Anal. Chem. **2002,** *74,* 1484−1492

A Procedure for Decreasing Uncertainty in the Identification of Chemical Compounds Based on Their Literature Citation and Cocitation. Two Case Studies

Boris L. Milman*

D.I. Mendeleyev Institute for Metrology, 19 Moskovsky pr., 198005 St. Petersburg, Russia

An identification procedure connected with selection of candidates for identification according to high values of their literature citation and cocitation is suggested. The citation rate of the compound is the number of different literature units, such as papers, patents, etc., recording its name. The cocitation rate is the number of such units mutually recording the names of two corresponding compounds or the compound and the sample/matrix. General citation of a chemical compound is assumed to be related to the prior probability of its being contained in a sample to be analyzed. This citation measures abundance and popularity of the compound. Cocitation of a compound with a known/identified constituent of a sample is related to their mutual similarity in structure or properties, origin, use, etc. This data processing method is validated by counting citations and cocitations for detected impurities in pure *n*-hexane and naphthalene, polycyclic aromatic hydrocarbons in waste gas, as compared with counts for similar or dissimilar compounds that are absent in the samples. The analytes are preidentified by combined gas chromatography and mass spectrometry techniques. A median and a mean value of citation and cocitation are always higher for the group of unambiguously identified compounds. A difference between identified and similar compounds in citations or cocitations may be rather insignificant, with combined evaluation of both indicators distinguishing these groups. Chemical dissimilarity results in a large difference in cocitation values.

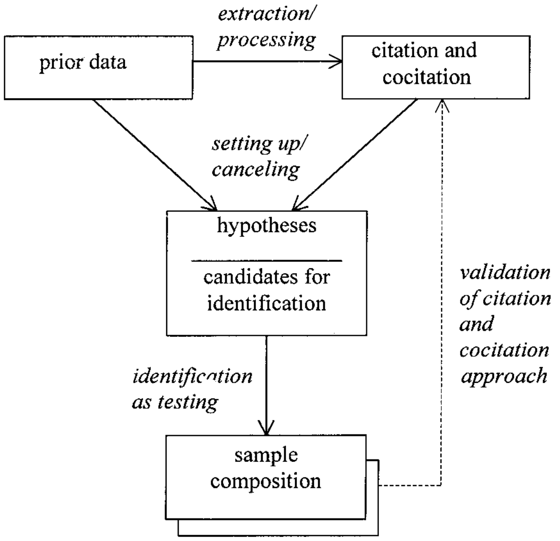

Figure 1. Involving prior data as direct operation or through extracted/processed information into an identification procedure that is the hypotheses test. Citation and cocitation data intermediately obtained from raw prior data lead to advanced hypotheses for sample composition. This operation is validated as the reverse one in this paper with the use of the identified constituents of the samples.

In recent years, particular emphasis has been placed upon the nature of the identification procedure in chemical analysis.[1−6] This procedure was considered by the author to be screening of identification hypotheses followed by experimental testing of each hypothesis and by the eventual chemical classification decision of an analyst as an expert.[2−4] Analogous concepts have been suggested by other researchers.[5,6] An identification hypothesis is an assumption that a particular chemical compound is present in a sample (matrix) to be analyzed. Search for prior data and estimation of a prior probability[1,3,4] of the presence of a compound in the sample provide setting up and canceling hypotheses before testing (Figure 1), thus leading to a series of promising candidates for identification.

Different ways and means for retrieving and processing prior information can be used.[1,3,4] A prior probability for a compound to be contained in a sample might be related to its abundance/spreading/popularity and similarity to a known constituent of this sample in structure, properties, origin, etc. Other conditions being equal, a more abundant or known chemical compound has a better chance than a rare compound of being detected in a random sample. Citation of a compound in chemical literature can be

* Fax +7 812 3 27 97 76. E-mail: bmilman@mail.rcom.ru.

[1] Ellison, S. L. R.; Gregory, S.; Hardcastle, W. A. *Analyst* **1998**, *123*, 1155−1161.

[2] Milman, B. L.; Konopelko, L. A. *CCQM/99−01 working document submitted to the 4th meeting of the Consultative Committee for Amount of Substance* Bureau international des poids et mesures, Sevres (France), 1999.

[3] Milman, B. L.; Konopelko, L. A. *Fresenius' J. Anal. Chem*. **2000**, *367*, 621−628.

[4] Milman, B. L.; Kovrizhnyh, M. A. *Fresenius' J. Anal. Chem*. **2000**, *367*, 629−634.

[5] Hartstra, J.; Franke, J. P.; De Zeeuw, R. A. *J. Chromatogr. B* **2000**, *739*, 125−137.

[6] Eriksson, J.; Chait B. T.; Fenyö, D. *Anal. Chem.* **2000**, *72*, 999−1005.

10.1021/ac010611p CCC: $22.00
Published on Web 02/22/2002

considered as an indicator of its abundance and related features. Another assumption that two compounds detected in the same matrix may have the same origin, similar structure/properties/use and, therefore, be simultaneously recorded, that is, cocited in the same scientific literature units (papers, patents, etc.), is also quite appropriate. However, to use citation and cocitation rates for identifying chemical compounds, the efficiency of this data method should be validated (Figure 1) using samples of known composition and comparing the citation and cocitation behavior of available vs absent compounds.

Following the identification of impurities in *n*-hexane and naphthalene in our laboratory, information searching in *Chemical Abstracts* (CA) for 1997 revealed that the identified impurities, as compared the with the compounds similar to them in mass spectra or gas chromatography (GC) retention indices (RI), are more highly cited in the chemical literature.[4] Furthermore, impurities were found to be rather frequently cocited together with the ground substances, that is, *n*-hexane and naphthalene.[4] Hence, the indicator of cocitation (coreference[4]) as compound-to-compound cocitation may be efficient for advancing or canceling the assumption associated with the presence of substances similar or dissimilar with respect to a known constituent/matrix in formulas, properties, etc.

This paper presents the first straightforward study of citation and cocitation of chemical compounds for analytical aims.[7] Considerably extending initial observations and conclusions,[4] this study makes it possible to suggest and validate a new identification procedure based on citation and cocitation estimation of the prior probability of the presence of an analyte in various samples.

