Evaluation of Automatically Generated Substructure Identification Rules from Tandem Mass Spectra

K. J. Hart,* A. P. Wade,‡ B. D. Nourse,* and C. G. Enke
Department of Chemistry, Michigan State University, East Lansing, Michigan, USA

Substructure identification rules for phenothiazine and barbiturate substructures were generated by using a new version of the Method for Analyzing Patterns in Spectra (MAPS) software. This software uses tandem mass spectra and known substructure content of reference compounds to provide "feature-combination" rules. A feature-combination is a series of tandem mass spectral features which are completely unique to compounds containing a specified substructure. The current reference databases contain over 11,000 daughter spectra of 100 compounds acquired at two different collision gas pressures (i.e., single- and multiple-collision conditions). The results of rule evaluation procedures are presented and include a comparison of the spectral features developed in rule generation to those identified in documented fragmentation pathways of the indicated substructure. Two potential sources of error due to spectral feature and substructure "cross-correlation" were identified. If errors occur, they can be detected by calculating cross-correlation coefficients and edited from the rules. A beneficial cross-correlation involving feature-combinations was also discovered. The rules obtained by using single- and multiple-collision data were further evaluated by applying them to tandem mass spectra of 20 test compounds (compounds not in the reference database). The results of these evaluations give a good indication of the utility of the rules for use in an automated structure elucidation system for tandem mass spectrometry data. *(J Am Soc Mass Spectrom 1992, 3, 169–180)*

A detailed evaluation of two rules generated for the Automated Chemical Structure Elucidation System (ACES) [1] was undertaken to illustrate that the rules generated by the Method for Analyzing Patterns in Spectra (MAPS) software (discussed in the preceding article [2]) make chemical sense, and that they are effective in predicting the presence of substructures in training set compounds and unknowns. A set of candidate structures is generated by the other ACES program modules using the presence of substructures identified by the MAPS rules for elemental and structural constraints [3–5]. Thus the reliability and recall of the MAPS substructure identification rules greatly affects the utility of the ACES system.

Three different rule evaluations were performed on the MAPS rules generated from a reference database of 100 compounds. A large number of the reference compounds used were regulated drugs because an important area of application of the tandem mass spectrometry (MS/MS) technique is in the analysis of pharmaceuticals. These analyses involve the screening of formulations for active drug components, impurities, and synthetic markers; structural analyses of new drugs; and quantitation of drug metabolites in biological fluids [6]. In that physiologically active drugs often have similar structures, MS/MS can be used to establish the structures of variants of more commonly encountered drugs [6]. MAPS rules for **phenothiazine** and **barbiturate** substructures are examined here in detail to illustrate the utility of the new MAPS software for generating reliable substructure identification rules for these compound classes. Note that names in bold type indicate a substructure definition such as that shown in Figure 1 in the preceding article [2] (i.e., **phenothiazine**: a ring system composed of 12 carbons, one nitrogen, one sulfur, bonded together as shown in the figure and no specific substituents or substitution pattern). In contrast, the compound "phenothiazine" has a molecular formula of $C_{12}H_9NS$ and all of the "free valences" on the structure shown in the figure are occupied by hydrogens.

One of the rule evaluations performed for this

*Present address: Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6365.
‡Present address: Chemistry Department, University of British Columbia, Vancouver, BC, Canada, V6T 1Y6.
Address reprint requests to Chris G. Enke, Department of Chemistry, Michigan State University, East Lansing, MI 48824.

1044-0305/92/$5.00

Received May 10, 1990
Revised September 25, 1991
Accepted September 25, 1991

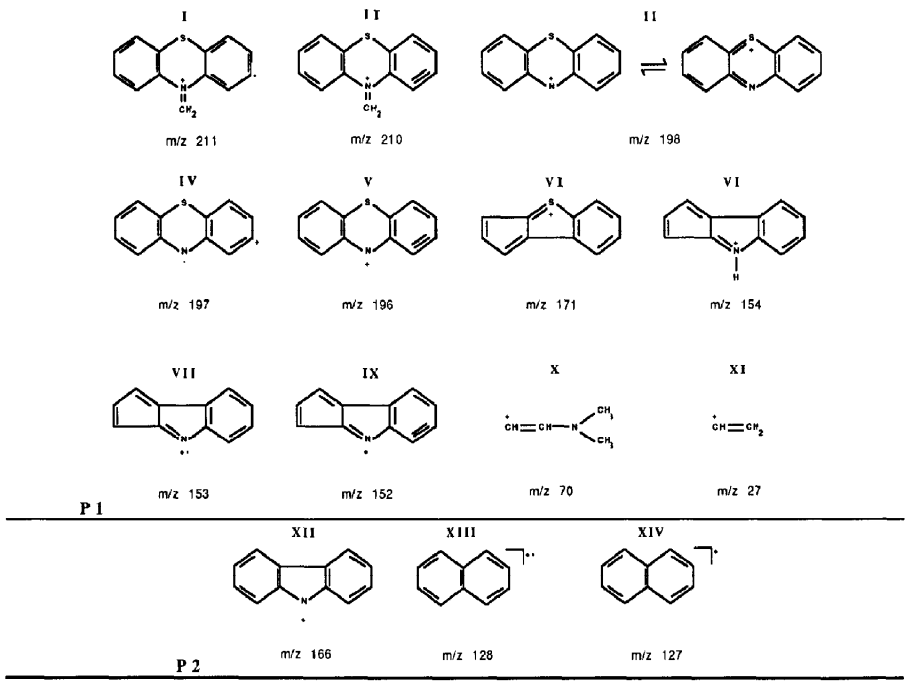


Figure 1. Structures for several fragment ions of phenothiazine compounds (P1, single-collision conditions/DB 1 data; P2, multiple-collision conditions/DB 2 data). Adapted from refs 9 and 11.

study used rule reliability and recall values, which are measures of rule accuracy and frequency of prediction, with respect to the reference database. The reliability and recall obtained for a rule depend on the initial feature uniqueness (Ui) and initial feature correlation (Ci) program parameters used in the generation of the rule. However, given that the reference database contains only 100 compounds, these evaluative parameters do not necessarily reflect the reliability and recall of a rule when the rule is used on unknowns. Thus, the reliability and recall of the rules with respect to the reference database were used mainly in the optimization of the new MAPS program. This discussion focuses on the remaining two evaluation procedures which test the accuracy of the MAPS rules. The first of these evaluations involved the comparison of the tandem mass spectral features contained in several MAPS rules to those contained in documented fragmentation pathways for the corresponding substructures. This evaluation confirms that the MAPS software is indeed selecting relevant features from within the MS/MS data space for inclusion in the substructure identification rules. The remaining evaluation used 20 "test compounds" (i.e., compounds not in the reference database) to estimate the reliability and recall of the MAPS rules for identifying substructures in unknown compounds.

