Generating and Counting Unbranched Catacondensed Benzenoids

RATKO TOŠIĆ* and MILOVAN KOVAČEVIĆ

Institute of Mathematics, Faculty of Science, University of Novi Sad, 21000 Novi Sad, Yugoslavia

Received May 28, 1987

The path code is a word on a six-letter alphabet that uniquely determines an unbranched catacondensed benzenoid system (UCBS). Necessary and sufficient conditions are given for a word over a six-letter alphabet to be the path code of a UCBS. All nonisomorphic UCBS with h hexagons are enumerated up to $h = 20$.

INTRODUCTION

The problem of "cell-growth" is a classical problem in mathematics.[1-5] When applied to hexagonal "animals", it has relevance to the studies of benzenoid (polycyclic aromatic) hydrocarbons, especially to their enumeration.[6-15]

In the present paper we consider polyhex graphs. Here a polyhex graph, also called a benzenoid or hexagonal system, is defined as in Gutman's review.[8] A polyhex graph corresponds to a network obtained by arranging congruent regular hexagons in the plane so that two hexagons are either disjoint or possess exactly one common edge. It is a strictly planar system that is simply connected (Figure 1). The last restriction is released in the definition of coronoids (Figure 2).

In the present work we consider a special class of benzenoid systems—unbranched catacondensed benzenoid systems (UCBS) A benzenoid system is catacondensed[8] if it does not posses any internal vertex. All other systems are referred to as pericondensed. A catacondensed benzenoid system is said to be unbranched if each of its hexagons has at most two neighbors (Figure 3). All other catacondensed benzenoid systems are referred to as branched (Figure 4).

Given a number of hexagons, how many benzenoid systems can be constructed? This general mathematical problem is still unsolved. In the present paper we report the new results of enumeration of UCBS, where the range of computation is extended with regard to the h value up to 20. Here h denotes the number of hexagons of a UCBS. Some previous results of enumeration of such systems are summarized in Brunvoll et al.,[7] Table I, where the range of computation has been extended to 12 for symmetrical UCBS and to 11 for unsymmetrical UCBS.

In the case of the benzenoid systems there is a bijection between the molecular graph and the dualist[15] obtained by the replacement of the hexagon centers with vertices. Two such vertices are connected by an edge if and only if they correspond to adjacent hexagons (Figure 5).

The dualist of a hexagonal system can be considered as a subgraph of a triangular lattice T obtained by tiling the plane by congruent regular triangles (Figure 6). It is easy to see that the dualist of a UCBS is a path in a triangular lattice (Figure 7).

CHARACTERIZATION OF A PATH IN THE TRIANGULAR LATTICE BY WORDS OVER A SIX-LETTER ALPHABET

Before defining the path code of a UCBS, we would like to remind the reader of some notions from formal languages and geometry.

The set of the first k natural numbers is denoted by N_k. Hence $N_k = \{1, 2, ..., k\}$. The set $A = \{a_1, a_2, ..., a_k\}$ is called the alphabet if for all $i \in N_k$, a_i are arbitrary symbols. The elements of A are then the letters of the alphabet, and A is a k-letter alphabet.

If $x \in A^n$, i.e., if $x = (x_1, x_2, ..., x_n)$ is an ordered n-tuple with components from A, we say that x is a word of length n over the alphabet A. For the sake of brevity we shall write x as $x_1x_2...x_n$, and we shall denote the length of x by $\|x\|$.

A subword of length m of the word $x_1x_2...x_n$ is any word $x_sx_{s+1}...x_{s+m-1}$, where $s \in N_{n-m+1}$ and $m \in N_n$.

The set of all words of finite length over the alphabet A will be denoted by A^*. We write $d_j(x) = r$ if the letter $j \in A$ occurs r times in the word $x \in A^*$.

Through every point of the triangular lattice one can set three straight lines such that every edge of the lattice is parallel to exactly one of these lines. Consider such three lines and direct them. The unit vectors of these lines are denoted by 0, 1, and 2 so that the angle between 0 and 1 is equal to $\pi/3$, whereas the angle between 0 and 2 is equal to $2\pi/3$. The oppositely directed unit vector will be denoted by 3, 4, and 5, respectively (Figure 8).