The new research topics are as follows: In the first case covering the impurities in the above-mentioned hydrocarbons and compounds similar to these impurities, their citation and cocitation are investigated for another time range in order to reveal the stability in differentiation of compounds according to these indicators. Then the potentiality of cocitation is additionally elucidated. A response of this indicator to the hardly probable copresence in the same sample of the hydrocarbons and the compounds of very different natures and uses (such as common pharmaceuticals) is studied. The cocitation of chemical compounds with the names of a series of common matrixes, that is, compound-to-matrix cocitation, is also evaluated.

The second case refers to polycyclic aromatic hydrocarbons (PAH) and some of their isomers unambiguously or ambiguously identified and unidentified in the waste gas of aluminum production. Different citation and cocitation rates are also estimated and discussed. Mass spectrometry (MS) and GC were used and are briefly outlined as the applied analytical technique.

EVALUATION OF CITATION AND COCITATION IN IDENTIFICATION PROCEDURE

General. In the identification procedure, some analytes usually remain unidentified or ambiguously identified. The former should be provided with supplemental identification hypotheses. The latter need the deletion of redundant candidates for identification. An estimation of citation and cocitation is intended for these purposes.

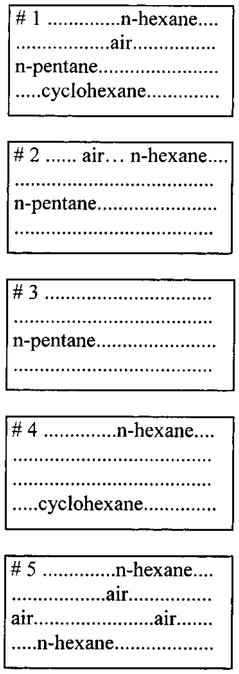
# 1 ....n-hexane....
..................air................
n-pentane........................
.....cyclohexane..............
# 2 ...... air… n-hexane....
....
n-pentane........................
....
# 3 ....
....
n-pentane........................
....
# 4 ....n-hexane....
....
....
.....cyclohexane..............
# 5 ....n-hexane....
..................air................
air.......................air.......
.....n-hexane...................

Figure 2. Example of an abstracts subset recording the names of chemical compounds and a matrix. Other words are omitted. According to definitions (Table 1), the citation rates for this abstracts subset are as follows: *n*-pentane, 3; *n*-hexane, 4; and cyclohexane, 2. The cocitations are: *n*-pentane-to-*n*-hexane, 2; *n*-pentane-to-cyclohexane, 1; *n*-hexane-to-cyclohexane, 2; *n*-pentane-to-air, 2; *n*-hexane-to-air, 3; and cyclohexane-to-air, 1.

First, before or after tentative identification resulting in unidentified analytes, the compounds selected on the basis of high citation and cocitation rates are considered as advanced candidates for identification. An initial list of candidates can be generated by the estimation of one or several types of cocitation (Table 1), for example, compound X-to-compound 0 and compound X-to-matrix cocitation. Here, compound 0 is the most cited among known or identified compounds in the sample/matrix, and compound X is a candidate for identification. The exact definitions of these indicators and the examples of their evaluation are given in Table 1 and Figure 2. A compound-to-compound indicator can be substituted by a compound-to-compound group indicator, with compound group referring to all known or identified compounds (Table 1).

A relatively short list of initial identification hypotheses does not need reducing to be tested by experimental techniques for finite time. For a rather long list, a selection of candidates for identification using high or above-threshold cocitation values should be performed. There may be many ways of estimating the threshold value. The cocitation of all identified compounds except compound 0 with this compound 0 can be estimated, and the lowest corresponding value or some parameter of distribution in cocitation, for example, a low quartile (see the next subsection) as the threshold value can be chosen. If there are no available known or identified substances, cocitation with matrix can be used, the threshold being derived from distribution in this indicator.

[7] So far, the use of citation and cocitation connected with bibliographic references and some kinds of words, such as the key words, terms, etc., was restricted by scientometrics, computer science, sociology, etc. See, for example: Garfield, E. C*itation Indexing* − *Its Theory and Application in Science, Technology, and Humanities*; John Wiley & Sons: New York, 1979. Smal,l H.; Sweeney, E. *Scientometrics* **1985**, *7*, 391−409. Law, J.; Bauin, S.; Courtial, J.-P.; Whittaker, J. *Scientometrics* **1988**, *14*, 251−264. Milman, B. L.; Gavrilova, Yu. A. *Scientometrics* **1993**, *27*, 53−74.

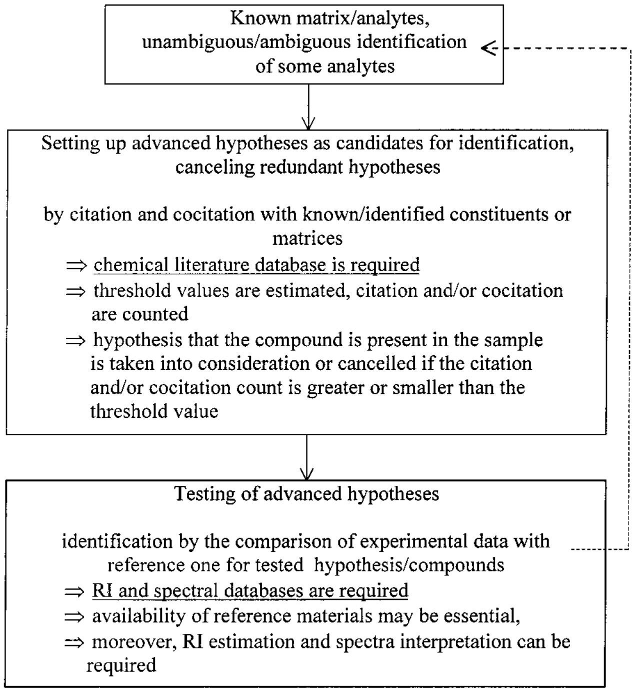
Known matrix/analytes, unambiguous/ambiguous identification of some analytes
Setting up advanced hypotheses as candidates for identification, canceling redundant hypotheses
by citation and cocitation with known/identified constituents or matrices
⇒ chemical literature database is required
⇒ threshold values are estimated, citation and/or cocitation are counted
⇒ hypothesis that the compound is present in the sample is taken into consideration or cancelled if the citation and/or cocitation count is greater or smaller than the threshold value
Testing of advanced hypotheses
identification by the comparison of experimental data with reference one for tested hypothesis/compounds
⇒ RI and spectral databases are required
⇒ availability of reference materials may be essential,
⇒ moreover, RI estimation and spectra interpretation can be required

Figure 3. Chart of citation- and cocitation-based algorithm for identifying the unknown constituent of the sample to be analyzed. The threshold is optional and depends on the analytical problem.

