

Maximum Common Substructures of Organic Compounds Exhibiting Similar Infrared Spectra

K. Varmuza,^{*,†} P. N. Penchev,[‡] and H. Scsibraný[†]

Technical University of Vienna, Laboratory for Chemometrics, A-1060 Vienna, Getreidemarkt 9/152, Austria,
and University of Plovdiv, Faculty of Chemistry, BG-4000 Plovdiv, Bulgaria

Received October 10, 1997

Information about the unknown chemical structure of an organic compound can be obtained by comparing the infrared spectrum with the spectra of a spectral library. The resulting hitlist contains compounds exhibiting the most similar spectra. A method based on the maximum common substructure concept has been developed for an automatic extraction of common structural features from the hitlist structures. A set of substructures is derived that are characteristic for the query structure. Results can be used as structural restrictions in isomer generation.

INTRODUCTION

The development of automated systems for structure elucidation and identification of organic compounds – based on computer-assisted interpretation of spectra – continues to attract the attention of spectroscopists and chemometricians. Besides NMR spectroscopic techniques and mass spectrometry the use of infrared (IR) spectral data plays an important role in structure elucidation.^{1,2} Computer-based approaches for the interpretation of IR spectra can be classified into three categories: (1) knowledge-based systems in which chemical expertise is encoded to assist in spectra interpretation^{3–5}; (2) pattern recognition methods based on multivariate data analysis, statistics, and neural networks^{6–9}; and (3) the most widely used technique, namely search in spectral libraries. Each of these approaches has its own advantages and limitations, but especially library search methods have demonstrated their usefulness in scientific and laboratory practice.^{3,10} Recently, a new approach for investigating the relationships between three-dimensional (3D) molecular structures and IR spectra has been described.¹¹

The primary result of a spectral library search is a hitlist containing a set of – typically ten to hundred – reference spectra (the *hits*) that are most similar to the spectrum of the unknown. If the unknown is a member of the library, then the correct answer often appears as the first hit that exhibits a significant larger similarity to the spectrum of the unknown than the others. In such cases, a more or less unambiguous identification of the unknown is possible. However, if the unknown compound is not contained in the spectral library, a more sophisticated interpretation of the hitlist is necessary, assuming that similar spectra indicate similar structures.¹⁰

Several approaches have been suggested for a computer-assisted evaluation of spectral hitlists with the aim to extract structural information about the unknown. These approaches can be roughly divided into three groups:

(1) The structures of the hitlist are characterized by a set of molecular descriptors that describe selected features of chemical structures. A statistical evaluation of these data may indicate those structural features that have a high probability to be present or to be absent in the structure of the unknown. This method has been implemented with a pre-defined set of descriptors in the mass spectrometric library search systems STIRS,¹² MassLib,^{13,14} and NIST.¹⁵ In the multispectral database system SpecInfo,¹⁶ a set of atom-centered fragments is derived from the hitlist structures and statistically evaluated to obtain structural information about the unknown.¹⁷

(2) Spectra and structures of a hitlist are represented by two corresponding matrices, **X** (containing spectral features) and **Y** (containing molecular descriptors). The relationships between the two matrices can be investigated by multivariate chemometric methods. Promising applications of principal component analysis and partial least squares mapping have been demonstrated for mass spectra¹⁸ and IR spectra.¹⁹

(3) The concept of maximum common substructures (MCS) among the hitlist structures has found only little interest up to now for the interpretation of hitlists. An early work reports an application in mass spectrometry.²⁰ More extensive use of the MCS approach has been described together with a cluster analysis of mass spectra.^{21,22} Furthermore, hitlists obtained by a spectral similarity search with ¹³C NMR spectra have been analyzed by a MCS algorithm that also included the prediction of ¹³C NMR signals²³; it was demonstrated that the obtained MCSs often explain main structural features of the tested unknowns. A similar approach is used in the present paper. Recently, a simplified MCS concept has been applied to results obtained by a new type of library search for IR spectra.²⁴

The aim of this work was to investigate whether a MCS approach can be successfully applied to hitlists obtained from library searches with IR spectra. The final purpose was to generate structural restrictions for a systematic structure elucidation applicable to compounds that are not contained in the library. A large spectral and structural library containing >13 000 entries served as the database. Auto-

* Corresponding author. E-mail: kvarmuza@email.tuwien.ac.at. FAX: +431-581-1915.

[†] Technical University of Vienna.

[‡] University of Plovdiv.

