Substructure Isomorphism Matrix

K. Varmuza* and H. Scsibrany

Laboratory for Chemometrics, Institute of Food Chemistry, Vienna University of Technology, A-1060 Vienna, Getreidemarkt 9/160, Austria

Received August 3, 1999

A substructure isomorphism matrix $n \times p$ contains binary elements describing which of the given p query structures (substructures) are part of the given n target structures (molecular structures). Such a matrix can be used to investigate the diversity of the target structures and allows the characterization and comparison of structural libraries. A quadratic substructure isomorphism matrix $n \times n$ is obtained if the same structures are used as molecular structures and as substructures; this matrix contains full information about the topological hierarchy of the n structures. A hierarchical arrangement of chemical structures is useful for the evaluation of results obtained from searches in structure databases.

INTRODUCTION

Multivariate data analysis methods play an important role in computer chemistry, especially in modeling structure−activity/property relationships, in cluster analyses of chemical structures, in similarity (or dissimilarity) searches with chemical structures, and in describing the diversity of structural libraries. The application of multivariate data analysis methods[1,2] requires a representation of chemical structures by a set of variables (molecular descriptors, features). A number of different approaches for defining molecular descriptors is available.[3−7] Widely used is the characterization of molecular structures by binary descriptors that indicate whether predefined substructures are part of a molecular structure or not. This type of descriptor has the advantage of a clear chemical background. The bit string containing a set of substructure-based descriptors is often called a "fingerprint" of the molecule.[8,9]

The values of p molecular descriptors for a set of n chemical structures form a matrix of size n times p; for binary descriptors that describe the presence or absence of substructures we call such a matrix a substructure isomorphism matrix. A special type of substructure isomorphism matrix is obtained if the same set of structures is used as molecular structures and as substructures.

METHOD

General. A substructure isomorphism matrix **S**(n,p) consists of n rows for n target structures (molecular structures) and p columns for p query structures (substructures), with matrix element $s(i,j) = 1$, if query structure j is part of target structure i, and $s(i,j) = 0$, otherwise (Figure 1). In this work chemical structures are defined by connection tables; stereoisomerism and three-dimensionality are not considered.

The software SubMat has been developed for automatic and fast computation of substructure isomorphism matrices:

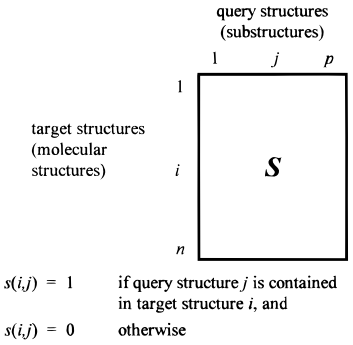


Figure 1. Substructure isomorphism matrix (**S**-matrix).

the input are two files containing structures in Molfile format, one file for the target structures, the other for the query structures; the output is an ASCII file containing the **S**-matrix in a user-selectable format. The output file is directly usable by some statistical or chemometrical software products. A number of tools allows a flexible definition of isomorphism; for instance the topology of atoms in the substructures can be optionally defined as "part of a chain", "part of a nonaromatic ring", or "part of an aromatic ring". Bond types available are single, double, triple, aromatic, some mixed types, and "query bond"; in the last case the bond type is not checked. Typical computation time on a PC 300 MHz is 4 s for $n = 1000$ target structures and $p = 100$ substructures (molecular masses of substructures were between 40 and 120; molecular masses of target structures were between 40 and 400); the size of the ASCII output file is 100 kB for this example. For $n = 100\,000$ and $p = 100$ a computation time of 13 min is necessary, and the output file has a size of 10 MB. Software SubMat is running under MS Windows 3.11/95/NT and is available from the corresponding author.

The relevance and utility of a **S**-matrix greatly depends on the used substructures. For many problems a set of general-purpose substructures is appropriate; an example in Results compares the structural contents of two spectral libraries using such a set of substructures. Another approach builds a set of substructures that are characteristic for the

* Corresponding author. Fax: +431-58801-16091. E-mail: kvarmuza@email.tuwien.ac.at.