For a case of great significance, for example, an enigmatic large-scale pollution, each unit cocitation with key words as the names of the matrix and identified compounds should be taken into account.

Moreover, the citation of candidates for identification is a natural constraint for their list. The candidates are removed from the initial list provided the citation rate is lower than some threshold value established from their distribution in citations or on the basis of the analyst's experience.

For the case of ambiguously identified analytes, data processing is a similar procedure. The rarely cited and cocited candidates for identification can be treated as redundant and, therefore, can be canceled.

The algorithm for the identification of an unknown on the basis of these statistical indicators is given in Figure 3. Although some kind of a special database would be ideal for this purpose, general chemical databases, such as CA, are suitable for citation/cocitation studies.

Case Studies. This data method is displayed and validated for the samples that are specially analyzed (see the next section). The first case concerns impurities in pure *n*-hexane and naphthalene (Table 2). The values of indicators estimated for them were compared with those for three reference groups of the compounds. The two groups contained substances of the same nature having similar chromatographic behavior and mass spectra (see ref 4 and Table 2). The third group consisted of 10 pharmaceuticals (Table 2), compounds containing more than one functional group, as the most popular analgesics and antibiotics advertised in Russian medical and pharmaceutical periodicals.[8]

The second case is PAH contained in waste gas, as compared with three reference groups of related compounds (see the next section).

Different individual citation and cocitation indicators are defined, and examples are given in Table 1. The citation/cocitation count was based on the *Chemical Substance Index* (CSI) to the printed version of CA. The classification of matrixes for the evaluation of compound-to-matrix cocitation is given in Table 3. The indicator of this type covers only the abstracts/entries associated with chemical analysis (Table 3). Only environmental and waste matrixes are considered for PAH as well-known pollutants. The issues of CSI for January−June 1987 and 1997 (*n*-hexane and naphthalene impurities, similar substances, and pharmaceuticals) and 1995 (PAH) were used as the specific information source. Data retrieval, counting citation and partly cocitation for *n*-hexane and naphthalene impurities and the compounds similar to both types of impurities (Table 2) in CA for 1997 have also been outlined in ref 4.

Compound distributions in statistical indicators were estimated by the following parameters: median, quartiles and mean values, standard deviation, and confidence range. Groups of analytes as unambiguously identified compounds and reference groups were compared in median and mean values, the significance level,[3] α, of the *t* test, and the fraction of compounds having the above-threshold citation/cocitation value. The characteristic calculation results are given here. The low quartiles for unambiguously identified impurities and constituents of the PAH mixture were conditionally considered as threshold values. Relative cocitation values (Table 1) for zero count of citation and cocitation are conditionally taken as zero values.

All statistical evaluations were carried out using standard commercial software.

EXPERIMENTAL SECTION

Impurities in Hexane and Naphthalene. Identification of impurities was described.[4] These compounds and the substances with mass spectra fit values or RI most similar to those for the impurities are listed in Table 2.

Polycyclic Aromatic Hydrocarbons. PAH were contained in the waste gas of Volgograd Aluminum Company (Volgograd, Russia). The gas containing dust was filtered with an AFAS−PAU-type sorption filter (Novopor Company, Novorossiysk, Russia) for 20 min. The analytes in the filter were ultrasonically extracted with *n*-hexane (extra pure grade, Criochrom Company, St. Petersburg, Russia). The extract was loaded onto a 1−2-g silica gel L (40/100-μm particles from Ecros Company, Saint Petersburg, Russia) column. The analytes were eluted with a 4:1 mixture of hexane and methylene chloride (Criochrom Company). This solution was analyzed by the GC/MS technique.

MS and GC data were acquired using a Fisons GC 8060 gas chromatograph and an MD 800 mass spectrometer operating under standard electron ionization conditions. Chromatographic columns were (I) a Hewlett-Packard HP-5MS, 5% phenyl-substituted methylpolysiloxane, 60 m length, 0.32 mm i.d., 0.1 μm film thickness; and (II) an Alltech AT-50, 50% phenyl-substituted methylpolysiloxane phase, 30 m length, 0.25 mm i.d., and 0.25 μm film thickness. All GC runs for coeluting (co-GC) with reference materials and evaluating RI were (I) 2 min at 40 °C, then 4 °C/min to 300 °C followed by an isothermal period of 1

[8] *Drugs and Their Promotion on the Russian Market*; Informational Bank for Pharmaceutical Production and Consumption: St. Petersburg, 1994.

Table 1. Citation and Cocitation Indicators

type	subtype	definition, explanation	example of evaluation
citation		number of different documents (paper, patent, etc.) recording the compound as the number of abstracts in abstract database/journal or the number of entries in index to them counted for the compound	Information retrieval in the *Chemical Substance Index* (CSI) to CA for January−June 1987 results in 403 entries for cyclohexane, i.e., the citation rate of this hexane impurity for this period is 403.
cocitation	compound-to-compound cocitation	number of different documents mutually recording two compounds as the number of the same abstracts in abstract database/journal or the number of entries in the index to the abstracts counted for the compounds pair	Examination of entries in this issue of CSI for hexane and cyclohexane reveals 113 different abstracts belonging to both compounds, as indicated by the same abstract numbers. Hence, the cyclohexane-to-hexane cocitation indicator is 113 in January−June 1987.
	relative compound-to-compound cocitation	compound-to-compound cocitation in reference to the citation rate of one of the compounds pair	The relative cyclohexane-to-hexane cocitation for cyclohexane is 113 divided by 403 (citation rate), i.e., 0.28.
	compound-to-compound group cocitation	sum of individual cocitations of the compound with every member of the group of compounds	9-Methylanthracene was not detected in waste gas (see similar compounds group, Table 6). Its cocitation with unambiguously identified compounds is as follows: 9, with anthracene; 5, with pyrene; etc; as indicated by the count of entries in CSI to CA for January−June 1995. The sum of individual citation values (27) is 9-methylanthracene-to-unambiguously identified compounds cocitation. In calculation within the same group, a self-cocitation is canceled.
	relative compound-to-compound group cocitation	compound-to-compound group cocitation in reference to citation of an individual compound	Relative 9-methylanthracene-to-unambiguously identified compounds cocitation is 27 divided by 15 (citation rate), i.e., 1.8
cocitation	compound-to-matrix cocitation	number of different documents related to the determination of the compound in the matrix as the number of abstracts in abstract database/journal counted for the pair of the compound and the matrix	For cyclohexane, the examination of the entries in CSI to CA for the first half of 1987 reveals three different abstracts associated with determination of this hydrocarbon in air, i.e., the sample of environmental type. Hence, the indicator of cyclohexane-to-environmental matrix cocitation is 3 for the time range
	relative compound-to-matrix cocitation	compound-to-matrix cocitation in reference to the citation of a compound	This indicator for cyclohexane and environmental matrix for the first half of 1987 is 3/403 = 0.0074, where 403 is the citation. See above.

