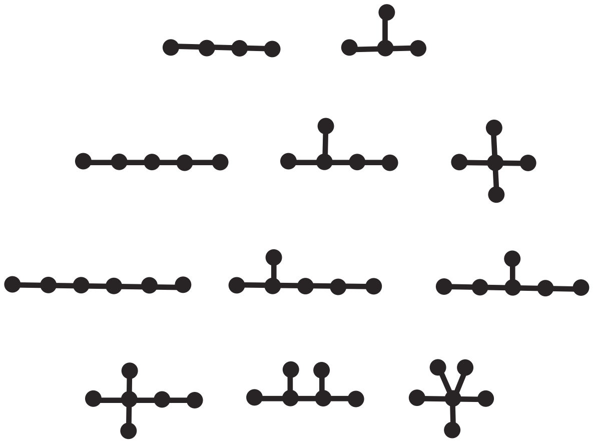

Fig. 4. Connected acyclic graphs with 4, 5, and 6 vertices. Such graphs are called trees.

The reason why we have introduced the concepts of graph and tree is because there is an obvious analogy between the structural formulas used in chemistry and graphs. In particular, to each alkane one can associate a tree, as illustrated in Fig. 5.

The way in which an alkane formula (from Fig. 1) is related to a tree (from Fig. 5) should be self-evident: every symbol for an atom (carbon or hydrogen) is replaced by a vertex; every symbol for a chemical bond is replaced by an edge. Then, in agreement with Definition 1, the graph representation of any alkane will necessarily be a tree. Such a tree is referred to as the *molecular graph* of the respective alkane.

This analogy between structural formulas and graphs has far-reaching consequences and is the basis of the so-called Chemical Graph Theory, a discipline of contemporary Mathematical Chemistry. Interested students should consult the textbooks [5–7].

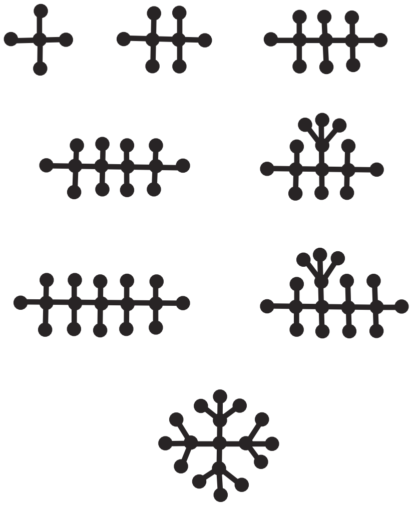

Fig. 5. Trees that in an obvious manner correspond to the structural formulas of alkanes, depicted in Fig. 1.

4. Some elementary properties of trees

In this section we prove three elementary properties of trees, which are necessary for deducing the formula C_nH_{2n+2}.

We start with another definition.

Definition 2. Let G be a graph and v its vertex. The number of first neighbors of the vertex v, that is, the number of vertices connected to v through an edge, is the *degree* of the vertex v, and is denoted by $\deg(v)$.

For instance, the vertices 1, 2, 3, 4, 5, 6, 7, and 8 of the graph from Fig. 2 have degrees 1, 1, 1, 6, 2, 2, 3, and 2, respectively. The trees shown in Fig. 5 have only vertices of degree 1 and of degree 4.

Theorem 1. *Any tree with at least two vertices has a vertex of degree one.*

Proof. First note that a tree with at least two vertices, since it is connected, cannot possess vertices of degree zero. Consequently, all vertices of a tree are either of degree one or of degree greater than one.

Let T be any tree. Choose in it any vertex, say v_1. If v_1 has degree one, Theorem 1 is satisfied. Therefore we need to consider only the case when $\deg(v_1) > 1$.

If v_1 has degree different than one, then $\deg(v_1) \geq 2$. Thus v_1 has at least two neighbors, say v_0 and v_2. If v_2 is of degree one, then Theorem 1 is satisfied. Therefore we need to consider only the case when $\deg(v_2) > 1$. Then v_2 has at least two neighbors, of which one is v_1 and another is, say, v_3. The vertex v_3 must be different from v_0, since in the opposite case T would contain a three-membered cycle.

If v_3 is of degree one, then Theorem 1 is satisfied. Therefore we need to consider only the case when $\deg(v_3) > 1$. Then v_3 has at least two neighbors, of which one is v_2 and another is, say, v_4. The vertex v_4 must be different from v_0 and v_1, since in the opposite case T would contain a four- or a three-membered cycle.

The above reasoning must stop at a certain point, because T has a finite number of vertices. This will happen when a vertex is encountered that has degree one. ■

Theorem 2. *Any tree with p vertices has $p - 1$ edges.*

Proof. We prove Theorem 2 by induction on the number p of vertices. The readers can check for themselves that for $p = 1, 2, 3$, the (unique) tree with p vertices has, respectively, 0, 1, and 2 edges. The validity of Theorem 2 can also be checked on the trees with 4, 5, and 6 vertices, depicted in Fig. 4.

Assume now that any tree with p_0 vertices, $p_0 \geq 1$, has $p_0 - 1$ edges. Let T be a tree with $p_0 + 1$ vertices.

According to Theorem 1, T must have a vertex v of degree one. If this vertex v is deleted from T, we obtain a tree T_0 with p_0 vertices. According to the induction hypothesis, T_0 has $p_0 - 1$ edges. Since the vertex v was connected to the tree T_0 by a single edge, the tree T must have exactly one edge more than T_0, i. e., $(p_0 - 1) + 1 = p_0$ edges.

