(unless somehow we wish to identify H_2 with the empty graph, whence one would presumably add a constant into our formal cluster expansions).

The elimination of H from larger fragments follows a similar development. Let G_A be a (connected induced) subgraph containing an A atom at a given position, and let i be the number of bonds this special A has internal to G_A. Then counting the additional bonds radiating from such an A in G_A in two different ways (as before), we find

$$n_\Gamma[G_A](v_A - i)n_A[G_A] = \sum_B n_\Gamma[G_{AB}]n_{G_A}[G_{AB}] + 2\sum_B n_\Gamma[G_{A:B}]n_{G_A}[G_{A:B}] \quad (A.2)$$

where G_{AB} and $G_{A:B}$ are (induced) graphs obtained from G_A by joining a B atom with a single or double bond. For the case that there are no H atoms in the fragment G_A, eq A.2 can be used to eliminate fragments G_{AH}. For the case that there is one H atom in G_A, eq A.2 can be used to eliminate G_{AH} in terms of G_A, which in turn are eliminated as indicated in the preceding sentence. The idea is similar if G_A contains two (or more) H atoms.

APPENDIX B. LINEAR NEAR-DEPENDENCE

The establishment of the general relation eq 3.2 most quickly follows through the use of known Möbius function theory.[10] This theory derives from the inversion of eq 3.1 to give

$$x(G) = \sum_{\gamma \in C(G)} \mu(\gamma, G)X(\gamma) \quad (B.1)$$

where $\mu(\gamma, G)$ is the *Möbius function* associated to the subgraph partial-ordering relation defined on $C(\Gamma)$. For this "poset" it turns out that[11]

$$\mu(\gamma, G) = \begin{cases} (-1)^{|G| - |\gamma|}, & G \in \varepsilon(\gamma) \\ 0, \text{ otherwise} \end{cases} \quad (B.2)$$

Now also, as is suggested upon substitution of eq B.1 into eq 2.1

$$\sum_{G \in C(\gamma)} \mu(\gamma, G) = \delta(\gamma, \Gamma) \quad (B.3)$$

But the G-sum over labeled graphs here may be broken into two parts; first, one over all "unlabeled" graphs (i.e., all chemical isomorphism classes), and second, one over all possible labelings available for the unlabeled graph, so that

$$\sum_{[G] \in C[\Gamma]} \sum_{G' \in [G]} \mu(\gamma, G') = \delta(\gamma, \Gamma) \quad (B.4)$$

Yet we also may introduce a sum over all $\gamma' \in [\gamma]$, with the right-hand side of eq B.4 remaining unchanged since it is nonzero only in the case that there is but a single term in the sum, so that

$$\sum_{[G] \in C[\Gamma]} \sum_{\gamma' \in [\gamma]} \sum_{G' \in [G]} \mu(\gamma', G') = \delta(\gamma, \Gamma) \quad (B.5)$$

Now the number of $\gamma' \in [\gamma]$ for which $\mu(\gamma', G')$ at a given G' takes a (fixed nonzero) value is just $m_{G'}[\gamma]$ of Section 3, so

$$\sum_{[G] \in C[\Gamma]} \sum_{G' \in [G]} m_{G'}[\gamma]\mu(\gamma', G') = \delta(\gamma, \Gamma) \quad (B.6)$$

Finally utilizing eq B.2 with the elimination of the G'-sum, we obtain the desired result of eq 3.2.

REFERENCES AND NOTES

(1) See, e.g., Trinajstić, N. *Chemical Graph Theory*; Chemical Rubber Co.: Boca Raton, FL, 1983.
(2) Klein, D. J. *Int. J. Quantum Chem.* **1986**, *S20*, 153-173.
(3) Smolenskii, E. A. *Russ. J. Phys. Chem. (Engl. Transl.)* **1964**, *35*, 700-702.
(4) Gordon, M.; Kennedy, J. W. *J. Chem. Soc., Faraday Trans. 2* **1973**, *69*, 484-504.
(5) Essam, J. W.; Kennedy, J. W.; Gordon, M.; Whittle, P. *J. Chem. Soc., Faraday Trans. 2* **1977**, *73*, 1289-1307.
(6) See, e.g., Sellwood, P. W. *Magnetochemistry*; Interscience Pub.: New York, 1956.
(7) Hameka, H. F. *J. Chem. Phys.* **1961**, *34*, 1996. Haley, L. V.; Hameka, H. F. *J. Am. Chem. Soc.* **1974**, *96*, 2020-2024. O'Sullivan, P. S.; Hameka, H. F. *J. Am. Chem. Soc.* **1970**, *92*, 25-32.
(8) Burnham, A. K.; Lee, J.; Schmalz, T. G.; Beak, P.; Flygare, W. H. *J. Am. Chem. Soc.* **1977**, *99*, 1836-1844.
(9) Randić, M. *Chem. Phys. Lett.* **1978**, *53*, 602-605, and private communication.
(10) See, e.g., Berge, C. *Principles of Combinatorics*; Academic Press: New York, 1971; Chapter 3.
(11) Domb, C. In *Phase Transitions and Critical Phenomena*; Domb, C., Green, M. S., Eds.; Academic Press: New York, 1974; pp 1-94.

Representation of Molecular Graphs by Basic Graphs†

MILAN RANDIĆ

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311

Received August 27, 1991

We consider the problem of the analytical representation of graphs, molecular graphs in particular, based on a sufficiently broad selection of invariants that permits analogies with analytical representations of vectors. Desirable features of basis invariants are discussed and illustrated on paths and matching paths. It appears that they form a sufficiently broad basis for representation of graphs and for discussion of their differences and similarities. The approach is illustrated on smaller alkanes. Next we consider an ad hoc set of several pairs of graphs which have the same count of selected invariants and, thus, may be suspected as counterexamples to the uniqueness of the prime codes. Finally, similarity for a set of monocyclic monoterpenes is considered. Factors influencing a measure of similarity are discussed. Normalization of similarity matrices is proposed when comparisons are based on the representation of graphs using a different number of components. The notion of an overall or global similarity based on all components of a sufficiently comprehensive set of invariants is suggested.

INTRODUCTION

Mathematical modeling in chemistry, physics, and biology often starts with graphs as the objects, the examination of which may answer the questions of interest in such studies. In contrast to computational approaches to chemistry, physics, and biology, conveyed by adopting a suitable vector basis, such as illustrated by the use of selected basis functions in molecular orbitals calculations, in chemical graph theory there are no

† Dedicated to Professor Manfred Eigen (Nobel Laureate in Chemistry, 1967).

0095-2338/92/1632-0057$03.00/0

similar quantities that will play a role of basis functions of a vector space. Such basic graphs, if the analogy with vectors can be shown, have yet to be selected, assuming that collectively their properties will suffice to characterize in enough *details adequately* an arbitrary graph or structure. Then we would be justified to refer to such a collection of graphs as a basis. It is the purpose of this article to suggest one such choice.

DESIDERATA

We have to qualify the attributes "details" and "adequate", which are mentioned as conditions that basis vectors have to satisfy. By "adequate" we mean that two nonisomorphic graphs when represented via the basis graphs should, in general, result in different characterizations; that is, the count of the selected basis invariants ought to be different. To have a unique representation is desirable of course. However, being aware of the difficulties in arriving at a unique characterization of graphs based on *invariants* alone, we will accept a more pragmatic route in demanding a high discrimination among graphs where coincidence (duplicates) are the exception rather than the rule. In particular we are interested in discriminating among graphs depicting alkanes up to a certain size N, with N being as large as possible. By "details", we mean a possibility to use the same invariants to characterize molecular fragments, components, or subgraphs and thus arrive at a characterization of local, rather than global, structural features. Finally, as a requirement we expect that basis graphs will lead to characterizations in which structurally similar objects will result in representations of an apparent similarity. If the invariants used to characterize graphs are of a direct and simple structural origin, then their use will result in a direct and simple structural interpretation of the results, which is of paramount importance in the analysis of the properties of molecules.

We will very briefly review selected properties of vectors, as these will offer general guidance to desirable features for basis descriptors for graphs. Familiar basis vectors (functions) are unit vectors for finite dimensional vector spaces, like vectors **i**, **j**, and **k** for the 3-dimensional Cartesian coordinate system; functions sin **kx** and cos **kx** used in the Fourier expansion of periodic functions; and the Gaussian basis functions widely used in molecular orbital calculations. The first two represent complete basis for a finite vector space and an infinite vector space, respectively, and the latter illustrates a pragmatic but incomplete basis for Hilbert vector space (discrete infinite number of components including continuum). In each case vectors (periodic functions in the case of Fourier decomposition and molecular wave functions in quantum chemistry) are represented as a string of numbers, each component characterizing the role of a single descriptor (basis function).

The most desirable feature of basis functions (vectors) is that they form a *complete* basis, so an arbitrary function (vector) can be fully represented, i.e., without loss of structural information. In such instances the object (vector, function) can be *reconstructed*, at least in principle; that is, the process can be reversed and from the knowledge of the parts, a full structure is re-generated. Reconstruction thus implies *uniqueness* for the representation, and conversely, a lack of uniqueness points to limitations of the basis.

Another desirable feature that simplifies computations and makes the interpretation of results possible is the orthogonality of the descriptors used. With the introduction of metric via scalar (dot) product in a vector space, one can orthogonalize basis functions if they were not already orthogonal. Of course, there are many alternative basis (or coordinate systems) to choose from, and while in some cases the choice is unimportant in other situations the choice may be critical and can make the difference between solving a problem or not.

All this is well known for vectors (functions), but how can such notions be extended to graphs? An open problem is that of identifying basis descriptors for graphs such that from the corresponding characterizations graphs can be reconstructed. It is believed that any *finite* list of simple graph invariants will not suffice, that is, will not lead to a unique list. It follows then, if this is true, that if we succeed in developing basis functions for graphs, the associated graph space will be similar to vectors of infinite dimensional space. The pragmatics then suggest a more modest goal: Design a finite basis (as in the case of the Gaussian basis function in quantum chemistry) to characterize graphs, and molecular graphs in particular, that will nevertheless result in unique representations for graphs up to a critical size N (N shows the number of vertices).

