CROATICA CHEMICA ACTA CCACAA **56** (3) 315—324 (1983)

YU ISSN 0011-1643
UDC 541
Original Scientific Paper

CCA-1380

Application of Graph-Based Chemical Nomenclature to Theoretical and Preparative Chemistry[1,*]

Alan L. Goodson

Chemical Abstracts Service, Columbus, Ohio 43210, U.S.A.

Received August 29, 1982

Development of graph-based systematic names containing mathematical descriptions of molecular graphs is described. Such names can be regarded as compact connection tables. Exploration of the use of graph-based systematic names for information storage and retrieval purposes, in substructure searching, and as an aid in pattern recognition, structure-activity relationships, drug design, etc., is discussed.

INTRODUCTION

Classical chemical nomenclature is really a collection of nomenclature systems rather than a single system. Efforts have been made to systematize and codify these nomenclature systems[2,3] but there is still no single, underlying philosophy linking these systems together. While the question of developing a new nomenclature system from first principles has been discussed by Fletcher[4] and by Fernelius et al.[5], no one, apart from Siboni and Perino[6] during the period 1930—5, has attempted to develop such a system until recently[7–10]. It now appears that such new nomenclature systems developed from first principles have potential applications in addition to the usual naming purposes, such as indexing. These potential applications are the subject of this paper.

Siboni and Perino's proposal is illustrated in Figure 1 for an inorganic, an acyclic organic, and a cyclic organic compound. Atoms and groups are represented by one through four letters and these morphemes are linked together in a prescribed manner to yield the name of a compound. Complex substances require the addition of locants which are never placed within the name but before, after, and below.

Verkade[6] referred to the Siboni-Perino proposal as a »play on letters« and as an »assembling nomenclature«. In fact, it appears that Siboni and Perino were devising a notation system.

CONNECTION TABLES

When chemical substances were first given names, little or nothing was known of their nature. But when computers began to be used for manipulation of chemical information about 25 years ago, the atomic nature of chemical

[*] Presented at *The IUPAC International Symposium on Theoretical Organic Chemistry,* held in Dubrovnik, Croatia, August 30—September 3, 1982.

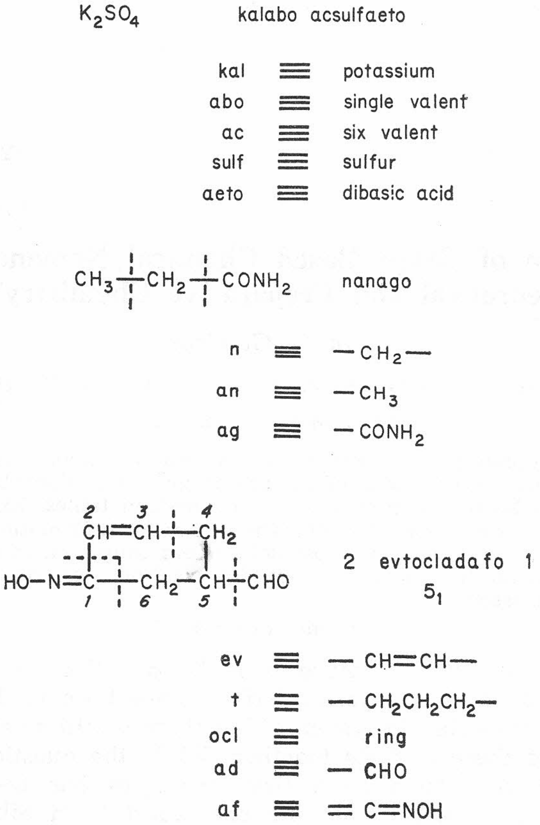


Figure 1. Examples of the Siboni-Perino nomenclature proposal.

substances was much better understood, so it was possible to devise a single, comprehensive technique for recording chemical structures on computer files.

This technique involves use of the connection table, which appears to have originated with Wheland[11], whose proposal is illustrated in Figure 2.