Experimental

MS/MS fragmentation maps using electron ionization (EI) for a variety of standard compounds were acquired using a Finnigan-Mat (San Jose, CA) TSQ-70 mass spectrometer to build two reference databases (DB): DB 1 using single-collision conditions (0.4 mtorr Ar) and DB 2 using multiple-collision conditions (1.2 mtorr Ar). There are two ways of monitoring collision gas pressure on the TSQ-70. The first method measures collision gas pressure directly by a convectron gauge plumbed directly into the second quadrupole casing. The second method indirectly measures collision gas pressure by leakage into the manifold region of the instrument where pressure is monitored by an ion gauge. In practice, it was easier to set the collision gas pressure to a specific pressure reading on the manifold ion gauge (2.0×10^{-6} and 6.0×10^{-6} torr for the two pressures listed for DB 1 and DB 2) owing to the rapid and relatively large fluctuations in the convectron gauge readout. Because the readings from these gauges are available through the instrument

software, an average value of the convectron gauge was easily calculated once the collision gas pressure was set (by using the manifold ion gauge reading). The collision energy was fixed at 30 eV (E_{LAB}). A fragmentation map consists of the primary mass spectrum and daughter spectra for each ion that had an intensity > 1% of the base peak in the primary mass spectrum. Daughter spectra were not collected for peaks < 1% of the base peak to decrease the scan time required to take a complete map and to avoid unsuitable daughter spectra due to insufficient signal. Data systems that allow for automated data collection for commercial triple quadrupole mass spectrometry instruments have become increasingly available over the last few years. This important development has spurred the acquisition of tandem mass spectra for inclusion in the MAPS reference databases.

It should be noted that each reference compound has an assortment of substructures so it is probable that one compound will contribute spectral features to several different substructure identification rules. The average number of substructures found in the reference compounds was 11 substructures. A total of 11,598 spectra of 100 compounds were acquired under single- and multiple-collision conditions of collision gas pressure.

Four experiments were performed to collect the MS/MS data of each standard. All of the standards used were solids and were introduced into the ion source via a direct insertion probe. The samples were volatilized by heating the probe tip according to a predefined temperature program. The first experiment determined the total ion current generated versus probe temperature. The results of this experiment were used to create a probe temperature program to volatilize the sample at a relatively constant rate. A second experiment was performed to collect the primary mass spectrum for the sample. An instument control procedure was used to set the scan mode, mass limits, scan rate, and multiplier voltage, calculate which ions are at least 1% of the base peak and append the masses of those ions into a user list. The procedure also estimates the total scan time for collection of daughter spectra for all of the masses in the user list. The third experiment used another instrument control procedure to set the instrumental parameters for each daughter spectrum. The ions selected in Q1 (parent ion) are read from the user list created in the previous experiment. The fourth experiment repeated the spectrum collection regimen with a higher collision gas pressure (set manually). Separate runs were made in these last two experiments to allow the collision gas pressure time to equilibrate. Instrument acquisition time for the daughter spectra for one pressure varied from 0.5 to 3 min. Of the 100 compounds that were included in the reference databases, 79 were obtained in Theta-Kits from the Theta Corporation (Newton Square, PA). The other 21 standards used were obtained from General Motors Research. All of these standards were phenols. The reference names, compound names, and Chemical Abstracts Service (CAS) (American Chemical Society), numbers for the standards used to create the reference databases are available as supplementary material (Table S1). The compound name, molecular weight, molecular formula, and CAS number for each of the test compounds are also available (Table S2).

Results and Discussion

Comparison of MAPS Rules to Documented Fragmentation Pathways

Several MAPS rules were examined to determine the correspondence of the spectral features found in the MAPS rules with those from documented fragmentation pathways. These pathways have been discovered through the use of high resolution mass spectrometry to determine the elemental composition of the various fragment ions and isotope labeling to assist in determining the mechanism of the fragmentation. MS/MS has also proven to be an effective tool in probing the ion chemistry of a variety of compounds [6]. The following sections provide a comparison of the spectral features found in MAPS rules for several substructures with those identified in the literature. The effect of varying the MAPS program parameters on rule content is explored and an additional evaluation parameter, cross-correlation, is introduced. The differences observed in the rules generated under single- and multiple-collision conditions are also reported.

Phenothiazine. There are 13 compounds in the reference database (DB 1) that contain the **phenothiazine** substructure shown in Figure 1 of the preceding article [2]. These compounds are clinically useful as antipsychotic drugs [7]. Several studies of the EI and chemical ionization (CI) mass spectra of these compounds have been published. One of these studies examined the primary mass spectra of newly synthesized 2-substituted 10-*N*-(aminoacyl) phenothiazines which possess antigastric-antiulcer and antidepressant activity [8]. Another of these studies used metastable ions, exact mass measurements, and deuterated derivatives to investigate the fragmentation pathways of phenothiazines [9]. The sites of protonation in phenothiazines have been determined using methane and ammonia CI mass spectra with high resolution mass measurements to confirm the empirical formulas of fragment ions [10].

An early study of phenothiazine derivatives used high resolution mass spectrometry to determine the fragmentation pathways in preparation of the analysis of phenothiazine metabolites [11]. The investigators in this study divided the fragmentations of phenothiazine compounds into three groups: (1) fragments representing the side chain, (2) fragments representing the intact phenothiazine ring system with part of the

side chain attached, and (3) fragments representing a partially fragmented ring system. Tandem mass spectral features due to all three types of fragmentations have been observed in the MAPS rules for the **phenothiazine** substructure, depending on the U and C values used in generating the rules (see subsection 2). The following discussion compares the spectral features contained in the MAPS rules for **phenothiazine** to those which were identified in the aforementioned studies.

The initial features obtained for the **phenothiazine** substructure (Ui = 40% and Ci = 70%) are shown in Table 1. The MAPS software provides an option to mask features with mass-to-charge ratios greater than the nominal mass of the substructure. This option was not utilized in this example. Thus, there are two spectral features (e.g., D 211.0) in this rule with masses larger than the nominal mass of **phenothiazine** (i.e., 198 u). Structures of several fragment ions of **phenothiazine** and one substituent fragment ion are provided in Figure 1. Pertinent fragmentation pathways for the **phenothiazine** substructure were gleaned from refs 9 and 11. These pathways are listed in Table 2. Note that the **phenothiazine** "PD" features listed in Table 1 correspond directly to the key fragmentation pathways outlined in Table 2. The feature number from Table 1 is provided in brackets for each of the corresponding fragmentations shown in Table 2. These results demonstrate the ability of the MAPS rule generation software to select spectral features arising from the indicated substructure from within the MS/MS data space for use in generating substructure identification rules. The feature-combination rule shown in Figure 2 was generated using many of the features shown in Table 1 except those with a mass-to-charge ratio greater than the nominal mass of the substructure. The reliability and recall of this rule were 100% with respect to the reference database.