Table 2. Different Groups of Compounds

group	compds
n-hexane impurities (hi)	cyclohexane, *n*-heptane, 2-methylbutane, methylcyclopentane, 2-methylhexane, 3-methylhexane, 2-methylpentane, 3-methylpentane, *n*-pentane
compounds similar to *n*-hexane impurities (hs)	2,2-dimethylbutane, 2,3-dimethylbutane, 2,3-dimethyl-2-butene, 2,2-dimethylpentane, 2,3-dimethylpentane, 2,4-dimethylpentane, 3,3-dimethylpentane, 2,2-dimethylpropane, 2-methyl-1-pentene
naphthalene impurities (ni)	benzo[*b*]thiophene, 1H-indene, 1-methylindane, 2-methylindane, 4-methylindane, 5-methylindane, 2/4/7-methylindene/1,2-dihydronaphthalene[a], 2/5/6-methylindene[a], 4/7-methylindene/1,2-dihydronaphthalene[a]
compounds similar to naphthalene impurities (ns)	benzo[*c*]thiophene, benzylcyclopropane, 1,4-dihydronaphthalene, 4-ethenyl-1,2-dimethylbenzene, 1-ethenyl-3-ethylbenzene, *n*-heptadecane, 1-methylindene, 1-methyl-4-(1-methylethenyl) benzene, (*E*)-2-phenyl-2-butene, 1,2-propadienylbenzene, 1,1A,6,6A-tetrahydrocycloprop[*A*]indene
pharmaceuticals (pharm)	ampicillin, aspirin, cefalexin, cefazolin, diclofenac, gentamicin, ibuprofen, indomethacin, penicillin V, piroxicam

[a] Ambiguous identification for the corresponding peak. These hydrocarbons are conditionally considered as one compound, a mean value of citation and cocitation being taken.

min; (II) 5 min at 30 °C, then 5 °C/min to 250 °C followed by an isothermal period of 90 min. Injection of a 1-μL solution, not less than two replicates for every analysis, was performed.

MS and GC subcriteria and integral criteria for identification are given in Tables 4 and 5. The candidates (hypotheses) for identification were the compounds covered by MS libraries. These compounds were divided into three groups, unambiguously or ambiguously identified, and unidentified (similar) compounds, as tested by integral identification criteria (Table 6). In counting citations and cocitations, a group of isomers of these compounds without available reference data (mass spectra, RI), that is, being outside the identification procedure, was incorporated. The members of this group were taken from the same issue of CSI that was used for data retrieval (see above). Because of a considerable number of PAH covered with mass spectra and RI data, the absence of reference data means that outsiders were

Table 3. Different Matrixes and Corresponding Key Words in CSI

matrix	words in "analytical" entries[a]
biomedical	bacteria, bile, blood, breath, saliva, tissue, urine
coal	coal, coal gases, coal tar
environmental	aerosols, air, environmental, dust, gases, sediments, soil, water
petroleum	gasoline, naphtha, oil, petroleum, petroleum products
pharmaceutical	capsules, dosage forms, drug, pharmaceuticals, suppositories, suspensions, tablets, transdermal systems
waste	combustion gas, exhaust, refuse, smoke, waste

[a] These entries contain words as follows: "analysis", "determination", "assay", etc.

Table 4. Subcriteria for PAH Identification in Waste Gas

technique	subcriterion
MS	fit to the reference mass spectrum in m/z value and the order of abundance for 3−5 most abundant peaks[a]
co-GC	coelution of the compound under identification together with the reference compound
RI-GC	fall of RI within ± 2 i.u. about reference data [9,11,12]

[a] It corresponds to reverse fit value of the experimental spectrum to the reference value for identified and unidentified compounds being ≥800 and <700, respectively.

Table 5. Criteria for PAH Identification

identification	criterion
unambiguous	unique MS + co-GC or MS + RI or MS identification for the chromatographic peak
ambiguous	plural MS (with no available reference RI) or MS + RI identification for the peak
no (similar compounds)	negative MS or RI test or unambiguous identification of the other substance based on two subcriteria for this chromatographic peak
outside the procedure	isomers of identified/unidentified compound referred to in CA for 1995 and with no GC and MS available reference data

detected rarely, if at all, in waste and environmental samples; therefore, they can be considered as another kind of unidentified compounds.

For MS identification, Wiley 6 and two NIST reference mass spectra libraries consisting of 223 615, 62 235, and 129 136 entries, respectively, were used. A mass value and an abundance of mass peaks, as well as a forward and a reverse fit value of an experimental mass spectrum to a reference spectrum as calculated by MassLab (Fisons) and AMDIS (NIST) software, were taken into consideration (Table 4).

The experimental Lee RI applied for identification were obtained by the use of naphthalene, phenanthrene, and chrysene (both columns), dibenz[*ah*]anthracene (column I) and benzo[*a*]pyrene (column II) as the reference PAH, with RI[9−11] equal to 200, 300, 400, 495, and 452, respectively. The RI[9,11,12] for other PAH were taken as the reference data. Most comparisons in RI were carried out between our data obtained for column I and the data for similar SE-52 and DB-5 phases.[9,11] In addition, the RI of the sulfur compounds for column II and DB-17 (ref 12) as another pair of similar phases, were matched.