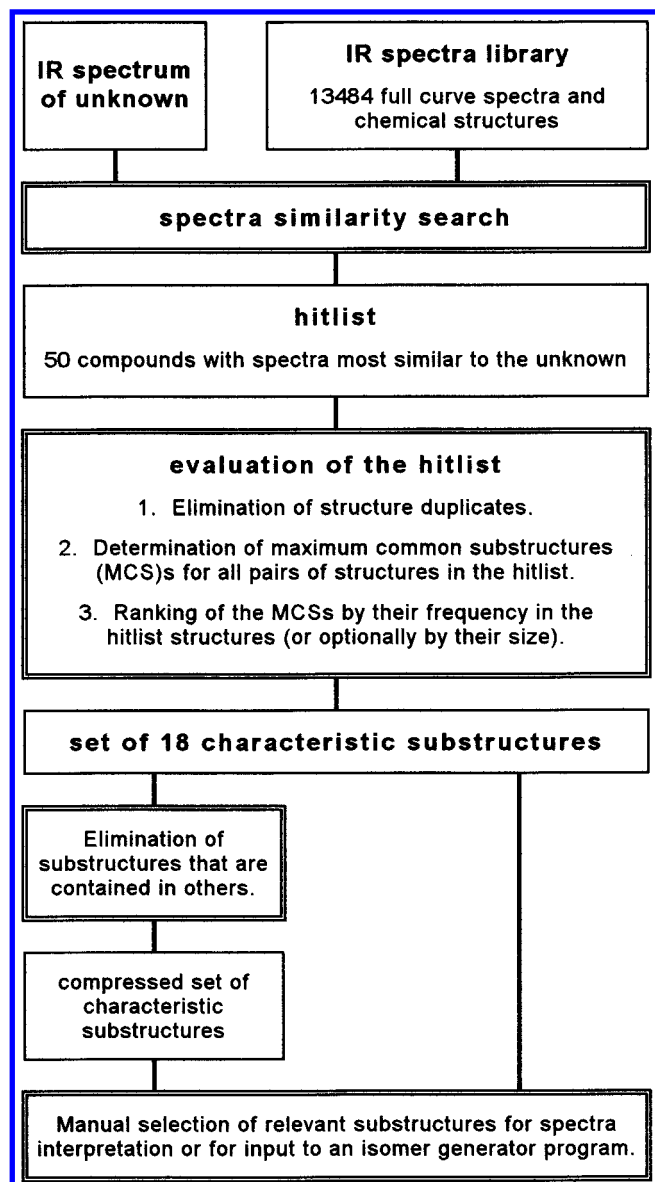


Figure 1. Scheme of the application of a maximum common substructure approach to evaluate hitlists from library search with the aim to obtain structural information. Single-line boxes denote data, others denote software or human interaction.

matic and exhaustive isomer generation by using the molecular formula and the structural restrictions were used for an objective evaluation of the method (Figure 1).

SPECTRAL LIBRARIES AND SOFTWARE

Hardware. All computations were performed on Pentium personal computers, 166 MHz, running under MS-Windows 3.11.

SpecInfo IR Library. SpecInfo¹⁶ is a multispectral database system, running on workstations. The IR database of this system contains 13 484 full-curve spectra together with chemical structures and was available for this work in JCAMP-DX format. The original spectral range is 400 to 4000 cm^{-1} , with a sampling interval of 1.93 cm^{-1} , corresponding to 1867 data points; the absorbance values are normalized to the range 0–999. Table 1 shows the distributions of molecular weights and of some compound classes in this database. The ranges of the most common elements

Table 1. Number of Compounds Per Molecular Weight Interval in the SpecInfo IR Database¹⁶

| molecular weight interval | all | no ring | any ring | benzene ring | alicyclic | only CHNO |
|---------------------------|-------|---------|----------|--------------|-----------|-----------|
| 18–100 | 396 | 286 | 110 | 6 | 94 | 331 |
| 101–200 | 5016 | 1235 | 3781 | 2195 | 1847 | 3401 |
| 201–300 | 4576 | 354 | 4222 | 3330 | 2238 | 2285 |
| 301–400 | 2520 | 75 | 2445 | 2083 | 1791 | 993 |
| 401–500 | 827 | 6 | 821 | 756 | 577 | 160 |
| 501–962 | 149 | 2 | 147 | 129 | 77 | 36 |
| sum | 13484 | 1958 | 11526 | 8499 | 6524 | 7206 |

are C_{0-50} H_{0-78} N_{0-10} O_{0-13} Br_{0-4} Cl_{0-7} F_{0-32} P_{0-3} S_{0-5} Si_{0-13} . The IR spectra, structural data, molecular formulas, and compound names were converted for use in the software products IRSS (for spectral library searches) and ToSiM (for substructure searches). IRSS uses the spectral range from 500 to 3700 cm^{-1} , with a sampling interval of 4 cm^{-1} , corresponding to 801 data points; the original spectral data were converted by a smoothing procedure based on weights from a normal distribution. The absorbances were transformed to absorbance units (AU) and scaled to the range 0–1 with a resolution of 8 bit.

IRSS. The IRSS program is for searching in libraries of IR spectra,^{25,26} and is run under MS-Windows. Seven different algorithms for the comparison of IR spectra are implemented: three methods for matching peaks, and four methods for comparing full spectral curves (as described in the *Methods* section). Furthermore, IRSS contains software tools for the import of IR spectra in JCAMP-DX format, for peak picking, and for an interactive analysis of IR spectra from mixtures based on multiple linear regression. Software IRSS is available from the authors.

ToSiM. The ToSiM program is run under MS-DOS and contains tools for the investigation of topological similarities in molecules, such as cluster analysis of chemical structures, determination of large and maximum common substructures (described in the *Methods* section), and determination of equivalent atoms and bonds in a molecule.²⁷ Import and export of structures via Molfile format is implemented. Software ToSiM is available from the authors.

MOLGEN. Version 3.1 of this software²⁸ was used under MS-Windows. MOLGEN computes complete sets of connectivity isomers for given brutto formulas. The construction of isomers is redundancy free, complete, and fast; it can be restricted by a goodlist and a badlist. The goodlist may contain overlapping substructures but also so-called *macro atoms* (nonoverlapping substructures possessing a maximum of 12 free valences). Furthermore, limits for ring size, valences, and the number of hydrogen atoms at C, N, and O atoms can be defined.