ANALOGY WITH VECTORS

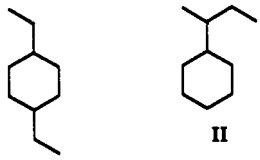
Useful and important properties of vectors include orthogonality, decompositions, projections (via scalar product), normalization, and transformation (with a change of the basis). While a string of numbers $(a_1, a_2, a_3 ..., a_k)$ representing a vector automatically satisfies the above computational features for a similar string of graph invariants $(i_1, i_2, i_3, ..., i_k)$ one has yet to define orthogonality, decomposition, projection, normalization, and transformations. The key concept that allows one to consider metrics is the scalar product. Thus, we have to find an analogous process that will assign to a pair of molecular descriptors a positive quantity that satisfies the axioms of metrics. This problem has recently been resolved.[1-4] The idea of orthogonality for molecular descriptors was outlined on the basis of linear regression between descriptors as the critical measure of the degree of their non-orthogonality. Orthogonality of the descriptors is interpreted as their mutual unrelatedness, that is, the lack of mutual correlation. If two descriptors D_i and D_j are uncorrelated, they are viewed as orthogonal, as each will give independently of the other in a multivariate regression a component that the other cannot account for. However, in practice as a rule various molecular descriptors are not only related but often highly interrelated causing multicollinearity,[5] a serious deficiency of multivariate analysis. If instead of the two descriptors D_i and D_j, which are interrelated (but not strictly collinear, in which case they strictly duplicate one another), we correlate one against the another, say D_j against D_i, then one can take beside D_i the residual Res (j,i), the part of D_j that D_i cannot reproduce, as the other orthogonal descriptor. With such a procedure we succeeded in extending the notion of orthogonality to structural descriptors. This fulfills one of the important desiderata for descriptors for the characterization of molecular graphs.

The other important characteristic for a basis is that of completeness, or more modestly, a fair spanning of the structure space. In quantum chemical computations on optimization of the molecular geometry using the minimum energy as the criterion, a deficiency of an incomplete basis can be judged by the numerical value for the computed energy. The best (lowest) energy can be compared to computations using alternative basis functions, and if lower it shows that a better basis was adopted. It is very important to have a criterion which reflects upon the properties (the "degree" of incompleteness) of a basis even though it does not tell one in which respect the basis is deficient, that is, in which "direction" one should make corrective steps.

Can we "measure" the degree of incompleteness of a collection of graph descriptors taken as a basis? This, of course, is very desirable not only when comparing the results of different analyses but also when designing a new basis. We argue here that it is possible to measure the degree of completeness of basis descriptors even if indirectly and without such a

powerful tool as the variation principle. We propose that such a measure is given by N, the size of graphs for which the particular basis assigns to two nonisomorphic graphs the same ordered set of descriptors. This condition can be somewhat relaxed by restricting such a requirement only for structures (graphs) of certain classes, such as acyclic structures (trees), molecular graphs (rather than graphs in general), etc.

The proposed criterion applies equally to (simple) quantum chemical calculations. Consider the well-known Hueckel molecular orbital calculations (HMO) for the two molecules



I II

1,4-divinylbenzene (I) and 2-phenylbutadiene (II) which somewhat surprisingly have all eigenvalues the same (orbital energies with the HMO model). Although the data on these molecules were available for a while,[6] the occurrence of a same set of eigenvalues for two different compounds was overlooked during the "golden age of HMO" and has only fairly recently been indicated by Živković.[7,8] Since then considerable study of these so-called isospectral graphs (isospectral as an adjective signifies the same graph spectra, as eigenvalues are referred to in mathematical literature) revealed that they are rather common.[9] A way to "correct" for this deficiency is to include in the hamiltonian some terms neglected in "the nearest-neighbor interaction" of Bloch,[10] which underlies the HMO model. Such an improvement of π-electron calculations was proposed by Pariser–Parr[11] and Pople[12] and proved to be a very successful model for π-electron polycyclic systems. One could have removed the degeneracy of the HMO eigenvalue problem by enlarging the basis (instead of refining the hamiltonian), for example, by introducing functions that have an increased range and discriminate the environment beyond the nearest neighbor. Such an alternative, which may be less attractive from chemical points of view, nevertheless shows that an occurrence of cases of "duplicate" computations is basis-sensitive. One may therefore characterize the limitations of isospectrality of HMO by determining, for each class of compounds of interest, the smallest number N for which it produces isospectral graphs. Overall, the smallest N for isospectral graphs is $N = 6$,[13] but for chemical interest we may take $N = 12$, which corresponds to the already-mentioned benzene derivatives (for stable molecules, and in case of radicals one finds $N = 11$[8]).

SINGLE DESCRIPTORS

To illustrate the limitations of a small basis, we will briefly review topological indices as graph descriptors.[14-16] We will consider only the graph invariant as a source for characterization of graphs excluding codes dependent on labeling, such as various canonical representations of graphs. Graph *invariants* are quantities that are independent of the choice of labels for vertices of a graph or the form of pictorial representation of a graph. Such quantities may be viewed as mathematical properties of molecules (graphs). Hence molecules have beside their physical, chemical, and biological properties also numerous mathematical properties. Several studies considered interrelatedness among topological indices[17-20] as well as the earliest occurrence of isocodal structures, that is, the smallest structures having a same count of selected invariants. In Table I there is a short list of selected topological indices which are given either as integers or alternatively as reals. The former are typically the outcome of an enumerative task and already, for relatively small values of N, often less than 10, show duplicate characterization for nonisomorphic structures. A weighting procedure typically extends the size of graphs with unique N. Some more general indices, such as the sum of all weighted paths, typically assign to different graphs different numerical values. By using the same weighting factors for bonds of different type as introduced with the connectivity index, one already gets an impressive discrimination among graphs, which justifies referring to this particular invariant as the molecular ID (identification) number.[21] Moreover, when the weights assigned to bonds of different bond type are based on prime numbers[22] one maximally reduces a chance for coincidences and thus increases the pool of structures with unique descriptors. For trees Knop and co-workers[23] reported the smallest pair of trees with the same prime number ID as having $N = 20$ vertices, one such pair among over half a million graphs!

Table I. Selection of Topological Indices

		ref
Topological Indices as Integers		
Hosoya's Z index	count of nonadjacent bonds	44
Wiener number W	sum of all distances	*a*
Topological Indices as Reals		
connectivity index χ	weighted bond types	26
Balaban's J index	weighted distances	*b*
molecular ID no.	sum of weighted paths	21
Kier's shape index κ	scaled connectivity difference	*c*
graph bond order P'/P	sum of bond orders for all bonds	*d*

[a] Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 2636. [b] Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399. [c] Kier, L. B. *Quant. Struct.—Act. Relat.* **1986**, *5*, 1. Kier, L. B. *Acta Pharm. Jugosl.* **1986**, *36*, 171. [d] Randić, M. *J. Math. Chem.* **1991**, *7*, 155.

A single descriptor hardly represents a basis. It is, however, important to recognize how much distinctive structural information can be condensed in a meaningful way into a single descriptor. By "meaningful" we mean that structurally *related* graphs show *similar* numerical values for the invariant. This property of the invariant is so essential for applications, particularly in structure–property–activity studies, that it ought to be incorporated when constructing a basis for graph representation. Since single descriptors show limitations obviously, a simple ad hoc combination of descriptors, such as a superindex,[24] will similarly, sooner or later, result in the occurrence of isocodal structures. The situation is somewhat better if one considers an ordered collection, a sequence, of descriptors. For structures to have duplicate descriptors we need a simultaneous coincidence in all the descriptors adopted. This is less likely to occur for smaller graphs. Because we would like to develop an analogy with vectors, an ad hoc ordered set of descriptors is not an attractive proposition. The individual parts in such a sequence will typically be conceptually unrelated, and so the interpretation of the results may be hindered. It is not obvious in such situations which components should be given priority. Ordering is required, for example, in an orthogonalization process. The situation is similar in the traditional QSAR (quantitative structure–activity relationship[25]) where various molecular properties are used as descriptors, and where often it is not apparent which property is to be given a priority. A way out of such a dilemma is to consider components which are size-dependent because then a natural ordering of the components follows. An illustration of such a natural order is illustrated by the Fourier expansion of periodic functions using trigonometric functions, where the frequencies of sine and cosine basis functions dictate an ordering for the components.

SIMPLE BASES

We may view the connectivity index[26] combined with "higher" connectivity indices[27] as an attempt to introduce a

basis of molecular descriptors for representations of molecular graphs. Such a basis facilitates comparative studies of structure–property and structure–activity. It permits use of the *same* descriptors for a different set of compounds or use of the *same* descriptors on the same compounds but different properties. Not surprisingly, however, such a basis is still rather limited. Two structures having a same bond-type decomposition and "extended bond" types will necessarily have the same connectivity indices. However, this situation will occur for relatively large graphs. The smallest trees having all weighted paths equal are among those that Knop and co-workers, using an exhaustive computer search, identified as having the same molecular ID numbers.[28] Even if we truncate the basis by limiting the number of higher weighted paths (or higher connectivity terms) to be used in analysis, such truncated basis would still cover molecules of different forms and considerable size reasonably well. Weighted paths[19] may be viewed as a modification of the "higher connectivities", the distinction being that in the higher connectivity indices the weighting factors are associated with atoms while similar weighting factors are assigned directly to bonds when counting (that is, calculating) weighted paths.[30] Weighted walks[31] are defined similar to higher connectivities and weighted paths except that now count is performed for walks, which is computationally not involved.

The limitations of the connectivity indices or weighted paths in applications using multivariate regression analysis are not so much in an early occurrence of duplicate but in a lack of adequate to account for some specific structural features involving branching of the molecular skeleton. Such limitations of the connectivity indices (disregarding limitations of graphs as a model for a chemical structure) are apparent from a need to extend the basis by explicitly considering subgraphs as the so called "path-cluster" and "cluster":

and in this way to enrich the structural information on parent structures. A deficiency of the connectivity indices is also reflected in the occasional slow convergence in multivariate regressions, by-passed by using in addition to selected connectivity indices their reciprocals.[32] Other functional forms for the weighting procedure[33] may in some applications lead to a simpler (for example, linear regression rather than a quadratic one) dependence of a property on the structural invariant.

In retrospect, the introduction of the connectivity indices and weighted paths opened an important direction in the study of structure–property–activity relationships. This is even if past applications were occasionally in part marred by ad hoc and unsystematic use of selected connectivity descriptors or omission of others, both induced by the multicollinearity problem, which equally plagued alternative QSAR schemes. With the recently outlined orthogonality procedure[1-4] for the multivariate analysis using molecular descriptors, much of the past results could be critically evaluated and updated. Such efforts, which desirable, will not elevate the limitations inherent to the simple basis, the connectivity indices, or the path numbers, which use *too few* components to be able to satisfactorily characterize a general graph or structure. Thus, the problem remains to find a better, that is, more comprehensive basis for characterization of chemical graphs and later chemical structures.