RESULTS AND DISCUSSION

Validation of the Procedure. The identification procedure based on literature citation/cocitation will be efficient and validated, provided a significant difference between the analytes and other compounds in these statistical values is evident. The validation procedure that needs samples of known composition involves the evaluation of a given type of statistical rates (Figure 1). The GC/MS analysis of *n*-hexane, naphthalene, and waste gas provided necessary analytical data. The significance of citation and cocitation data will be discussed step by step for each case and each statistical indicator.

Impurities in Hexane and Naphthalene: Citation. Figure 4 shows that the analyte citation is higher than that for similar compounds. However, the difference between the two groups as expressed by the median value; the significance, α; etc., is lower in the case of naphthalene (Figure 4b). According to this, the impurity citation threshold is surmounted by most similar substances connected with naphthalene impurities (Table 7). Unlike the hexane impurities and similar compounds, the two groups connected with naphthalene are not fairly differentiated by the citation indicator. Hence, the selection of candidates for identification based on citation value may lead to a rather large number of additional candidates for identification. An evident conclusion can be made that citation alone is not the only versatile parameter that can be used to set up advanced identification hypotheses. Then a comparison between compounds of different chemical natures in citation data seems useless.

Compound-to-Compound Cocitation. Figure 5 and Table 7 show that relative cocitation with hexane is very close to zero for impurities in naphthalene and compounds similar to them (predominantly aromatic hydrocarbons) and *absolutely foreign* compounds as pharmaceuticals. The only compound from these groups is cited above the threshold established for the hexane impurities. This indicator does not differentiate efficiently between hexane impurities and similar compounds. However, rather highly cited similar compounds connected with hexane are not highly cocited and vice versa. Eventually, no compound from this group is above the dual citation and relative cocitation threshold (Table 7).

In 1987, relative cocitation with naphthalene was rather high for only its impurities. A slight difference in this indicator is observed between the naphthalene impurities and three other groups in 1997 (Figure 5, Table 7). However, the percent of similar compounds of the naphthalene group above the threshold was lower than this value in the case of the citation indicator. Unlike the compound connected with hexane, relative cocitation with naphthalene differentiates reasonably well between its impurities and similar compounds.

[9] Lee, M. L.; Vassilaros, D. L.; White, C. M.; Novotny, M. *Anal. Chem.* **1979**, *51*, 768−773.

[10] Vassilaros, D. L.; Kong, R. C.; Later, D. W.; Lee, M. L. *J. Chromatogr.* **1982**, *252*, 1−20.

[11] Rostad, C. E.; Pereira, W. E. *J. High Resolut. Chromatogr., Chromatogr. Commun.* **1986**, *9*, 328−334.

[12] Mössner, S. G.; De Alda, M. J. L.; Sander, L. C.; Lee, M. L.; Wise, S. A. *J. Chromatogr. A* **1999**, *841*, 207−228.

Table 6. PAH Identification

formula	no. GC peaks	group of compounds: unambiguously identified	group of compounds: other	no. outsiders
$C_{12}H_8S$	1	dibenzothiophene	similar: naphtho[2,3-*b*]thiophene, 1-thiaphenalene	4
$C_{13}H_{10}$	1	9H-fluorene	similar: 1H-phenalene	4
$C_{14}H_{18}S$	1	acenaphtho[1,2-*c*]thiophene		1
$C_{14}H_{10}$	2	anthracene, phenantrene	similar: benz[*a*]azulene, diphenylethyne	5
$C_{15}H_{10}$	1	4H-cyclopenta[*def*]phenanthrene	similar: 4H-cyclobuta[*jk*]phenanthrene	
$C_{15}H_{12}$	4	1-methylphenanthrene	ambiguously identified: 1a,9b-dihydro-1H-cyclopropa[*a*]anthracene, 1a,9b-dihydro-1H-cyclopropa[*i*]phenathrene, 1,3-diphenylpropyne, 1-methylanthracene, 2-methylanthracene, 2-methylphenanthrene, 3-methylphenanthrene, 4-methylphenanthrene, 9-methylphenanthrene, naphtho[2,3-*b*]norbornadiene, 1-phenyl-1H-indene, 2-phenyl-1H-indene, 3-phenyl-1H-indene	16
			similar: 9-methylanthracene	16
$C_{16}H_{10}$	2	fluoranthene, pyrene	similar: 1,4-diphenylbutadiyne	8
$C_{16}H_{10}O$	7		ambiguously identified: 2-aceanthrenone, benzo[*k*l]xanthene, anthra[1,2-*b*]furan, anthra[1,2-*c*]furan, benzo[*b*]naphtho[2,1-*d*]furan, benzo[*b*]naphtho[2,3-*d*]furan, 1-hydroxypyrene, phenanthro[1,2-*c*]-furan, phenanthro[3,4-*c*]furan, phenanthro[9,10-*b*]furan	
$C_{16}H_{10}S$	3	benzo[*b*]naphtho[1,2-*d*]thiophene, benzo[*b*]naphtho[2,1-*d*]thiophene, benzo[*b*]naphtho[2,3-*d*]thiophene		3
$C_{16}H_{12}$	2	1-phenylnaphthalene	ambiguously identified: dibenzo[*a*,*e*]cyclooctene, 4,5-dihydroacephenanthrylene, 5,10-dihydroindeno[2,1-*a*]indene, 1,4-dihydro-1,4-ethenoanthracene, 1,9-dihydroperene, 4,5-dihydroperene, 1,4-diphenyl-1-buten-3-yne, 9-ethenylanthracene, 5-methylene-5H-dibenzo[*a*,*d*]cycloheptene, 1-(phenylmethylene)-1H-indene, 2-phenylnaphthalene, 4b,4c,8b,8c-tetrahydrocyclobuta[1″,2′:3,4;3″,4″:3′,4′]dicyclobuta[1,2:1′,2′]dibenzene, trycyclo[8.2.2.2(4,7)]hexadeca-[2,4,6,8,10,12,13,15]-octaene,	
$C_{17}H_{12}$	9	11H-benzo[*a*]fluorene, 1-methylpyrene	ambiguously identified: 7H-benz[*de*]anthracene, 7H-benzo[*c*]fluorene, 11H-benzo[*b*]fluorene, 2-methylpyrene, 4-methylpyrene	15
$C_{18}H_{10}$	1	benzo[*ghi*]fluoranthene	similar: cyclopenta[*cd*]pyrene, 1,8-diethynylanthracene	19
$C_{18}H_{12}$	3	benza[*a*]anthracene, benzo[*c*]phenanthrene, chrysene	similar: 3,4-dihydrocyclopenta[cd]pyrene, 1-(1H-indene-1-ylidene)-1H-indene, naphthacene, triphenylene	8
$C_{20}H_{12}$	7	benzo[*e*]pyrene, benzo[*a*]pyrene, perylene	ambiguously identified: benz[*a*]aceanthrelene, benz[*d*]aceanthrelene, benz[*j*]aceanthrelene, benz[*k*]aceanthrelene, benzo[*e*]fluoranthene, benzo[*j*]fluoranthene, benzo[*k*]fluoranthene	3
$C_{22}H_{12}$	6	benzo[*ghi*]perylene, dibenzo[*def*,*mno*]chrysene, indeno[1,2,3-*cd*]pyrene	ambiguously identified: indeno[1,2,3-*cd*]fluoranthene	21