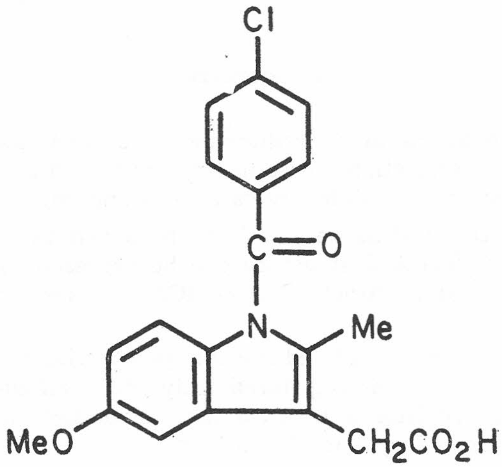
Ethylene Oxide

	C	C	H	H	H	H	O
C	"	1	1	1	0	0	1
C		"	0	0	1	1	1
H			"	0	0	0	0
H				"	0	0	0
H					"	0	0
H						"	0
O							"

Acetaldehyde

	C	C	H	H	H	H	O
C	"	1	1	1	1	0	0
C		"	0	0	0	1	2
H			"	0	0	0	0
H				"	0	0	0
H					"	0	0
H						"	0
O							"

Figure 2. Wheland's nongeometric statements of the structures of ethylene oxide and acetaldehyde.



IUPAC Name:

[1-(4-Chlorobenzoyl)-5-methoxy-2-methylindol-3-yl]acetic acid

Connection Table:

TOPOLOGY

ATOM NO	1	2	3	4	5	6	7	8	9	10
CONN		01		03	04	04			08	
ELEMENT	C	O	C	C	O	O	C	O	C	CL
BOND		−2		−1	−4	−4			−1	

RING CORRESPONDENCE LINK

RING ID 46T.150A.182

RING ATOM NOS	1	2	3	4	5	6
SUBSTANCE ATOM NOS	20	21	22	23	24	25

RING ID 333X.151D.57P

RING ATOM NOS	1	2	3	4	5	6	7	8	9
SUBSTANCE ATOM NOS	12	11	15	16	13	14	17	19	18

LINK GROUP 01-1 13, 01-1 20, 03-1 15, 07-1 17,
08-1 19, 10-1 25,

RING

RING IDENTIFIER NO = 46T.150A.182

ATOM NO	1	2	3	4	5	6
CONN		01	01	02	03	04
ELEMENT	C	C	C	C	C	C
BOND		*5	*5	*5	*5	*5

RING CLOSURE PAIRS 05*5 06,

RING

RING IDENTIFIER NO = 333X.151D.57P

ATOM NO	1	2	3	4	5	6	7	8	9
CONN		01	01	01	02	02	03	04	06
ELEMENT	C	C	C	C	N	C	C	C	C
BOND		*5	*1	*5	*1	*5	*2	*5	*5

RING CLOSURE PAIRS 05*1 07, 08*5 09,

Figure 3. A chemical structure, its IUPAC name, and its connection table.

Wheland's proposal contains much redundancy. For example, as is apparent in Figure 2, all of the connections for ethylene oxide and acetaldehyde appear in the first two rows; the remaining rows are redundant.

Development of this technique has been reviewed by Knop et al.[12] and by O'Korn[13]. Today, a chemical structure can be represented by a connection table which, at Chemical Abstracts Service (CAS), takes the form shown in Figure 3.

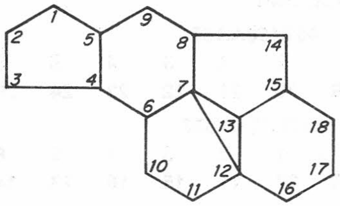
This technique has been refined to the point where all substances of known structure can be machine registered; only 3% of all substances recorded at CAS are manually registered because their structures are unknown, are partially known, or exceed system limitations.

GRAPH-BASED NOMENCLATURES

While the connection table that assigns unique, as opposed to unambiguous, numbering is a successful technique for computer manipulation of chemical structure information, manual derivation of such unique connection tables is tedious and prone to error. This raises the question of whether it would be beneficial to develop a connection table that assigns unique numbering and that can be derived manually without undue difficulty. If such a connection table can be developed, it may prove more efficient than existing methods. This question will now be explored.

There are conceptually two approaches to numbering and naming chemical graphs, viz., assembling fragments or tracing a path[9].

The first approach was taken by Taylor[10,14] and is illustrated in Figure 4. The name indicates that the structure is a saturated hydrocarbon containing a 3-membered ring, three 5-membered rings, and two 6-membered rings. In short, the name verbalizes the ring analysis: C_3—C_5—C_5—C_5—C_6—C_6.