10.1021/ci990267l CCC: $19.00
Published on Web 01/08/2000

Table 1. Principal Component Analysis of **S**-Matrices from Data Sets C7A and C7B: a, Number of Principal Components; v_a, Variance of Scores for Principal Component a; $v_{cum,a}$, Cumulative Variance of Principal Component Scores 1 to a (Variances in Percent of Total Variance)

	data set C7A		data set C7B	
a	v_a	$v_{cum,a}$	v_a	$v_{cum,a}$
1	40.3	40.3	29.3	29.3
2	27.5	67.9	18.3	47.6
3	14.6	82.5	12.9	60.5
4	9.6	92.1	9.0	69.5
5	7.9	100.0	6.5	76.0
6			5.9	81.9
7			3.8	85.6
8			3.1	88.8
9			3.0	91.8
10			2.1	93.9
11			1.5	95.4
12			1.3	96.8

molecular structures under investigation. Such a set of characteristic substructures can be automatically obtained for instance by the determination of maximum common substructures among the molecular structures.[10,11] The software SubMat provides full flexibility for the application of user-defined substructures.

PCA. Principal component analysis (PCA) of a **S**-matrix gives characteristic numerical measures for the sets of structures—independent from the sequences of the target and the query structures. From the variances of the principal component scores the rank of the **S**-matrix can be estimated. The rank can be considered as a measure for the number of factors necessary to describe the structural diversity (of course within the limits of the used substructures). The loadings of the principal components possessing the highest variances indicate those descriptors (substructures) which are most relevant for the investigated molecular structures. Finally, scatter plots with principal component scores as coordinates offer an interactive cluster analysis of the molecular structures.[2] The following example with two small sets of structures demonstrates an application of PCA to **S**-matrices; another example is contained in Results.

Data set C7A contains 20 isomeric structures with molecular formula C_7H_{12} possessing two double bonds but no allene substructure. These 20 dienes constitute a random subset from the 31 isomers[12] which are possible under the given restrictions. Data set C7B contains 20 monofunctional structures with molecular formulas $C_7H_x(N,O)_1$; 10 of them are aromatic or heteroaromatic, and the others are aliphatic or alicyclic. A set of 135 substructures covering a number of substance classes has been defined and used for the determination of the **S**-matrices. Table 1 contains the variances and the cumulative variances (in percent of the total variance) of the first 12 principal component scores. For set C7A four principal components are necessary to reach 90% of the total variance, while for set C7B nine principal components are necessary. This result demonstrates the larger structural diversity (higher rank) of structure set C7B.

Figure 2 shows the loading plots for the first and second principal components; because each point in these scatter plots corresponds to a substructure (descriptor, feature), the importance of the applied substructures for the two data sets can be evaluated. The number of relevant substructures (with nonzero variances of the descriptors) is only five for set C7A but 42 for set C7B. For the dienes the score of the first principal component is mainly influenced by the presence of conjugated double bonds (high positive value for b_1); consequently the loading for substructure C=CCCC=C must be—and actually is—negative because both substructures cannot be present simultaneously in the molecules of C7A. Furthermore, absence of the substructures CCCC and C(C)-CC is characteristic for the first principal component.

The loading plot for the more diverse set C7B shows groups of correlating features; these groups indicate the main structural factors. For instance group S1 contains five substructures with high positive loadings for the first principal component; all are aliphatic and characterize branches and chains built by carbon atoms. Group S2 consists of features for oxygen-containing functional groups, group S3 for substructures with a benzene ring, group S4 for nitrogen-containing functional groups, and group S5 for cyclohexane substructures.

Hierarchy. Interesting applications for the substructure isomorphism matrix arise if the same structures are used as target structures and as query structures. The resulting quadratic auto-**S**-matrix of size n times n contains information about which structures are part of others. This matrix describes the topological hierarchy of the investigated structures as is demonstrated by the following example; another example is contained in Results.