Pharmaceuticals are very seldom cocited with both classes of hydrocarbons (Figure 5, Table 7) and, therefore, are efficiently differentiated from other groups. Thus, the citation test alone is efficient in the case of a large difference in structure, origin, or use of compounds compared, this test having varying significance for rather similar chemical structures.

Compound-to-Matrix Cocitation. Table 8 shows a significant difference between groups considered in the fraction of compound cocited with common types of samples in analytical papers. Hexane impurities, alkanes, and cycloalcanes demonstrate more frequent cocitation with petroleum and environmental matrixes in comparison to naphthalene impurities and pharmaceuticals. The latter can be differentiated from naphthalene compounds and all other compounds by cocitation with biomedical and pharmaceutical matrixes, respectively (Table 8). In some cases, the difference between hexane impurities and similar compounds in cocitation is pronounced and that for substances related to naphthalene is less significant (Table 8).

PAH in Waste Gas. One-half of the main peaks refer to unambiguously identified compounds (see Table 6). The rest, 25 peaks, belong to some of 49 ambiguously identified hydrocarbons (Table 6).

For unambiguously identified compounds, citation, relative cocitation with other reliably identified substances, and relative cocitation with environmental- and waste-type matrixes is far greater than these three values for three other groups of compounds. In a comparison between unambiguously identified compounds and other groups, all of the α values were <0.01. Less than one-half of the compounds from these three groups exceeded the citation threshold established for unambiguously identified compounds (Table 9). Only $\sim$10−30% of the compounds, except those unambiguously identified, are above the compound-to-compound and compound-to-matrix cocitation thresholds. The selection by dual citation and cocitation diminishes this fraction still further, and only less than 15% of the other compounds were above the triple threshold (Table 9).

The group of ambiguously identified compounds with intermediate *identification degree* and the outsiders do not closely resemble any group connected with hexane/naphthalene. The former is assumed to have some analogy to unambiguously identified and unidentified compounds, since it contains both unresolved groups. Here, ambiguously identified compounds behave like the unidentified (similar) compounds (Table 9). As