PRIME BASIS

The higher connectivity indices and the weighted paths as the basis have one satisfactory property: a size-dependency. This allows components to be ordered in a natural way. This ordering property is lost when such bases are mended by inclusion of path-cluster subgraphs and other non-path subgraphs not only because of ambiguities of ordering components of the same size but also because of ambiguities in determining which non-path subgraphs to include and which not to. To improve the path basis we have to add subgraphs. Can this be done in some orderly fashion; that is, in a fashion that allows not only a natural ranking of the components but implies a well-defined structurally-based selection criterion for the inclusion/exclusion of additional graphs into the basis. We should dismiss as impractical the inclusion into a basis of *all* subgraphs of the graphs considered, such as implies in cluster expansions.[34-35] Beside computational complexities of such approaches when extended to a large graph, the inclusion of *all* subgraphs is counter to the notion of basis descriptors. A more restrictive selection of all graphs (subgraphs) of a certain class may result in a viable basis. This will critically depend on the properties of the selected graphs or subgraphs. For example, a set of Ulam's subgraphs, the subgraphs derived from a graph by erasing each time a single vertex, which if we assume the reconstruction conjecture[36] to hold would represent for individual graphs a *complete* basis, but it is likely to become overcomplete as the number of graphs considered increases, because in general it involves all graphs on $N - 1$ points. Caterpillars[37] defined as chains with branches of at most length one represents a well-defined class and may be considered as a basis. They form a subset of all trees, but as N increases the number of such graphs also increases fast. One could restrict the maximal valency in caterpillars to $a = 3$ (that is to caterpillars which represent binary trees) and consider the selected graphs as a basis. Rather than pursuing such direction, we will examine *path graphs* and their *subgraphs* as the basis for graph representation.

In mathematical literature there have been few recent papers that considered the problem of basis graphs, that is, the selection of graphs that may suffice to represent other graphs.[38-41] The seminal paper of Dewdney, in which he considered an analogy between certain factorings of graphs and basis for vector spaces which appeared 20 years ago,[42] posed a question:

> "Does there exist a set of graphs which behaves with respect to other graphs like the basis of a vector space?"

Dewdney proved the existence of such graphs, called "primal", most of which turned out to be disconnected. Restrictions on such graphs, like a request that in covering edges no same graph is used more than once, makes the above mathematical approaches of limited direct interest in chemistry. Nevertheless, the notion of primal graphs may stimulate development and a search for basic graphs that may suit chemical applications. At a recent graph-theoretical and combinatorial symposium we reported[43] on a characterization of graphs in which prime number labels were assigned to individual edges. Factors of an arbitrary number then in a unique way define which labeled edges will appear in a graph, which may have multiple edges, or loops or be disconnected. This work has lead to the idea that one way to enlarge the natural basis of *path graphs* is to consider all disjoint subgraphs of path graphs as an extension of the basis. Because of this tenuous connection with prime numbers and some parallel with primal graphs started by Dewdney and followed by Chinn (who generalized primal graphs to primary graphs), we will refer to the basis consisting of path graphs and all their subgraphs (connected and disconnected) as *prime* graphs (or prime paths, abbreviated PP). Such a basis implies partitioning of path graphs into disjoint components and enumeration of such parts individually.

In Figure 1 we listed smaller prime graphs with corresponding alphabetic labels and subscripts (used for convenience

Table II. Decomposition of Smaller Alkanes into the Prime Graphs Basis[a]

graph	prime decomposition	total no. of components
ethane	A_1	1
propane	$2A_1 + B_1$	3
butane	$3A_1 + 2B_1 + C_1 + C_2$	7
2-Me	$3A_1 + 3B_1$	6
pentane	$4A_1 + 3B_1 + 2C_1 + 3C_2 + D_1 + 2D_2$	15
2-Me	$4A_1 + 4B_1 + 2C_1 + 2C_2$	12
2,2-Me_2	$4A_1 + 6B_1$	10
hexane	$5A_1 + 4B_1 + 3C_1 + 6C_2 + 2D_1 + 6D_2 + E_1 + 2E_2 + E_3 + E_4$	31
2-Me	$5A_1 + 5B_1 + 3C_1 + 5C_2 + 2D_1 + 6D_2 + E_3$	27
3-Me	$5A_1 + 5B_1 + 4C_1 + 5C_2 + D_1 + 4D_2 + 2E_2 + E_4$	27
2,3-Me_2	$5A_1 + 6B_1 + 4C_1 + 4C_2 + 4D_2 + E_3$	24
2,2,3-Me_3	$5A_1 + 7B_1 + 3C_1 + 3C_2 + 3D_2$	21
heptane	$6A_1 + 5B_1 + 4C_1 + 10C_2 + 3D_1 + 12D_2 + 2E_1 + 6E_2 + 3E_3 + 4E_4 + F1 + 2F_2 + 2F_3 + 3F_4$	63
2-Me	$6A_1 + 6B_1 + 4C_1 + 9C_2 + 3D_1 + 13D_2 + 2E_1 + 4E_2 + 4E_3 + 2E_4 + F_3 + F_4$	55
3-Me	$6A_1 + 6B_1 + 5C_1 + 9C_2 + 3D_1 + 11D_2 + E_1 + 5E_2 + 2E_3 + 3E_4 + F_2 + F_3 + 2F_4$	55
3-Et	$6A_1 + 6B_1 + 6C_1 + 9C_2 + 3D_1 + 9D_2 + 6E_2 + 4E_4 + 3F_2 + 3F_4$	55
2,4-Me_2	$6A_1 + 7B_1 + 4C_1 + 8C_2 + 4D_1 + 14D_2 + 5E_3$	48
2,3-Me_2	$6A_1 + 7B_1 + 6C_1 + 8C_2 + 2D_1 + 10D_2 + 2E_2 + 2E_3 + 2E_4 + F_3 + F_4$	47
2,2-Me_2	$6A_1 + 8B_1 + 4C_1 + 7C_2 + 3D_1 + 12D_2 + 3E_3$	43
3,3-Me_2	$6A_1 + 8B_1 + 6C_1 + 7C_2 + D_1 + 8D_2 + 4E_2 + 2E_4 + F_4$	43
2,2,3-Me_3	$6A_1 + 9B_1 + 6C_1 + 6C_2 + 9D_2 + 3E_3$	39

[a] Components A, B, C, ... are illustrated in Figure 1. Successive letters indicate the presence of an additional vertex in the basis.

Table III. Decomposition of 18 Isomers of Octane into Prime Graph Basis

molecule	A_1	B_1	C_1	C_2	D_1	D_2	E_1	E_2	E_3	E_4	F_1	F_2	F_3	F_5	G_1	G_2	G_3	G_4	G_5	G_6	G_7	sum
octane	7	6	5	15	4	20	3	12	6	10	2	6	6	12	1	2	2	1	3	3	1	127
2-Me	7	7	5	14	4	22	3	10	8	7	2	4	6	9	0	0	1	0	0	2	0	111
3-Me	7	7	6	14	4	20	3	12	6	8	1	3	5	8	0	1	0	1	3	1	1	111
4-Me	7	7	6	14	5	20	2	10	5	8	1	6	4	11	0	0	2	0	0	3	0	111
3-Et	7	7	7	14	5	18	2	12	3	9	0	6	2	10	0	2	1	0	3	2	1	111
2,3-Me_2	7	8	7	13	4	20	2	10	6	6	0	2	4	7	0	0	1	0	0	2	0	99
3,4-Me_2	7	8	8	13	4	18	1	2	4	7	0	2	4	6	0	0	0	1	3	0	1	99
2-Me_3-Et	7	8	8	13	5	18	0	10	3	7	0	6	2	9	0	0	1	0	0	2	0	99
2,4-Me_2	7	8	6	13	5	23	2	8	8	5	0	2	4	6	0	0	0	0	0	1	0	98
2,5-Me_2	7	8	5	13	4	24	4	8	11	4	0	0	4	4	0	0	0	0	0	1	0	97
2,3,4-Me_3	7	9	8	12	4	20	0	8	7	4	0	0	4	4	0	0	0	0	0	1	0	88
2,2-Me_2	7	9	5	12	4	23	3	6	9	3	0	0	3	3	0	0	0	0	0	0	0	87
3,3-Me_2	7	9	7	12	4	19	1	8	4	5	0	2	2	6	0	0	0	0	0	1	0	87
3-Me_3-Et	7	9	9	12	3	15	0	12	0	7	0	3	0	6	0	0	0	0	3	0	1	87
2,2,3-Me_3	7	10	8	11	3	19	0	6	6	3	0	0	3	3	0	0	0	0	0	0	0	79
2,3,3-Me_3	7	10	9	11	2	17	0	8	4	4	0	0	2	4	0	0	0	0	0	1	0	79
2,2,4-Me_3	7	10	5	11	6	25	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	76
2,2,3,3-Me_4	7	12	9	9	0	18	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	64

only). Such graphs can be ordered first on the number of vertices involved followed by listing subpaths in a lexical decreasing order. The leading members are the paths of length k, labeled as A_1, B_1, C_1, ..., the successive letters indicate the increasing number of vertices. The last member of each alphabetical group represents the corresponding number $p(G,k)$ of Hosoya.[44] Since the number of paths in a graph can be obtained by a computer[45] and the numbers $p(G,k)$ are also known for many small graphs (being the coefficients of the acyclic or matching polynomial[46]), at least we have a way to check the count of novel invariants in some instances. The collection of subpaths for a chain having N vertices corresponds to the partition of the number N, excluding partitions involving one. One may consider extending the basis to include all partitions. Since partitions can be represented by Young diagrams for larger N, ordering of such partitions has to be specified by additional rules.[47,48] A convenient label for prime graphs may be based on writing disconnected path components as factors. Thus $C_2 = (p_1)^2$, $D_2 = p_1p_2$, $E_2 = p_1p_3$, $E_3 = (p_2)^2$, $E_4 = (p_1)^3$, etc.

In Figure 2 we illustrate the enumeration of prime basis components for a graph of 2-methylhexane. In larger graphs the enumeration of path graphs and their subgraphs becomes tedious and error-prone. Use of recursive relations similar to those referred to by Poshusta and McHughes[34] for use in cluster expansions will alleviate some of the difficulties of path and subpath enumeration and reduce errors in the count.