METHOD

Similarity of IR Spectra. Existing software³ uses a number of different measures for the similarity of two IR spectra. Our own experience and results reported in several papers indicate that IR spectra that are reduced to peaks yield less reliable results than full-curve spectra. In this work, for all searches, full spectra (containing 801 absorbance values between 500 and 3700 cm^{-1} , with a constant sampling interval of 4 cm^{-1}) were used and sequential searches through the entire library were always performed.

Four different measures (hit quality indices, HQI_1 to HQI_4) were applied to describe the similarity of IR spectra. All four hit quality indices range between 0 and 999 (the last value is obtained for identical spectra). Let N be the number of absorbance values in a spectrum (in this application N is equal to 801); A_k^U and A_k^R are the absorbances of the interval k in the spectrum of the unknown and in that of the reference (library) spectrum, respectively. All absorbance values are between 0 and 1.

Hit quality index HQI_1 is based on the sum of the squared absorbance differences, S_1 :

$$HQI_1 = 999(1 - S_1) \text{ with } S_1 = \sqrt{\sum_k (A_k^U - A_k^R)^2 / N} \quad (1)$$

Hit quality index HQI_2 is calculated from the sum of the absolute absorbance differences, S_2 :

$$HQI_2 = 999(1 - S_2) \text{ with } S_2 = \frac{1}{N} \sum_k |A_k^U - A_k^R| \quad (2)$$

Hit quality index HQI_3 is a normalized scalar product of two spectral vectors A^U and A^R , as shown in eq 3:

$$HQI_3 = 999S_3 \text{ with } S_3 = \frac{\sum_k A_k^U A_k^R}{|A^U||A^R|} \quad (3)$$

Hit quality index HQI_4 is based on the correlation coefficient, with A_m^U and A_m^R being the averaged absorbances in the spectrum of the unknown and the reference, respectively:

$$HQI_4 = 999(S_4 + 1)/2$$

$$\text{with } S_4 = \frac{\sum_k (A_k^U - A_m^U)(A_k^R - A_m^R)}{\sqrt{\sum_k (A_k^U - A_m^U)^2 \sum_k (A_k^R - A_m^R)^2}} \quad (4)$$

Maximum Common Substructures (MCS). The MCS of two chemical structures is the largest possible substructure that is present in two given structures. The software used, ToSiM,²² contains a tool for the determination of the MCS of two structures that are input by two-dimensional connectivity tables; the type of the MCS can be defined by some parameters. In this application, two substructures are considered to be identical (isomorphic) if all atoms (elements) and all bonds (single, double, triple, aromatic) can be matched. Optionally, a further restriction can be applied concerning the number of hydrogen atoms: two nonhydrogen atoms are considered to be identical only if the number of hydrogens bonded to them is equal.

The size of a substructure is defined by the number of nonhydrogen atoms. Only connected substructures are considered as a MCS; furthermore, a minimum number of nonhydrogen atoms in a MCS can be defined (with a default value of four). In the case where more than one MCS is possible, ToSiM only finds one of them. The algorithm applied in ToSiM is based on the generation of trees starting at selected atoms in each molecule.²²

The MCS is a measure and a description of the similarity of two structures. The MCS of a set of n structures, however, may be very small or may even not exist if an exotic structure is accidentally contained in the set. Furthermore, searches for the MCS of many structures are computationally very demanding. Therefore, the common structural properties of a set of structures are described by a set of characteristic substructures, each of them being the MCS of a pair of structures.^{21–23} Such a set of characteristic substructures is obtained as follows: In the first step for each of the $n(n-1)/2$ pairs of structures, the MCS is determined. In the second step for each MCS_{*i*} found in step 1, the number of occurrences, n_i (frequency), in the n structures is counted by applying substructure searches. In the third step, the MCSs are ordered by their decreasing ranking weight, R_i , as defined in eq 5. This ranking considers both the frequency and the size of the substructures; the different influences are determined by a user-adjustable factor f (ranging between 0 and 1):

$$R_i = (1 - f) n_i/n + f A_i/A_{\max} \quad (5)$$

where A_i is the number of nonhydrogen atoms in MCS_{*i*} and A_{\max} is the maximum number of nonhydrogen atoms in all n investigated structures. If factor f is zero, only the frequency counts for the ranking; if f is 1, only the size is considered; tests have shown that most informative results are usually obtained for values between 0 and 0.3. A set of substructures possessing the highest ranking weights can be automatically determined by the software ToSiM.

Generally, this approach cannot find the MCS of all n structures. However, in the case where the MCS from one pair of structures occurs in all n structures, no larger substructure common to all n structures can exist. The obtained set of characteristic, large, and frequently occurring substructures characterizes common and typical structural properties in the investigated set of compounds; the result is only less affected by outlier structures.

Evaluation of Library Search Hitlists by MCS. Hitlists from spectral similarity searches – each containing 50 hits – have been evaluated by the MCS approach just described. If the tested “unknown” is contained in the library, it appears as the first hit and has been removed from the hitlist; also structural duplicates in the hitlist have been removed. Tests showed that a reasonable size of the hitlist for an evaluation by the described MCS method is ~50 entries.