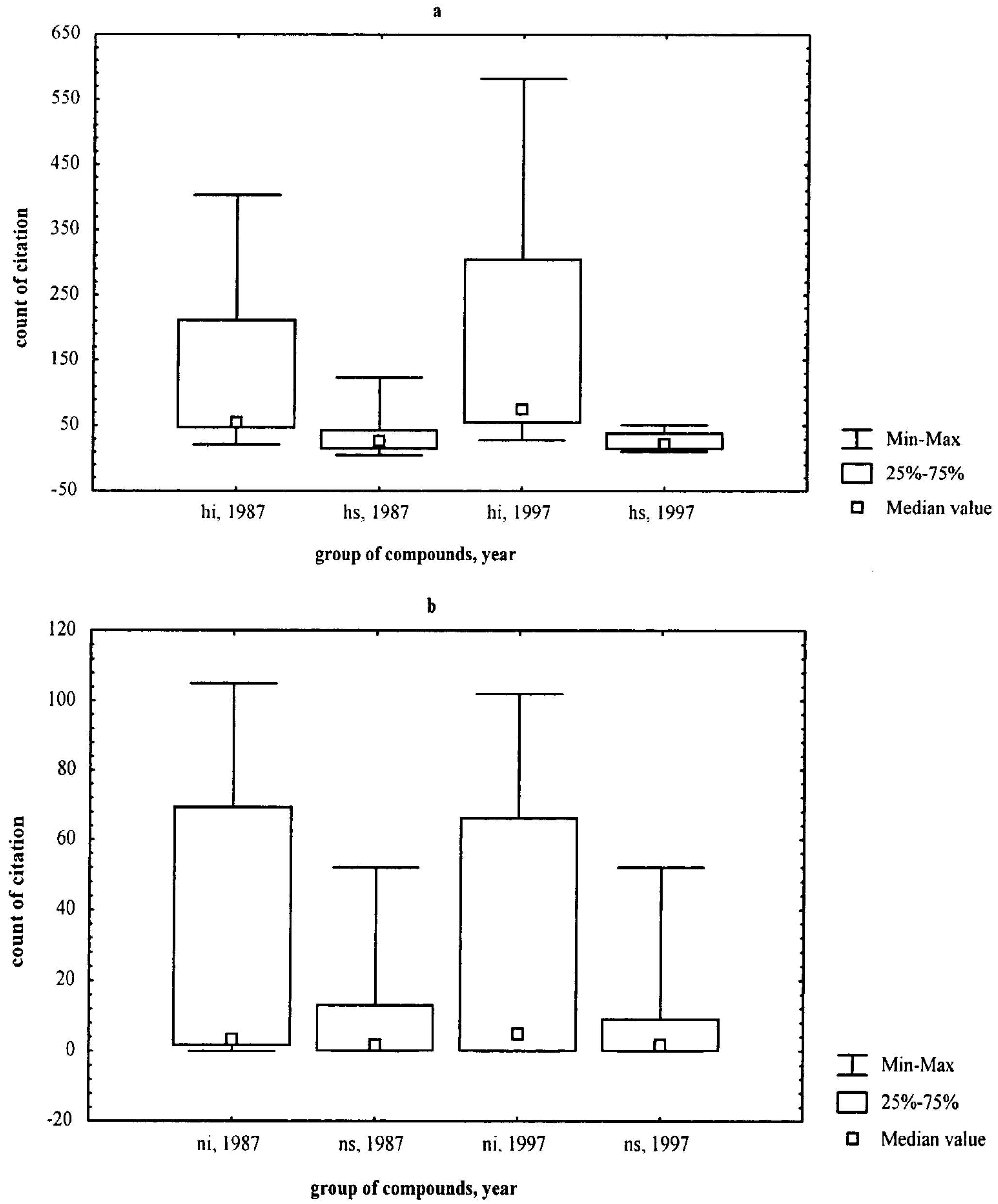

Figure 4. (a) Distribution of impurities in *n*-hexane and similar substances in citation counts in 1987 and 1997. The notations for the compound groups are given in Table 2. The significance levels, α, of the difference in count between the groups compared for 1987 and 1997 are 0.06 and 0.03, respectively. The low quartile values for impurities are 45 and 55, respectively. (b) Distribution of impurities in naphthalene and similar substances in citation counts in 1987 and 1997. The significance levels, α, of the difference in count between the groups compared for 1987 and 1997 are 0.15 and 0.11, respectively. The low quartile values for impurities are 2 and 0.3, respectively.

for outsiders, only a few compounds have rather high citation/cocitation rates and are worth inclusion in MS libraries.

General Citation and Cocitation Regularities. This study of citation and cocitation behavior of analytes in different samples and of compounds similar/dissimilar to them in formula, structure, mass spectra, chromatographic parameters, origin, or use makes it possible to establish some regularities. First, the analytes unambiguously detected in the samples are more highly cited than similar compounds; however, the difference in citation, for example, in the case of naphthalene, cannot be very significant (Table 7).

Second, compound-to-compound cocitation with the main constituent of the sample (*n*-hexane, naphthalene) or with the group of main constituents (the detected PAH) is higher for the detected compounds than for similar and the dissimilar or foreign compounds, that is, pharmaceuticals. Again, the difference in cocitation may be insignificant for similar compounds (belonging to *n*-hexane, Table 7). However, combined selection by citation and cocitation results in reasonable differentiation between identified and similar compounds. It is promising that cocitation differentiates between dissimilar compounds, that is, (1) alkanes and aromatic hydrocarbons and (2) both hydrocarbon groups and pharmaceuticals. The only exception is associated with compound-to-naphthalene cocitation in 1997 (Table 7). This can be avoided if the relationship of cocitation counts, for example, (compound-to-hexane)-to-(compound-to-naphthalene) cocitation is taken into account;[4] however, the time range for counting citation and cocitation should be properly chosen.

Table 7. Fraction of Compounds with Above-Threshold Values of Indicators for Different Groups[a]

			Cocitation[c] with				Citation[b] and cocitation[c] with			
	citation[b]		hexane		naphthalene		hexane		naphthalene	
group	1987	1997	1987	1997	1987	1997	1987	1997	1987	1997
impurities in *n*-hexane	78	78	78	78	0	33	56	56		
compounds similar to impurities	22	0	67	89	0	45	0	0		
impurities in naphthalene	82	82	11	0	82	73			55	55
compounds similar to impurities	64	73	0	0	9	27			9	27
pharmaceuticals			0	0	0	0				

[a] Percent. [b] Different citation thresholds established for the compounds of hexane and naphthalene groups. [c] Different cocitation thresholds established for cocitation with hexane and naphthalene. Since the low quartile is zero, the lowest nonzero value is conditionally taken as the threshold for cocitation with naphthalene in 1997.

Table 8. Fraction of Compounds Cocited with Different Matrixes for Different Groups Connected with *n*-Hexane and Naphthalene[a]

	biomedical		coal		environmental		petroleum		pharmaceutical	
group	1987	1997	1987	1997	1987	1997	1987	1997	1987	1997
impurities in *n*-hexane	100	89	56	0	100	100	100	78	22	0
compounds similar to impurities	0	11	0	0	22	56	56	33	0	0
impurities in naphthalene	9	18	27	18	27	27	18	9	0	0
compounds similar to impurities	0	9	9	0	9	9	0	0	0	0
pharmaceuticals	90	70	0	0	0	0	0	0	90	70

[a] Percent.

Table 9. PAH Fraction with Above-Threshold Values of Indicators for Different Groups[a]

		cocitation with			
group	citation	unambiguously identified compound	environmental and waste matrixes	citation and cocitation with unambiguously identified compound	citation and both cocitations
unambiguously identified	80	76	76	64	44
ambiguously identified	20	30	14	18	10
similar	43	21	21	21	14
outside the identification procedure	2	20	9	2	0

[a] Percent.

Third, the identified analytes differ from other compounds by higher cocitation with the names of the corresponding samples. Primarily, this refers to related matrixes, that is, the ground substances as hexane and naphthalene for impurities, drug dosage (tablets, etc.) for pharmaceuticals, environmental and waste samples for PAH. Compound-to-matrix cocitation can be used as an individual selection parameter or in combination with other indicators.

Identification with the Use of Citation/Cocitation Data. The revealed regularities prove that high citation/cocitation of a compound estimated by an appropriate method corresponds to its probable presence in a sample. This leads to the algorithm for identifying the unknown on the basis of these statistical values (Figure 3). The suggested strategy for the use of prior information resembles a mode of expert thinking and formalizes this essentially latent human operation/decision. Thus, it is common for an expert to take into account abundance (popularity) of different compounds when advancing different assumptions associated with sample compositions and to explain the simultaneous presence of some substances in a sample on the basis of their similarity in formula, structure, origin, properties, etc.