PRIME DECOMPOSITION OF SMALLER ALKANES

In Table II we show decompositions of smaller alkanes (up to heptanes), and in Table III we continue with similar decomposition of all octanes. In view that the enumeration of paths and subpaths is tedious and error-prone, it is worth noticing some regularities in the derived coefficients that show frequencies of occurrence of each component. For example, the hexane isomers coefficients of B1 and C2 add up to 10; for the heptane isomers the coefficients add up to 15; and for octanes the same sum is 21. Hence the sum of these components is given by the respective binomial coefficients (N being the number of all edges). Higher coefficients also show regularities for the sum of subgraphs having the same number of edges but the pattern is more involved. The last column in Tables II and III gives the number of components in the prime basis decompositions. A number of isomers show the same sum of the components, such as all 2-methyl- and 3-methylalkanes. Such regularities offer an additional check on the accuracy of the enumeration of the prime basis components.

As the size of alkanes increases, the number of components that can occur also increases. For octane isomers we need 21 prime graphs, for nonane isomers this becomes 28, and for

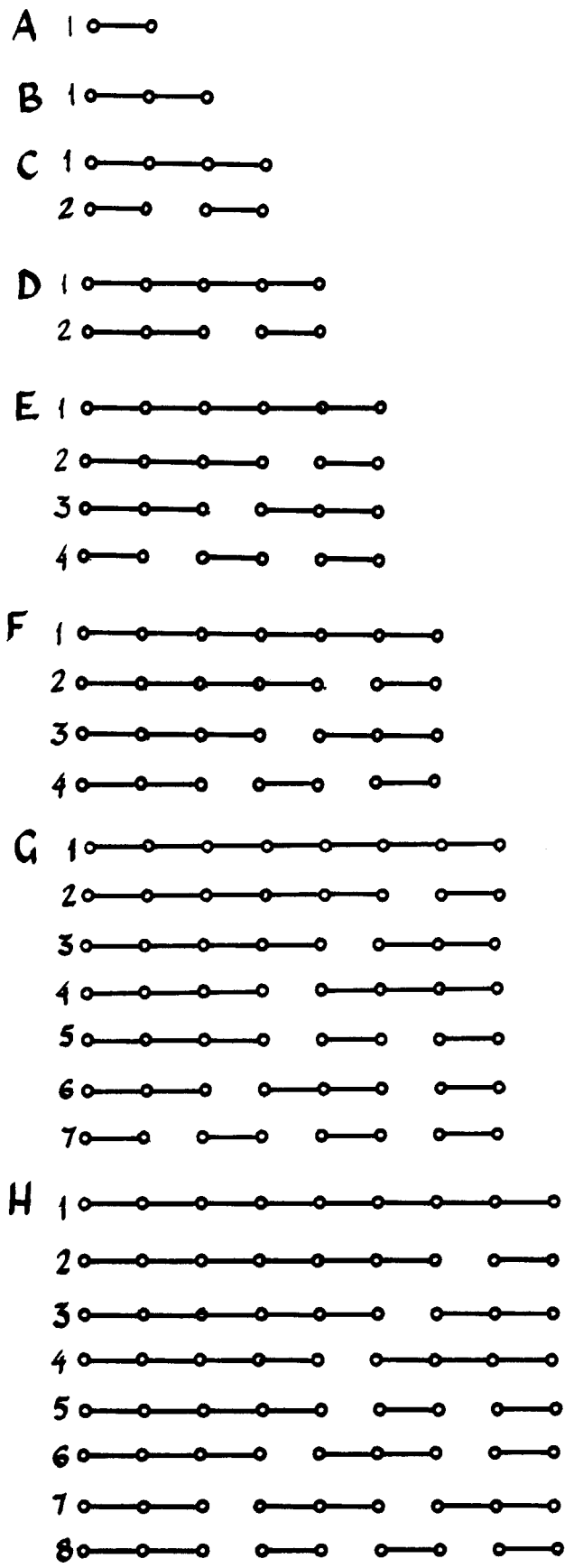


Figure 1. Prime graphs (up to n = 10 vertices).

decanes it is already 39. However, the growth is modest if one compares these numbers with the number of (connected) graphs having n = 8, 9, and 10 vertices. In practical applications one may truncate such extensive basis and confine attention to fewer (smaller) components, as has been the case with the application of weighted paths and connectivity indices.

Table IV. Graphs Examined for Possible Duplicate Prime Basis Decomposition[a]

Graph Distinguished by Path Sequence (Figure 3)	
same molecular ID no.	ID = 27.9964626 ref 28
A	14, 18, 20, 18, 16, 12, 6, 1
B	14, 18, 20, 19, 16, 14, 4
same molecular ID no.	ID = 29.6164833 ref 28
C	15, 23, 25, 27, 17, 10, 3
D	15, 23, 25, 25, 19, 8, 3
many same invariants	ref *b*
E	20, 40, 60, 87, 122, 153, 175, 195, 206, 166, 81, 20, 2
F	20, 40, 60, 87, 122, 153, 176, 195, 204, 165, 81, 20, 2
Graphs With Identical Path Sequences (Figure 4)	
same prime ID no.	total no. of paths = 23
G, H	19, 33, 45, 52, 35, 6
Slater's Pair	total no. of paths = 48
I, J	17, 30, 48, 42, 16

[a] Graphs of Figure 3 (top part of the table) can already be differentiated from the count of paths. Graphs of Figure 4 have the same path count but possess at least one component in the prime basis decomposition in which they differ. [b] Chalcraft, D. A. *J. Graph Theor.* **1990**, *14*, 341.

EXAMINATION OF SELECTED ISOCODAL GRAPHS

None of the alkanes in Tables II and III show the same decomposition in prime paths, but graphs having a duplicate signature may be expected among larger graphs. It is of interest therefore to examine graphs that have the same number of different invariants to see if an enlarged basis will eliminate duplicate characterization. In Figure 3 we illustrate this for selected graphs suspected as counterexamples for isomorphism, that is, suspected that they could lead to the same prime graph decomposition. Such graphs should be sought among various *isocodal* structures, structures having the same count for several graph invariants. The simplest cases are the so-called isospectral graphs, graphs having the same set of eigenvalues for adjacency matrices.[9] Next, one may consider graphs having the same distance sum sequence.[49] Graphs having the same ID number also should be examined,[50] in particular when the weighting factors in the path count are based on a prime number.[51] If two graphs differ in their path count, or in their matching polynomial [which counts the number of nonadjacent bonds $p(G,k)$ of Hosoya], this will suffice to show that they cannot present a counterexample to unique prime path decomposition. This is the case with the graphs of Figure 3 which show the different path counts. Path counts in these cases suffice therefore to indicate such graphs as nonisomorphic, which implies that such pairs of graphs have different components in a simple basis based on path counts. The graphs of Figure 3 therefore should be discarded as potential counterexamples to the uniqueness of the prime basis decomposition, and the search for counterexamples should be confined to graphs having the same path counts. However, in practice it is often easier to count the selected components of prime paths than testing first if the path counts are same. Particularly it may be easy to count prime path components that completely cover the graph, i.e., represent a spanning (disconnected) subgraph, because there are fewer of such and they often can be readily constructed. Use of the ALLPATH program[45] will facilitate the search for prime paths, not only by giving the count of paths but by allowing one to extend the count to selected prime path components by considering corresponding subgraphs of a graph examined.

In Table IV we give the path sequence for some of the graphs of Figure 4, a visual inspection of which is somewhat involved, particularly as a construction of spanning subgraphs is less apparent. In each such case it suffices to indicate a

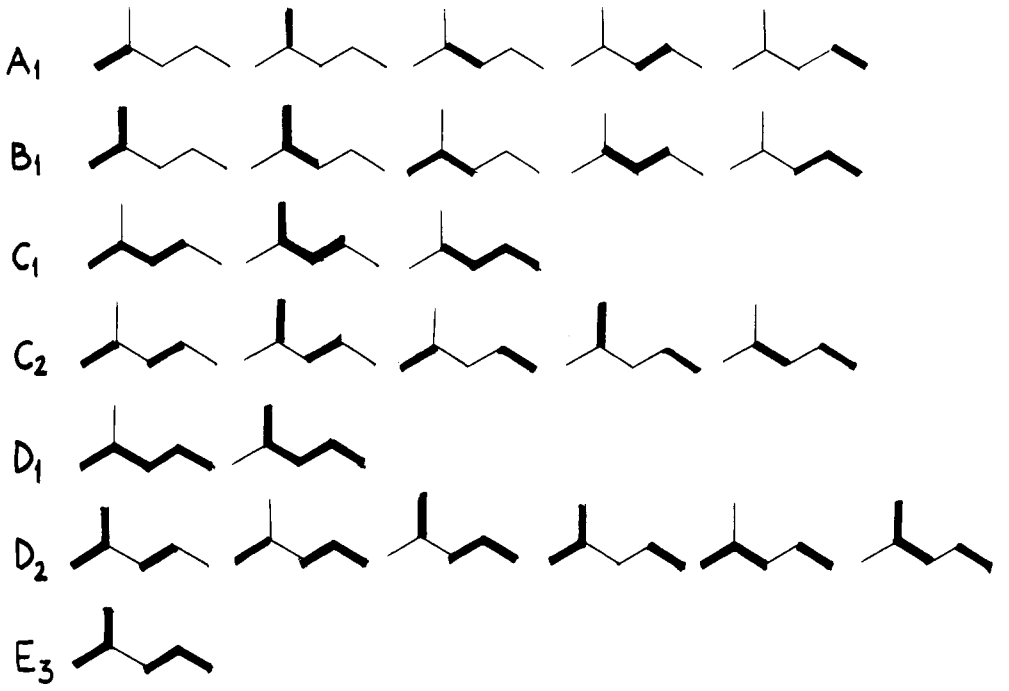


Figure 2. Illustration of the count of prime paths for 2-methylpentane.