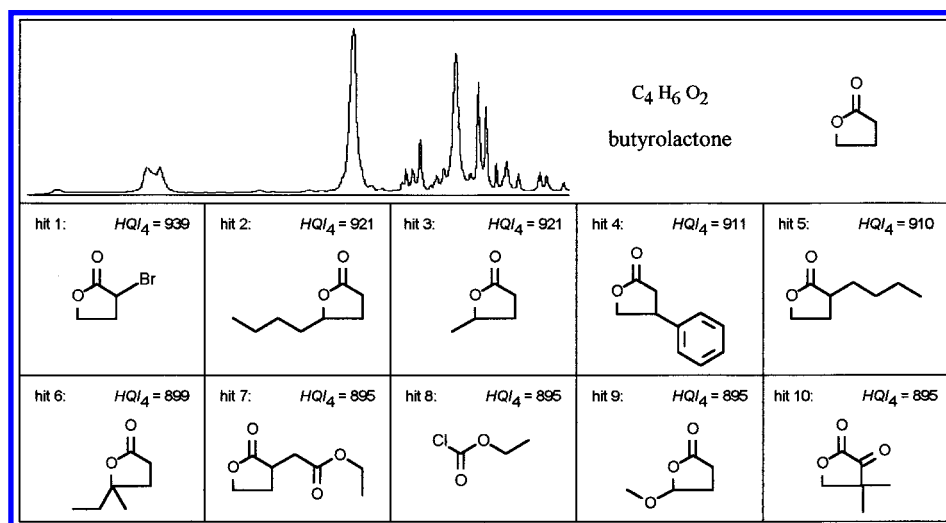
In addition to the application of single spectral similarity measures (i.e., HQI_1 to HQI_4) a combination of all four criteria has also been tested as follows: First, each similarity measure is used individually to generate a hitlist; then, a *combined hitlist* is determined as the intersection of the four hitlists and therefore only contains hits that are present in all four individual hitlists.

Figure 1 summarizes how the described MCS approach has been applied to determine typical common structural properties in hitlists. The set of characteristic substructures used usually contained 18 substructures ordered by their decreasing number of occurrences in the hitlist structures. Optionally, this list of substructures can be compressed by deleting those substructures that are contained in others. A number of tests showed that many (often all) of the obtained characteristic substructures are contained in the molecular

Table 2. Evaluation of the Statistical Significance for the Number of Occurrences of a Substructure i In a Hitlist of Size $n = 50^a$

| n_i | p_i | α | | | | | | | | |
|-------|-------|--|--------|--------|--------|--------|--------|--------|--------|--------|
| | | p (probability of the substructure in the library) | | | | | | | | |
| | | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| 5 | 0.1 | 0.5688 | 0.9815 | 0.9998 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 10 | 0.2 | 0.0245 | 0.5563 | 0.9598 | 0.9992 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 15 | 0.3 | 0.0001 | 0.0607 | 0.5532 | 0.9460 | 0.9987 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 20 | 0.4 | 0.0000 | 0.0009 | 0.0848 | 0.5535 | 0.9405 | 0.9986 | 1.0000 | 1.0000 | 1.0000 |
| 25 | 0.5 | 0.0000 | 0.0000 | 0.0024 | 0.0978 | 0.5561 | 0.9427 | 0.9991 | 1.0000 | 1.0000 |
| 30 | 0.6 | 0.0000 | 0.0000 | 0.0000 | 0.0034 | 0.1013 | 0.5610 | 0.9522 | 0.9997 | 1.0000 |
| 35 | 0.7 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0033 | 0.0955 | 0.5692 | 0.9692 | 1.0000 |
| 40 | 0.8 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0022 | 0.0789 | 0.5836 | 0.9906 |
| 45 | 0.9 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0007 | 0.0480 | 0.6161 |

^a The statistical risk α is the probability that a randomly selected hitlist contains n_i or more entries with substructure i in the molecule; α is the probability for rejecting the zero hypothesis although it is true; p is the probability of substructure i in the library; p_i is the probability of substructure i in the hitlist ($p_i = n_i/n$); α is rounded to four digits.

**Figure 2.** Hitlist obtained by a library search with the IR spectrum (3700 to 500 cm⁻¹) of butyrolactone as the unknown. The first 10 hits together with their value for hit quality index HQI_4 are shown.

structure of the unknown. Frequently, the characteristic substructures cover almost the entire molecule. In some cases, however, substructures are found that are not part of the unknown; these errors can be detected if additional information about the unknown is available (for instance a molecular weight range, the molecular formula, or presence/absence of substructures as obtained from spectral classifiers). The occurrence of false positives is not considered to be a severe problem in practice because structure elucidation always should apply different complementary techniques that allow cross checks of the results.

In addition to the relevance of an MCS for structure elucidation, the statistical significance of the number of occurrences in the hitlist also has to be taken into account. Let p be the probability of a substructure in the library. If the hitlist (size n) would be a random sample of the library (zero hypothesis), the probabilities $p(k)$ for having k compounds containing this substructure in a hitlist are given by the binomial distribution shown in eq 6:

$$p(k) = [n!/(k!(n-k)!)] p^k (1-p)^{n-k} \quad (6)$$

For an actual number n_i of occurrences of substructure MCS _{i} in a hitlist, a statistical risk α can be calculated by eq 7; α is the probability that a randomly selected hitlist contains n_i or more entries with substructure MCS _{i} . In other words,

α is the probability for rejecting the zero hypothesis although it is true:

$$\alpha = \sum p(k) \text{ with } k = n_i \dots n \quad (7)$$

Table 2 contains the values of α for probabilities $p = 0.1, 0.2, \dots, 0.9$, and numbers of occurrence $n_i = 5, 10, \dots, 45$ in hitlists of size 50. For example, if the probability of a substructure in the library is 0.3, the probability that a hitlist of size 50 accidentally contains 25 or more compounds possessing this substructure is 0.0024. Typical results for MCSs in hitlists that were obtained by IR spectra library searches had values of α of <0.001, which demonstrates a high statistical significance.