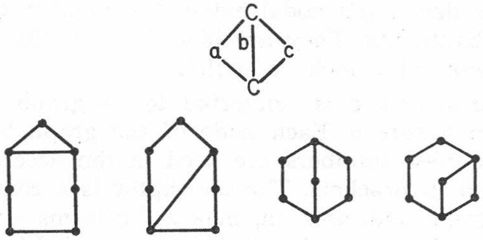
7(12)-Tria-1,6,8(13)-Ternipenta-4,12(15)-binihexalane

Ring Analysis: C_3-C_5-C_5-C_5-C_6-C_6

Figure 4. An example of Taylor's nomenclature proposal.

While the name is systematic and contains all the pertinent information needed, the information is divided between the word part, which indicates how many rings of each size the structure contains, and the number part, which indicates how these rings are fused together.

This scattering of information within a name is not really convenient and it would be better if the information were gathered together in one place and in one form.



Alter Name	Neuer Name
Norcaran	Bicyclo-[0, 1, 4]-heptan
—	Bicyclo-[0, 2, 3]-heptan
Norcamphan . .	Bicyclo-[1, 2, 2]-heptan
Norpinan	Bicyclo-[1, 1, 3]-heptan

Figure 5. Von Baeyer's nomenclature proposal.

An early example of the second approach was the proposal by von Baeyer[15] for bicyclic hydrocarbons (Figure 5). In principle, two carbon atoms are joined by three bridges: *a*, *b*, and *c*. The number of atoms in each bridge is listed in order of increasing size and the numbers are enclosed in brackets in the name.

The numbers inside the brackets are a mathematical description of the graph of each structure. In the first example, subtraction of 1 from the number of terms inside the brackets, which is 3, yields the number of rings, which is 2. In other words, the number of terms inside the brackets is equal to the number of rings plus 1. This being the case, the term »bicyclo« could be regarded as redundant, although such terms are needed, for example, for indexing purposes. Summing the numbers inside the brackets (i. e., 0+1+4) gives 5, and adding 2 for the bridgeheads yields 7, the number of nodes in the graph. This can also be restated as: the number of nodes in the graph is equal to 2 plus the sum of the terms inside the brackets. The Greek number name »hept« can therefore also be regarded as redundant.

The only information that this mathematical description does not provide is that the nodes represent saturated carbon atoms, which is indicated in the name, by convention, by the »an« ending.

Although von Baeyer's proposal, with minor modification to the format, has been generally adopted and extended[2a], it is not a comprehensive system because it is not applicable to acyclic structures.

This deficiency is overcome in a recent proposal by Lozac'h et al.[7] for what is known as nodal nomenclature. Like the connection table, nodal nomenclature assigns unique numbering to chemical structures, such as the one illustrated in Figure 3. (Classical nomenclature, by contrast, requires three independent numberings within the same structure, viz., one for the indole ring system, a second for the benzene ring, and a third for the acetic acid residue.) The same structure will be used to show how chemical structure information can be stored by using nodal nomenclature, so that a direct comparison can be made between it and classical nomenclature.

The first step in deriving a nodal name is to identify the atoms or groups to be treated as substituents. They are identified (in this case, as —Cl, =O, and —CO_2H) by means of a look-up table[8].

The rest of the structure is converted to its graph and numbered and named[7] as shown in Figure 6. Each node of the graph has been assigned a unique number and these numbers are used in the descriptor, i. e., the part of the name enclosed in brackets. The descriptor is a complete mathematical description of the graph and, as such, makes the terms »tricyclo« and »icosanodane« redundant at this point, though they are needed later for the complete name of the substance.

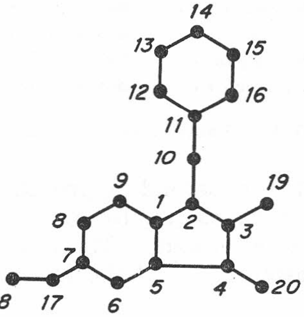


Figure 6. Tricyclo[(09.$0^{1,5}$)2:10(1)10:11(06)7:17(2)3:19(1)4:20(1)]icosanodane

If the descriptor is a complete mathematical description of the graph, what information can be deduced from it by inspection? First, there are 6 sets of parentheses inside the brackets, which indicates that there are 6 modules in the structure, a module being defined[7] as a cyclic or acyclic set of nodes treated as a separate entity. Two of the modules are cyclic, as indicated by the zeros immediately following the open parentheses, and the total number of rings in the cyclic modules is determined by summing the number of terms inside the parentheses; i. e., two in the first (09 and $0^{1,5}$) and one in the second (06) for a total of 3, which checks with »tricyclo« at the beginning of the name. The remaining 4 modules are acyclic. The number of nodes in the graph is determined by summing the numbers inside the parentheses; i. e., $9+1+6+2+1+1=20$, which checks with »icosa« after the descriptor.