Figure 3 shows seven molecular structures and the corresponding auto-**S**-matrix. Evidently all diagonal elements of the matrix are 1; however, in general the matrix is not symmetric. For example structure 1 (considered as a substructure with free valences instead of the H-atoms) is a part only of structure 4 (and of course of itself); structure 3 is contained in all seven structures; structures 4 and 7 are not contained in any other structure. The auto-**S**-matrix is a description of a directed graph that represents the topological hierarchy of the structures as shown in Figure 4. Root vertices of the hierarchy graph denote structures that are not part of any others; their column sum in the auto-**S**-matrix is 1. Leaf vertices of the hierarchy graph denote structures that do not contain any other structure; their row sum in the auto-**S**-matrix is 1. A detailed description of the construction of hierarchy graphs from auto-**S**-matrices will be described elsewhere.[13]

If the n structures are free from duplicates, then symmetric nondiagonal elements $s(i,j)$ and $s(j,i)$ in an auto-**S**-matrix either are both zero or have different values (0, 1 or 1, 0). The reason for this fact is because if structure j is contained in structure i, then i cannot be part of j—except i and j are identical. The sequence of the structures (of rows and of columns) in an auto-**S**-matrix can be rearranged so that the lower triangle only contains elements with value 0.

The auto-**S**-matrix describes structural relationships within a set of structures and thus characterizes one of the many aspects of structural diversity. An extreme composition of structures is present if none of the structures is part of any other; an example for this situation is a set of isomers from a given molecular formula. In this case the auto-**S**-matrix becomes a unit matrix, and the number of elements with value 1 in the matrix reaches the minimum of n. Another extreme is a series of structures in which each structure (except the first one) is contained in all previous ones (an example is the subset of the structures $4 > 1 > 6 > 3$ in