It is important to realize that the MAPS software does not utilize ion intensity in the rules. The only purpose that ion intensity plays in the software is in the selection of parent ions for daughter spectra (i.e., daughter spectra are acquired only for those parent ions that are at least 1% of the base peak in the primary mass spectrum of a compound). Relative parent ion intensity should not be misconstrued with relative importance for identifying a substructure. In fact, relative ion intensities may be misleading for substructure identification because the most diagnostic ions for a particular substructure may be among the weakest ions in the mass spectrum of some compounds. In fact, the mass spectra of phenothiazine compounds are often dominated by side chain fragmentations that do nothing to characterize the compound as a phenothiazine (e.g., m/z 58 or m/z 72 is often the base peak and a significant portion of the total ion current for most of the phenothiazines studied here). It is the daughter spectra of the relatively weak m/z 198, m/z 197, and m/z 196 parent ions (reflecting the intact phenothiazine ring system) that are most important in characterizing the presence of the phenothiazine substructure. No ions are discounted by the MAPS software unless the intensity of the ions fall below a specified threshold (usually 1%). If intense ions are not found in a rule, it is because they have little diagnostic value for the specified substructure (i.e., too many other compounds that do not have the specified substructure display the ions in their spectra). One possible disadvantage of this method may surface in the analysis of trace quantities of material. Usually the most intense ions possible are utilized for trace analysis to maximize the limit of detection for a chemical species. Thus, there is a sensitivity/specificity tradeoff in selecting a low parent ion threshold. Of course, it is possible to select a high threshold for parent ion selection and to utilize a

Table 1. Tandem mass spectral features selected for **phenothiazine**

Feature	[U, C]	Feature no.
D (211.0)	[40, 76]	{F1}
D (209.0)	[40, 76]	{F2}
D (198.0)	[41, 92]	{F3}
PD (198.0 → 171.0)	[76, 76]	{F4}
PD (198.0 → 154.0)	[92, 92]	{F5}
PD (197.0 → 196.0)	[68, 84]	{F6}
PD (197.0 → 153.0)	[90, 76]	{F7}
PD (196.0 → 152.0)	[90, 76]	{F8}
PD (70.0 → 27.0)	[45, 84]	{F9}

Ui = 40%, Ci = 70%; mass filter disabled.

Table 2. Correlation of fragmentation pathways and rule features for **phenothiazine**

Fragmentation path ($m/z \rightarrow m/z$)	Feature no.	Corresponding neutral loss
198 (III) → 171 (IV)	{F4}	(loss of HCN—27 u)
198 (III) → 154 (VII)	{F5}	(loss of CS—44 u)
197 (IV) → 196 (V)	{F6}	(loss of H—1 u)
197 (IV) → 153 (VIII)	{F7}	(loss of CS—44 u)
196 (V) → 152 (IX)	{F8}	(loss of CS—44 u)
70 (X) → 27 (XI)	{F9}	(loss of C_2H_5N—43 u)

```
IF    " PD (198.0 154.0)  and  D (198.0)              [100,92] "
OR    " PD (197.0 196.0)  and  PD (198.0 154.0)       [100,76] "
OR    " PD (197.0 196.0)  and  PD (197.0 153.0)       [100,76] "
OR    " PD (197.0 196.0)  and  PD (196.0 152.0)       [100,76] "
OR    " PD (198.0 171.0)  and  PD (198.0 154.0)       [100,76] "
OR    " PD (197.0 153.0)  and  PD (196.0 152.0)       [100,76] "

THEN substructure phenothiazine is present.

REL = 100% and REC = 100% with respect to reference database

        Ui = 40%, Ci = 70%, Cc = 70%, mass filter enabled
```

Figure 2. The MAPS (v.3) feature-combination rule obtained for the **phenothiazine** substructure.

database of CI daughter spectra to maximize the "sensitivity" of the MAPS rules but this has not, as yet, been attempted.

One additional point should be made regarding the **phenothiazine** rule. Most of the reference compounds with a **phenothiazine** substructure were mono- and di-substituted phenothiazines. Thus, it is possible that this rule will not identify the **phenothiazine** substructure in more highly substituted phenothiazines. Fragmentation of these more highly substituted phenothiazines to provide the key m/z 198, m/z 197, and m/z 196 parent ions may be unlikely since it involves the loss of all of the substituents from the intact ring system. Addition of phenothiazine compounds with greater than two substituents would alleviate this problem since a rule containing clauses indicative of the mono-, di-, and more highly substituted phenothiazines would be generated. The **barbiturate** rule, discussed below, is an example of a rule composed of groups of clauses indicative of a specifically substituted substructure.

Barbiturate. Barbiturates are another class of compounds that find use in pharmaceuticals and in the illicit drug trade [12]. There are 10 compounds that contain the **barbiturate** substructure in the reference databases with a variety of substituents at the 1, 3, and 5 positions. The spectral features from DB 1 that were selected by the MAPS program for use in generating a **barbiturate** substructure identification rule are listed in Table 3. Values of 50% and 30% were used for initial uniqueness and correlation, respectively. The mask which removes features with mass-to-charge ratio greater than the nominal mass of the substructure from the feature list was disabled so there are features in this list with masses greater than the nominal mass of the **barbiturate** substructure (128 u). Ion structures for the parent ions found in Table 3 are provided in Figure 3 except for the m/z 124 ion, which is very similar in structure to m/z 125, and for m/z 106 which is unassigned. These structures were obtained from the literature and published fragmentation pathways for barbiturates [12–17]. It should be noted that these structures should not be considered as indicative of the structures of all ions in the reference database with the given mass-to-charge ratio value due to the existence of isobaric ions in unit resolution mass spectra. For example, the m/z 203 barbiturate ions in the reference database appear to have at least three different ion structures (see Figure 3) based on the structures of the barbiturate standards. The ions shown in Figure 3 also demonstrate that the fragmentation of barbiturates are much more dependent on substituents than the phenothiazine compounds. But once again, it is apparent that the U/C selection criteria for tandem mass spectral features effectively limit the features used for rule generation to those that are directly related to the substructure of interest.