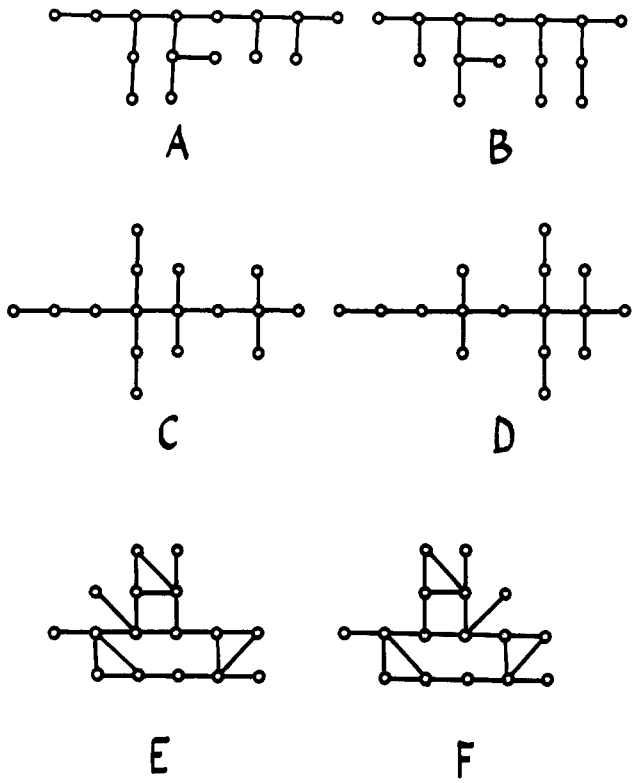


Figure 3. Illustration of several pairs of graphs with several invariants the same but showing different path counts.

single component in which the count differs for the two structures considered. The component of interest in many instances can be selected by inspection. Graph *H* in Figure 4 has a component $(p_3)^5$ which is absent in graph *G*, while graph *I* has the component $(p_2)^2(p_3)^3(p_5)$ which is absent in graph *J*. Therefore the two pairs will have different prime path decomposition. In all cases examined so far we found at least one component for which the count of the prime graphs is different. In particular it is significant that the prime decomposition is distinct for graphs *I* and *J*, the two graphs constructed by Slater[49] as a counterexample for a conjecture that a list of path sequences for all vertices in a graph may

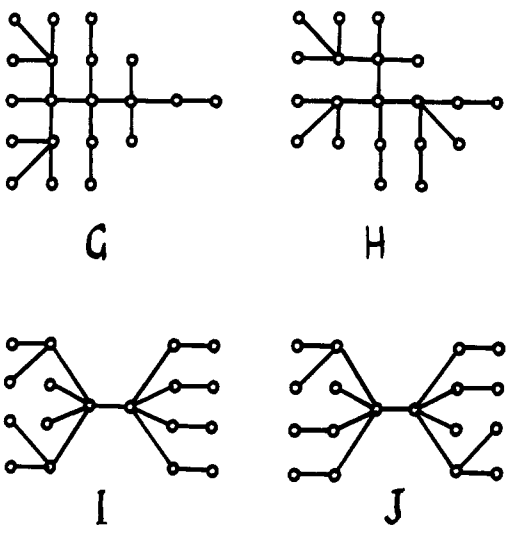


Figure 4. Illustration of pairs of graphs having the same count of paths, but having a different prime graph decomposition. Top right graph has a component $(p_3)^5$, while the top graph at the left can support at most four p_3 paths. Bottom left graph supports $(p_2)^3(p_3)^3(p_5)$, but the graph at right cannot be covered with the same subgraphs.

be unique.[51] Equally the only graph on $n = 20$ vertices for which weighted paths produce the same identification number even when individual edge types are weighted by a prime number also is discriminated against using the prime paths. These results are remarkable in view of the fact that we use for the characterization of graphs integers not reals, such as in weighted paths and the corresponding extension to weighted partitioned paths.

The proposed prime path basis is sufficiently broad to meet various needs in the modeling of structures by graphs. It is unlikely that duplicate characterizations will emerge for relatively small graphs. Equally, it is unpractical to launch an exhaustive search for isocodal systems. Such graphs, which may well exist, are more likely to be found either by accident or by design. A part of the difficulty is that for larger graphs, providing no counterexample emerges among the relatively small graphs, enumeration of subgraphs even if they are merely paths is quite an involved task so that when contemplated it ought to be for some valid reasons.

Table V. Similarity/Dissimilarity Table for 18 Isomers of Octane[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0	7.3	6.6	6.0	7.1	10.0	14.7	9.7	12.2	15.3	14.9	16.8	13.0	14.8	17.2	16.4	24.5	24.9
2		0	5.7	5.7	9.7	5.7	12.2	9.0	6.1	8.9	9.7	10.6	9.2	14.7	12.3	12.8	18.6	20.0
3			0	6.8	6.6	5.5	11.3	8.1	7.8	10.8	9.6	11.8	8.4	11.0	12.1	11.5	20.3	20.4
4				0	6.0	6.6	12.0	5.2	9.1	13.2	11.6	13.9	8.9	12.4	13.7	13.0	21.2	21.4
5					0	8.5	12.5	5.4	11.7	15.6	12.9	15.9	9.6	8.9	14.6	12.9	23.3	22.4
6						0	9.6	6.6	4.8	8.5	5.6	8.7	4.5	10.5	8.1	7.7	17.1	16.7
7							0	10.6	9.9	13.0	9.0	11.0	7.7	12.2	8.5	9.2	16.0	14.0
8								0	9.7	14.2	10.2	13.7	6.4	8.3	11.5	9.9	20.4	19.3
9									0	5.0	5.5	5.4	6.4	14.0	7.9	9.7	13.4	15.6
10										0	7.7	4.0	10.1	17.4	9.6	12.1	12.0	15.3
11											0	6.2	5.0	11.9	3.6	5.4	13.3	12.5
12												0	8.5	16.1	6.8	10.0	9.5	12.1
13													0	8.6	5.7	4.9	15.7	14.1
14														0	12.0	8.7	22.8	19.1
15															0	4.2	12.4	9.5
16																0	16.2	11.8
17																	0	10.9
18																		0

[a] The small entries in the table indicate more similar structures.

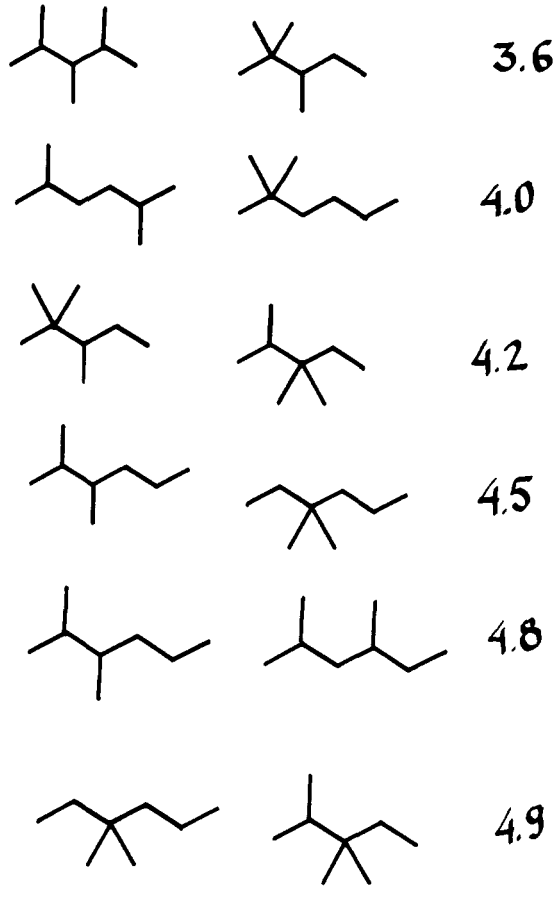


Figure 5. Several of the most similar octane isomers.

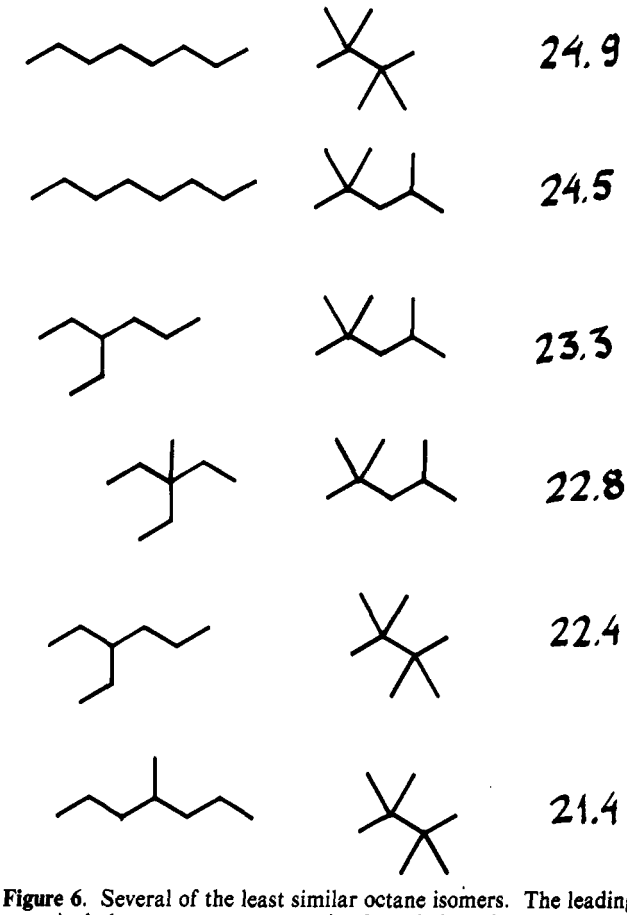


Figure 6. Several of the least similar octane isomers. The leading cases include a not-so-apparent pair: 3-methyl-3-ethylpentane and 2,2,3-trimethylpentane.

PRIME CODES AND SIMILARITY OF STRUCTURES

We do not claim uniqueness of the prime graph characterizations. Nevertheless the size N (the number of vertices) of the counterexamples to the uniqueness of the representation of graphs in prime basis is likely to be relatively large. Hence, prime basis is likely to perform well, approaching the behavior of a complete vector basis. However, there is another very important property that we would like basis functions to possess: apparently similar structures to be characterized by apparently similar codes. To see how this requirement is fulfilled in Table V we give the similarity (dissimilarity) matrix for the 18 octane isomers. In Figure 5 we show pairs of the most similar, and in Figure 6 are pairs of the most dissimilar octane isomers when the similarity is measured as the Euclidean distance for the structures represented by vectors in which the prime paths count signifies the component. Inspection of Figures 5 and 6 shows a very satisfactory correlation between the degree of similarity and apparent likeness or lack thereof. Occasionally one may detect a pair of isomers that are found similar (like 2,3,4-trimethylpentane and 3,3-dimethylhexane, similarity $S = 5.00$), which is not so apparent from their molecular graphs, or which appear less dissimilar than the quantitative measure suggests, like the pair 3-methyl-3-ethylpentane and 2,2,3-trimethylpentane with $S = 22.8$.

SIMILARITIES AMONG MONOTERPENES

We will illustrate the use of the prime paths characterization of molecular graphs on a set of monocyclic monoterpenes of

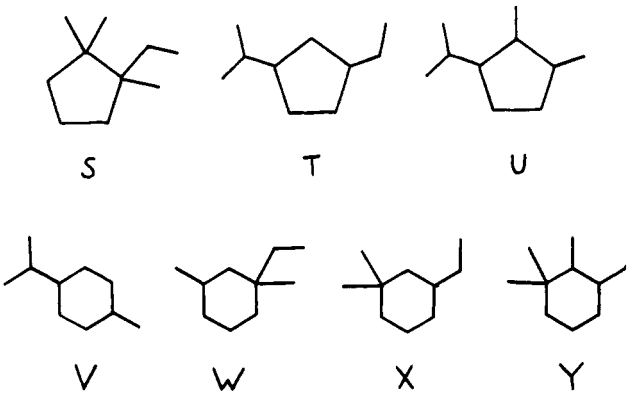


Figure 7. Carbon skeletons of the monoterpenes examined.