RESULTS

The main features, some applications, and limits of the described MCS approach for systematic structure elucidation are demonstrated and discussed next by three examples.

Example C₄H₆O₂. Figure 2 shows the first 10 hits obtained by a library search with the IR spectrum of butyrolactone as the unknown and using the hit quality index HQI_4 (based on the correlation coefficient). The butyrolactone ring can be easily identified as a characteristic substructure directly from the hitlist structures; it is contained in nine of the first 10 hits (if considering hybridization in eight hits).

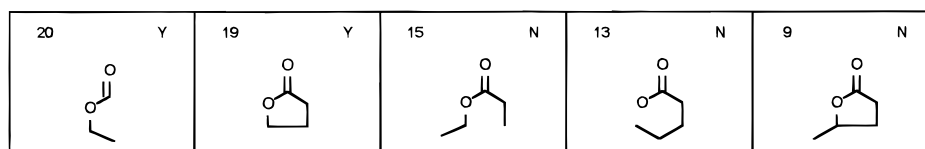


Figure 3. Five most frequent characteristic substructures obtained from the first 20 hits; library search was done with the IR spectrum of butyrolactone using hit quality index HQI_4 . For each substructure, the number of occurrences in the 20 hitlist structures is given. Key: (Y) substructure contained in the query structure; (N) substructure not contained in the query structure.

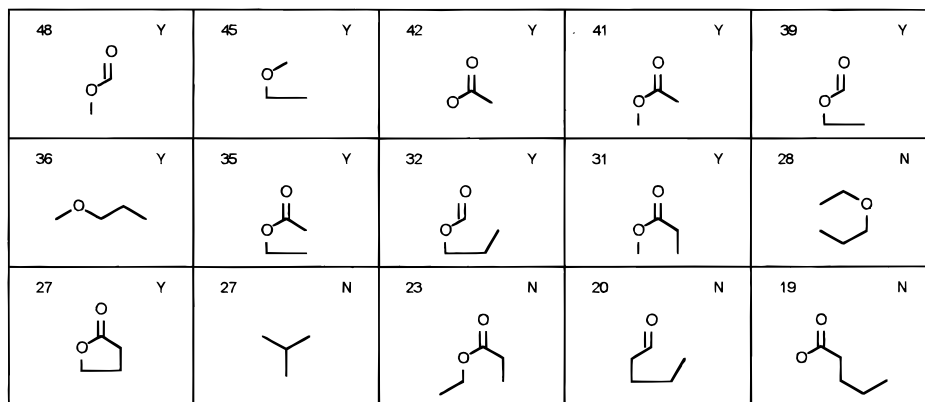


Figure 4. Fifteen most frequent characteristic substructures obtained from all 50 hits; library search was done with the IR spectrum of butyrolactone using hit quality index HQI_4 . For each substructure, the number of occurrences in the 50 hitlist structures is given. Key: (Y) substructure contained in the query structure; (N) substructure not contained in the query structure.

The structural information contained in a hitlist depends on the number of considered hitlist structures. This influence is demonstrated by a comparison of the results obtained from the first 20 hits and those obtained from all 50 hits.

The five most frequent characteristic substructures obtained from the first 20 hits are shown in Figure 3. The butyrolacton ring has been found as a characteristic substructure occurring in 19 hitlist structures. The substructure C—C—O—C=O is contained in all first 20 hitlist structures and therefore is a MCS of all investigated structures. Three other substructures that have been found are wrong because they are not part of the unknown.

For a comparison, the most frequent characteristic substructures obtained from all 50 hits are shown in Figure 4. The first nine substructures are all contained in the molecular structure of the unknown; in total, 10 are correct and five are wrong. If the molecular formula of the unknown is considered to be known, then four of the five wrong substructures are not relevant anymore because they contain too many carbon atoms. This example is representative for the general trend indicating that hitlists with ~50 compounds yield optimum results with the spectral library used.

The five most frequent characteristic substructures from Figure 4 have been used as structural restrictions for automatic isomer generation. The total number of isomers of $C_4H_6O_2$ is 263; if only the first substructure is requested to be present, the number of possible molecular structures is reduced to nine; if all five substructures are requested to be present, only one isomer (the correct butyrolacton) survives (Table 3).

Example $C_{14}H_{11}NO_2$. The IR spectrum of benzamide, N -(4-formylphenyl), shown in Figure 5a, has been searched in the spectral library by applying all four hit quality indices HQI_1 to HQI_4 separately. From each hitlist (each containing 50 compounds), 18 characteristic substructures have been derived; all except two are part of the query structure.

Table 3. Use of Characteristic Substructures from Figure 4 as Goodlist Restrictions In Isomer Generation for the Butyrolacton ($C_4H_6O_2$) Example

| substructure number j | number of isomers ^a | |
|-------------------------|--------------------------------|------------------------------------|
| | substructure j in goodlist | substructures 1 to j in goodlist |
| 1 | 9 | |
| 2 | 72 | 7 |
| 3 | 9 | 4 |
| 4 | 3 | 3 |
| 5 | 5 | 1 |

^a Without any restrictions 263 isomers are possible.