We consider the set of vectors $A = \{0, 1, 2, 3, 4, 5\}$ as an alphabet.[11-14] It is evident that every directed edge of the triangular lattice is in a one-to-one correspondence with an element of A. Hence we have defined a function that maps the set of all finite directed paths of the triangular lattice into the set A^* of all words over the alphabet A.

For example, Figure 9 presents a path on the triangular lattice connecting the vertices u and v. This path can be directed in two ways. It either starts at the vertex u and ends at the vertex v or vice versa. If this path is directed from the vertex u to the vertex v, then it corresponds to the word 011504505. If the same path is directed oppositely, we obtain the inverse path that corresponds to the word 232132443.

Two words that correspond to the opposite orientations of a given path of the triangular lattice are related in the following way. Let $x = x_1x_2...x_{n-1}x_n$ and $y = y_1y_2...y_{n-1}y_n$ be two such words. Then

$$x_1 - y_n = x_2 - y_{n-1} = ... = x_{n-1} - y_2 = x_n - y_1 = 3 \pmod 6$$

This means that by considering the letters of x as ordinary integers and by adding 3 modulo 6 to them, we arrive at a word that is equal to the word obtained by reading the letters of y from right to left.

CHARACTERIZATION OF THE PATH CODE OF UCBS

Let P be the dualist of a UCBS H (Figure 7). As we have already seen, P is a path in the triangular lattice. If x is the corresponding word of the path P, we shall say that x is the path code of the UCBS H. It is obvious that the path code of a UCBS with h hexagons is a word of the length $h - 1$, i.e., a word from the set A^{h-1}.

Two words $x,y \in A^n$ are said to be equivalent if they correspond to congruent paths, i.e., if they are path codes of two congruent UCBS.

Let s: $A \rightarrow A$ be the following mapping:

$$s = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$$

0095-2338/88/1628-0029$01.50/0

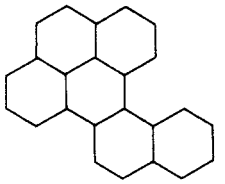

Figure 1.

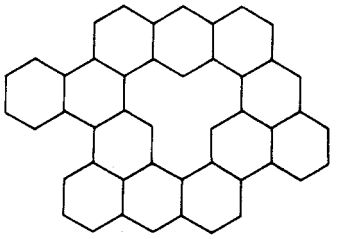

Figure 2.

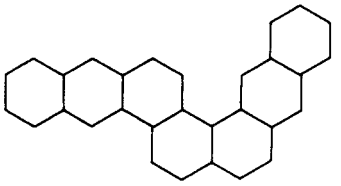

Figure 3.

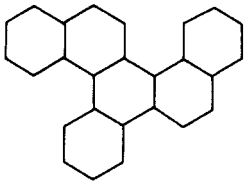

Figure 4.

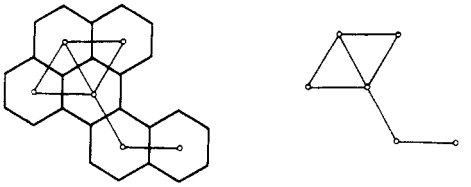

Figure 5.

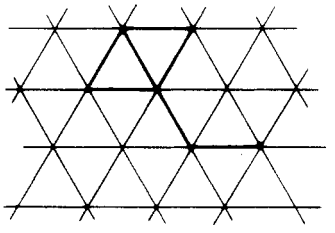

Figure 6.

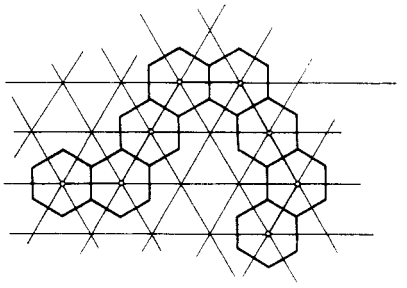

Figure 7.

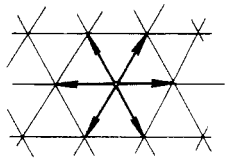

Figure 8.

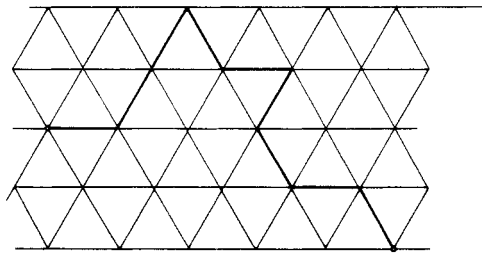

Figure 9.