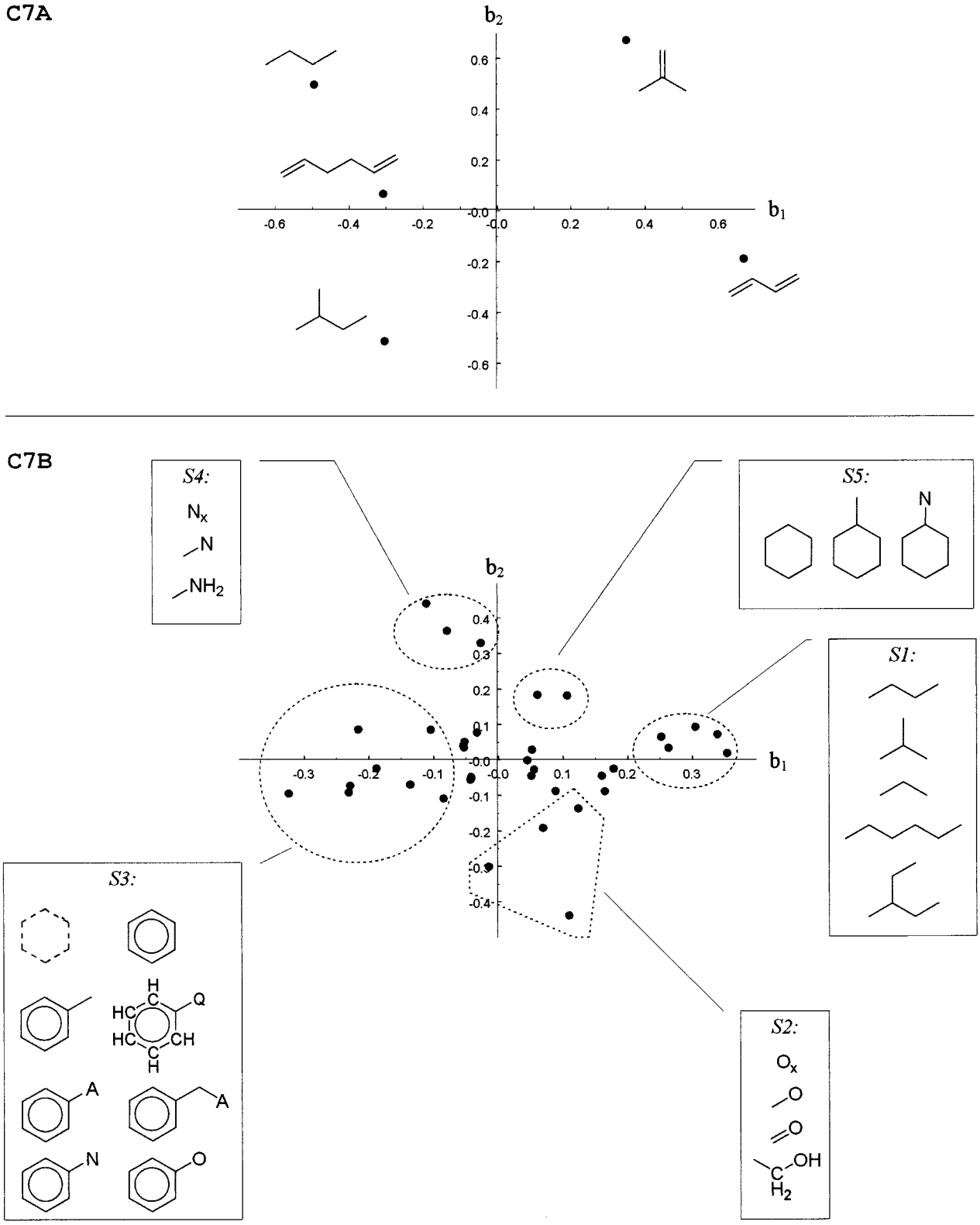
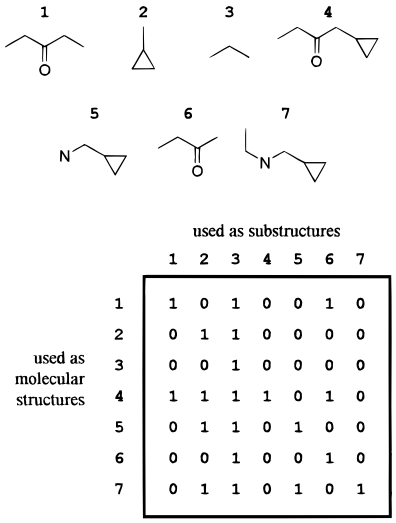


Figure 2. Loading plots from PCA of **S**-matrices from data sets C7A and C7B. b_1, b_2, loadings for first and second principal components (variances are given in Table 1); A, heteroatom; Q, non-H-atom; dashed line in structures denotes "any bond type".

Figure 4); in this case the auto-**S**-matrix becomes an upper triangular matrix, and the number of elements with value 1 in the matrix reaches the maximum of $n(n + 1)/2$.

RESULTS

Comparison of Libraries. The structural content of a spectral library heavily influences the result of spectra similarity searches which are standard techniques for the identification of compounds. All large spectral libraries available have not been built systematically and exhibit an uneven distribution of compound classes. An example demonstrates the application of **S**-matrices for finding characteristic structural differences between two spectral databases. Library A is part of the SpecInfo database;[14] it



	1	2	3	4	5	6	7
1	1	0	1	0	0	1	0
2	0	1	1	0	0	0	0
3	0	0	1	0	0	0	0
4	1	1	1	1	0	1	0
5	0	1	1	0	1	0	0
6	0	0	1	0	0	1	0
7	0	1	1	0	1	0	1

Figure 3. Substructure isomorphism matrix (auto-**S**-matrix) with the same structures used as molecular structures and as substructures.

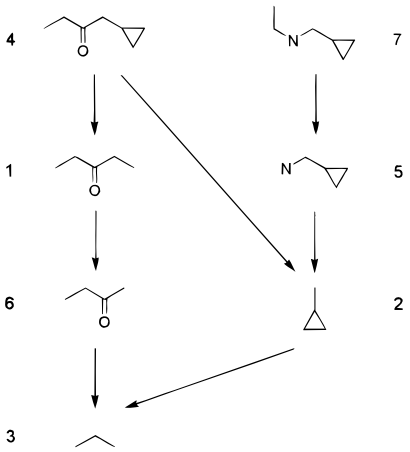


Figure 4. Topological hierarchy of seven chemical structures obtained from the auto-**S**-matrix given in Figure 3.