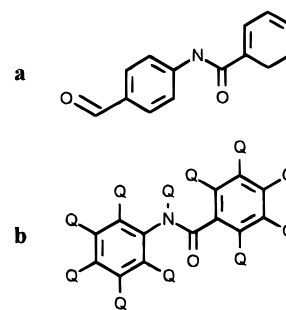


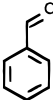
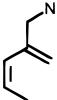




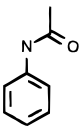


Figure 5. (a) Benzamide, N -(4-formylphenyl), example ($C_{14}H_{11}NO_2$). (b) Macro atom with 11 free valences (Q) as obtained from an evaluation of the IR library search hitlist by the described MCS approach. Although the macro atom covers most of the query molecule, 458 isomers are still possible if only valence rules are applied.

Removing all substructures that are contained in others resulted in four compressed sets of characteristic substructures, as displayed in Figure 6. The most informative results have been obtained by applying hit quality index HQI_4 which is based on the correlation coefficient: all five substructures are correct and their frequencies in the hitlist structures have the highest values. Application of the normalized scalar

| | | | | |
|------------------|---|---|---|---|
| HQI ₁ | 26 | Y | 25 | Y |
| |  | |  | |

| | | | | | | |
|------------------|---|---|---|---|---|---|
| HQI ₂ | 30 | Y | 26 | Y | 24 | Y |
| |  | |  | |  | |

| | | | | | | | | |
|------------------|---|---|---|---|---|---|--|---|
| HQI ₃ | 20 | N | 19 | N | 17 | Y | 16 | Y |
| |  | |  | |  | |  | |






| | | | | | | | | | | |
|------------------|---|---|---|---|---|---|--|---|---|---|
| HQI ₄ | 36 | Y | 30 | Y | 29 | Y | 25 | Y | 23 | Y |
| |  | |  | |  | |  | |  | |

Figure 6. Benzamide, *N*-(4-formylphenyl) example ($C_{14}H_{11}NO_2$); structure given in Figure 5a. Compressed sets of characteristic substructures obtained from hitlists that have been generated by the use of the four hit quality indices HQI_1 to HQI_4 , as defined in eqs 1–4. For each substructure, the number of occurrences in the 50 hitlist structures is given. Key: (Y) substructure contained in the query structure.

product (HQI_3) was less successful: two from the four substructures are wrong and their frequencies in the hitlist structures have the lowest values. Similar results have been obtained with other examples.

Although the resulting substructures shown in Figure 6 cover almost the entire molecular structure of the unknown, a direct use of them in the goodlist for isomer generation is not successful. The very high number²⁹ of isomers for $C_{14}H_{11}NO_2$ cannot be reduced to a manageable size just by using one of these substructures as a macro atom and the others in the goodlist. In this case, an interaction of the chemist and some assumptions are required for testing and utilizing the result. From the obtained characteristic substructures it is reasonable (and might be supported by other data or knowledge about the unknown) that the unknown molecular structure contains two benzene rings and two different carbonyl groups. Based on this hypothesis a macro atom as shown in Figure 5b can be constructed; thereby, the number of possible molecular structures is reduced to 458. Although this macro atom is almost identical with the skeleton of the query molecule, a relative large number of isomers still remains; the principal reason for this result is the great number of 11 free valences in the macro atom.

Browsing through the 458 candidate structures or applying a cluster analysis of the structures^{30,31} indicates that many of the generated structures contain ring systems that probably cannot be present in stable chemical compounds. Therefore, two further assumptions about the structure of the unknown have been postulated: (a) only six-membered rings were permitted to avoid a second bond between the two benzene rings, and (b) bridges between nonneighboring atoms of a benzene ring were forbidden if consisting only of two atoms. Considering these restrictions, only seven molecular candidate structures survive with the correct structure included.

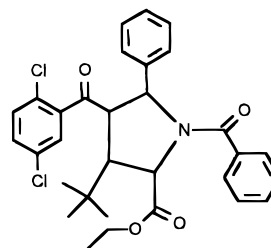


Figure 7. Proline, 1-benzoyl-4-(2,5-dichlorobenzoyl)-3-(1,1-dimethylethyl)-5-phenyl, ethyl ester, example ($C_{31}H_{31}NO_4Cl_2$).

Example $C_{31}H_{31}NO_4Cl_2$. The IR spectrum of a higher molecular weight compound, namely proline, 1-benzoyl-4-(2,5-dichlorobenzoyl)-3-(1,1-dimethylethyl)-5-phenyl, ethyl ester; Figure 7, is used to demonstrate the advantage of a joint use of different spectra similarity criteria. In the first step, separate library searches have been performed with the four hit quality indices HQI_1 to HQI_4 . In a second step, the intersection of the four hitlists has been built, resulting in a new hitlist containing 25 reference structures that are common to all four individual hitlists. In a third step, this new hitlist was used for the generation of a set of characteristic substructures by the MCS approach. The final fourth step generated the compressed set resulting in seven substructures (Figure 8a).