This completes the proof by induction. ■

Theorem 3. *If a tree T has vertices $v_1, v_2, \ldots, v_p$, then*

$$\deg(v_1) + \deg(v_2) + \cdots + \deg(v_p) = 2(p-1) \ . \tag{1}$$

Proof. As it often happens in mathematics, it is easier to prove a somewhat more general result: *If G is any graph with q edges, and $v_1, v_2, \ldots, v_p$ are its vertices, then*

$$\deg(v_1) + \deg(v_2) + \cdots + \deg(v_p) = 2q \ . \tag{2}$$

The degree of a vertex is just the number of edges that end at this vertex. In view of this, the left-hand side of (2) counts all edges of G. Because each edge ends at two vertices, each edge of G is counted two times, resulting in the right-hand side of (2).

Equation (1) is just a special case of (2), in which Theorem 2 has been taken into account. ■

5. Proving formula C_nH_{2n+2}

With the preparation done in the previous two sections it is now easy to prove the alkane formula.

Let the alkane considered possess n carbon atoms and h hydrogen atoms. The molecular graph of such an alkane is a tree with $n + h$ vertices, and thus (by Theorem 2) with $n + h - 1$ edges.

For chemical reasons the degrees of the vertices representing carbon atoms are equal to four, and the degrees of the vertices representing hydrogen atoms are equal to one. (Ask your chemistry teacher to explain you why this is so; this has something to do with the valency of carbon and hydrogen.)

Since there are n vertices of degree 4, and h vertices of degree one, the sum on the left-hand side of (1) is equal to $4 \times n + 1 \times h$, which results in:

$$4n + h = 2(n + h - 1) \ .$$

By simple rearrangements (which the reader should do himself), from this equality follows

$$h = 2n + 2$$

as required by the C_nH_{2n+2} formula. ■

6. History of alkane formula and alkane enumeration

The proof of the general validity of the formula C_nH_{2n+2} was first communicated in 1875 by the British mathematician William Clifford (see Fig. 6). In fact, Clifford obtained much more general results, of which the alkane formula is just a simple special case.

Fig. 6. William Clifford (1845–1879)

Fig. 7. Arthur Cayley (1821–1895)

Clifford's main contributions to mathematics fall in the area of geometry and algebra. He was a pioneer in the study of non-Euclidean geometry. His name is remembered in *Clifford-Klein spaces, Clifford algebras, Clifford numbers*, etc.

The great British mathematician Arthur Cayley (see Fig. 7) was the first who recognized the relation between the structural formulas in organic chemistry and the graphs.

Cayley is one of the most prolific mathematicians of all times. His main contributions are in the field of matrix theory, linear algebra, and group theory. He is also one of the pioneers of graph theory. By the way, Cayley proposed the name "tree".

In 1874 Cayley published a paper entitled "*On the Mathematical Theory of Isomers*", which represents the first serious chemical application of graph theory. In this paper he introduced the concept of molecular graph. Cayley's intention was to solve the problem of alkane isomers, that is to find a method by which the number $NI(n)$ of distinct isomers of formula C_nH_{2n+2} could be determined. He, however, did not succeed.

In the next 50 years the enumeration of alkane isomers remained an open problem. Numerous chemists and mathematicians tried to solve it, but without success. Eventually, in 1932 two American chemists Henry Henze and Richard Blair found a recursive procedure by which the numbers $NI(n)$ from Table 1 could be calculated.

The general solution of the enumeration problem was obtained in 1935 by the Hungarian mathematician George Pólya (see Fig. 8).

Fig. 8. George Pólya (1887–1985)

By means of the method discovered by Pólya (which nowadays is called *Pólya theory* and represents one of the cornerstones of modern combinatorics) it is possible to count the number of arbitrary objects, provided their symmetry is defined.

Returning to alkanes, it is worth noting that in 1932 there were no computers, and therefore Henze and Blair had to do their calculations by hand. At a certain point they made a numerical error, and therefore for larger values of n their $NI(n)$-values were incorrect. The correct values were obtained only in the 1980s, when the Henze-Blair procedure was repeated by using a (super)computer. This means that it took a little more than a century to enumerate the alkane isomers.

The readers of this article may be pleased to know that the number of possible isomers of $C_{50}H_{102}$ is

$$NI(50) = 1\,117\,743\,651\,746\,953\,270$$

and of $C_{80}H_{162}$,

$$NI(80) = 10\,564\,476\,906\,946\,675\,106\,953\,415\,600\,016.$$

REFERENCES

1. F. Harary, *Graph Theory*, Addison-Wesley, Reading, 1969.
2. R. Merris, *Graph Theory*, Wiley, New York, 2001.
3. D. Cvetković, S. Simić *Combinatorics and Graphs*, Računarski fakultet, Beograd, 2006 (in Serbian).

4. D. Stevanović, S. Simić, M. Ćirić, V. Baltić, *Discrete Mathematics – Elements of Combinatorics and Graph Theory*, Društvo Matematičara Srbije, Beograd, 2008 (in Serbian).
5. N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, 1983.
6. I. Gutman, O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer-Verlag, Berlin, 1986.
7. I. Gutman, *Introduction to Chemical Graph Theory*, Fac. Sci. Kragujevac, Kragujevac, 2003 (in Serbian).

Faculty of Sciences, Kragujevac University, Radoja Domanovića 12, 34000 Kragujevac, Serbia
E-mail: gutman@kg.ac.yu