Citation can be regarded as an indicator of abundance, popularity, and spreading for a compound. Cocitation can be considered as an indicator of chemical similarity for a pair of compounds. Unlike a personal expertise, which is not free of bias and lacunas, statistical processing of prior data can provide a total space of identification hypotheses covering a great body of data from different fields. Then the treatment of prior data by the suggested method does not require comprehensive chemical knowledge and could be performed by less experienced personnel. It can diminish identification uncertainty in routine determinations. This approach to prior data processing is also valuable for an experienced analyst determining an unknown in a sample with unfamiliar composition.

CONCLUSIONS

An identification procedure containing special processing of prior data extracted from general chemical literature/database, is suggested. To validate the procedure, two assumptions are proved in this study: First, analytes as compounds reliably detected in samples are more frequently cited in the literature than absent chemically similar compounds. Second, analytes are

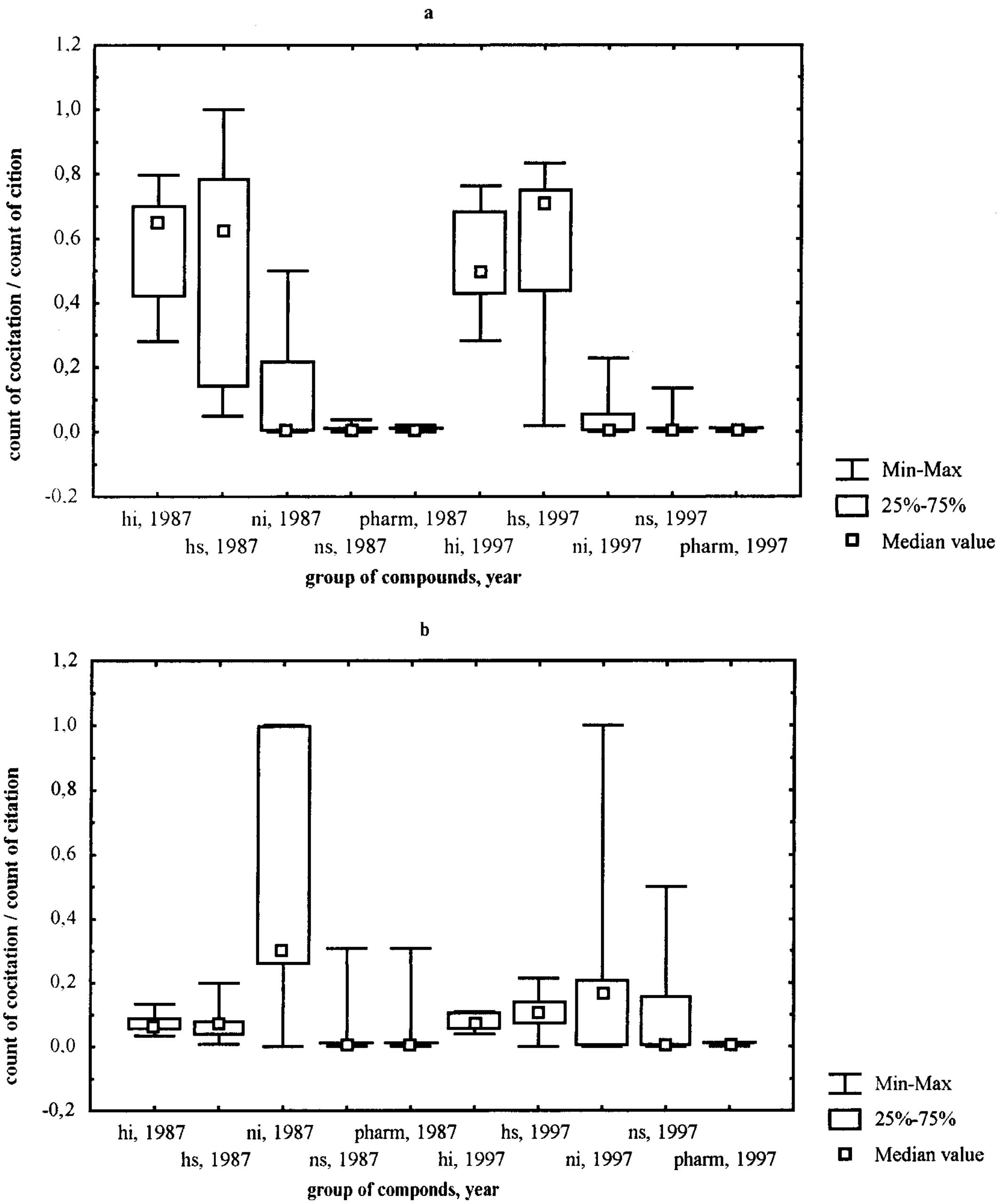

Figure 5. (a) Distribution of impurities, similar substances, and pharmaceuticals in relative compound-to-hexane cocitation in 1987 and 1997. The notations for compound groups are given in Table 1. The significance levels, α, of the difference in this indicator between hexane impurities and compounds being similar to them, are 0.75 and 0.60 in 1987 and 1997, respectively. The α values for the comparison between hexane impurities and other groups are ≤0.0001. The low quartile values for hexane impurities are 0.42 and 0.43, respectively. (b) Distribution of impurities, similar substances, and pharmaceuticals in relative compound-to-naphthalene cocitation in 1987 and 1997. The significance level, α, of the difference in this indicator between naphthalene impurities and other compound is ≤0.003 in 1987 and 0.1 ÷ 0.3 (with the exception of pharmaceuticals being 0.003) in 1997. The low quartile values for the naphthalene impurities are 0.26 and 0.00, respectively.

more frequently cocited with each other and corresponding matrixes than nondetected similar or dissimilar compounds. These rules can be used to set up and cancel hypotheses on the nature of unknown analytes in complex mixtures. The difference between the compared compounds in citation and cocitation count may be significant, for example, in the case of dissimilarity in structure, properties, use, etc., or insignificant. The latter case results in more candidates for identification as hypotheses being tested. The redundant hypotheses can be removed by counting more than one statistical indicator, for example, dual citation and cocitation. Screening of identification hypotheses set up on the basis of a citation-type prior data decreases identification uncertainty. Statistical processing of chemical literature units as reflected in abstracts journals/databases can be performed by workers with less expertise in chemistry, resulting in a more reliable routine identification. An experienced analyst can combine citation/cocitation data with a personal knowledge when identifying an unknown outside his regular field.

Received for review May 30, 2001. Accepted December 5, 2001.

AC010611P