Table 2. 2 × 2 Contingency Table for Presence and Absence of the Benzene Ring Substructure in Random Samples A and B from Two Spectral Libraries

	absent	present	sum
A	$n_{A0} = 175$	$n_{A1} = 325$	$n_A = 500$
B	$n_{B0} = 273$	$n_{B1} = 227$	$n_B = 500$
sum	$n_0 = 448$	$n_1 = 552$	$n = 1000$

contains 13 484 structures and the corresponding infrared spectra. Library B contains 60 909 structures and the corresponding mass spectra and is part of the NIST mass spectral database.[15] From each database a random sample containing 500 compounds has been selected. A set of 135 substructures covering a broad range of compound classes has been used to determine the **S**-matrices separately for random samples A and B. The number of substructures with nonzero variances of the descriptors was 122 for A and 130 for B. The variances of the principal components do not differ significantly for A and B; to reach 90% of the total variance, 43 and 46 principal components are necessary for A and B, respectively.

Structural differences between the two spectral databases are indicated by those substructures which differ significantly in their frequencies (column sums). The χ^2-test for 2 × 2 contingency tables[1] has been applied to find the descriptors (substructures) which are significantly more frequent in one of the two libraries. For example the column sums for the benzene ring substructure are 325 and 227, for A and B, respectively (Table 2). According to Pearson's χ^2-statistics the test quantity

$$\chi^2 = n\,(n_{A0}n_{B1} - n_{A1}n_{B0})^2/(n_A n_B n_0 n_1) \qquad (1)$$

is χ^2-distributed with 1 degree of freedom where n_{k1} is the column sum for random set k (k = A, B); n_k is the number of structures in random set k; $n = n_A + n_B$; $n_{k0} = n_k - n_{k1}$; and $n_0 = n_{A0} + n_{B0}$. For the benzene ring substructure χ^2 is 38.8, which indicates that the frequencies in the two libraries differ significantly (the lower limit of the test quantity is 6.63 for a maximum statistical risk of 0.01). Because χ^2 depends on the sizes n_A and n_B of the random samples, it is necessary to use equally sized random samples. In Table 3 substructures (in some cases elements) are listed which are dominant in one of the two databases. Also the probabilities of the substructures in the random samples are given to allow a practical oriented evaluation of the result found by the statistical test; considering only the value of χ^2 may be misleading because χ^2 is proportional to n (for $n_A = n_B$). Obviously the mass spectral database is influenced by many hydrocarbons; an explanation for this result is the content of data from early applications of mass spectrometry. On the other hand parts of the IR database have their origin in an organic synthesis laboratory whose activities are represented by many aromatic compounds and compounds with heteroatoms. The substructures found to be characteristic for the IR database exhibit a higher probability level than those found to be characteristic for the MS database.

Hierarchy of Chemical Structures. From a search for compounds containing the 2-norbornanone substructure a set of 21 molecular structures was obtained. The topological hierarchy of these structures was determined from the auto-**S**-matrix using the 21 structures as target structures and as

Table 3. Substructures and Elements Which Occur Significantly More Frequently in One of Two Spectral Libraries: χ^2, Test Quantity as Defined by Equation 1; p_{IR} and p_{MS}, Probabilities of Substructures in Random Samples from IR and MS Databases, Respectively

more frequent in IR database				more frequent in MS database			
	χ^2	p_{IR}	p_{MS}		χ^2	p_{IR}	p_{MS}
Cl	48	0.28	0.10	C_5 ring	40	0.01	0.11
Cl-substituted benzene	40	0.18	0.05	dimethylcyclohexane	38	0.01	0.09
benzene	39	0.65	0.45	CC(C)C(C)C(C)C	35	0.01	0.10
N	33	0.63	0.45	C(C)CC(C)C	29	0.03	0.12
para-substituted benzene	30	0.20	0.12	cyclohexane	25	0.04	0.13
C═N	28	0.26	0.12	condensed $C_6 + C_5$ rings	24	0.01	0.07
O	28	0.86	0.72	Si	21	0.01	0.06
N−O	25	0.13	0.04	O-substituted cyclohexane	20	0.01	0.06