Table 3. Tandem mass spectral features selected for the **barbiturate** substructure

Feature	[U, C]	Feature no.
PD (125.0 43.0 P1)	[100, 30]	{F1}
PD (141.0 80.0 P1)	[100, 30]	{F2}
PD (98.0 80.0 P1)	[85, 46]	{F3}
PD (98.0 27.0 P1)	[83, 38]	{F4}
PD (169.0 97.0 P1)	[83, 38]	{F5}
PD (141.0 73.0 P1)	[80, 30]	{F6}
PD (167.0 124.0 P1)	[80, 30]	{F7}
PD (203.0 132.0 P1)	[80, 30]	{F8}
PD (124.0 43.0 P1)	[80, 30]	{F9}
PD (98.0 28.0 P1)	[75, 46]	{F10}
PD (169.0 126.0 P1)	[71, 38]	{F11}
PD (155.0 112.0 P1)	[66, 30]	{F12}
PD (189.0 118.0 P1)	[66, 30]	{F13}
PD (106.0 51.0 P1)	[66, 30]	{F14}
PD (112.0 66.0 P1)	[66, 30]	{F15}
PD (54.0 39.0 P1)	[66, 30]	{F16}
PD (160.0 133.0 P1)	[57, 30]	{F17}
PD (79.0 39.0 P1)	[57, 30]	{F18}
PD (112.0 94.0 P1)	[57, 30]	{F19}
PD (98.0 44.0 P1)	[55, 38]	{F20}
PD (169.0 57.0 P1)	[55, 38]	{F21}
PD (80.0 52.0 P1)	[50, 46]	{F22}
PD (141.0 98.0 P1)	[50, 30]	{F23}
PD (55.0 28.0 P1)	[50, 30]	{F24}
PD (54.0 27.0 P1)	[50, 46]	{F25}

Ui = 50%, Ci = 30%. Mass filter disabled; USORT (sort by uniqueness) enabled.

The process of generating feature-combinations during rule generation further limits the spectral features because not all ions in the list of initial features are incorporated into the rules. For example, all of the feature-combinations generated for the **barbiturate** substructure with correlation (Cc) > 37% involve daughter ions of m/z 98. These combinations are listed in Table 4. The m/z 98 parent ion, as shown by structure XII in Figure 3, has lost most of the substituents that differentiate the barbiturate standards (e.g., allyl, ethyl, and phenyl groups). So the m/z 98 ion can be considered as the "lowest common denominator" among the barbiturate standards. One potential method for improving the **barbiturate** rules is to collect MS/MS data using alternative ionization conditions. It has been noted previously that the MAPS software does not limit the user to a particular set of operating conditions [5]. Rather, a reference database can be created using any set of MS/MS operating conditions (e.g., negative CI). In fact, it is a long-term goal to incorporate ancillary experiment recommendation capabilities into the ACES system.

Feature-combinations with correlations lower than 37% (not listed in Table 4) include several of the ions shown in Figure 3 with masses > 98 u (i.e., structures VI, VIII, X, and XI in Figure 3). These combina-

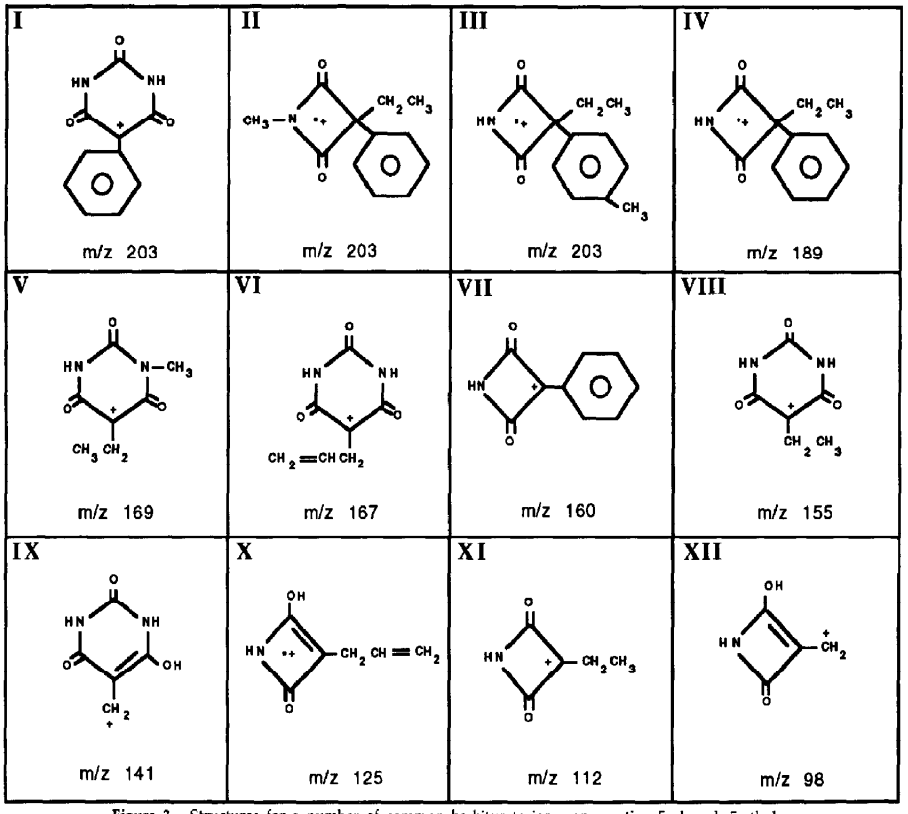


Figure 3. Structures for a number of common barbiturate ions representing 5-phenyl, 5-ethyl, and 5-allyl barbiturates.

tions have lower correlation because the features which comprise them are more specific to a particular subclass of barbiturates. For example, a feature-combination with a correlation of 23% (i.e., 3 out of 13 of the barbiturate standards) and involving m/z 203, m/z 189, and m/z 160 ions was obtained using the features listed in Table 3. All three of the barbiturate reference compounds which yield these MS/MS features were 5-phenyl barbiturates. Other feature-combinations also grouped a particular subclass of barbiturates (i.e., 5-ethyl and 5-allyl barbiturates). Once again, the feature-combinations involving m/z 98 were more generic in the barbiturate standards which contained this ion and therefore, yielded a larger correlation value. This information is currently being explored to determine if more specific information can be provided to the structure generator of ACES when feature-combinations with specific "cross-correlations" are found in the tandem mass spectra of unknown compounds. For example, if the **barbiturate** rule clause had a high degree of cross-correlation with the **phenyl** substructure (i.e., all of the reference compounds associated with the feature-combination had both the **barbiturate** and **phenyl** substructures) then the MAPS software could pass both of these substructures to the structure generator, possibly with a specific substitution pattern. The next section discusses

Table 4. MAPS feature-combination rule for the **barbiturate** substructure

Feature-combination (rule clause)	[U, C]
PD (98.0 28.0) and PD (98.0 55.0) and PD (98.0 70.0)	[100, 46]*
PD (98.0 80.0) and PD (98.0 27.0) and PD (98.0 70.0)	[100, 38]*
PD (98.0 80.0) and PD (98.0 27.0) and NL (111.0)	[100, 38]*
PD (98.0 80.0) and PD (98.0 28.0) and PD (98.0 70.0)	[100, 38]*
PD (54.0 27.0) and PD (98.0 55.0) and PD (98.0 70.0)	[100, 38]*

Rule clauses with Cc > 38% not shown. REL = 100% and REC = 92% with respect to reference database. Ui = 30%, Ci = 30%, Cc = 30%, MINF = 3; mass filter disabled.

cross-correlation which affects the reliability of the MAPS rules for compounds not in the reference database.