Table VI. Decomposition of Seven Monoterpene Skeletons of Figure 7 into Prime Graph Basis Use of Euclidean Distance in a 32-Component Representation of the Structures (Based on Prime Graph Decomposition)

		component structure						
		S	*T*	*U*	*V*	*W*	*X*	*Y*
1	A1	10	10	10	10	10	10	10
2	B1	16	13	14	13	14	14	15
3	C1	20	15	17	14	15	14	17
4	C2	29	32	31	32	31	31	30
5	D1	17	16	18	14	16	16	16
6	D2	64	65	66	67	67	69	68
7	E1	10	14	13	16	15	16	14
8	E2	52	55	54	54	51	48	53
9	E3	26	29	30	33	31	34	34
10	E4	28	37	33	36	33	32	29
11	F1	6	11	8	12	10	10	10
12	F2	27	39	37	34	33	33	26
13	F3	31	40	40	44	38	39	45
14	F4	58	76	70	77	70	72	61
15	G1	2	2	2	4	3	4	3
16	G2	8	18	13	22	15	18	6
17	G3	11	16	19	18	14	18	15
18	G4	5	9	8	10	7	6	10
19	G5	24	36	30	36	30	26	24
20	G6	27	38	38	45	34	40	30
21	G7	7	13	9	9	10	8	7
22	H2	2	10	4	8	5	4	2
23	H3	1	4	4	8	2	5	1
24	H4	2	4	4	4	3	2	4
25	H5	4	18	8	10	10	6	4
26	H6	15	25	22	28	21	19	17
27	H7	2	4	5	7	2	4	2
28	H8	9	20	12	15	15	11	9
29	13	0	2	1	2	0	0	0
30	17	1	7	2	3	3	1	1
31	19	1	3	2	3	1	1	1
32	I11	2	6	3	4	3	2	2

Figure 7. Five of the seven structures shown represent naturally-occurring compounds; two are computer generated.[52] The question considered by Smith and Carhart[52] was as follows: Can one discriminate between the two types, the naturally occurring and artifically constructed cases? and in particular: Can one identify, knowing the five naturally-occurring forms, which of the remaining computer-generated forms (among some 30 such considered initially) are likely to be a candidate form for a yet-uncovered natural monocyclic monoterpene?

This problem was resolved by Randić and Wilkins[53] successively by characterizing the graphs of Figure 7 by sequences of path numbers. Hence a collection of path graphs apparently form a suitable basis for discrimination among the similarity of compounds such as monoterpenes considered even though

Table VII. Similarity Table for Seven Monoterpenes of Figure 7 Based on Use of Prime Graphs as Their Representation

	S	*T*	*U*	*V*	*W*	*X*	*Y*	av
S	0	41.6	27.7	44.2	25.4	29.8	19.9	31.43
T		0	20.3	17.6	19.3	26.2	38.0	27.17
U			0	20.2	11.3	13.7	22.9	19.35
V				0	22.0	21.1	36.9	27.00
W					0	13.0	21.7	18.78
X						0	23.9	21.28
Y							0	27.22
								av of av 24.60

Table VIII. Similarity Table for Seven Monoterpenes of Figure 7 Based on Representation of Structures by Path Numbers Only

	S	*T*	*U*	*V*	*W*	*X*	*Y*
S	0	8.72	5.20	11.40	8.49	9.85	6.63
T		0	4.36	3.74	2.00	3.32	3.16
U			0	7.42	4.12	5.48	3.32
V				0	3.46	3.00	5.10
W					0	1.73	2.45
X						0	3.87
Y							0

	Path Sequences
S	10, 16, 20, 17, 10, 6, 2
T	10, 13, 15, 16, 14, 11, 2
U	10, 14, 17, 18, 13, 8, 2
V	10, 13, 14, 14, 16, 12, 4
W	10, 14, 15, 16, 15, 10, 3
X	10, 14, 14, 16, 16, 10, 4
Y	10, 15, 17, 16, 14, 10, 3

such basis is very limited. If we now enlarge the basis by including path graphs and all path subgraphs, the corresponding characterization will improve or perhaps it will even show that the earlier set-up similarity between the naturally-occurring structures was an artifact of using a small basis for representation of graphs?

We will introduce here an assumption of "overall or global similarity" as a constrast to the traditional use of similarity based on a comparison of selected components between different objects. When in a model one can identify the most relevant components for a particular property, then a clear comparison based on these components is going to be of utmost importance. If we have a complete (or almost complete) basis and represents objects of comparison within such a basis, then we can speak of an overall or global similarity, since *all* the components have been taken into account when the comparisons were made. Hence, if we do not specify the subset of components (invariant) that are of special interest, we may proceed with a comparison of objects using all components, and as a result we will get an indication of an overall similarity. We will illustrate the notion of a global or overall similarity on the selection of monoterpenes.

In Table VI we list all the 32 components that occur for the set of compounds considered. In Table VII we show the similarity matrix which represents Euclidean distances between the seven vectors of a 32-dimensional vector space. An inspection of Table VI suggests that the reported similarities and dissimilarities among the terpenes are upheld. The similarity and dissimilarity entries based on the use of only path numbers p_i (Table VIII) parallel the results based on a more complete representation of the structures. As we see structure *S* (the first row) is again associated overall with the largest entries, suggesting that this is the skeletal form the least similar to all others. Hence, if a search for a novel skeletal form for monocyclic monoterpenes is based on similarity with existing forms, then *S* is not likely to be the skeletal form for the unknown terpene. On the other hand, the entries corresponding to graph *W*, the other computer-generated possible skeletal form (suggested by a program of Smith and Carhart),

are found to be associated with the entries of the same size as those typifying similarities among known naturally occurring compounds. A close comparison between similarities based on prime paths and on path graphs shows many oscillations especially among the smallest entries in the table. The largest entries in the similarity table, however, remained unperturbed when a different number of components are considered, indicating a lesser sensitivity to basis size.

The present study confirms the earlier result that *W* rather than *S* (or any of the other computer-generated skeletal forms reported in ref 52) is most likely the form for an unknown monoterpene. Which of the two approaches, one based on paths only and the other based on prime paths, is more reliable and more trustworthy? Which approach better reflects the minor differences, particularly when one pays attention to details, even molecular fragments? If we are interested in identifying not the least similar, but the most similar pairs of skeletal forms, the two approaches, one based on prime paths and the other on paths alone, give different answers. When using all the 32 components to represent molecular graphs, we find the following as the most similar forms:

(*U*,*W*), (*W*,*X*), (*U*,*X*), (*T*,*V*), (*T*,*W*), (*S*,*Y*), (*U*,*V*), (*T*,*U*), etc.

But when the same skeletal forms are represented by path numbers p_k, we get the following ordering (in decreasing similarity):

(*W*,*X*), (*T*,*W*), (*W*,*Y*), (*V*,*X*), (*T*,*Y*), (*T*,*X*), (*U*,*Y*), (*T*,*V*), etc.

The most similar structures, the pairs (*U*,*W*) and (*U*,*X*), identified by prime graphs are not among the first half-dozen most similar cases when 7-dimensional vectors based on path numbers are used for characterization of the structures, but the pairs (*W*,*X*) and (*T*,*W*) have been indicated among the most similar by both schemes. Since the similarity paradigm is of more interest when examining objects that are the *most* similar rather than objects that are the *least* similar, the discrepancy between the two approaches deserves more attention. One should not be satisfied merely that two distinctive characterizations have lead to the overall same conclusion in discarding the *least* interesting structure from consideration, despite the fact that this information is of considerable importance and represents a very useful result. One should focus attention on the *most* similar structures and judge alternatives on how well such results correspond to available data.

Before we continue with a detailed analysis of various comparison of structures, let us emphasize that the discord between the results of similarity comparisons based on 32-component and 7-component vectors does not suggest that one approach is better than the other or more accurate. Both approaches are equally valid from mathematical and conceptual points of view. Similarity measure is not an *absolute* quantity, but a *relative* quantity, and it relates to specific attributes (components). Thus if we ask ourselves which is the most similar pair of structures when the comparison is based on *all* components of a prime representation, the answer is (*U*,*W*). Incidentally, the same pair looks almost the least similar (!), which shows how it is difficult to make judgements about *n*-dimensional quantities from their apparent projections in the 3-dimensional space. If on the other hand we want to identify the most similar pair based on *path* counts, not prime paths, then the answer is (*W*,*X*). In each case we asked a *different* question, and not surprisingly we received different answers.

In practical application, the difficult problem is that of identifying the pertinent structural features of interest for a comparison. Once such features are recognized the task is that of identifying suitable structural invariants. These would be the invariants that are sensitive to minor changes in the critical substructure that characterizes the features of interest. The construction of the similarity matrix is almost a trivial task. A typical structure–activity study will thus first involve a search for a pharmacophore. This is followed by selecting invariants of the compounds such that the variations in the immediate environment of the pharmacophore are suitably reflected in the graph characterizations. If a same compound, as is not uncommon, shows more than one activity, a quantification of such behavior may in each case call for a different active substructure and will in general result in a distinctive ranking of the compounds considered. The partial order will depend on what property and what standards are considered as a "target" system.

Table IX. Similarity Table for Seven Monoterpenes of Figure 7 Based on Use of First Six Prime Graphs (Prime Graphs up to 5 Vertices)[a]

	S	*T*	*U*	*V*	*W*	*X*	*Y*	av
S	0	1.52	1.06	1.92	1.48	1.89	1.20	1.51
T		0	0.75	0.68	0.55	0.99	1.04	0.92
U			0	1.20	0.68	1.06	0.72	0.91
V				0	0.60	0.72	1.06	1.03
W					0	0.51	0.60	0.74
X						0	0.78	0.99
Y							0	0.90

[a] Normalization factor = 4.42. No. of components = 6.

Table X. Similarity Table for Seven Monoterpenes of Figure 7 Based on Use of First Six Prime Graphs (Prime Graphs up to 6 Vertices)[a]

	S	*T*	*U*	*V*	*W*	*X*	*Y*	av
S	0	1.48	1.02	1.75	1.28	1.66	1.22	1.40
T		0	0.64	0.65	0.77	1.29	1.25	1.01
U			0	0.87	0.56	1.07	0.78	0.82
V				0	0.64	0.93	1.03	0.98
W					0	0.58	0.71	0.76
X						0	0.83	1.06
Y							0	0.97

[a] Normalization factor = 8.55. No. of components = 10.