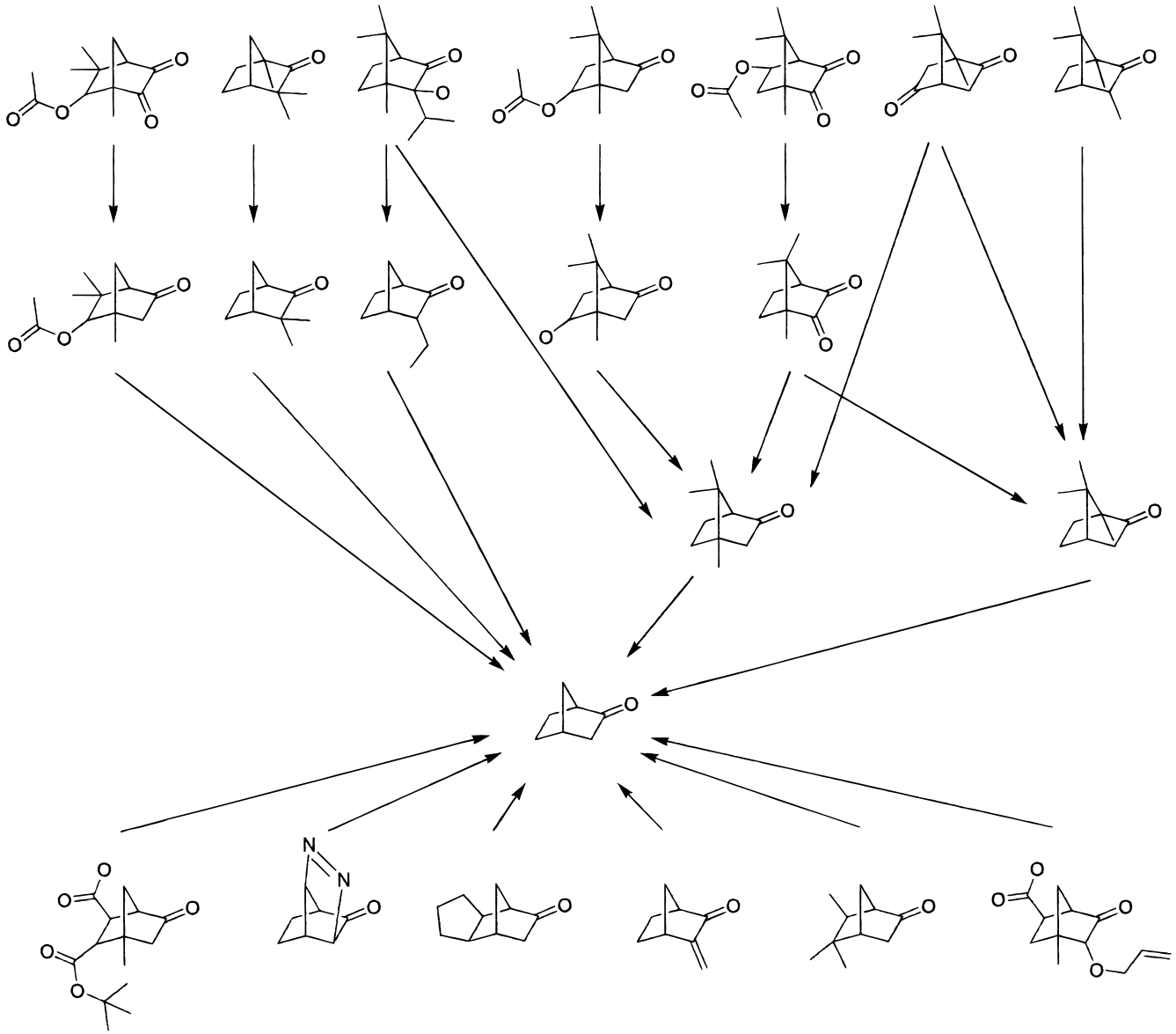


Figure 5. Topological hierarchy of 21 molecular structures all containing the 2-norbornanone substructure.