Cross-correlation

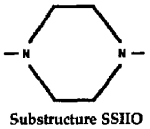
Cross-correlation is a potential source of error in using MAPS because of the empirical nature of the MAPS algorithm. The algorithm requires that any substructure be represented in the reference database in a number of different structural environments (i.e., with a variety of substituents) so that the common elements among the tandem mass spectra of a set of compounds with a specific substructure are fragments due to that substructure. Importantly, cross-correlations can be detected during the rule generation process and appropriate actions can then be taken to remove or at least reduce the cross-correlation.

One way to characterize cross-correlation during rule generation is to identify "feature cross-correlation." This method is most useful when very low correlation (Ci) values are used to select spectral features for rule generation. For example, the eight spectral features listed in Table 5 were obtained using Ui = 100% and Ci = 20% for the **phenothiazine** substructure. (Note that, whereas each of these features have 100% uniqueness, and therefore 100% reliability with respect to the reference database, substructure predictions based on the presence of only one spectral feature in the MS/MS spectra of an unknown are not recommended.) The list of compounds associated with a given spectral feature is retained by the MAPS program and can be reviewed by the user. Most of the features listed in Table 5 have correlations < 46% and are interesting because they are all produced by phenothiazine compounds with very similar substituents.

Table 5. High uniqueness, low correlation spectral features selected for **phenothiazine**

No.	Feature	(U, C)[a]	U(Fx) SS110 (%)
F1	PD (95.0 45.0)	[100, 46]	50
F2	PD (100.0 45.0)	[100, 23]	100
F3	PD (127.0 42.0)	[100, 23]	100
F4	PD (141.0 70.0)	[100, 31]	100
F5	PD (141.0 97.0)	[100, 23]	100
F6	PD (143.0 58.0)	[100, 23]	100
F7	PD (153.0 70.0)	[100, 23]	100
F8	PD (157.0 112.0)	[100, 23]	100

Initial conditions: UI = 100, CI = 20; mass filter enabled.
[a]With respect to the phenothiazine substructure
Uniqueness values calculated for each of the spectral features listed in the rule with respect to the "SS110" substructure are also provided.



Substructure SSIIO

Examination of the structures of the compounds listed suggest that the features listed in Table 5 are highly correlated with **phenothiazine** and a substructure bonded at the 10 position of **phenothiazine**. Feature cross-correlation can be determined quantitatively for each feature in Table 5 by listing any other substructures that have a high uniqueness value for these features. For example, the uniqueness values for the **SS110** substructure (defined as shown) are also provided in Table 5. Most of these features have 100% uniqueness for **phenothiazine** and for the unrelated **SS110** substructure. Thus, these features are ambiguous and should not be used in a rule that only identifies **phenothiazine.** These features may prove to be useful, however, in a new type of rule which provides "alternative" substructures. One remedy for feature cross-correlation is to set Ci high (e.g., 70% as in Figure 2) to ensure that cross-correlated features are not used for rule generation. Another remedy for this problem is to add a number of new standards to the reference database that include one, but not both, of the cross-correlated substructures.

Determining "substructure cross-correlation" is a more general method of identifying cross-correlation and is easily done prior to rule generation. A large cross-correlation of two or more substructures will likely cause spectral features from any of the cross-correlated substructures to be found in a rule for any of those substructures. If the cross-correlated substructures are not related (i.e., one is not a subset of the other), a false positive may result when the rule is applied to an unknown with one of these substructures but not the other.

The spectral feature "PD (70.0 27.0)" listed as F9 in Table 1 is an example of a potential interferent in generating a rule for **phenothiazine**. As the structures for the ions that comprise this feature (see Figure 1) indicate, this feature is likely due to a side chain. This side chain, $-CH_2CH_2N(CH_3)_2$, is defined in MAPS as **SS161.** A substructure cross-correlation factor can be calculated using eq 1,

$$XC(SS_j)_k = \frac{\text{number of compounds with } SS_j \text{ and } SS_k}{\text{number of compounds with } SS_j} \quad (1)$$

where SS_j and SS_k are substructures j and k, respectively. The substructure cross-correlation factor obtained for the **phenothiazine** and **SS161** substructures is 86% (i.e., six out of the seven **SS161**-containing compounds also contain **phenothiazine**).

The one case where the existence of a highly cross-correlated substructure does not impede generation of a reliable rule is when the cross-correlated substructure is, in fact, a substructure of the other. It is expected in this particular case, that the fragmentation of the smaller substructure be an integral part of

the fragmentation of the larger substructure. For example, the **phenylthiol** substructure also has a high XC value (i.e., 76%) for **phenothiazine**. The XC value for **phenothiazine** with **phenylthiol** is 100%. This cross-correlation does not pose a problem for generating a rule for **phenothiazine** since the **phenylthiol** substructure is an integral part of **phenothiazine**. The mask that removes features with a mass-to-charge ratio greater than the nominal mass of the substructure from the feature list and high value for Ci must be used, however, when generating a rule for the **phenylthiol** substructure to discriminate against spectral features due to **phenothiazine**. In any event, the cross-correlation factors for a given substructure with respect to all other substructures can be calculated by the MAPS software so substructures highly cross-correlated within the database are identified. For example, the cross-correlation factors for **phenothiazine** showed only one substructure which was badly cross-correlated with this substructure (i.e., **methyl**). All other substructures with high XC values were substructures of **phenothiazine**. For almost all the other substructures, the XC values were close to 0. This observation strengthens the theory that the database required to develop substructure identification rules will be much smaller, in terms of number of compounds, than that required for spectral matching since a relatively small number of compounds is sufficient to uniquely characterize the fragmentation of the substructure.