The numbers separated by colons, between the sets of parentheses, show how the modules are linked together. Each locant is unique, in contrast to classsical nomenclature; for example, locants 1—6 occur twice in the IUPAC numbering of Figure 3. Initially, the locant numbers in the descriptor increase steadily (i. e., from 2 to 10 to 11) and then drop (i. e., from 11 to 7, 17 to 3, and 19 to 4). These breaks in the numbering sequence indicate branching.

Information derived from the descriptor in this manner can be used in the form of screens during information retrieval.

The full nodal name[8] of the structure illustrated in Figure 3 is 14-chloro-10--oxo-2-aza-17-oxa-tricyclo[(09.$0^{1,5}$)2:10(1)10:11(06)7:17(2)3 : 19(1)4 : 20(1)]icosan(1,3--9,11-16)axene-2-carboxylic acid. »Chloro«, »oxo«, and »carboxylic acid« are substituent terms and »aza« and »oxa« are replacement terms. »An« replaces »nodane« and indicates saturation and (by convention) a hydrocarbon parent.

»(1,3-9,11-16)axene« defines the bonding, »axene«[8] indicating the maximum number of noncumulative double bonds.

Systematic names, then, can provide mathematical descriptions of chemical structures. This raises the question of whether a systematic name can also be used as a connection table, which will now be discussed.

The nodal name can be compared with the connection table (cf. Figure 3), as follows: the descriptor is comparable with the »atom number« and »connectivity« parts of the connection table. The substituent and replacement parts of the nodal name can be compared with the »element« part of the connection table, and the bonding parts of the nodal name and of the connection table are also comparable. It is because the nodal name and the connection table can be compared so directly that the nodal name can be regarded as a compact connection table.

APPLICATIONS OF GRAPH-BASED SYSTEMATIC NOMENCLATURE

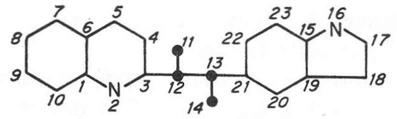
The most obvious application for any graph-based systematic nomenclature is for information storage and retrieval purposes. For example, it could be used more efficiently than classical nomenclature for machine naming of chemical structures and for indexing. However, once the similarity between nodal names and connection tables is recognized, a number of other potential applications suggest themselves. For example, because the descriptor is a mathematical description of the graph, it could be used to provide a form of structure index. Using the Ring Analysis Index of the CAS *Parent Compound Handbook* as a model, the graph shown in Figure 6 could be classified as having six modules, then as having two cyclic modules and four acyclic modules, and then the modules could be sorted, in turn, according to internal structure. This structure index could be used alone or as part of the full-name indexing referred to, above.

Nodal names can also be used for substructure searching. Classical names have already been used for this purpose by Dunn et al.[16] The techniques described by them for use with classical names (i. e., molecular formula screen, ring screen, nomenclature screen, and link search) can also be used with nodal names. However, nodal names also contain structural information. It, too, can be used in the form of screens, as discussed above, or as follows.

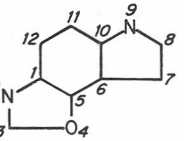
If a search is being made for the indole ring system in a file, the simplest way for it to occur would be in a form such as ». . . 2-aza . . . [$(09.0^{1,5})$ . . . an . . .« The module descriptor is $(09.0^{1,5})$, which describes the graph of the ring system. The skeletal structure is composed predominantly of carbon atoms, so the morpheme »an« would appear in the name after the descriptor. The nitrogen atom is in the 2-position of the indole ring system (when given nodal numbering) and because nodal numbering is unique, the 2-position of the indole ring system is also, in this case, the 2-position of the whole structure.