Table XI. Similarity Table for Seven Monoterpenes of Figure 7 Based on Use of First 14 Prime Graphs (Prime Graphs up to 7 Vertices)[a]

	S	*T*	*U*	*V*	*W*	*X*	*Y*	av
S	0	1.75	1.30	1.87	1.23	1.47	1.18	1.47
T		0	0.57	0.56	0.71	0.86	1.49	0.99
U			0	0.78	0.44	0.67	1.07	0.81
V				0	0.71	0.70	1.30	0.99
W					0	0.35	0.95	0.73
X						0	1.03	0.85
Y							0	1.17

[a] Normalization factor = 15.47. No. of components = 14.

COMPARISON OF COMPARISONS

Obviously the results of a comparison derived from a similarity matrix will depend, for a set of compounds considered and the selected characterization, on a selection of the components used in the similarity test. To get some insight into the variations of similarity data upon truncation of a basis set, we examined similarities among monoterpenes of Figure 7 using a smaller number of components. In Tables IX–XIV we collected information for the seven monoterpenes using from 6 to all 32 components of the prime representation. Because the entries in a similarity matrix in general will depend also on the number of components, being smaller when a smaller number of components are used (as can already be seen from Table VII with data on 32-dimensional vectors and Table VIII with similar data on 7-dimensional vectors) in order to ease comparisons of all such diverse data, they first ought

Table XII. Similarity Table for Seven Monoterpenes of Figure 7 Based on Use of First 21 Prime Graphs (Prime Graphs up to 8 Vertices)[a]

	S	T	U	V	W	X	Y	av
S	0	1.63	1.23	1.89	1.09	1.38	0.94	1.36
T		0	0.61	0.61	0.67	0.85	1.45	1.02
U			0	0.85	0.46	0.61	1.01	0.80
V				0	0.89	0.79	1.53	1.09
W					0	0.51	0.91	0.76
X						0	1.10	0.87
Y							0	1.16

[a] Normalization factor = 21.03. No. of components = 21.

Table XIII. Similarity Table for Seven Monoterpenes of Figure 7 Based on Use of First 28 Prime Graphs (Prime Graphs up to 9 Vertices)[a]

	S	T	U	V	W	X	Y	av
S	0	1.69	1.14	1.82	1.04	1.23	0.82	1.29
T		0	0.80	0.70	0.76	1.03	1.53	1.09
U			0	0.83	0.46	0.56	0.94	0.79
V				0	0.90	0.85	1.51	1.10
W					0	0.53	0.89	0.76
X						0	0.98	0.86
Y							0	1.11

[a] Normalization factor = 24.27. No. of components = 28.

Table XIV. Similarity Table for Seven Monoterpenes of Figure 7 Based on Use of All Prime Graphs (Prime Graphs up to 10 Vertices)[a]

	S	T	U	V	W	X	Y	av
S	0	1.69	1.13	1.80	1.03	1.21	0.81	1.28
T		0	0.83	0.72	0.78	1.06	1.54	1.10
U			0	0.82	0.46	0.56	0.93	0.79
V				0	0.89	0.86	1.50	1.10
W					0	0.53	0.88	0.76
X						0	0.97	0.86
Y							0	

[a] Normalization factor = 24.60. No. of components = 32.

to be normalized. We normalized each similarity matrix such that an average distance among the structure-points, corresponding to the individual monoterpenes, equals one. This is achieved by first deriving the average distance for each structure (the last entry in Table VII) and then averaging the so-obtained (seven) averages. In the case of all 32-components this overall average distance is 24.60 (for the distances that varied between 11.3 and 44.2 and the averages that varied between 18.78 and 31.43). The normalized distances are then obtained by scaling (that is, dividing) the actual distances (of Table VII) by 24.60. In Table XIV we show the so-derived normalized distances (similarities) based on the representation of structures with sequences using all the components of Table VI. The values greater than one now show less similar structures and those less than one indicate more similar structures. Tables IX–XIII show similar results derived from truncated vectors having 6, 10, 14, 21, and 28 components, respectively. The particular truncation corresponds to prime basis limited from 5 to 9 vertices, that is, to prime paths A-D, A-E, A-F, A-G, and A-H, respectively.

A close look at the individual entries in Tables IX–XIV is instructive and revealing. As the number of components increases (from Table IX to Table XIV) the range of the extreme values in the tables is reduced. The maximal values which always occurred for the pair (*S*,*T*), 1.52, 1.48, 1.75, 1.63, 1.69, and 1.69, apparently converge. The similar sequence of similarity parameters for other pairs show a similar regular change, with some oscillatory behavior when the number of components is smaller (Figure 8). Significantly, the oscillations (or the changes) are smaller as the number of components increases. These observations have important consequences if they are found to be generally true. It will allow one to use truncated representations of structure-vectors, since after some critical size the effect of additional components is not changing significantly the relative values for the similarities. The apparent convergence is a consequence of a smaller count of "larger" (i.e., those having many edges) subgraphs. However, a close look at the individual components for the seven structures shows that, although path numbers that associate with the intermediate path size are numerically the largest, the differences between the corresponding path numbers are spread throughout the sequence, except the very beginning of the sequences which is (for isomeric structures) rather alike.

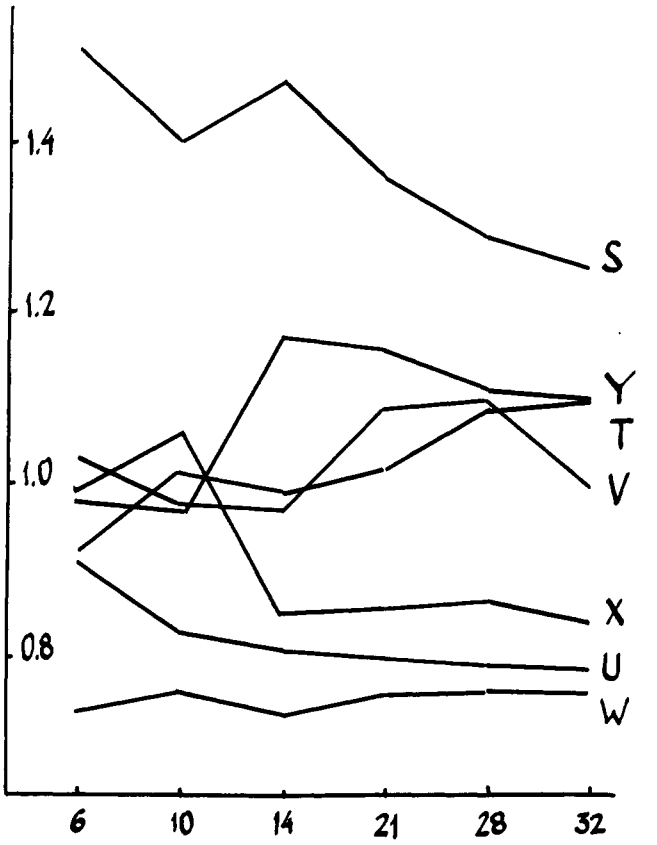


Figure 8. Variation of the relative similarity among monoterpenes with an increase in the number of components in prime graph representations.

GLOBAL SIMILARITY

As already stated and as implied in similarity comparisons, the outcome of such exercises are *relative* quantities. They only maintain a meaning in relation to the components used in a comparison. Hence, it is not only possible but also very likely that each time different components are used to characterize a structure, a different similarity ranking will result. In fact this is precisely the idea behind a search for active substructure in QSAR (quantitative structure–activity relationship) as illustrated on nitrosamines[54] and nitroarenes[55] in a search for molecular fragments responsible for the mutagenicity of these compounds.

If we have a *complete* basis and corresponding characterization for structures, then we can make a comparison by using all the components. Such an overall comparison will result in an *overall* or *global* similarity, where global indicates that all the information available on the structures was used. Such global similarity may be expected to be independent of the basis chosen and hence provides an *absolute* measure of similarity among objects considered. In practice, the problems of determining if a given basis is complete, overcomplete, or

incomplete are expected to be difficult. Hence, we suggest the use of the term *overall* similarity in a somewhat restrictive sense, as an atribute for a similarity based on all the components of a comprehensive basis. In particular we will use it with prime basis and comparisons based on representing graphs with the count of prime graphs that it contains. We expect that the overall similarity, because it includes comparison of all components, is going to better reflect the overall molecular features, such as size and shape of the molecules.

In applications of the similarity to structure–activity we may restrict comparison to selected molecular fragments, yet employ all the components of a prime graph decomposition. In this way we will get information on *local* similarity, that is, similarity among specified molecular fragment or molecular components. However the comparison will be based on the complete information on such fragments, not confined to the information based on selected few invariants. Such analysis applied to drug receptors as well as to drug matching cavities of receptors will produce quantitative characterizations for "docking" that one can visualize with the help of computer graphics. In this way one can perhaps improve some QSAR studies, adding a quantitative index to an otherwise qualitative pictorial matching model on computer monitors.

CONCLUDING REMARKS

We may conclude, from the information in Tables IX–XIV, that overall similarities based on larger basis vectors are more reliable. This is in the sense that when using more components, more attributes of a structure were taken into account in making comparisons. The oscillatory behavior of many entries when a smaller number of components are considered displays a hazard involved when a structure is characterized indiscriminately by a small number of invariants, unless, of course, one has definite reasons to restrict the comparisons to a few components. If we were to further increase the basis by inclusion of additional, linearly independent, descriptors, we could expect further minor changes. In an ideal case of a compete basis we would in this way obtain an absolute similarity index for a pair of structures. Observe that the differences already established with a smaller number of components persist even if they are modified and somewhat reduced. We may expect such a behavior in a general case. This expectation is based on the belief in the dogma of structure–property relationships that similar structures have similar properties, but not the same properties. Hence, similarity based on mathematical invariants, that can be viewed legitimately as mathematical properties of a structure, when the number of such properties is indefinitely increased will continue to introduce smaller differences and can only increase the dissimilarity regardless of how apparently similar the structures are. When such similarities are normalized, we will observe relative changes; similarity among some structures will increase and among other decrease as the number of descriptors is increased.