A number of program parameters in addition to Ui and Ci affect rule content (generally by limiting the features to be used by the feature-combination generator or the nature of the resulting feature-combination). These parameters include: Cc (specifies the minimum acceptable correlation of a feature-combination with a substructure), MINF (sets the minimum feature-combination length), MAXF (sets the maximum feature-combination length), MINCMPDS (specified the minimum number of compounds that must yield a given feature before the feature is used in a rule), and MINWSS (sets the minimum number of occurrences of a substructure in the reference database before a rule is generated). The next section discusses how instrumental conditions can affect rule content.

Rules Generated Using Multiple Collision Data

Substructure identification rules for the **phenothiazine** substructure were generated by using multiple-collision data (DB 2) as well as single-collision data. The uniqueness and correlation values of a given spectral feature in the reference database varies depending on the pressure regime chosen for data acquisition. In that the MAPS software uses these values to identify the starting features for feature-combination generation, the content of a MAPS rule may also be different from one generated by using another pressure regime. Consider, for example, the "PD" features in the rules for **phenothiazine** (single- and multiple-collision conditions) that have m/z 198 as the parent ion. These features are shown in Figure 4. Three PD features with a m/z 198 parent met the initial uniqueness and correlation minima of 30% by using the single-collision data. Note that four additional ions are observed under multiple-collision conditions and by using the same MAPS parameters. The structures for the ions labeled in Figure 4 are provided in Figure 1. The uniqueness and correlation for each of the PD features are given in brackets in Figure 4. Note that the very diagnostic feature, "PD 198 → 154," is less characteristic of the **phenothiazine** substructure under multiple-collision conditions than under single-collision conditions (U = 92 versus U = 68). However, the new features included in the initial rule by using multiple collision conditions are also quite diagnostic (e.g., "PD 198 → 45," U = 90; and "PD 198 → 166," U = 87).

The feature-combination rule obtained for Ui = 40%, Ci = 50%, Cc = 50%, mass-to-charge ratio mask enabled, and the multiple collision data are shown in Table 6. Several features observed only under multiple-collision conditions are highlighted in this rule in

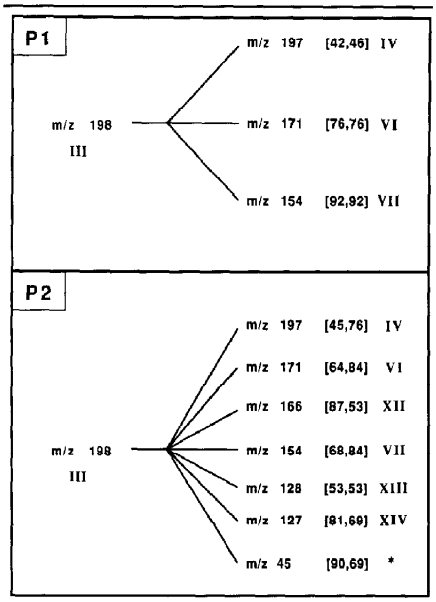


Figure 4. Daughter mass-to-charge ratio values observed in a MAPS rule from a m/z 198 parent ion using single- and multiple-collision conditions. Structures for these ions from Figure 1 are identified by roman numerals.

Table 6. MAPS feature-combination rule for **phenothiazine** generated by using multiple-collision MS/MS data

Feature-combination	[U, C]c
PD (197.0 17.0)	[100, 76]
PD(198.0 45.0) and PD (198.0 154.0)	[100, 69]
PD(199.0 167.0) and PD (198.0 154.0)	[100, 69]
PD(198.0 45.0) and PD (198.0 171.0)	[100, 69]
PD(199.0 167.0) and PD (197.0 196.0)	[100, 61]
PD(199.0 167.0) and PD (197.0 153.0)	[100, 61]
PD(196.0 45.0) and PD (197.0 196.0)	[100, 61]
PD(196.0 45.0) and PD (197.0 153.0)	[100, 61]
PD(196.0 45.0) and PD (196.0 69.0)	[100, 61]
PD(198.0 45.0) and **PD (198.0 127.0)**	[100, 53]
PD(198.0 45.0) and PD (199.0 155.0)	[100, 53]
PD(198.0 45.0) and PD (197.0 196.0)	[100, 53]
PD(198.0 45.0) and PD (197.0 153.0)	[100, 53]
PD(196.0 45.0) and PD (196.0 169.0)	[100, 53]
PD(198.0 166.0) and **PD (198.0 127.0)**	[100, 53]
PD(171.0 45.0) and PD (199.0 167.0)	[100, 53]
PD(171.0 45.0) and PD (199.0 155.0)	[100, 53]
PD(171.0 45.0) and PD (196.0 169.0)	[100, 53]
PD(171.0 45.0) and PD (197.0 196.0)	[100, 53]
PD(199.0 167.0) and PD (179.0 153.0)	[100, 53]

Ui = 40%, Ci = 50%, Cc = 50%; mass filter enabled.
Several features observed only under multiple-collision conditions are highlighted in bold-faced type.

bold-faced type (e.g., "PD 198 → 45"). Note that even though the rule content changed, 100% reliability and 100% recall were maintained for the **phenothiazine** substructure generated using the multiple collision data. There is one single-feature clause in this rule, "PD 197 → 170," which has 100% reliability in predicting the presence of the **phenothiazine** substructure within the reference database. Single-feature clauses are more likely to be unreliable when used with compounds not in the reference database. This result is not particularly surprising if one considers that a substructure identification based on one feature is not a result that an "expert" mass spectrometrist would find compelling. Generally a series of features is the most reliable indicator of the presence of a substructure. The rules shown here often have two or three features in each rule clause and it is expected that this length will increase as more compounds are added to the reference database. The length of the rule clauses will increase because more features will be required to achieve 100% uniqueness (reliability) within the reference database. The rate of increase, however, will decrease rapidly with increasing database size because all legitimate examples of the fragmentation of a given substructure will eventually be represented in the MAPS rules. Thus, the size of the reference database required to achieve this stabilization of rule content is probably much smaller than the size required for reliable spectral matching of unknowns.

Application of MAPS Rules to 20 Test Compounds

Several MAPS rules for the **phenothiazine** and **barbiturate** substructures were applied to the MS/MS data of 20 test compounds to estimate the reliability and recall of these rules for unknown compounds. The rules used in this test were generated using single-collision, multiple-collision, and both single- and multiple-collision reference data. In the latter case, a tag (e.g., P1) for each spectral feature was used to identify the collision conditions that were employed to obtain the feature.