If the indole ring system is not the first module to be described in the descriptor, as in the first example of Figure 7, the indole graph can be recognized by the $(09.0^{1,5})$ descriptor, as before, and the nitrogen can be located by summing the nodes in the preceding modules (i. e., $10+4=14$) and adding 2, the position of the nitrogen in the indole ring system. A »16-aza« term earlier in the name would confirm the presence of an indole ring system.

If the indole ring system is embedded in a larger ring system (as in the second example of Figure 7, which has a three-ring system), it would be



...2,16-diaza...[(010.$0^{1,6}$)3:12(4)13:21(09.$0^{1,5}$)...



...;2,9-diaza-4-oxa...[(012.$0^{1,5}0^{6,10}$)...

Figure 7. Use of nodal nomenclature for substructure search.

necessary to delete one ring to obtain the two rings of indole. If atoms 7, 8, and 9 are deleted, then, in effect, atoms 6 and 10 are no longer bridgeheads and the third term of the descriptor (i. e., $0^{6,10}$) is deleted. Further, since three atoms have been removed, the first term becomes 09 instead of 012. This yields the required descriptor for the indole ring system (i. e., 09.$0^{1,5}$) and there is a nitrogen atom at position 2. But there is also an oxygen atom at position 4, so the ring system is benzoxazole and is rejected.

If, instead, atoms 2, 3, and 4 of Figure 7 are deleted to yield two rings, and the second term in the descriptor (i. e, $0^{1,5}$) is deleted because atoms 1 and 5 are no longer bridgeheads, then the descriptor that remains is (09.$0^{6,10}$). If, further, 5 is subtracted from both locants (to reduce the first locant to 1, because nodal numbering of ring systems always begins at a bridgehead), then the descriptor for the indole graph, i. e. (09.$0^{1,5}$), is obtained. It remains to find a nitrogen at position 7 or 9, and »2,9-diaza« appears in the name before the descriptor to complete the identification of the indole ring system.

The prospect of using nodal nomenclature for substructure searching leads to the possibility of its use as an aid in pattern recognition, structure-activity relationships, drug design, etc., because such activities often require comparison of the properties of large numbers of related substances.

Elliott[17], for example, has stated that insecticidal activity of pyrethroids depends primarily on the shape of the molecule and chemical reactivity is of secondary importance. That being so, it would be possible to search for related structures on nodal descriptors alone. However, structure-activity relationships show that chemical reactivity is of prime importance in carbamate and phosphorus insecticides. Here, the emphasis would be more on the nature of the substituents (or functional groups) attached to a chemical structure and so the look-up table of substituents would be important. Elliott also stated that the high activity of the γ-isomer of hexachlorocyclohexane is a remarkable example of structural specificity, which implies that the precise shape of the molecule is critical for activity. Such stereospecificity in a flexible molecule would be obtained from the stereo descriptor of a nodal name.

In his discussion of pattern recognition, Wijnne[18] listed eighty-five 4-phenylpiperidines, two of which are shown in Figure 8, together with their nodal descriptors. Because of the numbering priority rules, the benzene ring in the first example is numbered first but the same ring is numbered last in the second example. In both examples, one position of the piperidine ring has two substituents (one of which is the benzene ring) and the locant is therefore repeated in each descriptor (7 in the first example and 13 in the second).

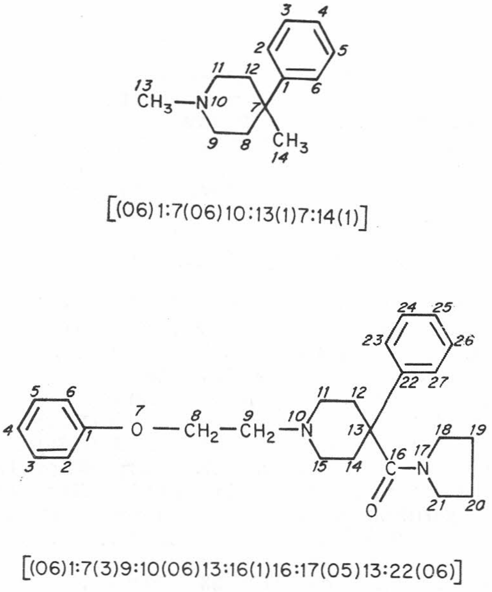


Figure 8. Some 4-phenylpiperidines and their nodal descriptors.