For comparison, the characteristic substructures that have been obtained by HQI_4 (based on the correlation coefficient) are also shown (Figure 8b). All these substructures are correct because they are part of the investigated structure; the number of occurrences in the hitlist is statistically significant for all substructures ($\alpha < 10^{-6}$). However, the substructures obtained by a joint use of the four hitlists provide more structural information than those deduced from the single hitlist: the former are larger and more unique

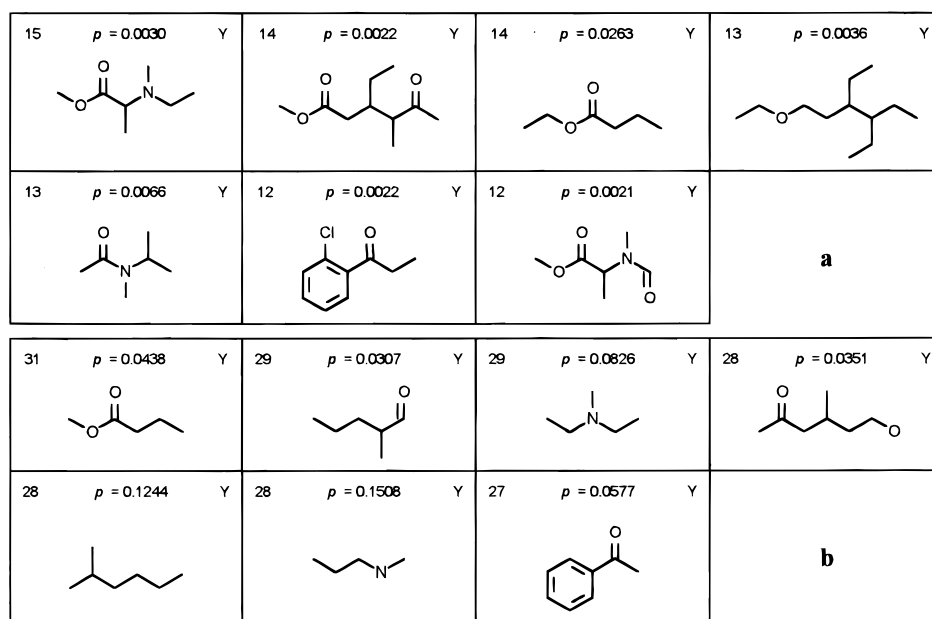


Figure 8. Proline, 1-benzoyl-4-(2,5-dichlorobenzoyl)-3-(1,1-dimethylethyl)-5-phenyl, ethyl ester, example ($C_{31}H_{31}NO_4Cl_2$), structure given in Figure 7. (a) Compressed set of characteristic substructures obtained from a joint use of four hitlists that have been generated by using hit quality indices HQI_1 to HQI_4 , as defined in eqs 1–4. The characteristic substructures were derived from the intersection (25 structures) of the four hitlists (each containing 50 structures). (b) Compressed set of characteristic substructures obtained from the hitlist (containing 50 structures) that has been generated by using only hit quality index HQI_4 (based on the correlation coefficient). For each substructure, the number of occurrences in the 25 or 50 hitlist structures, respectively, is given. Key: p probability of the substructure in the complete spectral library; (Y) substructure contained in the query structure.

(because the probability of occurrence in the complete spectral library is much smaller).

The substructures found to be characteristic are large building blocks of the query structure; however, a direct use of them as structural restrictions in isomer generation is not reasonable. In this case, a systematic and exhaustive structure elucidation would only be possible with additional structural data that allow the definition of large macro atoms with a small number of free valences.

CONCLUSION

A MCS approach has been described that allows the determination of a set of substructures from hitlists obtained by IR spectral similarity searches. These substructures are often characteristic for the molecular structure of the unknown. In some cases, they can be directly applied as structural restrictions for automatic isomer generation. In general, however, an interaction of the chemist is necessary to detect inconsistencies or wrong substructures.

The main factors influencing the result are: (1) the contents and size of the spectral library, (2) the applied spectral similarity measure, (3) the number of hitlist structures used for the determination of MCSs, and (4) the parameters of the MCS procedure. From the presented examples, some conclusions can be drawn regarding the influencing factors. The most powerful spectra similarity criterion tested is that based on the correlation coefficient. A parallel use of different similarity criteria and the intersection of the hitlist structures usually improves the result. The optimum size of hitlists is ~ 50 . The MCSs are preferably ordered by their frequency of occurrence in the hitlist structures. A compressed list of characteristic substructures is often more useful than many small substructures. The method can only be successful if several structures in the

library are similar to the unknown. If this is not the case a misuse can often be avoided because the hitlist then contains very different structures and the resulting “characteristic” substructures are present in only a few hitlist structures.

The resulting set of characteristic substructures has to be checked carefully and compared with available knowledge about the unknown. However, even if some of the substructures are wrong (that means they are not or not completely contained in the structure of the unknown), the result provides useful structural information that usually cannot be obtained directly from a visual inspection of the hitlist.

The described MCS approach is not limited by predefined molecular descriptors or compound classes; it is self-adapting to the type and complexity of the molecular structures contained in the hitlist.

ACKNOWLEDGMENT

We thank R. Neudert of Chemical Concepts (Weinheim, Germany) for providing the SpecInfo IR database. We are grateful to J. T. Clerc and E. Pretsch (ETH Zurich, Switzerland) for making this database available in an appropriate format. We also thank A. Kerber and R. Laue (University of Bayreuth, Germany) for providing the software MOLGEN.