There are a number of different sets of MAPS program parameters which yield substructure identification rules with 100% reliability and recall with respect to the reference database. Briefly, optimal rule reliabilities are obtained using the current reference database with initial uniqueness and correlation values greater than 30%. Also, the highest possible MINF value (minimum number of spectral features per rule clause) should also be used. In theory, the presence of a single rule clause (a feature-combination) in the tandem mass spectra of an unknown should be sufficient for a substructure prediction to be made since each rule clause has 100% uniqueness for the indicated substructure (with respect to the reference database). However, if a rule contains feature-combinations consisting of only two features then the substructure prediction may be based on the presence of only two spectral features in the tandem mass spectra of an unknown. Thus, the MINF parameter can be used to increase the number of spectral features which must be observed in the tandem mass spectra of an unknown before a substructure prediction is made.

The reliability and recall estimates (with respect to the reference database and with respect to the test compounds) obtained for the **phenothiazine** rules are listed in Table 7A. In this case, a MINF value of 2 and high values of Ui and Ci were used to generate the rules. A high degree of recall (i.e., 100%) was obtained for the rules generated by using the single-collision and combined single- and multiple-collision data while a recall of 76% was observed for the rule generated by using only multiple-collision data. Lower recall is often observed for rules generated by using multiple-collision data when high values for Ui and Ci and low values for MINF are used to generate the rules. A MAPS rule, not shown in Table 7A, was generated by using lower Ui and Ci values (i.e., 30% and 30%, respectively) and an increased MINF value (i.e., 4). The recall of this rule was 100%. This rule also identified all five of the **phenothiazine**-containing compounds among the 20 test compounds (REL = 100% and REC = 100% with respect to test compounds). This result tends to reinforce the idea that the rules should be optimized to include feature-combinations composed of several spectral features.

A reliability estimate of 100% was observed for the **phenothiazine** rules generated by using single-

Table 7. Rule reliability and recall values calculated for (A) **phenothiazine** rules, (B) **barbiturate** rules generated using a low MINF value, and (C) **barbiturate** rules generated by using a higher MINF value

A.

	P1		P2		P1 + P2	
	REL	REC	REL	REC	REL	REC
w/r/t RDB	100	100	100	76	100	100
w/r/t TC	100	100	80	80	100	100
info	7/7 f.	6 cl.	8/8 f.	5 cl.	10/15 f.	7 cl.

(Ui=40%, Ci=70%, Cc=70%, mass filter on, MINF=2, HITS=5, MAXSF=10,USORT selected)

B.

	P1		P2		P1 + P2	
	REL	REC	REL	REC	REL	REC
w/r/t RDB	100	92	100	84	100	76
w/r/t TC	50	100	60	100	60	100
info	30/56 f.	24 cl.	30/87 f.	26 cl.	30/129 f.	22 cl.

(Ui=30%, Ci=30%, Cc=30%, mass filter off, MINF=2, HITS=5, MAXSF=30,USORT selected)

C.

	P1		P2		P1 + P2	
	REL	REC	REL	REC	REL	REC
w/r/t RDB	100	92	100	69	100	92
w/r/t TC	100	100	75	100	100	100
info	56/56	0 cl.	87/87 f.	18 cl.	110/129 f.	20 cl.

(Ui=30%, Ci=30%, Cc=20%, mass filter off, MINF=4, HITS=5, MAXSF=110,USORT selected)

KEY:

w/r/t:	with respect to	f.:	# of initial features
RDB:	reference database	cl.:	# of rule clauses
TC:	test compounds		
info:	the ratio of the number of features used for generating feature-combinations to the number of features which meet the Ui/Ci criteria (sometimes not equal to one because of the MAXSF parameter)		
P1:	DB#1 data; single collision conditions		
P2:	DB#2 data; multiple collision conditions		
P1+P2:	combined databases; each feature is tagged with P1 or P2		
MINF:	minimum number of features in a feature-combination		
HITS:	maximum number of feature-combinations that identify only the same compounds as previously generated feature-combinations		
MAXSF:	maximum number of starting (initial) features		
USORT:	sort starting features by uniqueness		

collision data and the combined single- and multiple-collision data. One false positive was obtained by using the rule generated from the multiple-collision data (or four correct predictions out of five total predictions which translates into the 80% reliability figure listed in Table 7A). This rule incorrectly identified the presence of the **phenothiazine** substructure based on one out of five rule clauses hitting on a compound which did not have the **phenothiazine** substructure. All other predictions obtained using this rule were based on five out of five rule clauses hitting on a test compound. The latter predictions provide greater confidence that a correct prediction has been made since a larger number of spectral features were observed in the appropriate test compounds. Once again, the **phenothiazine** rule generated by using the increased MINF value and multiple-collision conditions correctly identified the five **phenothiazine** containing test compounds with no false positives.

It was noted earlier that the fragmentation of barbiturates was much more dependent on side chains than the phenothiazine compounds. Consequently, it was more difficult to obtain a reliable rule for the **barbiturate** substructure (see Table 7B). One problem involved the smaller mass ions which incorporate only a fraction of the original **barbiturate** substructure but often have the highest correlation with the **barbiturate** substructure. These ions can be produced from compounds which are not barbiturates. The reference database appears to have been too small to prune out all of the false feature-combinations which included these features.

The results obtained for another set of **barbiturate** rules are summarized in Table 7C. Note that these rules were generated by using a MINF value of 4. Once again, best rule application results are obtained by using the single- and the combined single- and multiple-collision data (REL = 100%, REC = 92% with respect to reference database and REL = 100%, REC = 100% with respect to test compounds). No false positives were observed by using these rules. One false positive for the **barbiturate** substructure was observed for the rule generated with only the multiple-collision data (or three correct predictions out of four total predictions, which translates into the 75% reliability figure listed in Table 7C). Significantly, this identification was based on only a single-rule clause firing. Also, the misidentified compound was primidone, a compound containing a substructure very closely related to the **barbiturate** substructure (i.e., lacking only one carbonyl functionality). This compound was included in the test compounds to test the ability of the MAPS rules to discern closely related substructures. The **barbiturate** rule generated using the combined data did not fire when applied to the tandem mass spectra of primidone.

Another method to increase the reliability of the MAPS rules for identifying substructures outside of the reference database is to increase the size of the reference database. A MAPS rule derived from a relatively large reference database (e.g., 1000 compounds) will have a greater inherent reliability than a rule derived from a smaller database (e.g., 100 compounds) because a greater number of false correlations will be eliminated from the rules. A larger reference database will also decrease the likelihood of cross-correlations since more substructures will be found in a greater number of structural environments.