Further, the nodal locant of the nitrogen atom of the piperidine ring is three locant numbers removed from the disubstituted position of the piperidine ring (i. e., from 7 to 10 and from 13 to 10). Such variation in numbering is similar to the situation encountered during substructure searching by connection table.

SUMMARY

Nodal nomenclature, in contrast to classical nomenclature, provides unique numbering for chemical structures which is of value, for example, when describing stereochemistry. It also provides mathematical descriptions of chemical graphs and this feature, together with the unique numbering, makes nodal nomenclature more amenable to computer manipulation than classical nomenclature. Nodal names are therefore more suitable for information storage and retrieval purposes, for example. However, the mathematical descriptions of chemical graphs can be used as screens and this raises the question, discussed here, of whether nodal names can be used as connection tables for use in substructure search, etc. This question will be explored in greater detail in future papers.

There are those who question the continued need for chemical nomenclature in a world increasingly dominated by computers with graphics capabilities. We must bear in mind that there is still, today, a use for such obsolete names as muriatic acid and that it is difficult to draw a realistic structure for, say, polyurethane. Hence, it seems likely that a need for chemical nomenclature will remain. We should therefore continue to strive to make chemical nomenclature serve our needs as efficiently as possible. Nodal nomenclature (with its potential for such uses as information storage and retrieval, machine naming, structure indexing, screen searching, and substructure searching) gives us hope that we may succeed.

REFERENCES

1. The work described here is part of the exploration by CAS of novel nomenclature systems. It does not imply any plans by CAS to change existing nomenclature practices in the foreseeable future.
2. IUPAC, *Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F, and H,* Pergamon Press, Oxford (1979): (a) *ibid,* Rules A-31, A-32, and B-14.
3. IUPAC, *Definitive Rules for Nomenclature of Inorganic Chemistry 1957,* Butterworths, London (1959).
4. J. H. Fletcher, *J. Chem. Doc.,* **7** (1967) 64.
5. W. C. Fernelius, K. Loening, R. N. Adams, *J. Chem. Educ.,* **48** (1971) 433.
6. For references to, and a review of, the work of G. Siboni and M. Perino, see P. E. Verkade, *Bull. Soc. Chim. Fr.* (Part II), (1978) 13.
7. N. Lozac'h, A. L. Goodson, W. H. Powell, *Angew. Chem.* **91** (1979) 951: *Angew. Chem. Int. Ed. Engl.* **18** (1979) 887.
8. N. Lozac'h, A. L. Goodson, in preparation.
9. A. L. Goodson, *J. Chem. Inf. Comput. Sci.* **20** (1980) 167.
10. A. L. Goodson, *J. Chem. Inf. Comput. Sci.* **20** (1980) 172.
11. G. W. Wheland, *Advanced Organic Chemistry,* Wiley, New York, p. 78 (1949).
12. J. V. Knop, I. Gutman, N. Trinajstić, *Kem. Ind.* **9** (1975), 505.
13. L. J. O'Korn, *Algorithms for Chemical Computations,* ACS Symposium Series No. 46, American Chemical Society, Washington D. C., p. 122 (1977).
14. F. L. Taylor, *Ind. Eng. Chem.* **40** (1948) 734.
15. A. Baeyer, *Ber. Dtsch. Chem. Ges.* **33** (1900) 3771.
16. R. G. Dunn, W. Fisanick, A. Zamora, *J. Chem. Inf. Comput. Sci.* **17** (1977) 212.
17. M. Elliott, *Chem. Ind. (London)* (1979) 757.
18. H. Wijnne, *Quant. Struct.-Act. Anal., Proc. Symp. Chem. Struct.-Biol. Act. Relat.: Quant. Approaches,* 2nd, Suhl, E. Ger., (1976). *Abh. Akad. Wiss. DDR, Abt. Math.,* (1978) 2N, 283.

SAŽETAK

Primjena kemijske nomenklature zasnovane na grafovima u teorijskoj i preparativnoj kemiji

A. L. Goodson

Opisan je razvoj na grafovima zasnovanih sistematskih imena koja sadrže matematički opis molekularnih grafova. Ova imena mogu se smatrati oblikom sažete tablice povezanosti. Diskutirana je upotreba ovih imena za svrhe skladištenja i vađenja informacija, u traženju podstruktura i kao sredstva u raspoznavanju obrazaca, odnosa strukture i aktivnosti, dizajniranju lijekova, itd.