query structures. Corresponding to the applied search 2-norbornanone is the only leaf vertex in the hierarchy graph (Figure 5), and this structure is part of all other structures. Among the 13 root vertices six only contain 2-norbornanone but no other structure. The topological hierarchy does not provide synthesis pathways; however, it gives a systematic arrangement of structures; this representation is often more informative than a list which for instance is ordered by substance names or molecular formulas.

CONCLUSIONS

The substructure isomorphism matrix (**S**-matrix) is a general and useful representation of the structural diversity in a set of chemical structures. Computation of an **S**-matrix only requires a substructure search algorithm which is a standard tool in computer chemistry. PCA of a **S**-matrix provides results that are capable of characterizing the main structural factors within a set of structures and can be used to compare structural libraries.

The quadratic auto-**S**-matrix is obtained if the same structures are used as molecular structures and as substructures. From this matrix the topological hierarchy of the chemical structures can be derived; this tool is useful for an evaluation of chemical structures resulting from database searches. A set of found structures can be arranged automatically in a pedigree scheme which supports the interpretation of search results.

ACKNOWLEDGMENT

We thank R. Neudert of Chemical Concepts (Weinheim, Germany) for providing the SpecInfo IR database and the late J. T. Clerc as well as E. Pretsch (ETH Zurich, Switzerland) for generating appropriate files from this database. We also thank A. Kerber and R. Laue (University of Bayreuth, Germany) for the software MOLGEN. We are grateful to W. Demuth (Laboratory for Chemometrics, Vienna) for method and software developments. Unknown reviewers contributed by constructive hints.

REFERENCES AND NOTES

(1) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. C. M.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics*, Part A; Elsevier: Amsterdam, 1997.

(2) Varmuza, K. Chemometrics: Multivariate view on chemical problems. In *The encyclopedia of computational chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, I. H. F., Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; Vol. 1, pp 346−366.
(3) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*; Wiley: New York, 1986.
(4) Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH: Weinheim, Germany, 1993.
(5) Randic, M. On characterization of chemical structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687.
(6) Todeschini, R.; Cazar, R.; Collina, E. The chemical meaning of topological indices. *Chemom. Intell. Lab. Syst.* **1992**, *15*, 51−92.
(7) Trinajstic, N. *Chemical graph theory*; CRC Press: Boca Raton, FL, 1992.
(8) Bayada, D. M.; Hamersma, H.; van Geerenstein, V. J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.
(9) Willet, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(10) Varmuza, K.; Penchev, P. N.; Scsibrany, H. Maximum common substructures of organic compounds exhibiting similar infrared spectra. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 420−427.
(11) Varmuza, K.; Penchev, P. N.; Scsibrany, H. Large and frequently occurring substructures in organic compounds obtained by library search of infrared spectra. *Vib. Spectrosc.* **1999**, *19*, 407−412.
(12) *MOLGEN: Isomer Generator Software*, Version 3.1; Institute for Mathematics II, University of Bayreuth: Bayreuth, Germany, 1998.
(13) Varmuza, K.; Demuth, W.; Scsibrany, H. Automatic determination of the topological hierarchy of chemical structures. Manuscript in preparation.
(14) *SpecInfo: Spectroscopic Information System*, Version 3.1; Chemical Concepts: Weinheim, Germany, 1996.
(15) *NIST Mass Spectral Database*, Version 4.0; National Institute of Standards and Technology: Gaithersburg, MD 1992.

CI990267L