REFERENCES AND NOTES

- (1) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Wiley: New York, 1986.
- (2) *Computer-Supported Spectroscopic Databases*; Zupan, J., Ed.; Ellis Horwood: Chichester, 1986.
- (3) Luinge, H. J. Automated Interpretation of Vibrational Spectra. *Vib. Spectrosc.* **1990**, *1*, 3–18.
- (4) Ehrentreich, F. Representation of Extended Infrared Spectrum-Structure-Correlations Based on Fuzzy Theory. *Fresenius' J. Anal. Chem.* **1997**, *357*, 527–533.

- (5) Curry, B. A Distributed Expert System for Interpretation of GC/IR/MS Data. In *Computer-Enhanced Analytical Spectroscopy*; Meuzelaar, H. L. C.; Ed.; Plenum: New York, 1990; Vol. 2, pp 183–209.
- (6) Klawun, C.; Wilkins, C. L. Joint Neural Network Interpretation of Infrared and Mass Spectra. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249–257.
- (7) Munk, M. E.; Madison, M. S. The Neural Network as a Tool for Multispectral Interpretation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231–238.
- (8) Novic, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counterpropagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454–466.
- (9) Ricard, D.; Cachet, C.; Cabrol-Bass, D. Neural Network Approach to Structural Feature Recognition from Infrared Spectra. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 202–210.
- (10) Clerc, J. T. Automated Spectra Interpretation and Library Search Systems. In *Computer-Enhanced Analytical Spectroscopy*; Meuzelaar, H. L. C.; Isenhour, T. L., Eds.; Plenum: New York, **1987**, 145–162.
- (11) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.
- (12) Kwok, K. S.; Venkataraghavan, R.; McLafferty, F. W. Computer-Aided Interpretation of Mass Spectra. III. A Self-Training Interpretive and Retrieval System. *J. Am. Chem. Soc.* **1973**, *95*, 4185–4194.
- (13) Zalfen, U. Entwicklung und Anwendung computerunterstützter Verfahren in der Massenspektrometrie. Dissertation, University of Bielefeld, Germany, 1991.
- (14) Henneberg, D.; Weimann, B.; Zalfen, U. Computer-Aided Interpretation of Mass Spectra Using Data Bases with Spectra and Structures. I. Structure Searches. *Org. Mass Spectr.* **1993**, *28*, 198–206.
- (15) Stein, S. E. Chemical Substructure Identification by Mass Spectral Library Searching. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 644–655.
- (16) *SpecInfo: Spectroscopic Information System*, vers. 3.1, 1996; Available from: Chemical Concepts, P.O. Box 100202, D-69442 Weinheim, Germany.
- (17) Bremser, W.; Grzonka, M. SpecInfo – A Multidimensional Spectroscopic Interpretation System. *Mikrochim. Acta II* **1991**, 483–491.
- (18) Varmuza, K.; Werther, W.; Henneberg, D.; Weimann, B. Computer-Aided Interpretation of Mass Spectra by a Combination of Library Search with Principal Component Analysis. *Rapid Commun. Mass Spectrom.* **1990**, *4*, 159–162.
- (19) Werther, W.; Varmuza, K. Exploratory Data Analysis of Infrared Data. *Fresenius' J. Anal. Chem.* **1992**, *344*, 223–226.
- (20) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. W. Molecular Structure Comparison Program for the Identification of Maximal Common Substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668–7671.
- (21) Scsibraný, H.; Varmuza, K. Common Substructures in Groups of Compounds Exhibiting Similar Mass Spectra. *Fresenius' J. Anal. Chem.* **1992**, *344*, 220–222.
- (22) Scsibraný, H.; Varmuza, K. Topological Similarity of Molecules Based on Maximum Common Substructures. In *Software Development in Chemistry*; Ziessow, D., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1993; Vol. 7, pp 77–87.
- (23) Chen, L.; Robien, W. Application of the Maximum Common Substructure Algorithm to Automatic Interpretation of ¹³C-NMR Spectra. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 934–941.
- (24) Ehrentreich, F. Derivation of Substructures from Infrared Band Shapes by Fuzzy Logic and Partial Cross Correlation Functions. *Fresenius' J. Anal. Chem.* **1997**, *359*, 56–60.
- (25) Penchev, P. N.; Sohau, A. N.; Andreev, G. N. Description and Performance Analysis of an Infrared Library Search System. *Spectrosc. Lett.* **1996**, *29*, 1513–1522.
- (26) Penchev, P. N.; Kochev, N. T.; Andreev, G. N. IRSS: A Program System for Infrared Library Search. *Compt. Rend. Acad. Bulg. Sci.* **1998**, in print.
- (27) Scsibraný, H.; Varmuza, K. ToSiM: PC-Software for the Investigation of Topological Similarities in Molecules. In *Software Development in Chemistry*; Jochum, C., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1994, Vol. 8, 235–249.
- (28) Wieland, T.; Kerber, A.; Laue, R. Principles of the Generation of Constitutional and Configurational Isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413–419.
- (29) Within 115 h computing time (Pentium 166 MHz), the software MOLGEN generates 2.1×10^9 isomers of C₁₄H₁₁NO₂ and roughly estimates that this number is ~0.3% of the total number of isomers.
- (30) Varmuza, K.; Scsibraný, H. Cluster Analysis of Chemical Structures Based on Binary Molecular Descriptors and Principal Component Analysis. In *Software Development in Chemistry*; Moll, R., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1995; Vol. 9, pp 81–90.
- (31) Varmuza, K.; Werther, W. Mass Spectral Classifiers for Supporting Systematic Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 323–333.

CI9700889