There are several questions that the notion of graph basis introduces that have yet to be considered, beside the already-mentioned problems of establishing the degree of completeness (or overcompleteness) of a basis. One desirable feature to consider is how to increase a ratio of signal to noise characteristics of representing graphs with prime basis (and alternative bases). One way to achieve this is by eliminating hidden linear dependencies among subgraphs. For example, each time a graph contains path p_k it also contains all components of p_k. It seems desirable to prune prime graph representation forms of such linearly dependent components, but that task is beyond the scope of the present paper. In addition, one should consider orthogonalized descriptors and in this way eliminate structural duplications present in several components. Additional tasks raised by this work include a search for recursive relations for primal paths, finding regularities in the decompositions among isomers, and finally search and construction of larger graphs (trees) having the same path counts. Such graphs may lead to a counterexample of the uniqueness of the representation of graphs by prime paths and may establish N, the smallest number of vertices for graphs with the same count for all prime path components. Alternatively, one may consider the problem of reconstruction of graphs from the list of prime path components. If reconstruction could be demonstrated, this would ensure the uniqueness of the prime graph decomposition.

SUMMARY

We introduced a novel extensive basis for representation of graphs that is based on the enumeration of paths and their subgraphs. For the first time in the study of molecular graphs, one can consider an expansion of graphs in relatively simple objects, paths, and their subgraphs, and in this way analyze the properties of graphs. The proposed prime basis appears satisfactory in that it apparently does not result in the same characterization for two nonisomorphic structures, unless N, the number of vertices, is relatively large. While actual enumeration of prime paths in a larger graph remains tedious (and error prone), the basis appears to offer satisfactory comparisons of graphs. Moreover, it allows one to consider as a valid approximation the notion of an absolute similarity of graphs, where "absolute" refers to "all" (linearly independent) graph invariants used in the characterization of a structure.

ACKNOWLEDGMENT

I thank Professor M. Razinger (University of Ljubljana, Slovenia) for careful reading of the manuscript and for suggesting valuable improvements in the presentation of the results. Also I thank Professor P. Z. Chinn (Humbolt State University, Arcata, CA) for sending a number of preprints on primal graphs prior to their publication.

REFERENCES AND NOTES

(1) Randić, M. *New J. Chem.* **1991**, *15*, 517.
(2) Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
(3) Randić, M. *Croat. Chem. Acta* **1991**, *64*, 43.
(4) Randić, M. *J. Mol. Struct. (Theochem)* **1991**, *233*, 45.
(5) Dillon, W. R.; Goldstein, M. *Multivariate Analysis*; Wiley: New York, 1984; pp 271–289.
(6) Coulson, C. A.; Streitwieser, A., Jr. *Dictionary of PI-Electron Calculations*; W. H. Freeman: San Francisco, 1965.
(7) Živković, T. Reported at Theoretical Chemistry Summer School, Repino, USSR, 1973.
(8) (a) Živković, T.; Trinajstić, N.; Randić, M. *Mol. Phys.* **1977**, *30*, 517. (b) Herndon, W. C. *Tetrahedron Lett.* **1974**, *8*, 671.
(9) Randić, M.; Barysz, M.; Nowakowski, J.; Nikolić, S.; Trinajstić, N. *J. Mol. Struct. (Theochem)* **1989**, *185*, 249.
(10) Bloch, F. *Z. Phys.* **1929**, *52*, 555.
(11) Pariser, R.; Parr, R. G. *J. Chem. Phys.* **1953**, *21*, 466.
(12) Pople, J. A. *Trans. Faraday Soc.* **1953**, *49*, 1375.
(13) Balaban, A. T.; Harary, F. *J. Chem. Doc.* **1971**, *11*, 258.
(14) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vol. II, Chapter 4.
(15) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Top. Curr. Chem.* **1983**, *114*, 21.
(16) Sabljić, A.; Trinajstić, N. *Acta Pharm. Jugosl.* **1981**, *31*, 189.
(17) Balaban, A. T.; Motoc, I. *MATCH* **1979**, *5*, 197.
(18) Bonchev, D.; Mekenyan, O.; Trinajstić, N. *J. Comput. Chem.* **1981**, *2*, 127.
(19) Razinger, M.; Chretien, J. R.; Dubois, J. E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23.
(20) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. *Math. Modeling* **1987**, *8*, 302.
(21) Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.
(22) Randić, M. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 134.
(23) Szymanski, K.; Müller, W. R.; Knop, J.; Trinajstić, N. *Croat. Chem. Acta* **1986**, *59*, 719.
(24) Bonchev, D.; Mekenyan, O.; Trinajstić, N. *J. Comput. Chem.* **1982**, *2*, 127.

(25) Hansch, C. *Acc. Chem. Res.* **1969**, *2*, 232.
(26) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(27) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1974.
(28) Szymanski, K.; Müller, W. R.; Knop, J.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 413.
(29) Randić, M. *Int. J. Quant. Chem., Quant. Biol. Symp.* **1984**, *11*, 137.
(30) Program ALLPATH was suitably modified so that weighted paths are enumerated. Listing of the program in BASIC (for Apple IIe computers) is available. Send request to: Graph Theory Center, Dept. of Mathematics & Computer Sci., Drake University, Des Moines, IA 50311.
(31) Trinajstić, N. *Croat. Chem. Acta* **1977**, *49*, 593.
(32) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(33) Randić, M.; Hansen, P. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 60.
(34) Poshusta, R. D.; McHughes, M. C. *J. Math. Chem.* **1989**, *3*, 193.
(35) Klein, D. J. *Int. J. Quant. Chem., Quant. Biol. Symp.* **1986**, *20*, 153.
(36) For a review see: O'Neal, P. V. *Am. Math. Monthly* **1970**, *77*, 35.
(37) Harary, F.; Schwenk, A. J. *Discrete Math.* **1973**, *6*, 359.
(38) Buhler, J. P.; Chinn, P. Z.; Richter, R. B.; Truszczynski, M. *Congr. Numerant.* **1988**, *66*, 5.
(39) Chinn, P. Z.; Richter, R. B.; Thoelecke, P. A. *Ars Comb.* **1989**, *26*, 45.
(40) Chinn, P. Z.; Richter, R. B.; Truszczynski, M. *Ann. N.Y. Acad. Sci.* **1989**, *576*, 118.
(41) Chinn, P. Z.; Yixun, L. *Decomposing Graphs into Primary Graphs*; (preprint) Humboldt State Univ.: Arcata, CA, 1990.
(42) Dewdney, A. K. *Aequat. Math.* **1970**, *4*, 326.
(43) Randić, M.; Oakland, D. O. *Congr. Numerant.*, in press.
(44) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(45) Randić, M.; Brissey, G. M.; Spencer, R. G.; Wilkins, C. L. *Comput. Chem.* **1979**, *3*, 5.
(46) Heilmann, O. J.; Lieb, E. H. *Commun. Math. Phys.* **1972**, *25*, 190.
(47) Ruch, E. *Theor. Chim. Acta* **1975**, *38*, 167.
(48) Muirhead, R. F. *Proc. Edinburgh Math. Soc.* **1906**, *24*, 45.
(49) Slater, P. J. *J. Graph Theory* **1982**, *6*, 89.
(50) Quintas, L. V.; Slater, P. J. *MATCH* **1981**, *12*, 75.
(51) Randić, M. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 101.
(52) Smith, D. H.; Carhart, R. E. *Tetrahedron* **1976**, *32*, 2513.
(53) Randić, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31.
(54) Randić, M.; Jerman-Blažič, B.; Rouvray, D. H.; Seybold, P. G.; Grossman, S. C. *Int. J. Quant. Chem., Quant. Biol. Symp.* **1987**, *14*, 245.
(55) Randić, M.; Grossman, S. C.; Jerman-Blažič, B.; Rouvray, D. H.; El-Basil, S. *Math. Comput. Modelling* **1988**, *11*, 837.

Topological Organic Chemistry. 4.[1] Graph Theory, Matrix Permanents, and Topological Indices of Alkanes

HARRY P. SCHULTZ*

Department of Chemistry, University of Miami, Coral Gables, Florida 33124

EMILY B. SCHULTZ† and TOR P. SCHULTZ‡

School of Forest Resources, Forest Products Laboratory, and Department of Chemistry, Mississippi State University, Mississippi State, Mississippi 39762

Received September 30, 1991

The permanents of the distance matrices that describe the structures of alkanes were calculated and demonstrated to be useful as molecular topological indices. Additionally, the easily derived products of the row sums of the distance matrices were observed also to be useful as molecular topological indices.

INTRODUCTION

Weiner,[2] Hosoya,[3] and Randić[4] are among many who have devised various ways of utilizing graph theory to numerically characterize chemical structures. Rouvray[5] has summarized numerous techniques for calculating topological indices. An earlier paper[6] of this series reported on the use of matrix determinants as potential topological indices. It was a project, also studied by Knop et al.,[7] that met with only partial success. This paper further extends the use of structure-descriptive matrices of alkanes as sources of single-sum numbers that serve as descriptors or codes of alkane structures; it reports the results of experiments involving permanents, matrix functions related to the determinants.

The adjacency and distance matrices that describe the structures of molecular graphs are square ($n \times n$) matrices. If $\mathbf{M} = (a_{ij})$ is any $n \times n$ matrix, the permanent of M is defined by

$$\text{per}(\mathbf{M}) = \sum_{\sigma} a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

where the summation extends over all one-to-one functions from $\{1, ..., n\}$ to $\{1, ..., n\}$. The sequence $a_{1\sigma(1)}, ..., a_{n\sigma(n)}$ is called a diagonal of M, and the product $a_{1\sigma(1)} \cdots a_{n\sigma(n)}$ is a diagonal product of M. Thus, the permanent of M is the sum of all the diagonal products of M.[8]

The definition of the permanent of M contrasts interestingly with the definition of the determinant of M:

$$\det(\mathbf{M}) = \sum_{\sigma} \text{sgn}(\sigma)\, a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

The two expressions differ only in the omission of the ± sign of the permutation in the expression for the permanent. Indeed, the permanent is frequently referred to as the plus, or positive, determinant. Therefore, the same Laplace expansion, which serves for the calculation of a determinant, is even more easily applied to the computation of a permanent, since no sign changes need be injected into the Laplace expansion. Experimentation demonstrated that the permanent of a square matrix was an invariant value, independent of the number sequence of the graph vertices. Trinajstić[9] discussed the permanent of the adjacency matrix and outlined several of its properties.

COMPUTATIONS

Molecular graphs, hydrogen-suppressed and with the interatomic carbon-carbon edge counts set at unity, were derived from the alkanes listed in Table I. The matrix permanents were computed on a Data General MV/7800XP minicomputer

† School of Forest Resources.
‡ Forest Products Laboratory and Department of Chemistry.

0095-2338/92/1632-0069$03.00/0