Implications for the Analysis of "True" Unknowns

One significant advance in MS/MS instrumentation will substantially enhance the applicability of the MAPS rules for "true" unknowns (mixtures which include analytes and contaminants). This advance involves the development of a tandem mass spectrometer capable of MS/MS "mapping" on a practical chromatographic time scale. A separation step is required because the ACES system must assume that *all substructure identifications for an unknown are due to one component*. This assumption limits current application of the MAPS rules to complete structure elucidation of purified unknowns or substructure screening of multicomponent unknowns. This last capability was first observed when one test compound was contaminated with the previously run test compound. A rule-

base of 23 MAPS rules was used to analyze this compound and a disappointing number of false positives (i.e., five false positives out of 11 total predictions) were observed. It was then realized that the misidentified substructures were all present in the previously run test compound. New tandem mass spectra were acquired to replace the contaminated spectra and no false positives were observed by using these spectra.

The time required to analyze an unknown is determined by the number and length of the MAPS rules. This process can take up to several minutes with the bulk of the analysis time being used to load the primary mass spectrum with all associated daughter spectra and generating a list of tandem mass spectral features. The number of daughter spectra generated for an unknown can vary over a large range (i.e., from just a few to over a hundred daughter spectra depending on the number and relative intensity of the fragment ions in the primary mass spectrum of the unknown). Thus, the analysis time can vary from seconds to minutes. There should not be a significant increase in analysis time as the reference database is expanded to refine existing rules, in contrast to the conventional spectral matching method, because the number of MS/MS features that are compared to those in the rules will not vary to a great extent. A significant increase in the number of rules in the MAPS rulebase, however, will add to the analysis time. This effect can be minimized by adding filters (based on molecular weight and elemental composition, if known) to the RULE program that will reduce the number of rules checked.

Conclusion

The use of an initial uniqueness and correlation value in the new MAPS program provides a convenient means of limiting the spectral features used for feature-combination rule generation to those features directly related to the indicated substructure. In cases where substituents play an integral role in the fragmentation of a substructure, initial correlation may need to be lowered to obtain a rule (e.g., the barbiturate substructure). Cross-correlation coefficients can be calculated to identify spectral features that may be due to an interfering substructure. Another cross-correlation coefficient for feature-combinations may prove useful for providing more specific information to a structure generator in an automated system but this option has not been incorporated into the software as of this writing.

Uniqueness and correlation values tend to decrease for features observed under single-collision conditions when multiple-collision conditions are employed. However, spectral features not observed under single-collision conditions often possess a high degree of uniqueness and correlation (e.g., the daughter ions of m/z 198 in phenothiazine compounds).

The application of rules for the **phenothiazine** and **barbiturate** substructures to 20 test compounds indicates that reliable substructure identification rules for unknowns can be generated using the MAPS software. Optimal uniqueness and correlation values tend to vary among substructures but it is generally unwise to use values < 30% (assuming at least 10 representative compounds are present in the reference database so any feature selected will be observed in at least three compounds). Also, a MINF value of at least 4 should be used when low values of Ui and Ci are used to ensure that any feature-combination in a rule, which may identify a substructure in an unknown, is comprised of at least four spectral features. Rules generated by using combined single- and multiple-collision data add a new dimension (i.e., collision gas pressure dependence) to the information used for structure elucidation. Lastly, a larger EI MS/MS reference database and perhaps a database of tandem mass spectra of negative ions (acquired using negative CI) will prove useful for increasing the variety and reliability of the MAPS rules used for structure elucidation of unknowns.

Supplementary Material

Supplementary material for this article is available in photocopy form from the office of the Editor-in-Chief (see inside front of journal for address). Requests must include complete title of article, names of authors, issue date, and page numbers. The supplementary material consists of the tables listed below.

Table S1. List of reference names, compound names, and CAS numbers for the standards used to create the reference databases

Table S2. Compound name, molecular weight, molecular formula and CAS number for each of the test compounds used to evaluate the MAPS rules.

Acknowledgment

The authors wish to thank Peter Palmer, Drake Diedrich, and Chris Weaver for their contributions to this project. This work is supported by National Institutes of Health grant GM-28254. Thanks are also due to Finnigan-MAT and Michigan State University for funds to purchase the VAXstation 3200.

References

1. Hart, K. J. Ph.D. thesis; Michigan State University: East Lansing, MI, 1989.
2. Hart K. J.; Palmer, P. T.; Diedrich, D. L.; Enke, C. G. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 159-168.
3. Hart, K. J.; Enke, C. G. In *Proceedings of the Symposium on Chemometrics and Intelligent Laboratory Automation*, Canadian Chemical Conference, Victoria, BC. *Chemometrics and Intelligent Lab. Syst.* **1990**, *8*, 293-302.

4. Enke, C. G.; Wade, A. P.; Palmer, P. T.; Hart, K. J. *Anal. Chem.* **1987**, *59*, 1363A-1371A.
5. Hart, K. J.; Enke, C. G. In *Computer-Enhanced Analytical Spectroscopy*, vol. 3; Jurs, P., Ed.; Plenum: New York, 1992, in press.
6. Busch, K. L.; Glish, G. L.; McLuckey, S. A. *Mass Spectrometry/Mass Spectrometry: Techniques and Applications of Tandem Mass Spectrometry*; VCH Publishers: New York, 1988.
7. Windholz, M., Ed. *The Merck Index*; Merck & Company: Rahway, NJ, 1983; Merck Index numbers 7688, 5862, 9492, 64, 1844, 7044, 1485, 5755, 5826, 7723, 7691, 3696, 9202.
8. Morosawa, S.; Kamal, S.; Dandiya, P. C.; Sharma, H. C. *Org. Mass Spectrom.* **1982**, *17*, 309-314.
9. Hallberg, A.; Al-Showaier, I.; Martin, A. R. *J. Hetecocyl. Chem.* **1984**, *21*, 841-844.
10. Flurer, R. A.; Busch, K. L. *Org. Mass Spectrom.* **1988**, *23*, 118-128.
11. Gilbert, J. N. T.; Millard, B. J. *Org. Mass Spectrom.* **1969**, *2*, 17-31.
12. Falkner, F. C.; Watson, J. T. *Org. Mass Spectrom.* **1974**, *8*, 257.
13. Watson, J. T.; Falkner, F. C. *Org. Mass Spectrom.* **1973**, *7*, 1227.
14. Thompson, R. M.; Desiderio, D. M. *Org. Mass Spectrom.* **1973**, *7*, 987.
15. Grutzmacher, W. F.; Arnold, W. *Tetrahedron Lett.* **1966**, 1365.
16. Dilli, S.; Pillai, D. N. *Aust. J. Chem.* **1975**, *28*, 2765.
17. Soltero-Rigau, E.; Kruger, T. L.; Cooks, R. G. *Anal. Chem.* **1977**, *49*, 435.