We define the mappings b, c, and e of A^n into A^n in such a way that for arbitrary $x = x_1x_2...x_{n-1}x_n \in A^n$

$$b(x) = b_1b_2...b_{n-1}b_n \quad \text{where } b_i = x_{n+1-i} + 3 \pmod 6$$

$$c(x) = c_1c_2...c_{n-1}c_n \quad \text{where } c_i = x_i + 1 \pmod 6$$

$$e(x) = e_1e_2...e_{n-1}e_n \quad \text{where } e_i = s(x_i)$$

for $i = 1, 2, ..., n$.

Now the following statement can be proved.

Theorem 1. Two words $x,y \in A^n$ are equivalent if and only if

$$y = (c^k b^r s^t)(x)$$

where $k \in \{0, 1, 2, 3, 4, 5\}$, $r,t \in \{0, 1\}$, and for any mapping f, f^0 is identical mapping.

Proof. The mappings c, e, and b correspond to a rotation through angle $\pi/3$, to mirror symmetry, and to taking the edges of the corresponding path in the reversed order, respectively. On the other hand, two congruent paths can be obtained one from the other by composition of these transformations.

The following statement gives the necessary and sufficient condition for a word from A^* to be the path code of a UCBS.

Theorem 2. A word $x = x_1x_2...x_n \in A^n$ is the path code of a UCBS if and only if for every subword y of x, $|d_0(y) + d_5(y) - d_3(y) - d_2(y)| + |d_1(y) + d_2(y) - d_4(y) - d_5(y)| > 1$.

Proof. The "only if" part of theorem 2 is elementary.

In the case $\|y\| = 2$, the condition of the theorem implies that either $x_{i+1} = x_i \pmod 6$ or $x_{i+1} = x_i + 1 \pmod 6$ or $x_{i+1} = x_{i-1} \pmod 6$. Generally, the condition of theorem 2 guarantees that any two hexagons corresponding to nonadjacent vertices of a dualist are nonadjacent.

COMPUTATIONAL RESULTS

Theorems 1 and 2 enable one to generate all nonisomorphic UCBS with h hexagons by recognizing and generating all nonequivalent words of the length $h - 1$ over the alphabet $A = \{0, 1, 2, 3, 4, 5\}$ satisfying the conditions of theorem 2. We made the generation of all such words much easier by taking always as a representative of an equivalence class a word beginning with 0 and having 1 as the first letter different from 0. In such a way a computer program has been written for the generation and enumeration of UCBS with h hexagons.

Table I shows the results of enumeration up to $h = 20$. The unbranched catacondensed benzenoid systems are classified

Table I. Number of Different Types of Unbranched Catacondensed Benzenoids

h	d	m	c	d	total unbranched
1	1	0	0	0	1
2	1	0	0	0	1
3	1	1	0	0	2
4	1	1	1	1	4
5	1	4	1	4	10
6	1	3	4	16	24
7	1	12	4	50	67
8	1	10	13	158	182
9	1	34	13	472	520
10	1	28	39	1406	1474
11	1	97	39	4111	4248
12	1	81	116	11998	12196
13	1	271	115	34781	35168
14	1	226	339	100660	101226
15	1	764	336	290464	291565
16	1	638	988	837137	838764
17	1	2141	977	2408914	2412033
18	1	1787	2866	6925100	6929754
19	1	6025	2832	19888057	19896915
20	1	5030	8298	57071610	57084939

according to their symmetries into dihedral, d, D_{2h} (regular hexagonal, D_{6h}, for $h = 1$); mirror-symmetrical, m, C_{2v}; centrosymmetrical, c, C_{2h}; and unsymmetrical, u, C_s.

A listing of the computer program in PASCAL is available on request to the authors.

REFERENCES AND NOTES

(1) Klarner, A. D. "Some Results Concerning Polyominoes". *Fibonacci Q.* **1965**, *3*(1), 9-20.
(2) Golomb, S. W. *Polyominoes*; Scribner, New York, 1965.
(3) Harary, F.; Read, R. C. "The Enumeration of Tree-like Polyhexes". *Proc. Edinburgh Math. Soc.* **1970**, *17*, 1-14.
(4) Lunnon, W. F. "Counting Polyominoes" in *Computers in Number Theory*; Academic: London, 1971; pp 347-372.
(5) Lunnon, W. F. "Counting Hexagonal and Triangular Polyominoes". *Graph Theory Comput.* **1972**, 87-100.
(6) Brunvoll, J.; Cyvin, S. J.; Cyvin, B. N. "Enumeration and Classification of Benzenoid Hydrocarbons". *J. Comput. Chem.* **1987**, *8*, 189-197.
(7) Balaban, A. T., et al. "Enumeration of Benzenoid and Coronoid Hydrocarbons". *Z. Naturforsch., A: Phys., Phys. Chem., Kosmophys.* **1987**, *42A*, 863-870.
(8) Gutman, I. "Topological Properties of Benzenoid Systems". *Bull. Soc. Chim., Beograd* **1982**, *47*, 453-471.
(9) Gutman, I.; Polansky, O. E. *Mathematical Concepts in Organic Chemistry*; Springer: Berlin, 1986.
(10) Tošić, R.; Doroslovački, R.; Gutman, I. "Topological Properties of Benzenoid Systems—The Boundary Code". *MATCH* **1986**, No. 19, 219-228.
(11) Doroslovački, R.; Tošić, R. "A Characterization of Hexagonal Systems". *Rev. Res. Fac. Sci.-Univ. Novi Sad, Math. Ser.* **1984**, *14*(2) 201-209.
(12) Knop, J. V.; Szymanski, K.; Trinajstić, N. "Computer Enumeration of Substituted Polyhexes". *Comput. Chem.* **1984**, *8*(2), 107-115.
(13) Stojmenović, I.; Tošić, R.; Doroslovački, R. "Generating and Counting Hexagonal Systems". *Proc. Yugosl. Semin. Graph Theory, 6th, Dubrovnik 1985*; pp 189-198.
(14) Doroslovački, R.; Stojmenović, I.; Tošić, R. "Generating and Counting Triangular Systems". *BIT* **1987**, *27*, 18-24.
(15) Knop, J. V.; Müler, W. R.; Szymanski, K.; Trinajstić, N. *Computer Generation of Certain Classes of Molecules*; Association of Chemists and Technologists of Croatia: Zagreb, 1985.

SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules

DAVID WEININGER

Medicinal Chemistry Project, Pomona College, Claremont, California 91711

Received June 17, 1987

SMILES (Simplified Molecular Input Line Entry System) is a chemical notation system designed for modern chemical information processing. Based on principles of molecular graph theory, SMILES allows rigorous structure specification by use of a very small and natural grammar. The SMILES notation system is also well suited for high-speed machine processing. The resulting ease of usage by the chemist and machine compatability allow many highly efficient chemical computer applications to be designed including generation of a unique notation, constant-speed (zeroeth order) database retrieval, flexible substructure searching, and property prediction models.

INTRODUCTION

The first step in the formalization of chemistry is to name a chemical compound. This requires an unambiguous and reproducible notation for the simplest atom to the most complicated structure. All other chemical information procedures follow from the fundamental process of chemical nomenclature. Consequently, the improvement of chemical notation has been an ongoing endeavor, amply documented in the pages of this journal.

Wiswesser[1] described the historical development of chemical nomenclature from the beginning of chemistry as a rudimentary science to the start of the computer era. When computers opened up the processing and storage of chemical information, this depended on the description of chemical structure. Morgan[2] developed a technique for generation of unique machine description from which followed the CAS (Chemical Abstracts Service) ONLINE search system.[3] Other important advances were the application of graph theory to chemical notation[4] and chemical substructure search systems.[5-7] The uniqueness of chemical information has also been considered from a theoretical point of view.[8]

With the introduction of computers, line notation[9] became widely used in chemical nomenclature because computers can process linear strings of data with relative ease. Line notation serves as the basis of the International Union of Pure and Applied Chemistry (IUPAC) notation system.[10] It is described by Read[11] in general terms in relation to graph-theoretical concepts. Read lists 12 attributes considered desirable in a chemical coding system. In 1983 many of these attributes were incompatible with each other. In the few intervening years, however, advances in computer technology have accelerated so that they are overcoming the incremental increase of chemical information. Computer technology and chemical knowledge are now at a point where it is possible to store all of the extant chemical information on existing hardware. The chemical information problem of just getting a machine to store information is largely historical. Current and future efforts must be directed to building highly efficient systems

0095-2338/88/1628-0031$01.50/0